Secure Data Integration: A Formal Concept Analysis Based Approach

Mokhtar Sellami, Mohamed Mohsen Gammoudi, and Mohand Said Hacid

ISETK Kef Tunisie, ISAMM Manouba Tunisie, UCBL Lyon 1 France
sellamimokhtar@yahoo.com, momogammoudi@gmail.com,
mshacid@bat710.univ-lyon1.fr

Abstract. Integrating and sharing information, across disparate data sources, entail several challenges: autonomous data objects are split across multiple sources. They are often controlled by different security paradigms and owned by different organizations. To offer a secure unique access point to these sources, we propose two-step approach based on Formal Concept Analysis. First, it derives a global vision of local access control policies. Second, it generates a mediated schema and the GAV/LAV mapping relations, while preserving the local source properties such as security.

Keywords: Access Control, Data Integration, Formal Concept Analysis.

1 Introduction

Data Integration [7] aims at providing a unique access interface to distributed data sources. This involves several issues: heterogeneous data objects, owned by different organizations, are often controlled through different access control paradigms. Heterogeneity needs the definition of global schema and the mapping. The mapping between the global schema and each local schema can be delineated through one of the prominent approaches [7] (i.e. GAV (Global As View), LAV (Local As View) or GLAV (Global As View)). These data-centric approaches aim at solving the data heterogeneity, query processing, and optimization problems. These approaches don't focus on the security aspects (i.e. availability, confidentiality, and integrity) which are major issues. Hence, access control aims at preventing unauthorized users from accessing sensitive data [4]. Data integration security is an 'open' issue since each source defines its own access control policies. Thus, the integration of the various security policies derives a representative policy to manage access to the whole data sources. To tackle these issues, we propose a two-step approach based on FCA (Formal Concept Analysis) theory [8]. It starts by combining the local policies to generate a synthesis policy at the mediator level. Then, it generates a mediated schema from the global policy. Finally, a mapping between the global schema and the local schemas is performed either by GAV or LAV. The use of FCA is justified by their sound mathematical foundations. FCA is a renowned formalism in data analysis and knowledge discovery because of its usefulness in important domains of knowledge discovery in databases (KDD) [11].

H. Decker et al. (Eds.): DEXA 2014, Part II, LNCS 8645, pp. 326–333, 2014.

Thus, we focus on three issues; i) a policy-centric approach: by investigating the global schema derivation according to the global policy (i. e. taking into account access control policies as the key to define visible parts of the sources to be integrated). ii) The preservation of the local source policies: an access control, enforced at the mediator level, has to preserve the local access control policies. iii) a mapping language-independent approach: by deriving mapping relations based on the different mapping languages (GAV, LAV).

This paper is organized as follows. Section 2 discusses the state of the art on information security and policy integration. Section 3 presents our FCA-based solution for secure data integration. The last section is devoted to the conclusion and future work.

2 Related Work

Information integration security is a challenging process, especially in enforcing access control to data in distributed environment [4, 5]. The authors in [6] present an approach which enforces rules and conditions expressed by privacy policies in the case of Hippocratic databases. Enforcing privacy policies does not require any modification of existing database applications. It is fulfilled by rewriting queries. For instance, a query Q is transformed to Q' in such a way that its result complies with the cell-level disclosure policy P. There is no query modification needed in our approach. This is due to the generation global schema according to the global policy. The authors in [9] propose an approach to integrate data using GAV while taking into account the authorization policies. This approach identifies the combination of virtual relations that could lead to the no preservation of the local authorization while ensuring no conflicts arise at mediator level. Its drawback is that it relies on the GAV as assumption to be applied. Moreover, this work lacks flexibility as it doesn't cover other access controls models and other integration approaches (i.e. LAV, GLAV).

