An Efficient Privacy Preserving in MSN using Improved Decision Tree Algorithm

Neetu Raghuwanshi
Dept. of CSE
Samrat Ashok Technological Institute,
Vidisha (M.P.), India

Abhishek Mathur
Assistant Professor,Dept. of CSE
Samrat Ashok Technological Institute,
Vidisha (M.P.), India

ABSTRACT
Here in this paper, efficient privacy preservation over Mobile Social Networks is implemented to secure attacks over Mobile Social Networks. The Existing methodology implemented for the Friending Mobile Social Networks is efficient which provides an efficient computation of Data and privacy from unauthorized users. Here an efficient Decision Tree based algorithm is implemented using Partition of Data using Some Partition based algorithm and then classify data using an ID3 algorithm. The Experimental results when performed on Social Network Dataset the proposed methodology gives efficient results in comparison.

Keywords
Mobile Social Network, Decision Tree, Privacy Preservation, ID3, Classifier, Facebook.

1. INTRODUCTION
World Wide Web (WWW) is the world with rapid and continuous growth in all aspects. It is a data repository which is massive, huge, diverse, dynamic and unstructured. This repository is used as an information repository for the purpose of knowledge reference. The challenges that are faced by web are in the form of large, semi-structured web pages and also the information on web is likely to be diverse in meaning, quality of the information extracted and the conclusion of the knowledge is obtained from extracted information [1]. Thus for the appropriate perceptive and analysis, the data structure of the Web plays an important role for efficient Information Retrieval.

Web mining can be explained a mechanism that categorizes the web pages and internet users in accordance with the contents of the web page and the behavior of the user adopted in the past on the internet. Web Mining is considered as an application of data mining technique. It is generally used to find and retrieve information from the WWW automatically [2].

Social network advances to understand social interaction which is needed to be first visualized and then investigated through the properties of the relations between the units and not upon unit properties itself.

In a social network, there exist different types of relations which may be in singular or combination form with the network configurations and network analytics.

While a social networking service provides a platform for building the social networks or social relations by the users and among the

users who share interests, activities, backgrounds, relations etc. social network service helps each user to maintain its profile that contains his or her social links and information about other additional services [3].

Therefore Social networks generally enable the users to create a public profile and maintain a list of users for sharing connections and views and even cross the connections inside the social system. Social network services are web based services facilitating the user to interact over the Internet that may be in the form of e-mail servicing and instant messaging. Social network even allows multiple information and communication tools in the form of mobile connectivity, photo, video, sharing, blogging etc.

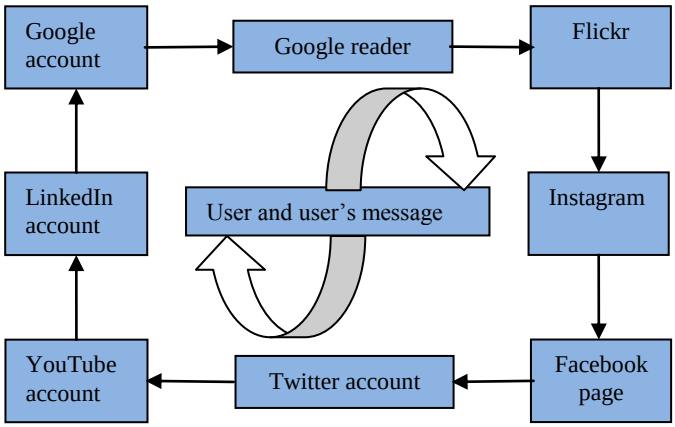

Figure 1: Architecture of OSN

Social networking concept is greatly unique in its own way as users can collectively identify others if they are inappropriate, unoriginal or fake and also in social networks users do not compartmentalize their life by having only one social account. Multiple communities in the social network are held together and sometimes recognized by a common interest.

The users may possess a hobby for which the community members may be passionate, have a common goal, project, similar lifestyle, geographical location, profession, common interest etc. Thereby in social networks, there are generally two types of users those exhibits and have different influence and different behavior [4].

A Social network provides base over the internet for maintaining the social associations among the users and helps the users to search other users that may have alike types of interests. It also provides platform for publishing the content and provide knowledge which is provided or generated by other users and also shared, authorized and approved by other users [5].

