Hamming Distance Approach to Reduce Role Mining Scalability

Nazirah Abd Hamid[1], Siti Rahayu Selamat[2], Rabiah Ahmad[3], Mumtazimah Mohamad[4]
Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM), Melaka, Malaysia[1, 2, 3]
Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin (UniSZA), Terengganu, Malaysia[1, 4]

***Abstract*—Role-based Access Control has become the standard of practice for many organizations for restricting control on limited resources in complicated infrastructures or systems. The main objective of the role mining development is to define appropriate roles that can be applied to the specified security access policies. However, the mining scales in this kind of setting are extensive and can cause a huge load on the management of the systems. To resolve the above mentioned problems, this paper proposes a model that implements Hamming Distance approach by rearranging the existing matrix as the input data to overcome the scalability problem. The findings of the model show that the generated file size of all datasets substantially have been reduced compared to the original datasets It has also shown that Hamming Distance technique can successfully reduce the mining scale of datasets ranging between 30% and 47% and produce better candidate roles.**

Keywords—Role-based Access Control; role mining; hamming distance; data mining

I. Introduction

Role mining techniques exploit the existing user-permission assignment (UPA) to define roles that are suitable to the policies of an organization. The UPA usually involves big scale of data, which usually makes role mining difficult to process. The input data should contain at least a set of users (U), permissions (P), and user permission assignment relation (UPA) which is commonly depicted in a Boolean matrix form. A number of studies, such as in [3], [4], [5] and [6], have found that pre-processing phases are significant to be implemented, especially to simplify the scalability of the UPA matrices that work as the input to the role mining process. Furthermore, the authors also have discussed that pre-processing steps can be divided into two major classes: data cleansing and data selection. Data cleansing can be defined as a method to decrease the noise that resides in the UPA matrices, while data preparation ensures that the input data is suitable to be executed by role mining algorithms.

From the perspective of RBAC, when an organization's existing UPA includes numerous permissions and users, or when the size of UPA becomes excessively large, it increases the likelihood of the UPA to contain overlapping permissions and an unnecessary number of roles. To address these challenges and to uncover roles that are of high quality and meaningful, data mining techniques, specifically clustering techniques, are required. Eventually, it can produce reliable and optimal candidate roles that would be forwarded to the next stage or phase. In addition, the migration from traditional ACM to RBAC model usually would not result in the best RBAC states either they are too complex or not scalable enough to be passed on the next phase.

Based on the above discussion, there is a main question from RBAC's standpoint on how to overcome the scalability problem in existing system of an organization. So, the main objective of this research is to discover an approach, specifically a data mining technique in pre-processing stage, such as Hamming Distance, k-Nearest Neighbor (k-NN) or deep learning algorithm as in [7] to be implemented to rearrange the matrix to overcome the huge scale of UPA as input data to produce more scalable, optimal and accurate datasets that can be used into the next phase. Furthermore, the selection of data mining methods for this research must be suitable for binary or Boolean data type.

The output of pre-processing stage is optimal candidate roles. A candidate role in a RBAC system contains a set of permissions that is connected to a user-to-role assignment that can be visualized as a row in matrix PA, a column in matrix UA, and a user is permitted with permissions if he/she is appointed with a role that includes the designated permissions. This stage is the most meaningful process because this phase usually produces a big pool of candidate roles [4], [8], [9], therefore appropriate techniques are needed to recover optimal candidate roles by exploiting appropriate data mining techniques or heuristics algorithms.

This paper proposes an approach that can manage the conversion from traditional and existing ACM in an organization to RBAC that contains large data since the large data can be complicated and unscalable and consist of redundant data with immoderate permissions and roles. The approach has utilized data mining technique particularly Hamming Distance, prior to the role mining process that can cluster a more accurate RBAC system.

The remainder of the paper is structured as follows. Section II presents a background study of this work, specifically on clustering techniques. Section III discusses the general methodology developed for the proposed approach; Section IV elaborates the experimental results and discussion on the proposed approach; lastly Section V sums up the research with conclusion and future works.

