Intelligent Automation & Soft Computing
DOI:10.32604/iasc.2020.011731
Article

Tech Science Press

A Pursuit of Sustainable Privacy Protection in Big Data Environment by an Optimized Clustered-Purpose Based Algorithm

Norjihan Binti Abdul Ghani[1], Muneer Ahmad[1], Zahra Mahmoud[1] and Raja Majid Mehmood[2,*]

[1]Department of Information Systems, Faculty of Computer Science & Information Technology, Universiti Malaya, Kuala Lumpur, 50603, Malaysia
[2]Information and Communication Technology Department, School of Electrical and Computer Engineering, Xiamen University Malaysia, Sepang, 43900, Malaysia
*Corresponding Author: Raja Majid Mehmood. Email: rmeex07@ieee.org; rajamajid@xmu.edu.my
Received: 26 May 2020; Accepted: 07 July 2020

Abstract: Achievement of sustainable privacy preservation is mostly very challenging in a resource shared computer environment. This challenge demands a dedicated focus on the exponential growth of big data. Despite the existence of specific privacy preservation policies at the organizational level, still sustainable protection of a user's data at various levels, i.e., data collection, utilization, reuse, and disclosure, etc. have not been implemented to its spirit. For every personal data being collected and used, organizations must ensure that they are complying with their defined obligations. We are proposing a new clustered-purpose based access control for users' sustainable data privacy protection in a big data environment. The clustered-purpose based access control significantly contributes to handling the personal data for stated, unambiguous, and genuine purposes. The proposed algorithm picks specific records from the sample space. It ensures the sustainability and utilization of data for intended purposes by validating the existing privacy tags, assigning new privacy tags based on a clustered-purpose based approach. The proposed method equally ensures the security and sustainable privacy aspects of existing as well as new personal data managed inside large databases repositories. The comparative analysis of significant results presents the outperformance of the proposed algorithm as compared to existing non-purpose based conventional methods of sustainable privacy preservation. The proposed algorithm clusters the large datasets in a big data environment and allows only authorized access to users. The current study is limited to purpose-based access control based on privacy tags. However, future research can also consider other types of privacy protection scenarios in a shared environment.

Keywords: Sustainable privacy; access control; purpose; role attribute; access purpose; intended purpose; clustering data

1 Introduction

We can observe an exponential growth of data in the recent era. The information technology interfaces human-computer interaction in the best possible ways but suffers equally to ensure the privacy and security

of sensitive information of users. Though we can notice very sophisticated systems that collect massive amounts of personal data, store and manage the data accordingly, still optimal preservation of privacy is an open optimization problem at organizational levels. Therefore, ensuring security and privacy has become quite a challenge. As this challenge has yet to be tackled, still quite a few people fear to share their information online or otherwise.

The central concept of security is always protecting the integrity, confidentiality, and availability. With the development of online data sharing and the advancement of information technology, data security became an increasingly important issue. Data are vulnerable to exposure by several factors such as cyber-attack, data combinations, or end-user tracking. Protection on these rouges is possible using technologies that could enhance privacy (PETs) and provide optimized security procedures. Such could also deal with all other kinds of data and privacy protection employing new tools for personal protection of data through offline and online transactions [1].

The amount of data that an organization has helps its management make strategic decisions. Business intelligence talks about the power of data over anything else. Data analytics based on the individual also yields unintended conclusions, for example, the case of a father finding out about his teenage daughter's pregnancy through a personalized promotion directed to the daughter via Amazon data analytics. There are specific security threats involved in the utilization of big data that emerge from public repositories (migration of data to the cloud and its sharing with the public users) [2].

Most organizations have come up with a guideline or regulations about protecting consumer/user's privacy to ease people about sharing their information. The problem arises when the responsible party fails to comply with its rules. This concern is amplified when the data used within an organization and is shared across the platform among their companies. Even worse, data is now available to purchase via certain vendors, for example, Amazon. According to Yang et al. [3], the conventional procedures that grant through authorized access to authorized data suffer since a decrypted access to a patient's medical data hinders the timely treatment and may cause outsourcing the sensitive information.

Restricting access to sensitive data or clustering the specific data based on users' tags can be an effective solution to control access to data centres. It is noticeable that access to sensitive data should be equipped with necessary security requirements in addition to efficient and flexible management, insertion, and retrieval of data. The security and privacy requirements should be implemented through the organizational policies for granting access and control of sensitive users' data [4]. As privacy policies closely relate to the purpose of the data usage compared to the actions performed on the transactions, the conventional access control models are not suitable to be used in achieving privacy protection. Hence, Byun et al. [5] introduced the concept of purpose as an essential component in models implementing access control to protect privacy. The idea of this access control is the use of "purpose" as the basis of access control policy.

Similarly, Byun et al. [6] also proposed similar models based on privacy-related access control as a synonym to Ghani et al. [7–9]. According to Byun et al. [6], traditional access control is not appropriate in ensuring privacy protection due to its focus on the object that performs a particular action on a specific object or transaction. Nevertheless, when users' privacy is the primary consideration, a trustworthy policy is required that could bind the data object with the specific purpose of access. Based on the output of this study, to ensure data privacy, the concept of intent should be considered, and a suitable metadata mechanism should be developed for having consideration of privacy-related access control criteria. Therefore, an approach based on purpose is introduced. The purpose is classified into two types:

- Intended purpose: A policy that recognizes the deliberate type of access of data.
- Access purpose: A policy that recognizes the purpose of data access bound with the intention.

Ethically and professionally, the organizations collecting sensitive data of users should prior inform the users about the purpose and intention for seeking the information. Besides, such organizations should also notify the users in the context of exposing or forwarding this sensitive information to other entities for other purposes. The privacy of users, though, can be ensured in this manner, but mostly the users are not willing to allow organizations to access and spreading sensitive information for specific purposes. In such a model, organizations may lose the chance to seek data from users. Kabir et al. [8] enhanced this model by proposing:

- Allowed intended purpose: Any access to data is permitted for a specific use defined by the data provider
- Prohibited intended purpose: Any access to information is not enabled for any particular purpose specified by the data provider.

2 Related Works

This section discusses the previous works in data clustering and purpose-based access control. Tab. 1 shows the following existing data clustering techniques.

