International Journal of Research In Science & Engineering e-ISSN: 2394-8299
Volume: 3 Issue: 2 March-April 2017 p-ISSN: 2394-8280

ACCESS CONTROL BASED PRIVACY PROTECTION MECHANISM IN KNOWLEDGE EXTRACTION: A Study

Mala Dutta[1], Varsha Gaur[2]
Assistant Professor, Department of Computer Enginnering, Institute of Enggineering & Technology, Khandwa Road,Indore,Indore
mdutta@ietdavv.edu.in
Lecturer, Departmemt of Information Technology, Prashanti Institute of Technology & Science, Gram Gangedi near Mahaveer Tapobhumi, ,Ujjain
gaurvarsha16@gmail.com

ABSTRACT

The privacy is primary requirement of growing technology. Maintaining Isolation over sensitive data in public environment is a big challenging task. It becomes more complex when data set becomes very large and number of users reaches to huge figure. Access Control principle help to classify the users according to rights and permission. Integration of Access Control model with knowledge extraction process is proposed to achieve privacy over sensitive data.

Sensor nodes are one of the major utility to control the complete home appliance and can be known as smart home controller. Data Analytics may used to explore pattern and explore more relevant a useful use cases. Privacy preservation also plays crucial factor during the whole phenomena. The complete solution generates a need to improve the capability of access control model. The research paper has considered education domain as platform to develop solution to maintain the privacy during the analytics process on smart home data. Subsequently, access control model will help to categorize and priorities the database attributes and data according to access. It will help to maintain data privacy during web mining.

Keyword: Access Control, Privacy Protection, Sensitive Data

1. INTRODUCTION

The enhancement in technology is changing the practice of human. Development of industry without computer and use of computer without internet is a joke today. Internet based services and applications are rapidly emerging and increases demand to upgrade applications and existing solutions. Internet based large storage and services is known as cloud computing. Cloud computing services often rely on specific systems such as Hadoop Map Reduce, an open source proposed by Google. Map Reduce is being adopted by many academic researchers for data processing in different research areas, such as high-end computing, data intensive scientific analysis, large scale semantic annotation and machine learning.

As the rate at which we generate data increases, we find a greater and greater need to handle voluminous amounts of data within traditional machine learning algorithms. As a general rule of thumb, the more examples you can provide to a machine learning algorithm, the better it will be able to perform. The ability to quickly and efficiently process large amounts of data is necessary in order to effectively scale learning algorithms to match the growth of data available. Here we explore algorithms to keep privacy during knowledge extraction from large dataset.

With the increase in the volume of data, the demand for cluster computing has grow as problem sets become 50 larger and more interest develops in the field. Along with this, time and speed has become the major factor in computing such data. This huge volume of data available needs to be organized and managed so as to facilitate proper understanding. Therefore the need of clustering comes into existence. Clustering analysis has been an emerging research issue in data mining due to its variety of applications. It's expensive use in wide variety of applications, including image processing, computational biology, mobile communication, medicine and economics, has lead to the popularity of this algorithm.

Data mining is an interdisciplinary subfield of computer science. It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.

A big challenge of preserving privacy and security in cloud computing is that developers and users wanted to spend as little effort and system resources on security as possible. Therefore, motivation of this research is how to design a system that satisfies below demands.

1. Enables efficient distributed computations
2. Provides privacy's enhancement to results.
3. Supports a friendly usability that users can write

1.1 Access Control Mechanism

Access Control is primary principle of information security and specifies "Who Can Access What". Implementation of Access Control mechanism will filter out complete user access to framework data and avoid data leakage. It will also help to classify the request and response according to user rights.

To implement the access control, a list of services and users are expected. Access Control Matrix will give relation between users and services. It will classify the all user into categories and also services as same. The complete phenomena will help develop structured security plan to implement privacy protection mechanism using Access Control.