II. Related Work

The fast growth of internet technologies has created extreme escalation of data gathering, storing, and analysis of large datasets and data mining can be described as a method to obtain informative patterns from these datasets. One of a common data mining task that is appropriate for role mining is clustering. Generally, clustering can be expressed as a process of combining items that have the same attributes, specifically in role mining. Clustering is the act of grouping similar users and permissions to produce a common set of roles. Clustering techniques have been comprehensively reviewed in many applications, such as the pattern recognition field. According to the authors, clustering technique offers many advantages as the following features [10], [11]:

1) Firstly, it can be utilized as a pre-processing technique to obtain related groups within the datasets.

2) Secondly, it can decrease the cost that involved in data mining technology.

3) Thirdly, it is effective to get information regarding the properties of the datasets.

Many clustering techniques have been proposed for different datasets, and most of the conventional clustering algorithms are unsuitable for handling categorical datasets, such as in role mining. Researchers [12] have introduced a statistical method using Hamming Distance (HD) to cluster categorical datasets. In their research, HD vectors has been utilized to generate clusters for each iteration until no notable clusters can be produced. The proposed method has performed significantly better than the other algorithms. Furthermore, the application of clustering techniques in role mining has been discovered by [13], [14] and according to the authors, the dataset would be segregated into clusters based on the same characteristics that eventually would decrease size of the dataset significantly.

Moreover, the authors also have discovered, based on the simulations and analysis conducted, that the application of the Bayesian model can successfully cluster datasets that contain categorical data [15]. Most recently, [16] have effectively proposed a model to cluster categorical datasets using a mixture of distributions based on HD. Additionally, according to the authors, the role mining scales were huge that could produce results that were not easily interpreted; hence the authors have adapted the basic role mining concept into clustering problem with the application of HD to rebuild the original matrix into a compressed matrix [17].

In recent years, there has been an increasing amount of literature on role mining techniques in Role-based Access Control (RBAC). However, numerous existing role mining algorithms in RBAC do not provide any appropriate approach to overcome the huge scale of existing UPA in an organization that may contain overlapping permissions that can lead to inaccurate and too many roles that eventually overwhelmed the existing system and burdened the administrator. Researchers [18] have proposed a technique based on frequent pattern mining. However, this technique has constructed a large number of potential permission sets. Additionally, researchers [19] have discussed a feasible solution based on a constraint satisfaction problem. However, this solution still produced quite a large of number of users and permissions. Therefore, the rest of this paper will discuss on the methodology, results, and discussion on Hamming Distance approach to reduce role mining scalability.

III. Materials and Methods

This section presents the detailed specification of an approach or phase to restructure the huge scale of role mining input data, namely user-permission assignment (UPA), that exists in a form of Boolean matrix to produce optimal and accurate candidate roles. This approach discusses the application of data mining technique, specifically Hamming Distance (HD), to reduce the scalability of UPA input data. The process begins by grouping users with the same permissions and considering them as a user group. Then each user group can be depicted as different user clusters. The output of this process is a less complex matrix UPA.

A. Datasets

The dataset that has been used in this research is the benchmark access control datasets, as shown in Table I and the datasets comprise of the numbers of users |U|, the number of permissions |P|, the size of user-permission assignment |UPA|, number of roles |R| and density that can be described as the number of entries equivalent to one with the respect to its size in an unrestricted setting. The Apj dataset was acquired from the network access control rules used in Hewlett Packard (HP) and the profile was obtained from the Cisco firewalls and used to authenticate the users with the related network access [20]. Furthermore, the healthcare dataset was collected from the US Veteran's Administration [21]. Additionally, the firewall1 and firewall2 datasets were gained from Checkpoint firewalls and lastly the domino dataset came from a set of user and access profiles for a Lotus Domino server [20].