Table 1: Related works of data clustering

Technique	By	Method
CURE (Clustering using representative)	[10]	• The individual clusters can be demonstrated with the data points that are supposed to be encapsulated within the domain of clusters. These data points could shrink towards the center to help to reduce the distance between points and within the cluster space also.
k-Mode algorithm	[11]	• A well-known partition clustering algorithm. • Works by employing a mode of data points under consideration • Tries to reduce the cost function similar to other clustering algorithms • Robust to deal with outliers and works fine for numerical attributes of data.
ROCK (Robust Clustering using Links)	[12]	• Belongs to the domain of agglomerative clustering algorithms. • Similar to other agglomerative approaches, it employs the links strategy for quantifying the similarity. • Scalability depends on the sample size
k-Histogram	[13]	• Suitable for categorical data and is considered as an extension of k-means • Dynamic updates the clustering process and works at the histogram concept that should be used in place of mean concept.
DBSCAN	[14]	• Famous clustering algorithm base on the density of data points within the domain and suppresses the noise (outliers in data).
Fuzzy rule-based clustering algorithm	[15]	• Unsupervised clustering is achieved by employing supervised classification approaches. • Fuzzy rules are exploited to identify the essential clusters in data space.

(Continued)

Table 1 (continued).

Technique	By	Method
Squeezer	[16]	• It deals with categorical data in contrast to numerical data. • It comprises of two types of data structures in its implementation. • Produces high-quality cluster result and good scalability
Herd clustering	[17]	• Inspired by the human mobility pattern and the herd behavior from the real world. • Clusters are formed by the moving particles, which are represented by the data instances.

We can notice several studies aiming at protecting and preserving the privacy of users employing the concept of "purpose" for seeking an access-control related to a particular policy.

Table 2: Summarized previous works for PBAC

Technique	By	Method
Platform for privacy preferences (P3P)	[18]	• In this way, the website can encode the data in a specific format called P3P and ensures the preservation of users' information accessible to legitimate people only.
Hippocratic databases	[19]	• These databases contain specific policies and authorization access patterns/ways to seek sensitive information of users for particular purposes.
Strawman	[20]	• It also proposes a purpose-based access control aligned with specific access policies.
Hippocratic databases	[21]	• It also proposes a method of implementing a privacy policy in Hippocratic databases. • It emphasizes that access and exposure of data is granted only to legitimate entities and enlists the purpose of accessing sensitive users' data. • The proposed method introduces models based on granular level limited access and disclosure to users' data and implements the ideas employing the query modification method.
Granular level access control model	[22]	• It introduced a new notion of validity, conditional validity.
	[20]	• Proposes and implements the access control mechanisms at the granular level by consideration of concepts of transformation from RDBMS to privacy preservation levels.

Table 2 (continued).

Technique	By	Method
Purpose-based access control	[23]	• It employs VDM to ensure privacy preservation through sophisticated mechanisms. • The model defines and implements the entities listed in the PBAC aligned with the corresponding privacy preservation specifications.
	[5–6]	• Proposes a model that ensures the privacy protection of users. • The model entities correspond to the policies highlighted for purpose-based access to data. • Since the approach reflects the purpose of accessing and disclosure of data so it is considered to contribute in this direction.
Enterprise privacy authorization language (EPAL)	[24]	• Byun et al. [6] • IBM develops a language that aids in describing the privacy policies at the enterprise level. • The policies are listed in hierarchies reflecting the data-categories associated with specific purposes of data access. • The implementations of concepts aid with actions and obligations as defined in the policy set.
User authentication and data authorization	[7]	• Proposes a model that ensures user authentication and data authorization for safer access to users' data. • Implements the authorization policies for purpose-based access and disclosure of data.
Attribute level access control aligned with the purpose-based privacy policy	[25]	• Proposes a model that considers the attribute-level access control and ensures the purpose-based access to sensitive data.

Tab. 2 provides a summarization of models noted in the literature that are built at a purpose based grant of access and disclosure control to users' sensitive data. In recent years, the organization is paying much attention to access and disclose the users' data with a purpose-based access. It is also noted that users are much concerned while allowing specific data access controllers to grant and disclose data for specific purposes [26]. Thus, it is crucial to consider these two aspects related to the quality of data and the privacy of data while achieving the purpose-based access to users' data. The new models should encapsulate and implement the two concepts to their spirits [5,7,27–29].

3 Methodology

An achievement of sustainable privacy preservation is mostly very challenging in a resource shared computer environment. This challenge demands a dedicated focus on the exponential growth of big data. Despite the existence of specific privacy preservation policies at the organizational level, still the

protection of a user's data at various levels, i.e., data collection, utilization, reuse, and disclosure, etc. have not been implemented to its spirit. For every personal data being collected and used, organizations must ensure that they are complying with their defined obligations. We are proposing a new clustered-purpose based access control for users' sustainable data privacy protection in a big data environment. The clustered-purpose based access control significantly contributes to handling the personal data for stated, unambiguous, and genuine purposes.

The general architecture of the proposed purpose-based access model is shown in Fig. 1. Users' sensitive data is commonly managed by organizational servers manipulating data employing either local or cloud equipped resources. The organization mostly protects the credentials and sensitive data using many security and privacy tools. In a broader context, the organizational policymakers prepare and implement the data privacy of its stakeholders.

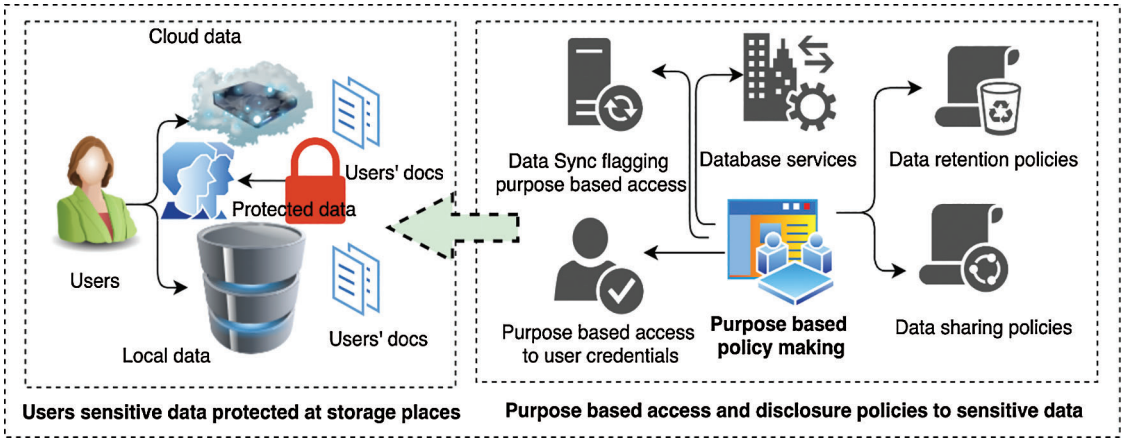

Figure 1: The general architecture of the proposed purpose-based access model

Contrary to conventional data access, archive, retention, and sharing policies, the proposed architecture incorporates the essential aspect "the sustainable purpose of access," ensuring that purpose-based data access and disclosure as a core component. The purpose-based access confirms the intentions of proper and appropriate usages of data for specifically defined purposes. It has been keenly noticed that satisfaction level, agreement, and trust of users towards purpose-based access of data authenticates the implementation of this architecture as compared to existing conventional data access architectures.

