Asian Journal of Information Technology 13 (8): 416-421, 2014
ISSN: 1682-3915

Graphical Data Mining using Hybrid Role Engineering Approach

R. Suguna, Divya Mohandass and R. Poorni
Department of CSE, SKR Engineering College, Chennai, India

Abstract: Role-Based Access Control (RBAC) is an effective mechanism that provides resources by reducing the risk of resource allocation when the designed roles are business-driven. Data mining represents an essential tool for role engineers. Graphical data mining, formally defines the problem by introducing a metric for the quality of the visualization. Using the RBAC approach, the user-permission assignments of a given set of roles are defined which provides an interface and allows visual elicitation of roles even in the absence of predefined roles. Besides being rooted in sound theory, this is supported by extensive simulations run over real data. Result of the visualization confirm the quality of the proposal and demonstrate its viability in supporting role engineering decisions.

Key words: Access controls, data and knowledge visualization, mining methods and algorithms, viability, assignments

INTRODUCTION

ACCESS control is the process of mediating requests to data and services maintained by a system determining which requests should be granted or denied. Significant research has focused on providing formal representation of access control models. Role-Based Access Control (RBAC) has become the norm in most organizations due to its simplicity. Each and every role has a set of permissions and the user has roles on responsibilities. To implement a RBAC System, it is important to devise a complete set of roles. This design task, known as role engineering has been defined as the richest part of a RBAC-oriented project. Recently, there has been an increasing interest in using automated role engineering techniques. All of them seek to identify de facto roles embedded in existing access permissions. Despite much research dedicated to the design of Role Mining algorithms, existing techniques deal with three main practical issues: meaning of elicited roles, noise within data and correlations among roles (Colantonio *et al*., 2012).

Meaning: Organizations are unwilling to deploy roles they cannot fully understand. However, automatically elicited roles often have no connection to business practice. Some Role Mining algorithms seek to identify a minimal set of roles which also minimize the resulting system complexity (Di Vimercati *et al*., 2007). Yet, it is legitimate to ask whether automated techniques can overcome and replace the cognitive capacity of humans. To gain greater flexibility, other algorithms propose a complete list of roles, so role designers can manually select the most relevant ones. However, there could be a loss in the complete view of data due to the typically large number of candidate roles and unavoidable exceptions.

Noise: Exceptionally or accidentally granted permissions can hinder the role mining task. To deal with this problem, a number of methods have recently been proposed. However, they usually require tuning several parameters that greatly affect algorithm performance and the quality of results (Frank *et al*., 2009). Further, proposed noise models do not always fit real cases, especially when exceptions are legitimate and cannot be avoided.

Correlations: The identification of relationships among roles (e.g., similarities, permission set inclusion, etc.) can further ease the identification of roles. But most of the existing Role Mining algorithms do not provide analysts with any correlation information. One possibility is to compute hierarchical relationships, based on permission set inclusion, after role elicitation (Colantonio *et al*., 2008a, b). However, such relationships do not always reflect the actual senior-junior hierarchy from a business perspective.

LITERATURE REVIEW

Ferraiolo *et al*. (2001) have provided single authoritative definition of RBAC by unifying ideas from a base of frequently referenced RBAC Models, commercial products and research prototypes. The proposed approach greatly differs from

Corresponding Author: R. Suguna, Department of CSE, SKR Engineering College, Chennai, India

(Di Vimercati *et al*., 2007; Ferraiolo *et al*., 2001): first, it adopts a different visualization cost metric that is more suitable for role engineering which overcomes the incompatibility with the core of their theory. Secondly, it also allows obtaining a matrix representation without resorting to any existing Mining algorithm. The role engineering problem was first illustrated by Coyne (1995) through a top-down perspective which has rules that specify conditions under which access is granted or denied. After that, several algorithms explicitly designed for role engineering purposes were proposed by Molloy *et al*. (2009), presented a comparative study by categorizing Role Mining algorithms into two classes based on their outputs; Class 1 algorithms output a sequence of prioritized roles while Class 2 algorithms output complete RBAC states. According to the provided business information indices, minability and similarity used to measure the expected complexity of analyzing the outcome of bottom-up approaches (Colantonio *et al*., 2009). A Visualization algorithm has been proposed that extends existing graph sorting algorithms to offer a good matrix visualization of previously defined hypergraphs which can be mapped to the role concept in the RBAC terminology (Colantonio *et al*., 2009; Frank *et al*., 2009). Frank *et al*. (2008) give a possible way to build a matrix representation of user-permission relationships. However, this construction is limited to the special case of non overlapping roles far from being general and optimal according to the definition. Moreover, it is not applicable to generic role mining approaches. Kuhlmann *et al*. (2003) first introduced the term role mining, tries to apply existing data mining techniques to elicit roles from access data. In general, this approach can be considered a complement for all the existing role engineering methodologies and tools. Indeed, it allows for an effective, viable and intuitive way to evaluate and select roles generated by other approaches. Colantonio *et al*. (2010a, b) recently addressed the problem of analyzing the role mining complexity by also proposing a way to reduce it by decomposing the data into smaller subsets. A Branch and Bound algorithm for mining large tiles (Geerts *et al*., 2004) (that is regions of database consisting solely of ones) is introduced. Instead of finding large tiles, this approach focuses on the problem of visually representing tiles. As for visual representation of mined data, a few visualizers have been proposed in the current literature and most of them are not explicitly designed for binary data.