Simpler access control models often cannot adequately meet the complex access control requirements that such relationships require, and so more granular, powerful, dynamic models and mechanisms are needed to address these new realities. In short, increasingly complex data access and sharing requirements drive the need for increasingly complex access control models and mechanisms. It is shown in figure 1.1.[Source: PVM Survey]

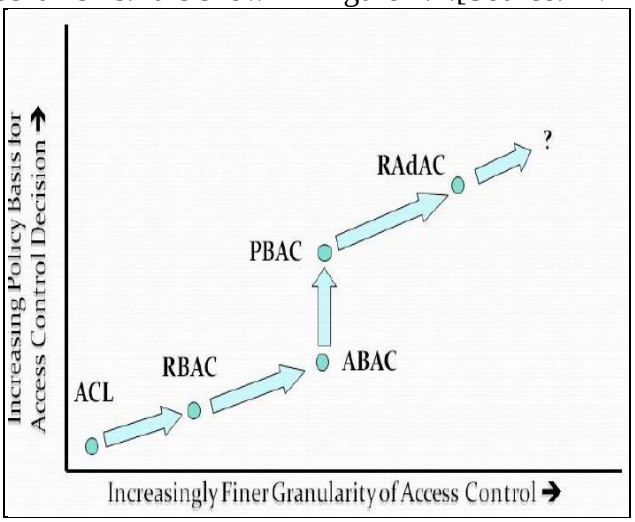

Figure 1.1: Types of Access Control

Access Control Lists (ACLs) are the oldest and most basic form of access control. The concept of an ACL is very simple: each resource on a system to which access should be controlled, referred to as an object, has its own associated list of mappings between the set of entities requesting access to the resource and the set of actions that each entity can take on the resource. For example, each file on a file system might have an associated data structure that holds the list of users that the operating system as a whole recognizes, along with a flag which indicates whether each user may read, write, execute, delete, or modify the file (or some combination of these). Whenever a user tries to perform any of these actions on the file, the operating system checks the file's ACL and determines whether the requested action appending data to the file, for example is allowed. If the action is allowed for that user, the data is appended; if not, the append operation fails.

Role-based Access Control (RBAC) is a newer access control model than the ACL paradigm. Unlike ACLs, access to a resource is determined based on the relationship between the requester and the organization or owner in control of the resource; in other words, the requester's role or function will determine whether access will be granted or denied.

Role-based Access Control addresses some of the shortfalls of the ACL model, while presenting some new and interesting opportunities. For example, one limitation of the ACL model is that it treats every user as a distinct entity with distinct sets of permissions for each resource. This means that ACLs are resource-focused. ACLs have to be set for each resource (or group of resources) separately, a cumbersome process when large groups of resources are involved, or when different people need to be able to access different resources. The fact that ACLs are generally set a by resource's owner and are not always centrally managed only complicates matters, since a fair amount of coordination and planning has to be done to ensure that the correct people have the correct access to the correct resources. In short, the largest single pitfall of the ACL model is that it has limited scalability at the enterprise level.

Attribute Based Access Control (ABAC) is an access control model wherein the access control decisions are made based on a set of characteristics, or attributes, associated with the requester, the environment, and/or the resource itself. Each attribute is a discrete, distinct field that a policy decision point can compare against a set of values to determine whether or not to allow or deny access. The attributes do not necessarily need to be related to each other, and in fact, the attributes that go into making a decision can come from disparate, unrelated sources. They can be as diverse as the date an employee was hired, to the projects on which the employee works, to the location where the employee is stationed, or some combination of the above. One should also note that an employee's role in the organization can serve as one attribute that can be (and often is) used in making an access control decision.

Policy-based Access Control (PBAC) is an emerging model that seeks to help enterprises address the need to implement concrete access controls based on abstract policy and governance requirements. In general, PBAC can be said to be a harmonization and standardization of the ABAC model at an enterprise level in support of specific governance objectives. PBAC combines attributes from the resource, the environment, and the requester with information on the particular set of circumstances under which the access request is made, and uses rule sets that specify whether the access is allowed under organizational policy for those attributes under those circumstances. In an ABAC-only model, the attributes required to gain access to a particular resource are determined on a local level and can vary greatly from one organizational unit to the next.