Policy integration approaches aim at specifying policies by more than one policy authors and integrating them to check their compliance with the global requirements. In [1, 10], the authors describe an algebra for composing access control or privacy policies when different enterprises cooperate. Hence, the policy algebra is modeled as a composition language. Rao [12] propose algebra for fine grained integration of XACML policies. The former supports complex integration requirements using the defined algebraic operations. Nevertheless, these approaches were not designed to take into account data integration property. They aim at providing users of one system with access to data of another system, but do not consider how access to combined data, provided by different systems, should be enforced. The integration is still related to the link between the policy elements.

3 Formal Concept Analysis (FCA)

FCA is a branch of mathematical order theory [8], or more precisely a branch of lattice theory that has emerged during the 1980s.

Definition 1. Formal Context. is a triple K = (G,M, I) where G called objects and M are called attributes and I=G × M is a binary relation. We say that an object g has attribute m if g and m are in relation I (denoted by gIm).

Definition 2. Derivation Operators. Let K =(G,M,I) be a formal context and A ⊆ G be a set of objects. We define A'={ m ∈ M | ∀ g ∈ A:gIm } i. e. A' is the set of all attributes that all objects in G share. Analogously, let B ⊆ M be a set of attributes. We define B'={ g ∈ G | ∀ m ∈ B:gIm } ; i. e. B' is the set of those objects that have all attributes from B.

Definition 3. The concept lattice of a context (G; M; I) is a complete lattice in which infimum and supremum are given by:

$$\bigwedge_{t\in T}(A_t, B_t) = \left(\bigcap_{t\in T} A_t, \left(\bigcup_{t\in T} B_t\right)\right)$$
$$\bigvee_{t\in T}(A_t, B_t) = \left(\left(\bigcup_{t\in T} A_t\right), \bigcap_{t\in T} B_t\right)$$

Definition 4. Partial order relation between concepts: Let (A1, B1) and (A2, B2) two formal concepts. (A1, B1) << (A2, B2) if and if A1⊆A2⇔B2⊆B1 (A1, B1) is said a sub concept and (A2, B2) is said a super concept.

Definition 5. Galois lattice: the set of formal concepts Ordered by a partial order relation ≪ is said a Galois lattice.

4 A FCA-Based Secure Data Integration Approach

The proposed approach takes as input a set of source schema with its policies. This will be incrementally integrated based on the following 2 steps:

1. Step 1: Global Policy Generation: it starts by translating the schemas and policies to formal contexts. Then, first, it identifies the preserved-rule set of each attribute individually; second, it detects the possible attribute combination to identify the adequate rules that must be added to control this kind of combinations.
2. Step 2: Global Schema and Mapping Generation: it extracts the mediated schema according to the inferred policy. Then, it derives the GAV or LAV mapping relation between the global schema and the local schemas. Finally, it translates the global policy using specific access control model.

Our approach is illustrated through relational data integration as a reference framework. We conventionally assume that:

Data Model: 3 Relational Data Sources use the same attribute definition: S1: Admission: (SSN, AdmissionDate, Department), S2: Disease (SSN, DoctorID, Diagnosis) and DepartmentDoctor(DoctorID, Department) and S3: "Patient"(SSN, AdmissionDate, Department Sex).

Access Control Models: ABAC (Figure 1-a), VBAC (Figure 1-b) and Flat RBAC (Figure 1-c) using the same profile name to define the access control rules.

Mediator Level. Global Schema and Global policy aren't defined yet.