The proposed approach uses the BicOverlapper tool that integrates on a set of well-known visualization techniques that represent gene data information on different levels. However, typical representations for gene data such as repeating rows and columns of the analyzed matrix are confusing or not suitable for role mining.

PROPOSED SOLUTION

In the proposed system to address the aforementioned issues an approach referred to as graphical data mining is defined. Abstract user-permission patterns (i.e., RBAC roles) are managed as graphical patterns. The rationale behind this approach is that graphical representations of roles can actually amplify cognition, leading to optimal analysis results. Figure 1 gives the system architecture that shows the interface between the application user and the graphical view (Geerts *et al*., 2004). It offer a graphical way to effectively navigate the result of any existing Role Mining algorithm, showing at glance what it would take a large amount of data to explain. Moreover, it allows identifying meaningful roles within access control data by providing a visual representation. Visualization of the user-permission assignments is performed in such a way to isolate the noise allowing role engineers to focus on relevant patterns, thereby enhancing their cognition capabilities. Further, correlations among roles are shown as overlapping patterns, hence providing an intuitive way to discover and utilize these relations. Even though visual approaches sometimes raise some skepticism, they are generally considered to be highly beneficial when used to gain an overview of the underlying data set. In fact, a proper representation of user-permission assignments allows role designers to gain knowledge and conclude meaningful roles from both IT and business perspectives. The user permission or the access control is used as graphical pattern that leads to the optimal analysis result (Di Vimercati *et al*., 2007). Graphically, identification of

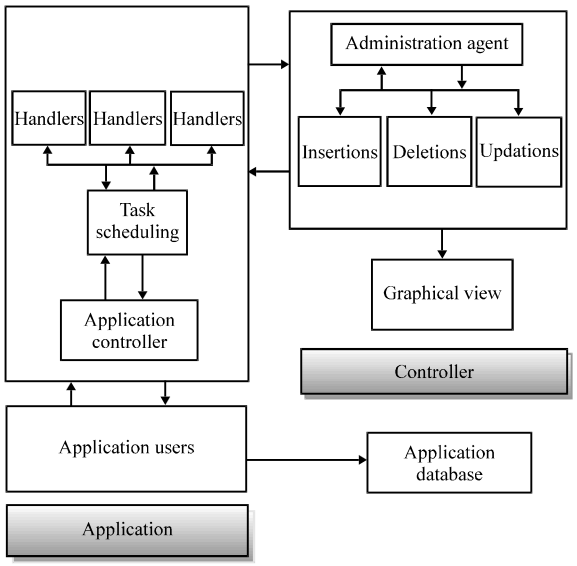

Fig. 1: System architecture of graphical data mining

meaningful roles within access control data can be performed by using the ADVISER and EXTRACT algorithms. Sorting is done to the user permission in order to reduce the cost of fragmentation and thus suitable for real time applications.