Social networks enable the present internet generation to maintain interaction with the technology and its usage as well as with other people. OSN's can be well thought-out as a combination of technological, economical and social drives

that are capable of fulfilling the need of the users for building social networks, relations etc. over the internet or the web [6].

2. LITERATURE SURVEY

In 2012 Yaping Li et al [7] Presented Enabling Multilevel Trust in Privacy Preserving Data Mining. Privacy conserving data processing (PPDM) addresses the matter of developing correct models concerning mass knowledge while not access to specific data in individual knowledge record. A wide studied perturbation-based PPDM approach introduces random perturbation to individual values to preserve privacy before knowledge area unit printed. Previous solutions of this approach area unit restricted in their inexplicit assumption of single-level trust on knowledge miners and MLT-PPDM permits knowledge homeowners to come up with otherwise discomposed copies of its knowledge for various trust levels. The primary problem lies in preventing the information miners from combining copies at completely different trust levels to collectively reconstruct the initial data a lot of correct than what's allowed by the information owner [7].

All assumption and expand the scope of perturbation-based PPDM to construction Trust (MLT-PPDM) and also the additional trusty an information jack is, the less rattled copy of the info it will access. Preventing such diversity attacks is that the key challenge of providing MLT-PPDM services. Here address this challenge by properly correlating perturbation across copies at totally different trust levels and prove that this resolution is powerful against diversity attacks with regard to privacy goal. That is, for information miners World Health Organization have access to AN impulsive assortment of the rattled copies, this resolution stop them from conjointly reconstructing the first information additional accurately than the most effective effort exploitation a person copy within the assortment. This resolution permits a information owner to come up with rattled copies of its data for impulsive trust levels on demand. This feature offers information house owners most flexibility [7].

In 2008 by Bart Kuijpers et al. [8] proposed the complexity analysis, in which case the earlier evaluation method is the more efficient and give an algorithm for privacy preserving ID3 over horizontally partitioned data involving more than two parties. For grid partitioned data, here discuss two different evaluation methods for preserving privacy ID3, that is, first merging horizontally and increasing vertically or first merging vertically and next developing horizontally with the help of these concept the complexity analysis of both algorithms shows that it is more efficient to first merge data horizontally and further develop it vertically than the other way around [8].

In year 2012, MS Ramya proposed Partial Information Hiding in Multi-Level Trust Privacy Preserving Data mining. The Multi-Level Trust in Privacy-Preserving Data Mining when integrated with partial information hiding methodologies help to find the right balance between maximum analysis results and keep the inferences that disclose private information about organizations or individuals at a minimum. Thus random rotation based data perturbation and K-anonymity are incorporated with MLT-PPDM to significantly enhance the data accuracy and to prevent the leakage of the sensitive data [9].

In 2012 by Alka Gangrade and Ravindra Patel gives the concept about the two layer protocol uses an Un-trusted Third Party (UTP) and explains how to build privacy preserving two-layer decision tree classifier, where database is horizontally partitioned and communicate their intermediate results to the UTP not their private data. In this protocol, an UTP allows well-designed solutions that meet privacy constraint and achieve suitable performance and finally proposed a new classifier using two-layer architecture that enables SMC by hiding the identity of the parties' attractive part in the classification process using UTP. Further they may describe that intermediate result is calculated by every party individually and send only intermediate result to UTP not the input data. During the communication among UTP and all party final result is carried out. It requires less memory space. Also provides fast and easy calculations. Using this protocol, classification will almost secure and privacy of individual will be maintained. Additional development of the protocol is estimated in the sense that for joining multi-party attributes using a trusted third party can be used [10].

They [10] addressed the issue related to privacy preserving data mining in a distributed manner. In particular, they also focus on privacy preserving two-layer decision tree classifier on horizontally partitioned data. The objective of privacy preserving data classification is to build accurate classifiers without disclosing private information in the data being mined. The performance of privacy preserving techniques should be analyzed and compared in terms of both the privacy protection of individual data and the predictive accuracy of the constructed classifiers.