2. LITERATURE REVIEW

Antorweep Chakravorty, Tomasz Wlodarczyk, Chunming Rong [1] propsoed a solution for reliably concealing privacy and ensuring security for analytics of smart home sensor data. The presented approach maintained the data utility by not transforming the stored data. Rather based on cryptographic techniques, we replace the personal/quasiidentifiers of collected sensor data with hashed values before storing them into a de-identified storage. Separate identifier dictionary storage, with hashed and actual identifier values was also maintained as a point of reference for re-introduction of identifiers. We proposed using heuristic-based k-anonymized algorithms based on the end-users privacy level, requirements and authorization on the identifier dictionary storage. The hashed identifiers from outputs of any data processing job on the de-identified store were replaced with their respective k-anonymized value, thus preserving privacy of any presented/shared results.
Access Control is classified into two categories based on convention and context techniques. A brief review of these techniques is follow as.

A. Conventional Access Control Models

Discretionary Access Control (DAC) permits the granting of access control privileges to be left to the discretion of the individual users [11]. A DAC mechanism allows users to grant access to any of the objects under their control. As such, users are said to be the owners of the objects under their control. ACLs (access control lists) and owner/group/other access control mechanisms are by far the most common mechanism for implementing DAC policies. Mandatory Access Control (MAC) is "A means of restricting access to objects based on the sensitivity (as represented by a label) of the information contained in the objects and the formal authorization (i.e. clearance) of subjects to access information of such sensitivity" [12].

In MAC, access control policy decisions are made by a central authority, not by the individual owner of an object, and the owner cannot change access rights. An example of MAC occurs in military security, where an individual data owner does not decide who has a Top Secret clearance, nor can the owner change the classification of an object from Top Secret to Secret.

Role-based access control (RBAC), access decisions are based on the roles that individual users have as part of an organization [13]. The process of defining roles should be based on a thorough analysis of how an organization operates and should include input from a wide spectrum of users in an organization. Permissions (i.e., performing operations on objects/resources) are defined for roles instead of individual users. Once a user takes a specific role, the role privileges are assigned to the user. Active roles of each user are stored in an associated session. RBAC is said to be the defect industry standard in access control.

B. Context-Aware Access Control

Although RBAC can satisfy access control requirements of most organizations, it is not suitable for context-aware applications where context information supply crucial access control parameters. Hence, a plethora of work, as we discuss next, has attempted to augment RBAC with context information. In GRBAC model [3], context information is considered as the environmental role, which an application needs to posses in order to perform context-dependent tasks. Such a definition leads to large number of roles in an access control system, as there might be potentially many environmental states that are relevant to an application. Gaia [2] defines three different role categories, corresponding to system-wide roles, active space roles, and application roles, and a mapping between them. Context-based constraints that limit a role's visibility to specific geographic areas are presented as part of the GEO-

RBAC model [4]. Similarly, the GTRBAC model [5] provides mechanisms for enabling and disabling of roles based on temporal constraints. The DRIVE authorization and access control approach [6] extended the RBAC model by making role activation function dynamic. They use the notion of context control to differentiate
between activated roles and predefined roles for a particular user. Seon-Ho et al. [7] introduced the notion of context role to a traditional RBAC. The context role represents environment state of the system by mapping context-roles and context information.