Data clustering plays an essential role in data mining due to its ability to work on a large amount of data [30–37]. The literature cites several existing clustering algorithms, namely hierarchical clustering, DBSCAN, k-means, and k-medoid algorithms work under different scenarios. Although the clustering has widely been applied for clustering documents, very few citations could be noticed for purpose-based clustered access and disclosure to users' sensitive data. We have implemented this architecture by proposing clustered purpose-based data access.

Fig. 2 depicts the schematic flow of clustered proposed purpose-based access control. The steps of algorithms are,

1. Pick a record vector $\boldsymbol{R}$ of sample space $\boldsymbol{N}$ that contains $\boldsymbol{R_i}$ sub-vectors such that $i \leq N$, let us assume that these sub-vectors are represented as $\boldsymbol{R_i} = (r_{1i}, r_{2i}, r_{3i})$, where $i \in \{1, 2, 3, N\}$.
2. Define purpose-based tags $\boldsymbol{T}_g = (t_{1g}, t_{2g}, t_{3g}, t_{4g})$, where $g \in$ {corresponding set of organizational policies determined by purpose-based policymakers according to users' sensitive data}.

3. Identify documents based on $\boldsymbol{T}_g$ represented by vectors $\mathbf{D}_k = (d_{1k}, d_{2k}, d_{3k})$, where $k \in \{1, 2, 3, 4\}$ (access levels in the hierarchy of access granted for purpose-based access).
4. Compute the similarity between documents contained in the vector $\mathbf{D}_k$ given by the Manhattan similarity metric:
$$S(\boldsymbol{D_k}) = \sum_{t=1}^{n_{ij}} \left|\boldsymbol{D_{ti}} - \boldsymbol{D_{tj}}\right|, \ (i \neq j) \tag{1}$$
5. Establish a similarity matrix based on Step (4) by assigning each document to the cluster that has the closest similarity as defined in (2).
6. Output the similarities of $\boldsymbol{S}m$ documents as cluster sets such that $1 < m < N$ as ruled per (2) and (3).

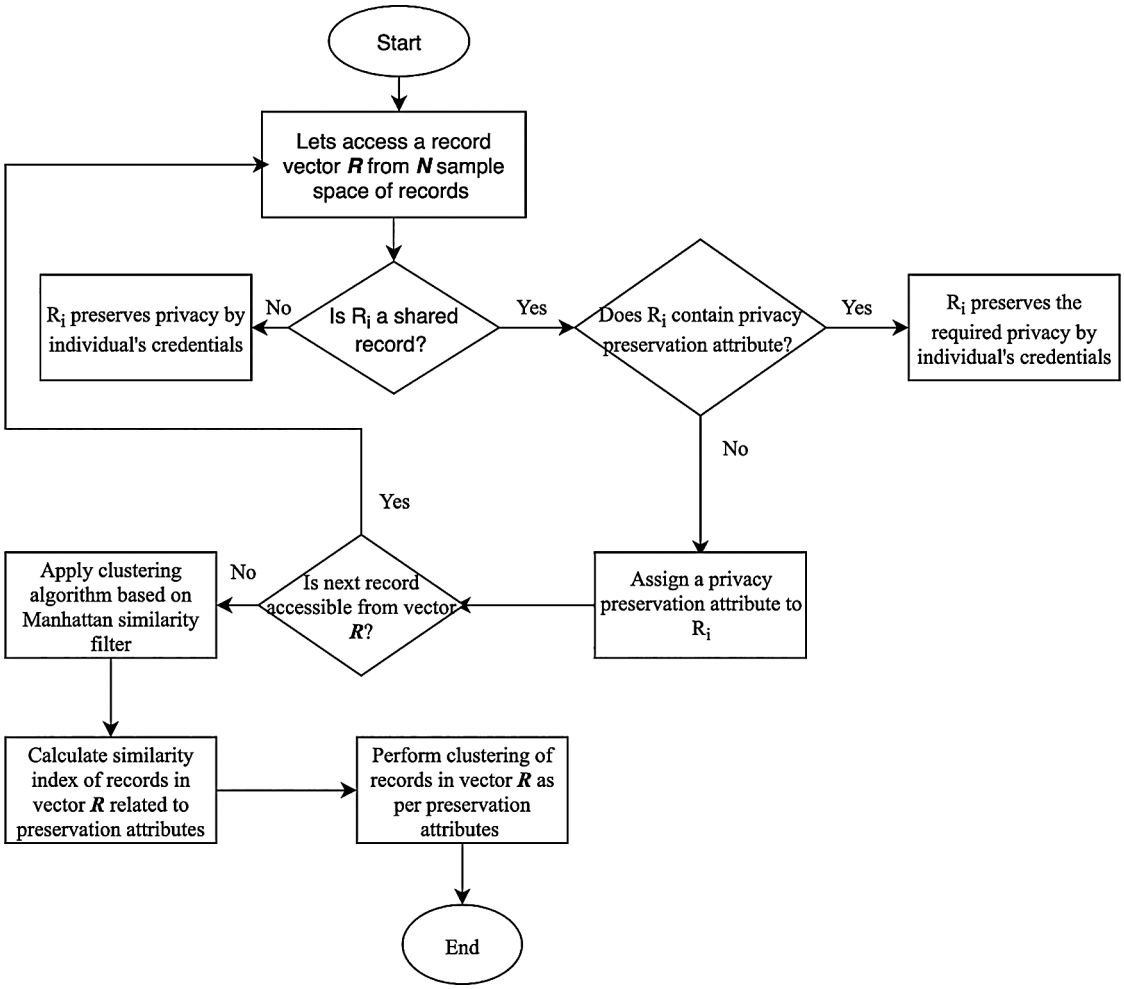

Figure 2: Schematic representation of the proposed purpose-based access model

The process starts by selecting a vector of users' records randomly from a sample space repository (with the concept of non-duplicate records for the next fetch of records from sample space). We devise a filter that validates the existence or non-existence of a purpose-based access tag of individual records. The records with the non-existence of purpose-based access tags are then assigned the tags defined by the organization in a purpose-based access policy. Once tags are assigned, we select a seed to start building a cluster, subsequently by selecting and adding more records to the cluster such that the record added incurs the

least information loss within the cluster. The algorithm determines the clusters having a proximity relationship with the neighboring clusters based on the similarity index score. The "purpose aware" semantic similarity identification is achieved through employing the Manhattan similarity index.