The role mining objective is to analyze access control data in order to elicit a set of meaningful roles that simplify RBAC management (Molloy *et al*., 2009; Colantonio *et al*., 2009, 2008). To this aim, various business information can be analyzed (Molloy *et al*., 2009; Colantonio *et al*., 2009) but for user-permission assignments minimal data set are required. A natural representation for this information is the binary matrix where rows and columns represents users and permissions, respectively and each cell is on when a particular user has a certain permission granted. User permission assignments will be done in order to avoid resource allocating issues. User permission assignment is the process of mediating requests to data and services maintained by a system, determining the request access to the application data. Privileges will be assigned to application users for achieving data access controls. Abstract user-permission patterns (i.e., RBAC roles) are managed as graphical patterns. These patterns are usually referred to as tiles. It demonstrates that it could be easier to find more patterns if users and permissions were reordered. With the access permission generated an interface text file will be used to provide graphical representation where a color for the roles will be assigned which can update automatically. This is used for the generation of heat map images which provides roles along with their corresponding assignments. A visual representation can highlight potential exceptions within data in an effective manner and a textual role representation reports on information about role-user and role-permission relationships in a less communicative fashion than a graphical representation. It works well for a larger subset of data by restricting the analysis to smaller subsets of data that are homogeneous with respect to some business-related information (Frank *et al*., 2008; Guo *et al*., 2008) (e.g., partitioning users by department, job title, cost center, etc.).

IMPLEMENTATION

By leveraging on the observations made in the study, the approach is implemented with the help of a banking application. The employees and the users of the application will be allocated with a set of permissions which is represented using the binary matrix. The details of few employees registered in the database are shown in Table 1.

The text file interface for the above dataset will be generated as follows with the color assigned in their corresponding role using their color index value (Table 2).

The same procedure is applied for customers with the details of account number and their corresponding loan information. Since, the color assignment is given as a text file interface it is highly fast and reliable.

By taking the interface file as an input the microarray heat-map image has been generated as shown in Fig. 2 by indicating the assigned color as the user-permissions. A viable, Fast Heuristic algorithm called ADVISER (Access Data Visualizer) (Colantonio *et al*., 2012) is used for role representation. Given a set of roles, this algorithm is able to provide a compact representation of them. In particular, it reorders rows and columns of the user-permission matrix to minimize the fragmentation of each role. Despite being relatively simple, it provides a good though not necessarily optimal solution. In particular, its running time is O(nX(|ROLES|+log n)) where, n = max {|USERS|, |PERMISSIONS|}. As a heuristic, ADVISER is based on some intuitions, summarized in the following:

- Introducing a gap in the visualization of large roles (namely, those roles that involve many users and permissions) increases (7) more than introducing gaps on smaller roles. Hence, larger roles should be better represented
- The more fragments in the visualization of a role, the higher the role visualization cost
- Changing the order of user without permissions only affects the number of gaps between columns and so do permissions

Table 1: User permission assignment

User name	Password	Address	E-mail	Role
br_123	br_123	Chennai	br@gmail.com	Branch manager
ac_2925	ac_2925	Chennai	ac_2925@gmail.com	Account management
suresh_5244	suresh	Chennai	suresh@gmail.com	Account management
rajesh_2076	rajesh	Chennai	rajesh@gmail.com	Loan management
sam_1250	sam	Chennai	sam@gmail.com	Loan management
sathish_9922	sathish	Chennai	sathish@gmail.com	Account management
priya_0	priya 1	Chennai	priya@gmail.com	Cashier
thomas_7843	thomas	Chennai	thomas@gmail.com	Cashier
xyz_9877	xyz	Chennai	xyz@gmail.com	Account management

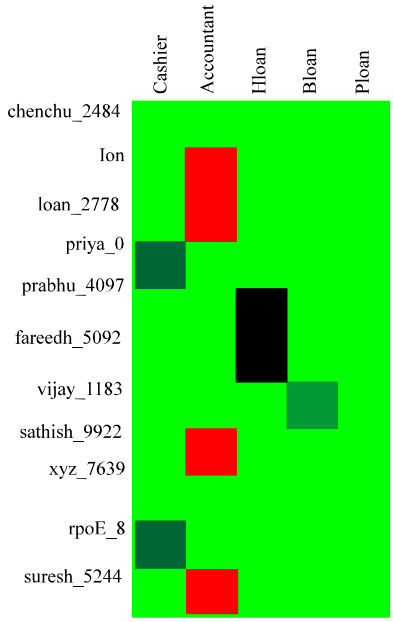

Fig. 2: Heat-map representation for the employee database

Table 2: Color assignment to the employee database

Names	Cashier	Accountant	Hloan	Bloan	Ploan
priya_0	258	0	0	0	0
rajesh_2076	0	0	400	0	0
xyz_9877	0	1000	0	0	0
sam_1250	0	0	400	0	0
sathish_9922	0	1000	0	0	0
thomas_7843	258	0	0	0	0
ac_2925	0	1000	0	0	0

When it is required to identify roles through visual analysis, a natural question is how a role set for ADVISER should be made in order to facilitate this task. An approach is to first compute all possible closed permission sets and later trying to best represent them. A permission set is closed when no proper supersets of permissions possessed by the same users exist. Examples of algorithms that compute such patterns are given (Vaidya *et al*., 2007; Colantonio *et al*., 2008; Fekete *et al*., 2008). Closed permission sets provide a compressed representation of all possible permission combinations that can be found within users. Closed permission sets are roles in RBAC terminology.