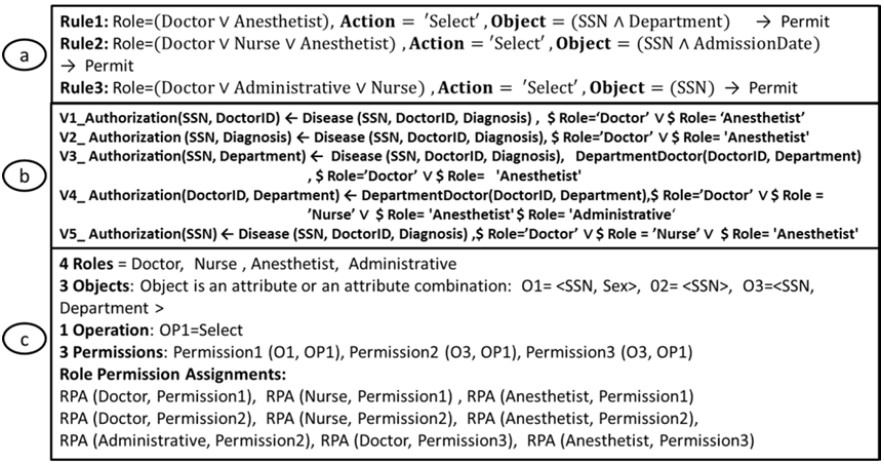

Fig. 1. A snapshot of the local access control policies

4.1 Step 1: Global Policy Generation

The generation of the global policy involves three stages: i) extracting formal contexts; ii) deriving a preliminary access rule set of each individual attributes; iii) identifying the possible attribute combinations and deriving the associated rules that complete the global policy to avoid answering queries of illegal users playing on this combination at mediator level.

Policy and Data Context Extraction. This step begins by identifying the similarities between attributes using existing matching techniques [2]. Then, it respectively generates the formal contexts: Data Context and Policy Contexts (table 1) based on the following definitions.

Definition 6 (Flat-RBAC Policy as a Formal Context (A, B, I) Given a Flat-RBAC Policy P and a transformation function σ_{RBAC}, a formal context is obtained as follows:

$$\boldsymbol{\sigma_{RBAC}(P)}\begin{cases} if\ R_i\ : Role\ and\ Att_j\ \in O_j |\ O_j : Object\ \ then\ \boldsymbol{A} = R_i\ \cup Att_j \\ if\ P_k : Permission\ \ |\ (k \leq t\ )\ \ then\ \boldsymbol{B} = Rule_k \\ if\ \ \exists\ RPA\ (R_i, P_k)\ and\ O_j\ \in P_k\ then\ \ \boldsymbol{I} = 1\ \boldsymbol{else}\ \ I = 0 \end{cases}$$

Definition 7 VBAC Policy as a Formal Context (A, B, I). Given a VBAC Policy P and a transformation function σ_{VBAC}, a formal context is obtained as follows:

$$\boldsymbol{\sigma_{VBAC}(P)}\begin{cases} if\ C_i\ : Constraints\ and\ Att_j : Attributes\ \ then\ \boldsymbol{A} = C_i\ \cup Att_j \\ if\ VH_i : Virtual\ Authorization\ then\ \boldsymbol{B} = Rule_k \\ if\ \ C_i \in VH_k and\ Att_j\ \in VH_k\ then\ \boldsymbol{I} = 1\ \ \ else\ \ \ \boldsymbol{I} = 0 \end{cases}$$

Definition 8 (ABAC Policy as a Formal Context (A, B, I). Given an ABAC Policy P and a transformation function σ_{ABAC}, a formal context is obtained as follows:

$$\boldsymbol{\sigma_{ABAC}(P)}\begin{cases} if\ R_i\ : Role\ and\ Att_j : Attribute\ \ then\ \boldsymbol{A} = R_i\ \cup Att_j \\ if\ Rule_k :\ Rule\ \ \ Then\ \boldsymbol{B} = Rule_k \\ if\ \ \ R_i \in Rule_k\ \ and\ Att_j\ \in Rule_k\ then\ \ \ I = 1\ else\ I = 0 \end{cases}$$

Table 1. KP1: Policy Context that represents the first policy[1]

	Source: Admission			Access Control Constraints				
	SSN	AdmissionDate	Department	Doctor	Nurse	Administrative	Pharmacist	Anesthetist
Rule 1	1	0	1	1	0	0	0	1
Rule 2	1	1	0	1	1	0	0	1
Rule 3	0	1	0	1	1	0	0	1