TABLE I. Real World Datasets

Dataset	**|U|**	**|P|**	**|UPA|**	**|R|**
Apj	2044	1164	6841	453
Domino	79	231	730	20
Firewall1	365	709	31951	64
Firewall2	325	590	36428	10
Healthcare	46	46	1486	14

B. Measurements

This approach needs data mining technique such as Hamming Distance to rearrange the matrix to overcome the huge scale of UPA as input data to produce more scalable, optimal and accurate datasets. Moreover, the selection of data mining methods for this research must be suitable for binary or Boolean data type. For this research, the successful implementation of such technique can be determined by size of generated files and the size should be reduced (smaller size) compared to the original dataset

C. Hamming Distance Approach

An effective transition from a conventional access control model to a role-based model needs to define an appropriate dataset that enables to capture the security policies of an organization. The complexity of the RBAC system can be quantified by parameters such as number of roles, permissions, hierarchy size, constraints, and user-permission assignment (UPA). Although the process may seem uncomplicated to accomplish when the roles can be defined from the beginning, for an organization with existing user-permission assignments, this procedure can be complicated to produce stable candidate roles, especially when the existing UPA contain a huge number of permissions and users. This enormous UPA may deteriorate the functionality of the RBAC system and become challenging to handle appropriately.

For a RBAC system, role mining is a method that can be implemented to cluster or group users who have the same or comparable permissions and role mining can be utilized to create various roles with these permissions. Users are commonly given roles with numerous duplicate permissions and this method can simplify the management and maintenance of RBAC system. Therefore, in order to reduce the complexity of the generated roles, it is necessary to cluster users who share the same attributes. However, implementing traditional role mining techniques resulting enormous mining scales and burdens the administration of the systems due to the miscellany of permissions and users [22].

Therefore, a pre-processing technique is needed to reduce the scalability of the UPA to produce more precise candidate roles. For this paper, basic role mining has been converted into a clustering problem using the Hamming Distance (HD) approach and basic role mining can be defined as the following:

Definition 1. Given a set of users (U), a set of permissions (PRMS), a user permission assignment (UPA), a set of roles (ROLES), a user-to-role assignment (UA), and a role-to-permission assignment (PA), 0-consistent with UPA and minimizing the number of roles, k.

Hamming Distance (HD) approach has been applied to decrease the scales of UPA. This approach can decrease the size of initial dataset by grouping users with the same permissions in the existing UPA to generate an initial set of roles. As input data of RBAC model is user-permission assignments (UPA) in a form of Boolean matrix, it can be observed that the matrix is at the same length, and it is a common practice to calculate the distance between two separate but similar length vectors applying Hamming Distance calculation. Generally, the approach to find the distance can be done by calculating the number of positions between two similar length vectors namely Distance (x, y).

IV. Experimental Results and Discussion

This section presents the results and discussion on the pre-processing model as discussed in previous sections. The algorithms have been implemented in Python 3.6 through Visual Studio Code and tested on a MacBook Air running macOS Monterey Version 12.6.1 on Apple M2 and CPU having 8 GB memory. Five real-world datasets have been widely employed in literature to evaluate the framework to analyze the performances of various unconstrained role-mining heuristics.

The effectiveness of pre-processing model that works to restructure role mining input data (UPA) that exists in the form of a Boolean matrix to produce optimum candidate roles can be demonstrated in the following subsection. The measurement that has been used is size of generated files and the size should be reduced (smaller size) compared to the original dataset. More precisely, the size of the generated files can be expressed as clustered size that can determine how well UPA are clustered. The dataset that has been used in this research is the benchmark access control datasets, as shown in Table I and the datasets comprise of the numbers of users |U|, the number of permissions |P|, the size of user-permission assignment |UPA| and number of roles |R| in an unrestricted setting.

Fig. 1 describes the model of pre-processing approach to generate candidate roles or can be described as optimal candidate role set identification, and this model can be divided into three main steps. In the first step, rows in the original dataset have been applied with Hamming Distance formula to find a distance value between two identical length rows. Furthermore, in the second step, based on the values of Hamming Distance, the generated dataset is divided into partitions according to similar clusters of users. Lastly, the generated dataset is rearranged in the third step to produce a meaningful smaller set of users that signify each cluster. Thus, these steps can be viewed as processes that can reduce the scalability of UPA, resulting in a compressed dataset containing final candidate roles or optimal candidate to be used in the next phase.