Various proximity based adaptable interpersonal associations are generated to relationships between any two entities or to facilitate a customer to determine entities with synchronized outline contained by of a convinced division. A testing responsibility in these requests is to make sure the safety measures the members' outlines and entity hobbies. Here author outlines new devices when given a preference outline put together by a user that follow a man with synchronizing outline in decentralized multi-bounce adaptable interpersonal associations. The schemes are safety measures defending: no members' profile and the suggested inclination outline are representation. The schemes set up a protected communication channel between the inventor and coordinating clients when the synchronize customer is originate. [11] The methodical test shows that the scheme is safe, defense safeguarding, understandable and creative both in communication and calculation. Extensive appraisals utilizing real interpersonal association data and genuine structure implementation on go forward cells show that the schemes are fundamentally more effectual than obtainable understandings.

3. PROPOSED METHODOLOGY

The proposed methodology implemented here consists of following phases:

- Take an input dataset from which some meaningful information can be extracted.
- Now "On Demand" of the untrusted third party the dataset can be partitioned vertically into 'N' parties.
- Each of the party contains a set of attributes with their respective classes.
- Computation of Information Gain by each of the party and send to UTP.
- UTP on the basis of information Gain will select the attributes having information gain and the remaining attributes with less information gain can be removed from the dataset.

- Now clustering is done for each of the party on the basis of classes available.
- Finally decision tree is generated from the available clustered dataset.

3.1 On Demand Vertical Partition

Input Layer – Input layer comprises of all the parties that are involved in the computation process. They individually calculate the Information Gain of each attribute and send Intermediate result to UTP. This process is done at every stage of decision tree.

Output Layer – The UTP exists at the 2nd layer i.e. the computation layer of our protocol. UTP collects only intermediate results from all parties not data and calculate the total information gain of each attribute. Then find the attribute with highest information gain and then create the root of decision tree with this attribute and send this attribute to all parties for further calculation. This process is also done at every stage of decision tree.

3.1.1 Informal Algorithm

3.1.1.1 Input Layer

1. Party individually calculates Expected Information of every attribute.
    - The input dataset taken here is first divided into a number of parties. Here parties are the various users who can calculate dataset attributes. The Information describes here is the impact of particular class in the dataset and is given by:

$$I(y,n) = -\frac{y}{(y+n)} log\left(\frac{y}{y+n}\right) - \frac{n}{(y+n)} log\left(\frac{n}{y+n}\right) \quad (1)$$

   Where, I is the information to be computed for the classes 'y' and 'n'.

2. Party individually calculates Entropy of every attribute.
    - After calculating the Information of dataset by each of the party. Entropy is computed based on the classes and attributes. The entropy can be computed using:

$$E(A) = \sum_{i=1}^{v} \frac{p_i+n_i}{p+n} I(p_i, n_i) \quad (2)$$

3. Party individually calculates Information Gain of each attribute.
    - Finally after the calculation of the Entropy of each of each of the attribute Gain of each of the attribute is computed on the basis of parties. The Information Gain Computed here computes the dependency factor of attribute in the whole dataset.
4. Calculation of information gain from Han and Kamber and Pujari.
5. Assume there are two classes, *P* and *N*.
6. Let the set of examples *S* contain *p* elements of class *P* and *n* elements of class *N*.
7. The amount of information, needed to decide if an arbitrary example in *S* belongs to *P* or *N* is defined as:

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n} \quad (3)$$

8. Assume that using attribute A set *S* will be partitioned into sets $\{S_1, S_2, \ldots, S_v\}$.
9. If S_i contains p_i examples of *P* and n_i examples of *N*, the entropy, or the expected information needed to classify objects in all subtrees S_i is:

$$E(A) = \sum_{i=1}^{v} \frac{p_i+n_i}{p+n} I(p_i, n_i) \quad (4)$$

10. The encoding information that would be gained by branching on A.

3.1.1.2 Output Layer

1. All party send Information Gain of each attribute to the UTP
2. UTP compute the sum of Information Gain of all parties of all attributes (Total Information Gain ( )).
3. UTP find out the attribute with the largest Information Gain by using Max Information Gain( )
4. Create the root with largest Information Gain attribute and edges with their values, and then send this attribute to all parties at Input Layer for further development of decision tree.
5. Recursively do when no attribute is left.