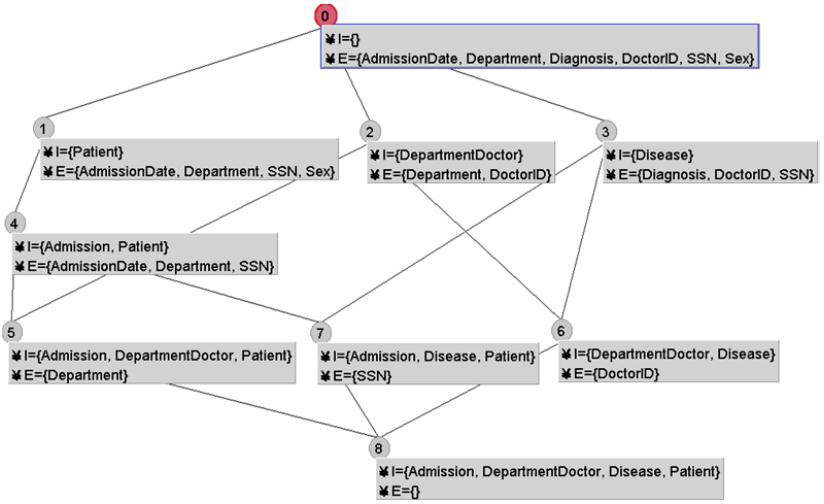

Fig. 2. Concept Data Lattice L^D of local sources and their attributes

The Figure 2 displays the Data Lattice L^D. It describes the local sources and theirs attributes. We can detect that the two attributes 'SSN' and 'AdmissionDate' come from two sources 'Patient' and 'Admission'. Indeed, the attribute 'Department' similarly exists at both sources 'Patient' and 'DepartmentDoctor'.

Preserved Rule Extraction: It considers each attribute individually in the Data Context K^D and extracts the corresponding access control rule set. Then, it identifies the shared profiles by all sources which have access to this attribute. Hence, the algorithm 1 takes as input the Data Context K^D and the Policy Contexts K^{Pi} (see table 1). For each attribute from K^D, it regroups the access control rules (line 2) of the attribute from K^{Pi} , and it respectively splits the obtained Attribute Policy Context on two matrices Attribute Matrix AM and Access Constraints Matrix CM (line 3,4). Then, the different rules obtained from the CM are combined, according to a **supremum** definition, to extract the profiles which must be derived at the global level (line 5-6). For instance, the preserved rule like '**Doctor, Nurse → SSN'** is made up of the profiles 'Doctor' and **'Nurse'**, and the attribute "SSN".

The algorithm 1 doesn't focus on the attribute combinations that can appear at the mediator level. So, we apply the following step to detect them and to retrieve rules that control this kind of combinations.

[1] A unique formal context is presented due to paper length requirements.

```
Input: Data Context K^D, Policies Context K^Pi
Output: a Preliminary Set Preserved Rules PR
1: foreach Attribute Att_i in K^D
2:  K^P_Atti =Extract rules associated to Att_i from K^Pi
3:  if (|K^P_Atti | > 2 )
4:     K^C_Atti=BinaryDecompose(K^P_Atti)
5:     if (supremum(K^C_Atti )# ∅)
6:        R^SA_G  = supremum( K^C_Atti) → Atti (*)
7:     endif
8:  else R^SA_G=ExtractRule(K^P_Atti)
9:  endif
10:   P^R ← P^R ∪ R^SA_G
11:endfor
```

Algorithm 1. Preserved Rule Extraction Algorithm

Attribute Combination Detection: It starts by detecting the intersection areas (Fig 3). Thereafter, it retrieves the shared attributes between sources that can be used for possible combinations. Thus, a combination between Department Doctor and Disease is ensured through DoctorID. A possible combination is {Department, Diagnosis}. By reapplying the algorithm 1, we obtain the rule "Doctor → SSN, Diagnosis" that control this kind of combination.