Generally, looking for records that are not directly next to the first cluster will result in a longer wait compared to looking for the closest record to build the second cluster since we need to find a degree of similarity that satisfies the purpose-based policy criteria also. Therefore, the distance of the next record is based on the distance function that can be determined and changed by the system administrator as per change or update in a purpose-based access policy.

It is viable that with the addition of an outlier in a cluster, the information loss ratio increases since the outliers occur in data samples regardless of the similarity. The records are now stored in the database along with the tag of the cluster. The amount of data that can be accessed by the user would depend on their role or the purpose of their search. For example, in a situation, an entity accessing the database would not have a need to access the entire database. Instead, the cluster of matching tags (with the notion of purpose-based access) will be reachable. This will enhance the privacy preservation of users' sensitive data for illegitimate access.

The methodology ensures that restricting access to users' sensitive data for specific data access is based on users' tags so that to provide an effective solution to control access to data centres. Besides, it is taken care that access to sensitive data should be equipped with necessary security requirements in addition to efficient and flexible management, insertion, and retrieval of data. The security and privacy requirements are implemented through the organizational policies for granting access and control of sensitive users' data. As privacy policies closely relate to the purpose of the data usage compared to the actions performed on the transactions, the conventional access control models are not suitable to be used in achieving privacy protection. Hence, the concept of purpose as an essential component in this proposed model for implementing access control to protect privacy.

3.1 Significance of the Proposed Methodology

There are different access control mechanisms in the cloud environment, e.g., discretionary access control, mandatory access control, and role-based access control mechanisms [37–38]. Based on the particular organization needs and cloud environment, the access control policy opts. The national institute of standards and technology (NIST) is considered as an institution that ensures that organizations are adopting the standard procedures for the secure execution of their operations [39]. As per the principle of NIST, the least privilege grants granular access to users according to defined attribute-based policy.

Fig. 3 describes an example of role-based access control in an organization. The diagram is an example drawn using Microsoft draw.io web application [40]. The role-based access control is defined in a static manner as shown in once the organization finalizes its policy. Contrary to this type of access control, we have proposed a purpose-based access control that clusters and grants access to the purpose-based data in a dynamically semantic manner that purpose-based access is customized when the organizational policy is changed. Similar research has also been carried out by Lo et al. [41] that grants role-based access to users in a cloud environment.

Contrary to the defined conventional access control mechanisms, the proposed purpose-based access control signifies these essential objectives,

1. Clusters the purpose-based users' data as per defined attributes
2. Dynamic purpose-based access control as per change in organizational policy
3. Authorizes access to users according to the purpose-based clusters where the user's authorization exists.

Fig. 4 glimpses the mechanism that clusters the organizational data on semantic understanding of corporate policies and attributes defined for the purpose-based access.