3.1.1.3 Assumptions

The following assumptions have been set

1. UTP computes the final result from the intermediate results provided by all parties at every stage of decision tree.
2. UTP computes attribute with highest information gain and send to all party at every stage of decision tree.
3. UTP has the ability to announce the final result of the computation publicly.
4. Each party is not communicating their input data to other party.
5. The communication networks used by the input parties to communicate with the UTP are secure.

3.1.2 Formal Algorithm

Input Layer

- Define P_1, $P_2 \ldots P_n$ Parties (Vertically partitioned).
- Each Party contains R set of attributes A_1, $A_2 \ldots A_R$.
- C the class attributes contains c class values C_1, $C_2 \ldots C_c$.
- For party P_i where i = 1 to n do
- If R is Empty Then
- Return a leaf node with class value
- Else If all transaction in $T(P_i)$ have the same class Then
- Return a leaf node with the class value
- Else
- Calculate Expected Information classify the given sample for each party P_i individually.

- Calculate Entropy for each attribute (A_1, A_2… A_R) of each party P_i.
- Calculate Information Gain for each attribute (A_1, A_2,…., A_R) of each party P_i
- Calculate Total Information Gain for each attribute of all parties (Total Information Gain ( )).
- $A_{BestAttribute}$ ← MaxInformationGain( )
- Let V_1, V_2… Vm be the value of attributes. $A_{BestAttribute}$ partitioned P_1, P_2,…., P_n parties into m parties
- $P_1(V_1)$, $P_1(V_2)$, …., $P_1(V_m)$
- $P_2(V_1)$, $P_2(V_2)$, …., $P_2(V_m)$
- . .
- . .
- $P_n(V_1)$, $P_n(V_2)$, …., $P_n(V_m)$
- Return the Tree whose Root is labelled $A_{BestAttribute}$ and has m edges labelled V_1, V_2… V_m. Such that for every i the edge Vi goes to the Tree
- NPPID3(R – $A_{BestAttribute}$, C, ($P_1(V_i)$, $P_2(V_i)$, …., $P_n(V_i)$))
- End.

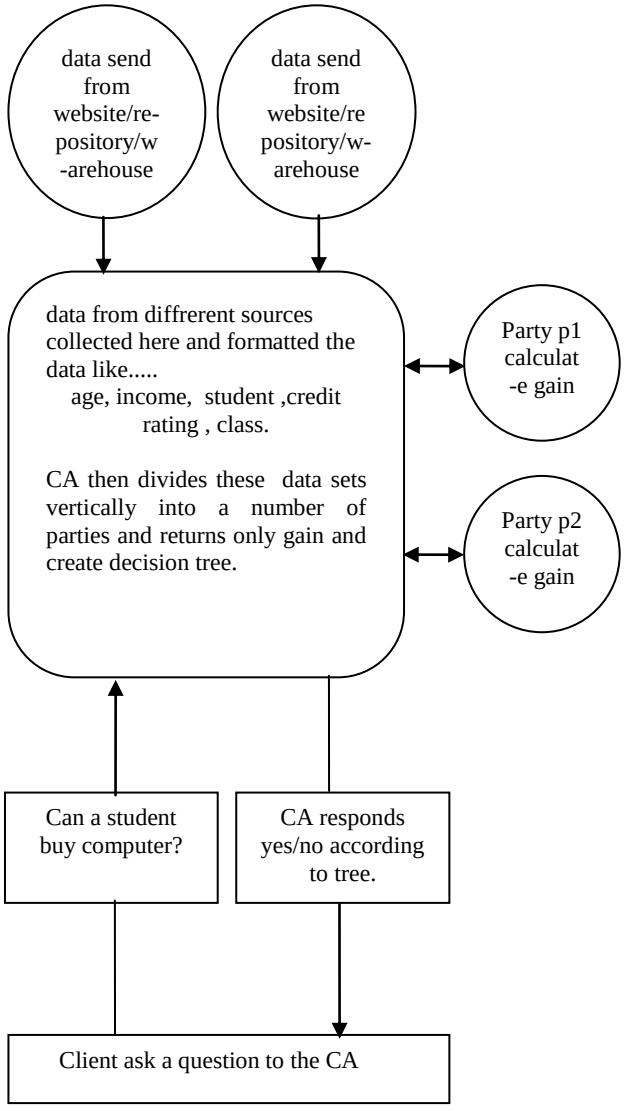

Figure 2: An Example of partitioned ID3 Algorithm

4. RESULT ANALYSIS

The Table Shown below is the analysis and Comparison of Candidate User Proportion on the number of attributes in the Dataset. The Proposed methodology provided high ratio of Candidate User Proportion in comparison with the Existing Privacy Preservation algorithm.