CR-RBAC (Context, Rule and Role-Based Access Control) [8] integrated the RBAC model with business rule management. It has both the role-based static character and context-rule based dynamic property. The model mends the weakness of the static management existing in the role based access authority of the RBAC model. The TMAC (Team-based Access Control) model [9] was developed to provide an access control model for collaborative activities accomplished by teams of users. Thus, the TMAC introduces the notion of "teams" as a collection of users in specific roles. Teams collaborate with one another with the objective of accomplishing a specific task or goal. Users are assigned to teams, where each team will encompass the set of roles. Each user's access for team resources depends on his current role and current team. The Context-based Team-based Access Control (CTMAC) [10] extends this idea to allow context information, like time and location, to tailor to fine-grained, flexible and context-based access control policies.

Juntaowang et al [8], developed density-based detection methods based on characteristics of noise data where the discovery and processing steps of the noise data are added to the original algorithm. Preprocessing the data improves the clustering result significantly and the impact of noise data on k-means algorithm is decreased. First a pretreatment is made with the data to be clustered to remove the outliers using outlier detection method based on LOF, so that the outliers cannot participate in the calculation of the initial cluster centers, and excluded the interference of outliers in the search for the next cluster center point. We secondly apply fast global k-means clustering algorithm on the new data set which is generated previously. Fast global k-means clustering algorithm is an improved global k-means clustering algorithm by Aristides Likes

3. PROBLEM DOMAIN

"Privacy is a state in which one is not observed or disturbed by other people" Privacy protection policy is an approach to isolate the sensitive information from unauthorized access. The complete work concludes that MapReduce Framework does not consist proposed security policy and suffering with data leakage problem.

Subsequently, Security threat attack is also possible and malicious framework may give open system access to unauthorized user. Furthermore, Airavat Solution is not efficient solution and does not perform well. The complete phenomena generate a problem to implement security policy with MapReduce Algorithm.

To balance the competing goals of a permissive programming model and the need to prevent information leaks, the untrusted code should be confined. Traditional approaches to data privacy are based on syntactic anonymization, i.e., removal of "personally identifiable information" such as names, addresses, and Social Security numbers. Unfortunately, anonymization does not provide meaningful privacy guarantees. High-visibility privacy fiascoes recently resulted from public releases of anonymized individual data, including AOL search logs and the movie-rating records of Netflix subscribers.

The datasets in question were released to support legitimate data-mining and collaborative-filtering research, but naive anonymization was easy to reverse in many cases.

These events motivate a new approach to protecting data privacy. One of the challenges of bringing security to cloud computing is that users and developers want to spend as little mental effort and system resources on security as possible. Completely novel APIs, even if secure, are unlikely to gain wide acceptance. Therefore, a key research question is how to design a practical system that (1) enables efficient distributed computations, (2) supports a familiar programming model, and (3) provides precise, rigorous privacy and security guarantees to data owners, even when the code performing the computation is untrusted.

The problem in the web mining arises when confidential information is derived from released data by unauthorized users. This problem is commonly known as the "database inference" problem. Recent advances in web mining field have lead to increased concerns about privacy. There are mainly three issues which arise in data mining process are as follows:

3.1 Privacy Issues

The concerns about the personal privacy have been increasing enormously recently especially when internet is booming with social networks, e-commerce, forums, blogs…. Because of privacy issues, people are afraid of their personal information is collected and used in unethical way that potentially causing them a lot of troubles. Businesses collect information about their customers in many ways for understanding their purchasing behaviors trends. However businesses don't last forever, some days they may be acquired by other or gone. At this time the personal information they own probably is sold to other or leak.

3.2 Security issues

Security is a big issue. Businesses own information about their employees and customers including social security number, birthday, payroll and etc. However how properly this information is taken care is still in questions. There have been a lot of cases that hackers accessed and stole big data of customers from big corporation such as Ford Motor Credit Company, Sony… with so much personal and financial information available, the credit card stolen and identity theft become a big problem.

3.3 Misuse of information/inaccurate information

Information is collected through data mining intended for the ethical purposes can be misused. This information may be exploited by unethical people or businesses to take benefits of vulnerable people or discriminate against a group of people.