By feeding ADVISER with closed permission sets a matrix visualization that seeks to contextually best depict all identifiable patterns is provided. However, the number of closed permission sets is often too large when compared to the number of users and permissions (Vaidya *et al*., 2006). Hence, leading to long running time and huge memory footprint. To reduce the overall problem complexity, a probabilistic algorithm called EXTRACTS (Exception-Tolerant Role Actualizer) (Colantonio *et al*., 2012) which helps to extract the data from a larger dataset

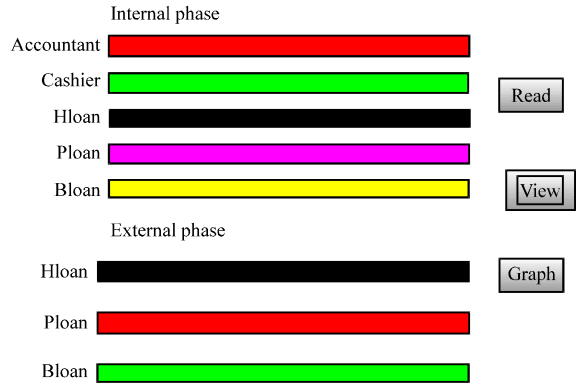

Fig. 3: Graphical mining using color assignment

along with its corresponding graph. It generates a list of pseudoroles used to feed ADVISER in lieu of closed permission sets. By computing such pseudoroles it takes just O(k|UP|) where k is a tunable parameter and UP is the user permission of the algorithm.

A naive approach to generate all pseudo roles is to scan all assignments in (u, p)€UP and identifying the corresponding pseudo role by computing users (p) and perms (u). During the scanning process, whenever an already existing pseudo role is generated, its frequency is updated. This intuitive and simple algorithm has a running time O(|UP| log |UP|). Assuming that UP is ordered, the search for all the users possessing p and all the permissions assigned to u can be executed in O(log |UP|) and this must be done for all assignments in UP. In the worst case, it generates |UP| pseudo roles, i.e., a different pseudo role for each assignment. Hence, searching and updating the frequency requires O(log |UP|), for instance by storing pseudo roles in a self-balancing binary search tree. Although, this algorithm is quite efficient still better results can be obtained.

In particular, a very fast Randomized algorithm to generate pseudo roles called EXTRACT is used. In order to extract the data from the generated heatmap images shown in Fig. 3, EXTRACTOR algorithm is used which selectively identifies the data with their corresponding graphical patterns. The algorithm is as follows:

- Approximate the frequencies of pseudo roles by sampling k times a relationship in UP uniformly
- Generate the corresponding pseudo role
- For each pseudo role count the number of times it has been generated

Hence, the overall computational complexity is O(k log |UP|).

CONCLUSION

Visualizing user-permission assignments in an intuitive graphical form makes it possible to simplify the role engineering process. The proposed representation of data allows role designers to gain insight, draw conclusions and ultimately design meaningful roles from both IT and business perspectives. It offered a formal description of the visual role mining problem. Second, construction of binary matrix representation of user-permission relations has been demonstrated. An efficient probabilistic tool EXTRACT produces approximate patterns that can be used in conjunction with ADVISER to obtain high-quality visualization results. The quality of the results produced by EXTRACT seems to be appreciable. Finally, extensive applications over real and public data confirm that this approach is efficient, both in terms of computational time and result quality. The contributions other than being useful for role engineering can have interesting applications in other fields as well. Another application that could benefit from this approach is the well-known market basket analysis. In general, whenever there is a need to analyze data represented as a binary matrix, this approach can introduce benefits.

As for future research, the solutions can be extended in several directions such as data filtering, zooming algorithms or approximated representation of data. Further, the problem of coping with very large data sets would deserve a deeper analysis.