Table 1. No. of Candidate User Proportion

	Candidate User Proportion	
Attribute Number	**Existing Work**	**Proposed Work**
0	0.53	0.62
1	0.34	0.38
2	0.21	0.26
3	0.16	0.22
4	0.08	0.15
5	0.05	0.09
6	0.01	0.04

The Table Shown below is the analysis and Comparison of Number of Candidate Profile Keys on the number of attributes in the Dataset. The Proposed methodology provided constant Number of Candidate Profile Keys in comparison with the Existing Privacy Preservation algorithm.

Table 2. No. of Candidate Profile Keys

	No. of Candidate Profile Keys	
Attribute Number	**Existing Work**	**Proposed Work**
1	3	1
2	4	1
3	4	1
4	4	1
5	2	1
6	1	1

Here in the given table comparison of Efficiency of existing and proposed methodology is given on the basis of Computation and Communication and Transmission. The Proposed Methodology shows better Performance in Comparison.

Where, H is the SHA-256 hashing Operator.
M is the operation or Computation required.
P1 & P2 are the parties as a constant.

Table 3. Comparison of Efficiency

Measures	**Existing Work**	**Proposed Work**
Computation	7H+6M+€(P1)	2M+€(P1 || P2)
Communication (KB)	0.7(P1)	0.23 (P1 || P2)
Transmission	1 broadcast with <100 (#candidate unicast)	Not Required

The Figure Shown below is the analysis and Comparison of Candidate User Proportion on the number of attributes in the Dataset. The Proposed methodology provided high ratio of Candidate User Proportion in comparison with the Existing Privacy Preservation algorithm.

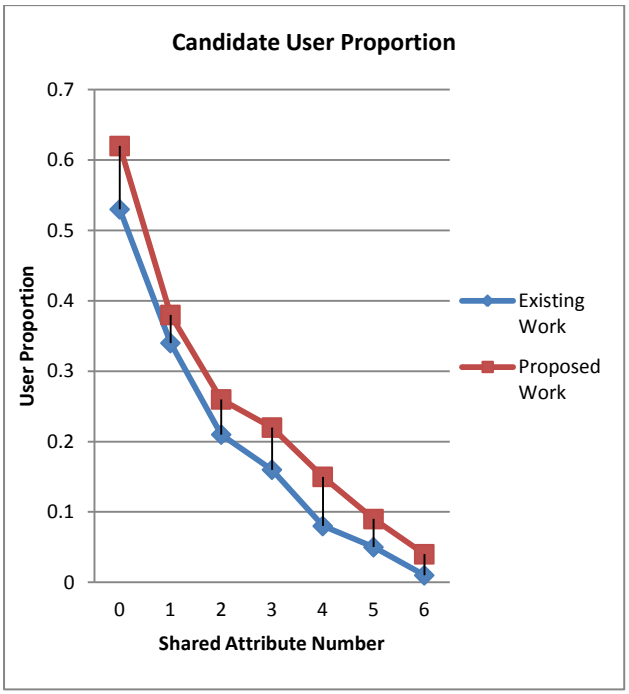

Figure 3: Comparison of Candidate User Proportion

The Figure Shown below is the analysis and Comparison of Number of Candidate Profile Keys on the number of attributes in the Dataset. The Proposed methodology provided constant Number of Candidate Profile Keys in comparison with the Existing Privacy Preservation algorithm.