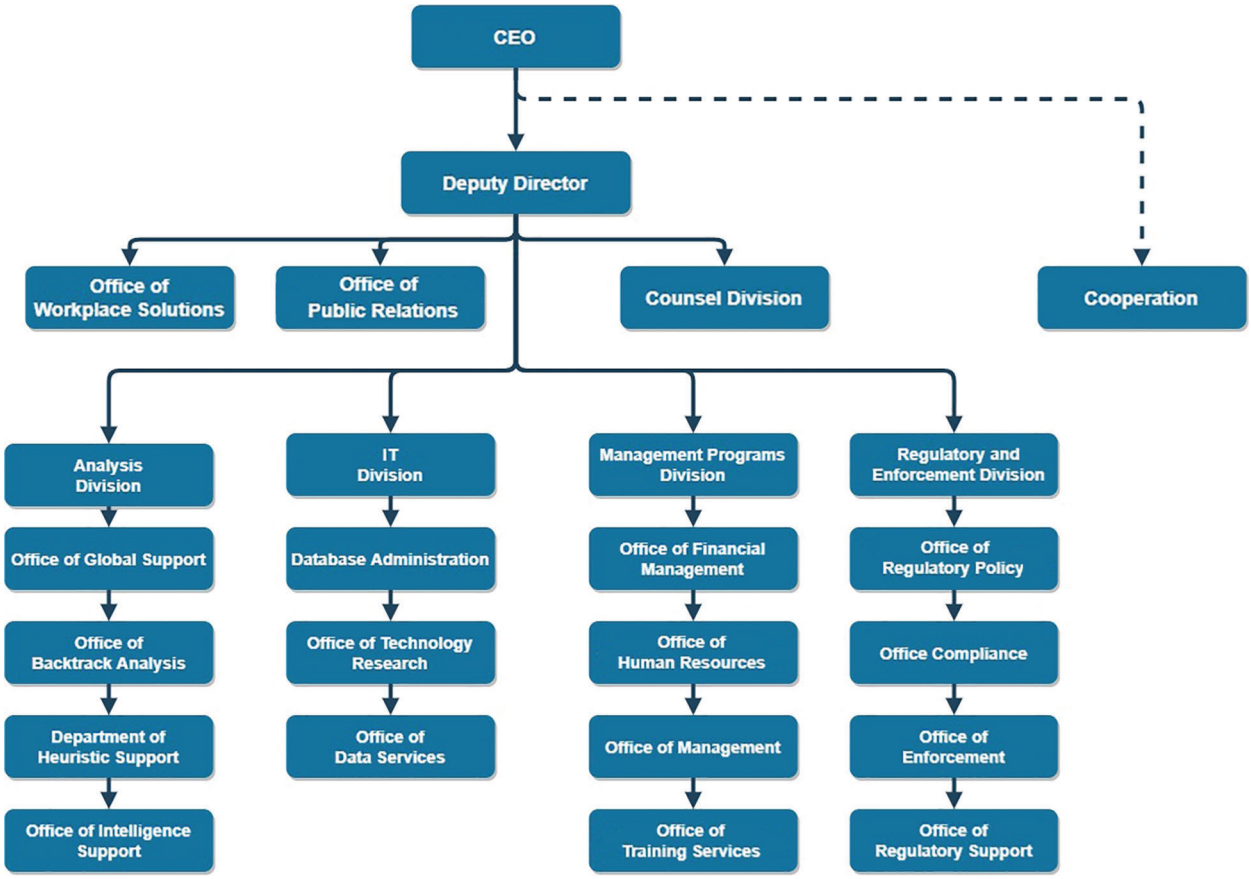

Figure 3: Example of role base hierarchy of employees in an organization

For instance, an organization comprises of users of different departments whose data is managed and controlled by a cloud environment. The organizational policies change from time to time to reflect the regulatory plan and users' requirements for data access. Let's assume that the users' data is randomly scattered, and there has been no semantic understanding assigned to access control to information. At one instance of time, the organization defines a defined policy on how to control the access given to the data. The proposed mechanism clusters the data as per organizational attributes given to users according to access policies. Later at some other instance of time, if the regulatory policy changes, the purpose-based access control is also customized as per needs.

4 Results and Discussion

The performance of the proposed clustered purpose-based access algorithm was evaluated with a non-purpose based scenario employing different sets of data. For simulation, we considered six datasets generated from Wisconsin Benchmark datasets [42] and four datasets from UCI machine repository datasets [43]. The simulated performance of the proposed algorithm was measured with the Wisconsin datasets, while UCI datasets were used for performance validation. Tab. 3 describes the datasets employed for comparative analysis between purpose-based access control and non-purpose based mechanisms.

The goal of these experiments was to investigate the performance of algorithms. The datasets were analyzed using python 3.0 with a Jupiter notebook. From the datasets, we have created two scenarios, i.e., clustered purpose-based access to users' records and non-purpose based access. We present here an

example that describes the purpose tree and its implementation using metadata structure for purpose-based access control to data.

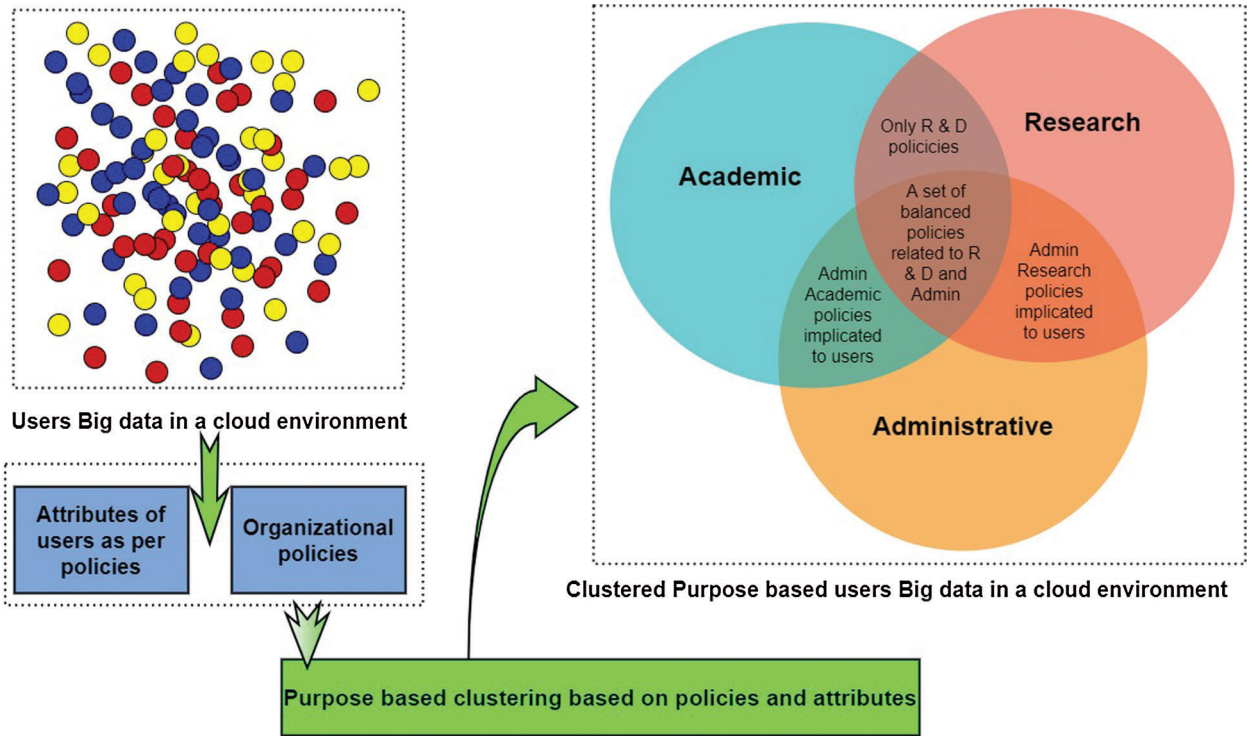

Figure 4: Purpose based clustering and access control based on defined attributes and policies

Table 3: Description of datasets for performance evaluation of algorithms

Repository	Dataset	Numeric attributes	String attributes	Customized purpose-based attributes	Length (number of instances)
Wisconsin	1	2	5	5	1200
	2	3	7	4	1700
	3	3	6	5	1450
	4	2	4	4	1700
	5	3	4	5	1300
	6	4	4	4	1400
UCI–7	7	5	7	3	2900
UCI–8	8	3	5	4	500
UCI–9	9	3	4	5	1500
UCI–10	10	3	3	4	2000

Fig. 5 describes a typical organizational structure that can be represented as a purpose tree for manipulation of purpose-based access control to users' data. In terms of metadata structure, we can describe it as,