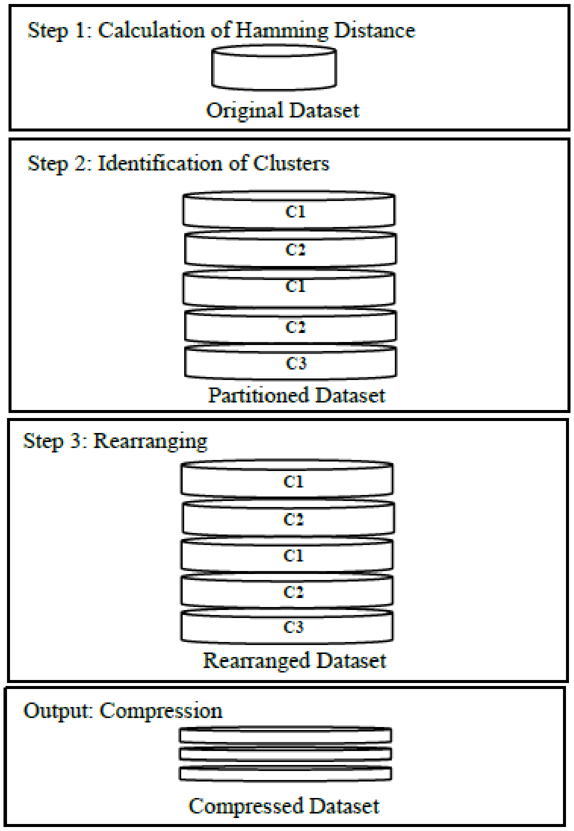


Fig. 1. Role selection and assignment phase.

Table II compares the original files size and generated files size that have been applied with Hamming Distance computation. Meanwhile, Fig. 2, Fig. 3, Fig. 4, Fig. 5 and Fig. 6 show the contrast between both sizes that are represented in form of a graph. Each graph can be designated by blue and red bar, the blue bar signifies the original file size, and correspondingly, the red bar symbolize generated file size, which are computed by Hamming distance computation. Furthermore, Fig. 7 shows the comparison of file sizes between both original datasets and generated datasets in the directory of the computer.

TABLE II. ORIGINAL VS EXTRACTED FILES SIZE

Dataset	Original File Size	Generated File Size
Apj	144 KB	68 KB
Domino	15 KB	6 KB
Firewall1	671 KB	273 KB
Firewall2	765 KB	319 KB
Healthcare	31 KB	10 KB

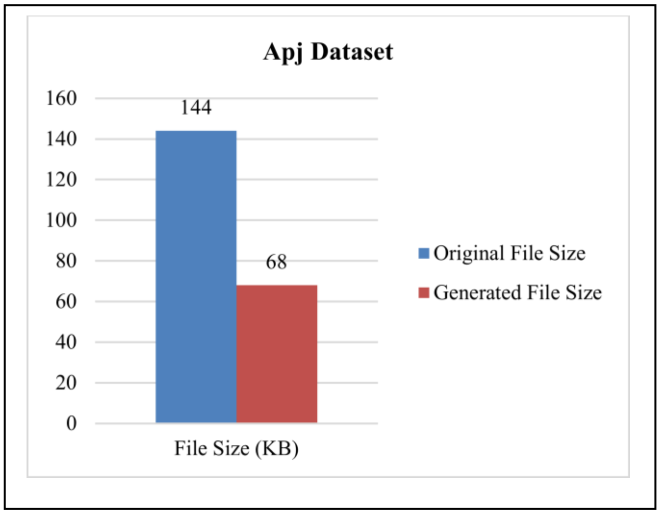


Fig. 2. Initial vs. extracted file size comparison of apj dataset.