4. SOLUTION DOMAIN

Data mining opens new threats to privacy and information security if not done or used properly. The main problem is that to hide sensitive information, including personal information, fact or even patterns which are generated by any algorithm of data mining from the others. In order to focusing on privacy preserving association rule mining, the simplistic solution to address the problem of privacy is presented. To overcome the security problem work concludes a need of role based access control model to provide security in Web Database Privacy algorithm for frequent item set mining and association rule learning over transactional databases may use to maintain privacy during mining. The wok concludes that proposed solution will maintain privacy as per access control rules. So wok will be the hybrid solution of data mining with security.Research suggests that tapping the potential of this data can benefit businesses, scientific disciplines and the public sector – contributing to their economic gains as well as development in every sphere. The need is to develop efficient systems that can exploit this potential to the maximum, keeping in mind the current challenges associated with its analysis, structure, scale, timeliness and privacy. There has

been a shift in the architecture of data-processing systems today, from the centralized architecture to the distributed architecture.

Access Control is primary principle of information security and specifies "Who Can Access What". Implementation of Access Control mechanism will filter out complete user access to framework data and avoid data leakage. It will also help to classify the request and response according to user rights.

To implement the access control, a list of services and users are expected. Access Control Matrix will give relation between users and services. It will classify the all user into categories and also services as same. The complete phenomena will help develop structured security plan to implement privacy protection mechanism using Access Control.

The complete solution will implement to avoid security attack is shown in Figure 2.

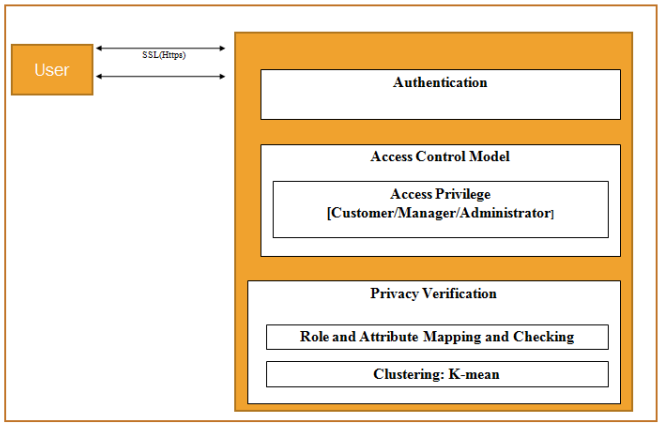

Figure 1.2: Proposed System Architecture

Following salient features will be achieved by implementing the proposed solution.

1. Proposed solution will consider a large Super Market dataset as input to retrieve important information.
2. A K-mean algorithm will be implemented in application for clustering the data and efficiently retrieve credential information.
3. Role Based Access Control Model with Access Matrix will be implements for proper access and rights classification to define "Who can Access What"
4. Security Model with Service Component will be configured to process and execute the complete application in efficient manner.
5. Performance will be measured after searching and information retrieval operation in term of computation & execution time, memory overhead and privacy management

CONCLUSION

The complete work concludes that proposed work will not only suggest a solution to implement access control mechanism with proposed security model but will help to achieve better performance in large data set. A super market dataset will be considered to observe the performance of proposed solution and evaluate computation and memory time period.

REFERENCES:-

1. Antorweep Chakravorty, Tomasz Wlodarczyk, Chunming Rong "*Privacy Preserving Data Analytics for Smart Homes*" published in IEEE Security and Privacy Workshops, 2013

2. R. Sandhu, E. Coyne, et. al., “Role-Based Access Control Models,” IEEE Computer, vol.29, no.2, pp.38-47, Feb. 2003
3. S. Moncrieff; S. Venkatesh, et. al., "Dynamic Privacy in a Smart House Environment," IEEE International Conference on Multimedia and Expo, pp.2034-2037, Jul. 2007
4. R. Bayardo, R. Agrawal, “Data Privacy Through Optimal k- Anonymization,” 21th International Conference on Data Enginnering, pp.217-228, Apr. 2005