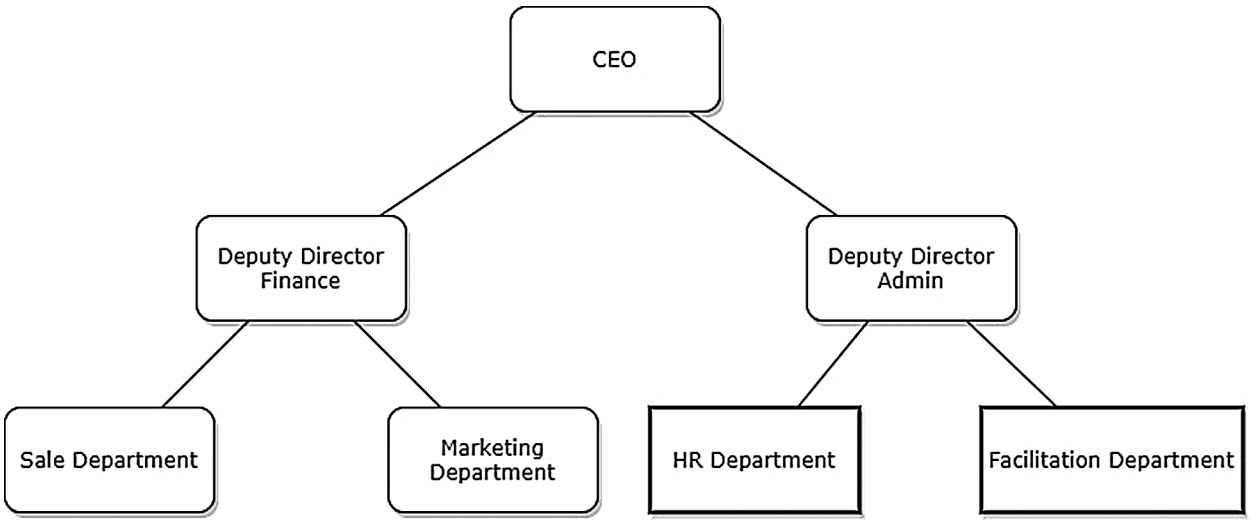

Figure 5: Purpose tree of an organization

For instance, Tab. 4 describes the metadata structure of the process tree defined in Fig. 5. A process ID represents each node in the process tree. Parent nodes of the process contain the reference of their children. We further assign a purpose-based access control ID for each node that later describes the access level of a particular node in the process tree, e.g., Process P_01 has access control to processes PBCA_01 to PBCA_07. This policy of purpose-based access control is governed by an individual organization and is not generally applicable to every scenario. With changing the organizational plan, the access control can either dynamically or manually be changed. Based on policy labels/tags, the purpose-oriented clustering algorithm clusters the data using semantic consideration of labels as a distance measure between data nodes.

Table 4: Metadata structure of process tree

Process ID	Process Name	Parent of process	Purpose based access control ID	Purpose based access level
P_01	CEO	None	PBAC_01	L_01 = {PBAC_01 to PBAC_07}
P_02	DDF	CEO	PBAC_02	L_02 = {PBAC_02, PBAC_04, PBCA_05}
P_03	DDA	CEO	PBAC_03	L_02 = {PBAC_03, PBAC_06, PBCA_07}
P_04	SD	DDF	PBAC_04	L_03 = {PBAC_04}
P_05	MD	DDF	PBAC_05	L_03 = {PBAC_05}
P_06	HRD	DDA	PBAC_06	L_03 = {PBAC_06}
P_07	FG	DDA	PBAC_07	L_03 = {PBAC_07}

The performance evaluation of purpose-based access control is measured in terms of the query control mechanism [5,6,9]. We investigate here the number of records and the time (in seconds) required by the query in fetching the desired data against purpose-based and non-purpose based access mechanisms. Close observation at simulations statistics, we noticed that the proposed purpose-based access algorithm carries only the intended records from the sample space of records. At the same time, the existing non-purpose based scenario brought all the records.

Fig. 6 presents the comparative analysis of access scenarios in terms of numbers of records fetched by the two approaches. It is evident that the proposed CBPA carries only intended records from the users' records space and thus reduces the space complexity involved in seeking all the records.

Similarly, we also observed the notable performance achievement of the proposed algorithm in terms of seek time for accessing a purpose-based number of records. We can vet that the proposed algorithm outperforms in reducing the time complexity involved in fetching the users' records.

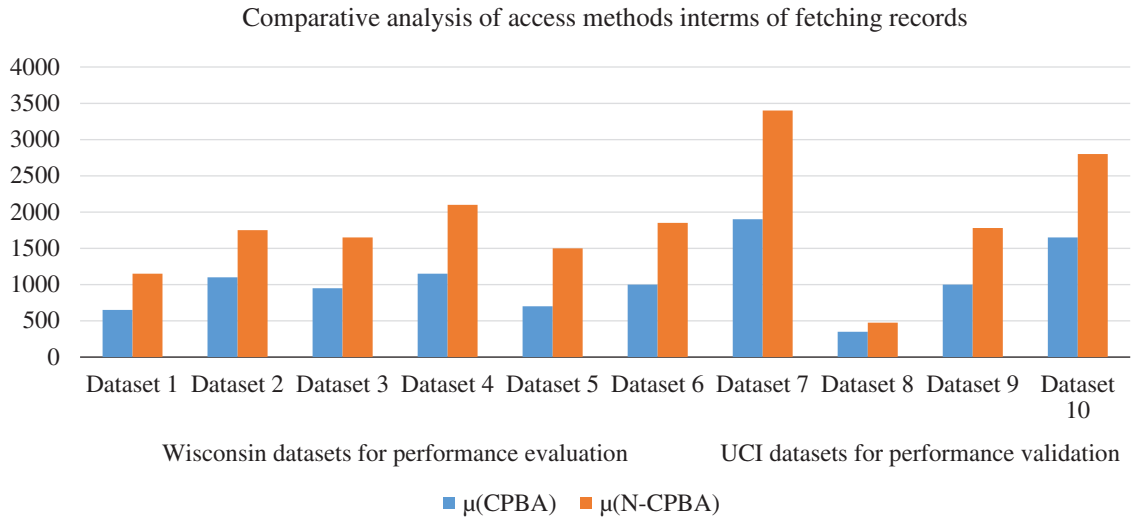

Figure 6: Comparative analysis of different access scenarios

Fig. 7 depicts the comparative analysis of two approaches in terms of seek time. We can observe that non-purpose based access takes comparatively longer access time as compared to purpose-based access. Hence, the time complexity involved in seeking users' data is significantly reduced with the proposed CBPA algorithm.