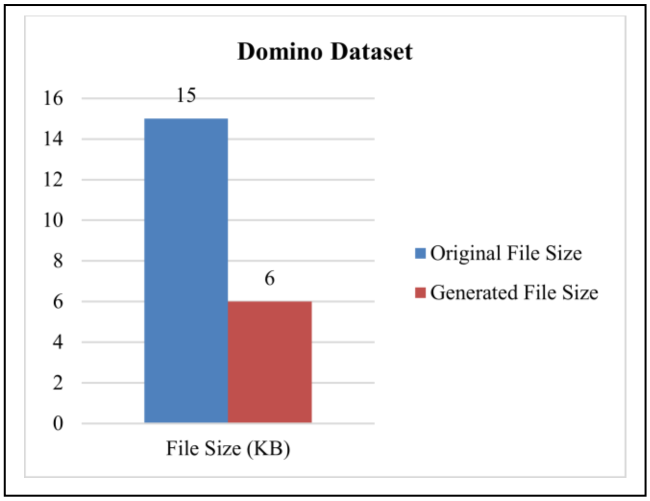


Fig. 3. Initial vs. extracted file size comparison of domino dataset.

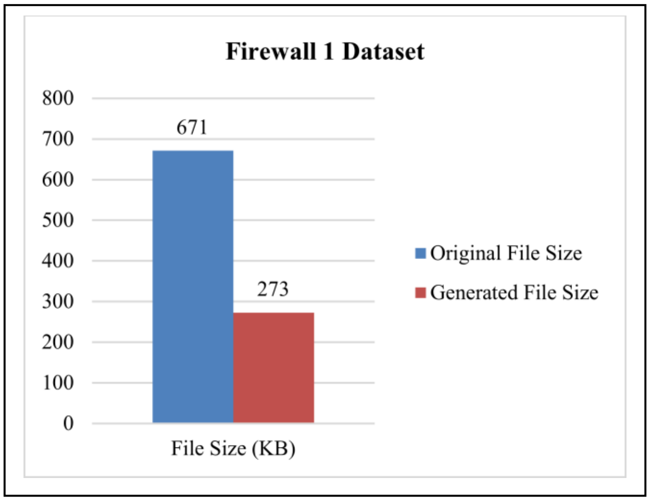


Fig. 4. Initial vs. extracted file size comparison of firewall 1 dataset.

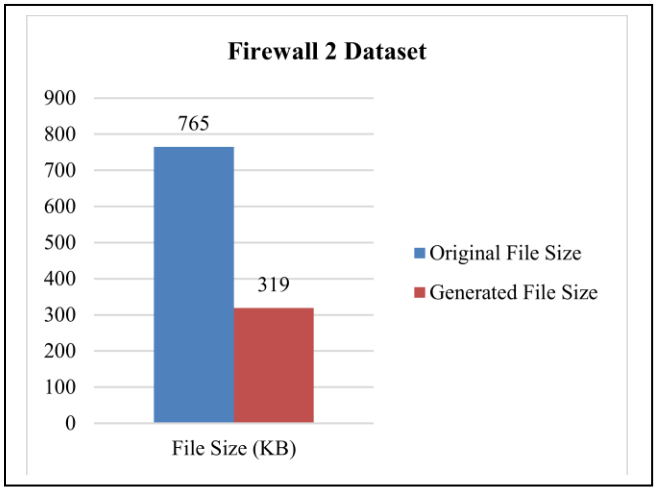


Fig. 5. Initial vs. extracted file size comparison of firewall 2 dataset.

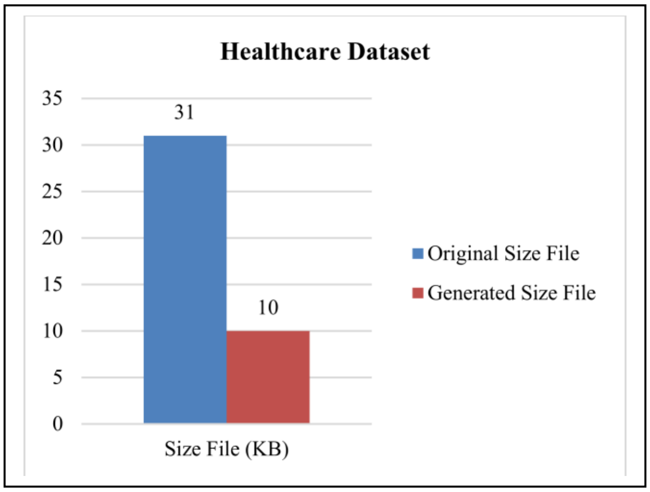
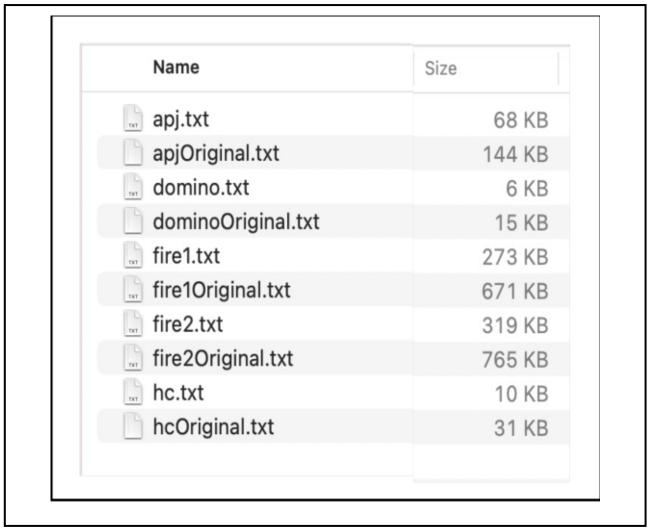