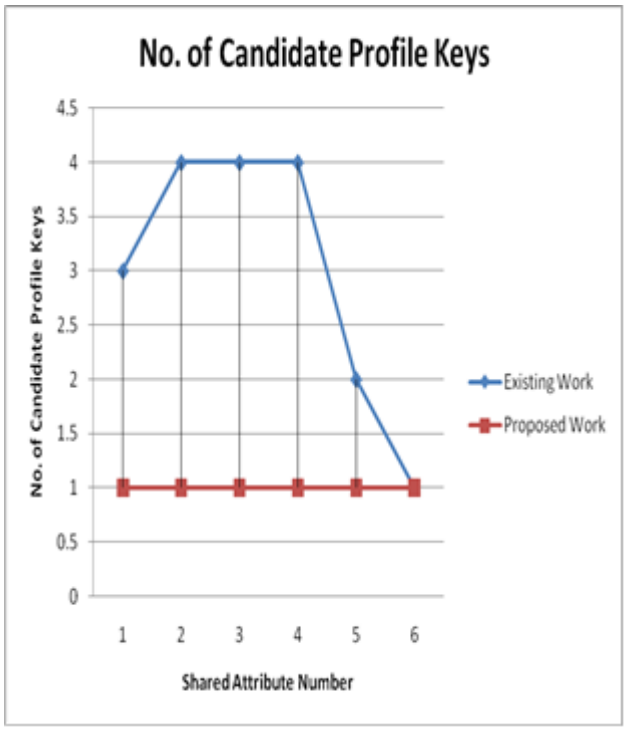

Figure 4: Comparison of No. of Candidate Profile Keys

5. CONCLUSION

The Proposed methodology implemented here for the Privacy Preservation over Mobile Social Networks for the Security of Unauthorized users. The Proposed Methodology implemented is based on the concept of Privacy Preservation using Third Party Computation by Partitioning the Data for number of parties and then Generate Decision Tree using ID3 algorithm. The Proposed Methodology implemented provides efficient Computation of Dataset and also provides privacy among users.

6. REFERENCES

[1] J. S. Park, S. Kim, C. Kamhoua and K. Kwiat "Towards Trusted Data Management in OSN Services" World Congress on Internet Security, IEEE- 2012, 978-1 - 908320-04/9 pp. 202-203.

[2] L. Kagal, T. Finin, M. Paolucci, N. Srinivasan, K. Sycara and G. Denker "Authorization and Privacy for Semantic Web Services" IEEE Intelligent Systems 2004, 1541 - 1672/04, pp. 50-56.

[3] Soumen Chakrabarti, Byron Dom, David Gibson, Jon Kleinberg, S. Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins, "Hypersearching the web", Scientific American, 1999.

[4] Liaoruo Wang, Tiancheng Lou, Jie Tang and John E. Hopcroft "Detecting Community Kernels in Large Social Networks", 2011.

[5] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel and Bobby Bhattacharjee, ―Measurement and Analysis of Online Social Networks, ACM, 2007.

[6] Walter Willinger, Reza Rejaie, Mojtaba Torkjazi, Masoud Valafar and Mauro Maggioni, ―Research on Online Social Networks: Time to Face the Real Challenges‖, 2009.

[7] Yaping Li, Minghua Chen, Qiwei Li, And Wei Zhang "Enabling Multilevel Trust In Privacy Preserving Data Mining" , Ieee Transactions On Knowledge And Data Engineering, Vol. 24, No. 9, September 2012.

[8] Bart Kuijpers, Vanessa Lemmens, Bart Moelans," Privacy Preserving ID3 over Horizontally, Vertically and Grid Partitioned Data",avrxi-0803.155v1[cs.db], 11 march 2008.

[9] Ramya, M. S. "Partial Information Hiding in Multi-Level Trust Privacy Preserving Datamining." Bonfring International Journal of Software Engineering and Soft Computing 2, no. Special Issue Special Issue on Communication Technology Interventions for Rural and Social Development, pp.11-15, 2012.

[10] Alka Gangrade, Ravindra Patel "Privacy Preserving Two-Layer Decision Tree Classifier for Multiparty Databases", International Journal of Computer and Information Technology, ISSN No: 2277 – 0764, Volume 01, Issue 01, September 2012.

[11] Lan Zhang , Kebin Liu , Taeho Jung and Yunhao Liu "Message in a Sealed Bottle: Privacy Preserving Friending in Mobile Social Networks" IEEE Transactions On Mobile Computing, Vol. 14, No. 9, September 2015.