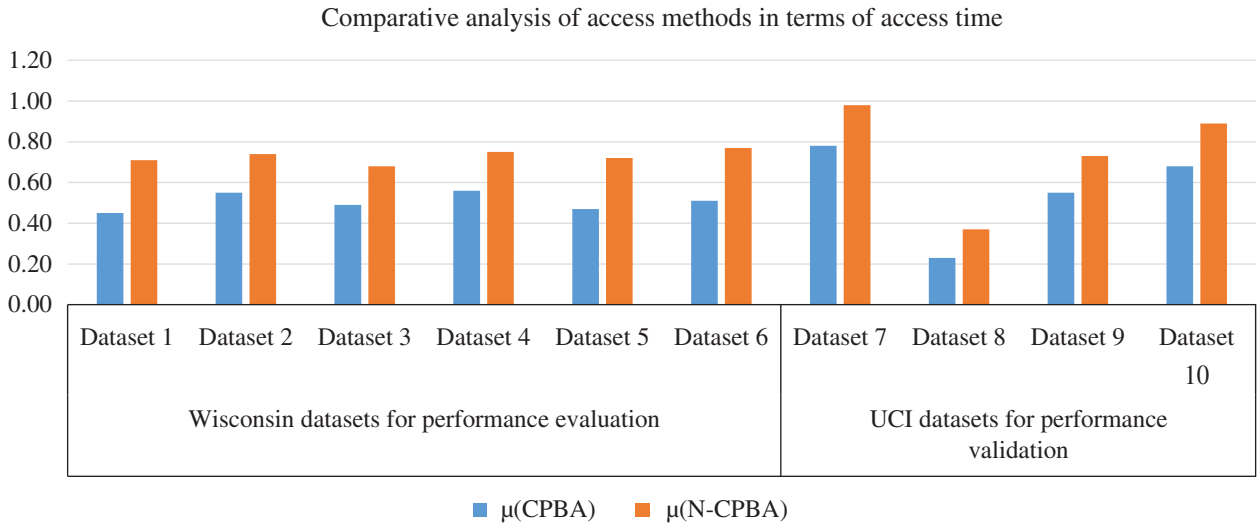

Figure 7: Comparative analysis in terms of seek time

4.1 Pursuit of Privacy Protection in Big Data Environment

The proposed algorithm ensures the sustainable privacy preservation to users' sensitive data for stated, unambiguous, and genuine purposes. The sustainability is achieved by validating the existing privacy tags and assigns new sustainable privacy tags based on non-privacy preserved data aiming clustered-purpose based approach. In this way, the proposed method equally ensures the security and sustainable privacy aspects of existing as well as new personal data managed inside large databases repositories.

5 Conclusion

Sustainable privacy preservation (especially in a shared computer environment) is quite challenging and requires careful access to users' sensitive data. This paper presented a new clustered-purpose based access control for users' sustainable data privacy protection in a big data environment. The clustered-purpose based access control significantly contributed to handle the personal data for stated, unambiguous, and genuine purposes. The proposed algorithm clusters and seeks access to users' records by validating the existing privacy tags and assigns new privacy tags based on non-privacy preserved data aiming clustered-purpose based approach. In this way, the proposed method equally ensures the security and privacy aspects of existing as well as new personal data managed inside large databases repositories. The comparative analysis of results reveals the outperformance of our cluster-purpose based access algorithm as compared to conventional non-purpose based access algorithms towards sustainable privacy presentation to users' sensitive records. The current research study assumes that the organizations have defined access policies that serve as inputs to the proposed model to cluster the data based on purpose-based tagging and access. The study is also limited to purpose-based access control based on privacy tags. However, future research can also consider other types of privacy protection scenarios in a shared environment.

Acknowledgement: Conceptualization, Norjihan Abdul Ghani; Data curation, Zahra Mahmoud and Raja Majid Mehmood; Formal analysis, Norjihan Abdul Ghani and Zahra Mahmoud; Funding acquisition, Raja Majid Mehmood; Methodology, Muneer Ahmad; Project administration, Norjihan Abdul Ghani; Resources, Raja Majid Mehmood; Writing–original draft, Zahra Mahmoud; Writing – review & editing, Muneer Ahmad.

Funding Statement: This work was supported by the Xiamen University Malaysia Research Fund (XMUMRF) (Grant No. XMUMRF/2019–C3/IECE/0007) and Fundamental Research Grant Scheme [FP072–2015A] funded by the Ministry of Education, Malaysia.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] The European Union Agency for Cybersecurity, "Cybersecurity," 2020. [Online]. Available: https://www.enisa.europa.eu/topics/data-protection.

[2] M. V. Rijmenam, *Think Bigger: Developing a Successful Big Data Strategy for Your Business*, 1st ed., Sydney, Australia: AMACOM, 2014. [Online]. Available: https://www.amazon.com/Think-Bigger-Developing-Successful-Strategy/dp/0814434150.

[3] Y. Yang, X. Zheng, W. Guo, X. Liu and V. Chang, "Privacy-preserving smart IoT-based healthcare big data storage and self-adaptive access control system," *Information Sciences*, vol. 479, pp. 567–559, 2019.

[4] S. Fugkeaw and H. Sato, "Scalable and secure access control policy update for outsourced big data," *Future Generation Computer Systems*, vol. 79, no. 1, pp. 364–373, 2018.

[5] J. Byun, E. Bertino and N. Li, "Purpose based access control of complex data for privacy protection," in *Sym. on Access Control Model and Technologies*, Stockholm, Sweden, pp. 102–110, 2005.

[6] L. Byun and N. Li, "Purpose based access control for privacy protection in relational database systems," *VLDB Journal*, vol. 17, no. 4, pp. 603–619, 2008.

[7] N. A. Ghani, H. Selamat and Z. M. Sidek, "Credential purpose-based access control for personal data protection," *Journal of Web Engineering*, vol. 14, no. 3, pp. 346–360, 2015.

[8] M. E. Kabir and H. Wang, "Conditional purpose based access control model for privacy protection," in *Proc. the Twentieth Australasian Conf. on Australasian Database*, Darlinghurst, Australia, pp. 135–142, 2009.

[9] H. Wang, L. Sun and E. Bertino, "Building access control policy model for privacy preserving and testing policy conflicting problems," *Journal of Computer and Systems Sciences*, vol. 80, no. 8, pp. 1493–1503, 2014.

[10] S. Guha, R. Rastogi and K. Shim, "CURE: An efficient clustering algorithm for large databases," *Information Systems*, vol. 26, no. 1, pp. 73–84, 2001.

[11] Z. Huang, "Extensions to the k-Means algorithm for clustering large data sets with categorical values," *Data Mining and Knowledge Discovery*, vol. 2, no. 3, pp. 283–304, 1998.

[12] S. Guha, R. Rastogi and K. Shim, "Rock: A robust clustering algorithm for categorical attributes," *Information Systems*, vol. 25, no. 5, pp. 345–366, 2000.