Fig. 6. Initial vs. extracted file size comparison of healthcare dataset.



Name	Size
apj.txt	68 KB
apjOriginal.txt	144 KB
domino.txt	6 KB
dominoOriginal.txt	15 KB
fire1.txt	273 KB
fire1Original.txt	671 KB
fire2.txt	319 KB
fire2Original.txt	765 KB
hc.txt	10 KB
hcOriginal.txt	31 KB

Fig. 7. Comparison of original and generated file size.

TABLE III. REDUCED FILES SIZE

Dataset	Original File Size	Generated File Size	Reduced Size (%)
Apj	144 KB	68 KB	47.2%
Domino	15 KB	6 KB	40.0%
Firewall1	671 KB	273 KB	40.7%
Firewall2	765 KB	319 KB	41.7%
Healthcare	31 KB	10 KB	32.3%

Significantly, based on Table II and Fig. 2 to Fig. 7, the generated file size of all five datasets substantially have been reduced compared to the initial or original datasets showing that Hamming Distance (HD) approach is successfully can be utilized to reduce the mining scale of datasets and eventually can produce better candidate roles. Thus, the three steps as in Fig. 1 can effectively recognize as processes that can reduce the scalability of UPA and resulting a compressed dataset that contains final candidate roles to be used in the next phase. Table III displays the percentage of reduced file size and indicates that HD enables to compress the original dataset to become a smaller generated dataset between 32% to 47%. The Apj dataset has disclosed the highest percentage of 47.2%. In the meantime, the Healthcare dataset has shown the lowest percentage of 32.3%.

V. CONCLUSION AND FUTURE WORKS

In conclusion, the generated file size of all five datasets has been significantly reduced compared to the original using Hamming Distance approach. The process begins by grouping users that have the same permissions and considering them as a user group, and then each user group can be depicted as different user clusters. The output of this process is a less complex UPA matrix.

For future works, two directions can be considered. The first direction is to consider other possible data mining techniques to be combined with role mining technique, particularly clustering techniques that can produce more accurate candidate role sets, and the second direction is to explore role-engineering optimization potential in other applications or environments such as in Internet of Thing (IoT) environment.

REFERENCES

[1] I. Molloy, H. Chen, T. Li, Q. Wang, N. Li, E. Bertino, S. Calo, and J. Lobo, "Mining roles with multiple objectives," ACM Transactions on Information and System Security, vol. 13(4), pp. 1–35, 2010.

[2] B. Mitra, S. Sural, J. Vaidya, and V. Atluri, "Migrating from RBAC to temporal RBAC," IET Information Security, vol. 11(5), pp. 294–300, 2017.

[3] L. Fuchs, and S. Meier, "The role mining process model - underlining the need for a comprehensive research perspective," Sixth International Conference on Availability, Reliability and Security, pp. 35-42, 2011.

[4] S. Das, B. Mitra, V. Atluri, J. Vaidya, and S. Sural, "Policy engineering in RBAC and ABAC," From Database to Cyber Security, pp. 24–54, 2018.