REFERENCES

Colantonio, A., R. Di Pietro and A. Ocello, 2008a. Leveraging lattices to improve role mining. Proceedings of the IFIP 20th World Computer Congress, September 7-10, 2008, Milano, Italy, pp: 333-347.

Colantonio, A., R. Di Pietro and A. Ocello, 2008b. A cost-driven approach to role engineering. Proceedings of the ACM symposium on Applied Computing, March 16-20, Fortaleza, Ceara, Brazil, pp: 2129-2136.

Colantonio, A., R. Di Pietro, A. Ocello and N.V. Verde, 2009. A formal framework to elicit roles with business meaning in RBAC systems. Proceedings of the 14th ACM symposium on Access control models and technologies, June 3-5, 2009, Stresa, Italy, pp: 85-94.

Colantonio, A., R. Di Pietro, A. Ocello and N.V. Verde, 2012. Visual role mining: A picture is worth a thousand roles. Knowl. Data Eng., 24: 1120-1133.

Colantonio, A., R. Di Pietro, A. Ocello and N.V. Verde, 2010a. Taming role mining complexity in RBAC. Comput. Secur., 29: 548-564.

Colantonio, A., R. Di Pietro, Ocello and N.V. Verde, 2010b. ABBA: Adaptive bicluster-based approach to impute missing values in binary matrices. Proceedings of the 2010 ACM Symposium on Applied Computing, March 22-26, 2010, Sierre, Switzerland, pp: 1026-1033.

Coyne, E.J., 1995. Role engineering. Proceedings of the 1st ACM Workshop on Role-Based Access Control, November 30-December 2, 1995, Gaithersburg, Maryland, USA., pp: 15-16.

Di Vimercati, S.D.C., S. Foresti, P. Samarati and S. Jajodia, 2007. Access control policies and languages. Int. J. Comput. Sci. Eng., 3: 94-102.

Fekete, J.D., J.J. Van Wijk, J.T. Stasko and C. North, 2008. The Value of Information Visualization. In: Information Visualization, Fekete, J.D., J.J. van Wijk, J.T. Stasko and C. North (Eds.). Springer, Berlin, Germany, ISBN: 978-3-540-70956-5, pp: 1-18.

Ferraiolo, D.F., R. Sandhu, S. Gavrila, D.R. Kuhn and R. Chandramouli, 2001. Proposed NIST standard for role-based access control. Trans. Inform. Syst. Secur., 4: 224-2743.

Frank, M., A.P. Streich, D. Basin and J.M. Buhmann, 2009. A probabilistic approach to hybrid role mining. Proceedings of the 16th ACM conference on Computer and Communications Security, November 9-13, 2009, Chicago, Illinois, USA., pp: 101-111.

Frank, M., D. Basin and J.M. Buhmann, 2008. A class of probabilistic models for role engineering. Proceedings of the 15th ACM Conference on Computer and Communications Security, October 27 to October 31, 2008, Alexandria, VA, USA., pp: 299-310.

Geerts, F., B. Goethals and T. Mielikainen, 2004. Tiling databases. Proceedings of the 7th International Conference on Discovery Science, October 2-5, 2004, Padova, Italy, pp: 278-289.

Guo, Q., J. Vaidya and V. Atluri, 2008. The role hierarchy mining problem: Discovery of optimal role hierarchies. Proceedings of the Computer Security Applications Conference, December 8-12 2008, Anaheim, CA, pp: 237-246.

Kuhlmann, M., D. Shohat and G. Schimpf, 2003. Role mining-revealing business roles for security administration using data mining technology. Proceedings of the 8th ACM Symposium on Access Control Models and Technologies, June 2-3, 2003, Villa Gallia, Como, Italy, pp: 179-186.

Molloy, I., N. Li, T. Li, Z. Mao, Q. Wang and J. Lobo, 2009. Evaluating role mining algorithms. Proceedings of the 14th ACM Symposium on Access Control Models and Technologies, June 3-5, 2009, Stresa, Italy, pp: 95-104.

Vaidya, J., V. Atluri and J. Warner, 2006. RoleMiner: Mining roles using subset enumeration. Proceedings of the 13th ACM conference on Computer and Communications Security, October 30-November 3, 2006, Alexandria, Virginia, USA., pp: 144-153.

Vaidya, J., V. Atluri and Q. Guo, 2007. The role mining problem: Finding a minimal descriptive set of roles. Proceedings of the 12th ACM symposium on Access Control Models and Technologies, June 20-22, 2007, France, pp: 175-184.