[13] H. Zengyou, X. Xiaofei, D. Shengchun and D. Bin, "K-histograms: An efficient clustering algorithm for categorical dataset," *arXiv*, *cs/0509033*, 2005.

[14] H. Zengyou, X. Xiaofei and D. Shengchun, "K-ANMI: A mutual information based clustering algorithm for categorical data," *Information Fusion*, vol. 9, no. 2, pp. 223–233, 2008.

[15] E. G. Mansoori, "FRBC: A fuzzy rule-based clustering algorithm," *IEEE Transactions on Fuzzy Systems*, vol. 19, no. 5, pp. 960–971, 2011.

[16] S. Elavarasi, J. Anitha and J. Akilandeswari, "Survey on clustering algorithm and similarity measure for categorical data," *ICTACT Journal on Soft Computing*, vol. 4, no. 2, pp. 715–722, 2014.

[17] K. C. Wong, C. Peng, Y. Li and T. M. Chan, "Herd clustering: A synergistic data clustering approach using collective intelligence," *Applied Soft Computing*, vol. 23, pp. 61–75, 2014.

[18] World Wide Web Consortium (W3C), "Platform for privacy preferences (P3P)," 2016. [Online]. Available: www.w3.org/P3P.

[19] R. Agrawal, J. Kiernan, R. Srikant and Y. Xu, "Hippocratic databases," in *28th Int. Conf. on Very Large Databases (VLDB)*, *Hong Kong, China*, pp. 143–154, 2002.

[20] R. Agrawal, P. Bird, T. Grandison, J. Kiernan, S. Logan *et al.,* "Extending relational database systems to automatically enforce privacy policies," in *21st Int. Conf. on Data Engineering*, Tokyo, Japan, pp. 1013–1022, 2005.

[21] T. Grandison, C. Johnson and J. Kiernan, "Hippocratic databases: Current capabilities and future trends, " in *Handbook of Database Security*, 2nd ed., Davis, USA: Springer, pp. 409–429, 2008.

[22] S. Rizvi, A. O. Mendelzon, S. Sudarshan and P. Roy, "Extending query rewriting techniques for fine-grained access control," in *Int. Conf. on Management of Data and Sym. on Principles Database and Systems*, Paris, France, pp. 551–562, 2004.

[23] N. Yang, H. Barringer and N. Zhang, "A purpose-based access control model," in *Third Int. Sym. on Information Assurance and Security*, New York, USA, pp. 143–148, 2007.

[24] World Wide Web Consortium (W3C), "The enterprise privacy authorization language (EPAL)," 2003. [Online]. Available: www.zurich.ibm.com/security/enterprise–privacy/epal.

[25] M. Amini and F. Osanloo, "Purpose-based privacy preserving access control for secure service provision and composition," *IEEE Transactions on Services Computing*, vol. 12, no. 4, pp. 604–620, 2019.

[26] A. Gregory, "Data governance—protecting and unleashing the value of your customer data assets," *Journal of Direct, Data and Digital Marketing Practice*, vol. 12, no. 3, pp. 230–348, 2011.

[27] OASIS, "Core and hierarchical role based access control (RBAC)," 2014. [Online]. Available: http://www.oasis-open.org.

[28] OASIS, "XACML-3.0-RBAC, Extensible access control markup language 2.0," 2014. [Online]. Available: http://www.oasis-open.org.

[29] OASIS, "XACML-3.0-RBAC, Privacy policy profile of xacml v2.0," 2014. [Online]. Available: http://www.oasis-open.org.

[30] H. Gulati and P. Singh, "Clustering techniques in data mining: A comparison," in *2nd Int. Conf. on Computing for Sustainable Global Development (INDIACom)*, New Delhi, India, pp. 410–415, 2015.

[31] W. T. Zhao, P. Li, C. Z. Zhu, D. Liu and X. Liu, "Defense against poisoning attack via evaluating training samples using multiple spectral clustering aggregation method," *Computers, Materials & Continua*, vol. 59, no. 3, pp. 817–832, 2019.

[32] L. L. Zhou, F. Tan, F. Yu and W. Liu, "Cluster synchronization of two-layer nonlinearly coupled multiplex networks with multi-links and time-delays," *Neurocomputing*, vol. 359, pp. 264–275, 2019.

[33] S. S. Li, T. J. Cui and J. Liu, "Research on the clustering analysis and similarity in factor space," *Computer Systems Science and Engineering*, vol. 33, no. 5, pp. 397–404, 2018.

[34] K. Gu, L. H. Yang and B. Yin, "Location data record privacy protection based on differential privacy mechanism," *Information Technology and Control*, vol. 47, no. 4, pp. 639–654, 2018.

[35] C. Y. Yin, X. K. Ju, Z. C. Yin and J. Wang, "Location recommendation privacy protection method based on location sensitivity division," *Journal on Wireless Communications and Networking*, vol. 2019, no. 1, pp. 566, 2019.

[36] C. Y. Yin, L. F. Shi, R. X. Sun and J. Wang, "Improved collaborative filtering recommendation algorithm based on differential privacy protection," *Journal of Supercomputing*, vol. 76, no. 7, pp. 5161–5174, 2019.

[37] C. Y. Yin, B. Zhou, Z. C. Yin and J. Wang, "Local privacy protection classification based on human-centric computing," *Human-Centric Computing and Information Sciences*, vol. 9, no. 1, pp. 33, 2019.

[38] K. Punithasurya and S. Jebapriya, "Analysis of different access control mechanism in cloud," *International Journal of Applied Information Systems*, vol. 4, no. 2, pp. 34–39, 2012.

[39] A. R. Khan, "Access control in cloud computing environment," *ARPN Journal of Engineering and Applied Sciences*, vol. 7, no. 5, pp. 613–615, 2012.

[40] Microsoft Drawing Application, "Draw.io application," 2020. [Online]. Available: https://www.microsoft.com/en-my/p/drawio-diagrams/9mvvszk43qqw?activetab=pivot: overviewtab.

[41] N. W. Lo, T. C. Yang and M. H. Guo, "An attribute-role based access control mechanism for multi-tenancy cloud environment," *Wireless Personal Communications*, vol. 84, no. 3, pp. 2119–2134, 2015.

[42] D. Bitton, J. D. DeWitt and C. Turbyfill, "Benchmarking database systems a systematic approach," in *Proc. the 9th Int. Conf. on Very Large Data Bases*, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 8–19, 1983.

[43] UCI Machine Learning Repository, "Center for machine learning," 2020. [Online]. Available: https://archive.ics.uci.edu/ml/index.php.