[5] H. Kiwan, and R. Jayousi, "Dynamic user-oriented role based access control model (DUO-RBAC)," Conference Business Intelligence & Big Data, pp. 281-290, 2018.

[6] H. Lu, X. Chen, J. Shi, J. Vaidya, V. Atluri, Y. Hong, and W. Huang, "Algorithms and applications to weighted rank-one binary matrix factorization," ACM Transactions on Management Information Systems, vol. 11(2), pp. 1–33, 2020.

[7] Y.M. Alwaqfi, M. Mohamad, A.T. Al-Taani, and N. Abd Hamid, "A novel hybrid DL model for printed arabic word recognition based on GAN," International Journal of Advanced Computer Science and Applications, vol. 14(1), 2023.

[8] H. Lu, J. Vaidya, and V. Atluri, "An optimization framework for role mining," Journal of Computer Security, vol. 22(1), pp. 1–31, 2014.

[9] H. Lu, Y. Hong, Y. Yang, L. Duan, and N. Badar, "Towards user-oriented RBAC model," Journal of Computer Security, vol. 23(1), pp. 107–129, 2015.

[10] R. Vijay, P. Mahajan, and R. Kandwal, "Hamming distance based clustering algorithm," International Journal of Information Retrieval Research (IJIRR), vol. 2(1), pp. 11-20, 2012.

[11] V.E. Mirzakhanov, "Value of fuzzy logic for data mining and machine learning: a case study," Expert Systems with Applications, vol. 162, pp. 1-35, 2020.

[12] P. Zhang, X. Wang, and P.X.K. Song, "Clustering categorical data based on distance vectors," Journal of the American Statistical Association, vol. 101(473), pp. 355–367, 2006.

[13] A. Colantonio, R. Di Pietro, A. Ocello, and N.V. Verde, "Visual role mining: a picture is worth a thousand roles," IEEE Transactions on Knowledge and Data Engineering, vol. 24(6), pp. 1120-1133, 2011.

[14] N.V. Verde, J. Vaidya, V. Atluri, and A. Colantonio, "Role engineering: from theory to practice," Proceedings of the Second ACM Conference on Data and Application Security and Privacy, pp. 181-192, 2012.

[15] M. Ye, P. M., Zhang, and L. Nie, "Clustering sparse binary data with hierarchical bayesian bernoulli mixture model," Computational Statistics and Data Analysis, vol. 123, pp. 32–49, 2018.

[16] E. Filippi-Mazzola, R. Argiento, and L. Paci, 2021, "Clustering categorical data via hamming distance," Book of Short Papers, Pearson, pp. 752–757.

[17] W. Sun, X. Yuan, and H. Su, 2021, "Role-engineering optimization with user-oriented cardinality constraints in role-based access control," International Journal of Network Security, vol. 23(5), pp. 845–855, 2021.

[18] Z. Dana, R. Kotagiri, E. Tim, and Y. Trevor Yann, 2008, "Permission set mining: discovering practical and useful roles," 2008 Annual Computer Security Applications Conference (ACSAC), pp. 247–256, 2008.

[19] H. J. Jafar, T. Hassan, T. Hakim, H. Ehsan, and S. Shehab, "Towards a general framework for optimal role mining: a constraint satisfaction approach," Proceedings of the 20th ACM Symposium on Access Control Models and Technologies, pp. 211–220, 2015.

[20] A. Ene, W. Horne, N. Milosavljevic, P. Rao, R. Schreiber, and R.E Tarjan, R. E., "Fast exact and heuristic methods for role minimization problems," Proceedings of the 13th ACM Symposium on Access Control Models and Technologies, pp. 1-10, 2008.

[21] B. Mitra, S. Sural, J. Vaidya, and V. Atluri, "A survey of role mining," ACM Computing Surveys," vol. 48(4), pp. 1–37, 2016.

[22] W. Sun, and H. Su, H., "Role-engineering optimization with mutually exclusive permissions constraints and permission-to-role cardinality constraints," International Journal of Innovative Computing, Information and Control, vol. 17(4), pp. 1373–1390, 2021.