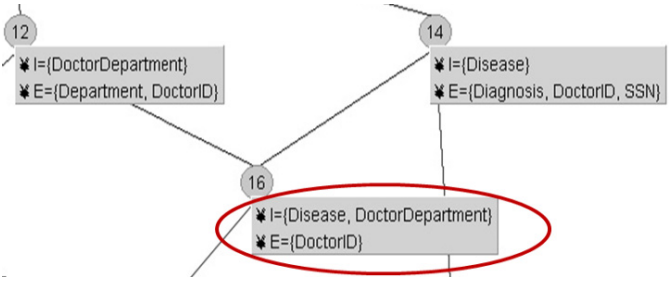

Fig. 3. Example of Attribute Combination

4.2 Step 2: Global Schema and Mapping Generation

In this step, we use the global policy to extract the attributes that belongs to the global schema. The step consists of global lattice generation, mapping derivation, and policy translation.

Global Lattice Generation: We use the work in [3] which presents a method of imposing constraints while extracting formal concepts. Virtual relations must contain the attributes used in the global policy at mediator level. Accordingly, we consider these visible attributes as constraints to generate the global schema.

Definition 9 Visible Attribute: Suppose that 'A' is an attribute. 'A' is a visible attribute, if 'A' has a global authorization rule that governs access to this attribute.

The Global Lattice is made up of a list of interesting concepts used in the global schema and mapping generation. It is composed of these concepts.

─C1:<Intent={SSN}, Extent={Admission, Disease, Patient}>
─C2:<Intent={SSN,Department,AdmissionDate},Extent ={Admission,Patient}>
─C3:<Intent={Diagnosis, DoctorID,SSN} ;Extent={Disease}>
─C4:<Intent={Department,DoctorID} ;Extent={ DoctorDepartment}>

Mapping Generation: It is performed using the GAV or LAV assumptions. The Figure 4-a (-b) describes the steps of the virtual relation generation and the GAV (LAV) mapping derivation using views as conjunctive rules [7]. First, it starts by building the virtual relation **G(X)** from the Concept Intent. Second, a conjunctive query **Q(X)** is built over the local sources (Concept Extent). Finally, a GAV mapping **M=G(X) ⊆ Q(X)** is generated (Fig.4-a-(3)). Whilst, the LAV mapping (Fig.4-b-(3)). **Mi=Si(X) ⊆ G(X)** is generated for each local source **S(X)** where Q(X) is conjunctive query over the global schema. For each lattice concept, the steps are performed while ensuring minimality [2] (i.e. no redundant relations appear in the global schema).

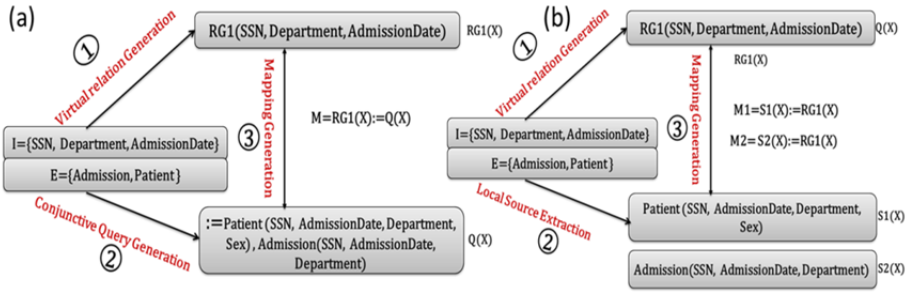

Fig. 4. GAV/LAV Mapping Generation Steps

Policy Transformation: Finally, the global policy obtained in step 1 (see section 4.1) is translated into a real access control policy. For each rule in the Global Policy, it generates a Global Authorization View. The global virtual relation which contains the attribute is a query part, and the profiles of the rule are the constraint part.

```
Input: P^G : Global Policy rules, G:Global Schema
Output: Policy^G: VBAC Policy
1: for each rule R_i ∈ P^G do //R has the form R=Cri-> Atti
2:  for each m_j∈ M do //m_i has the form G_i(Att):-S_1,..S_n
3:   if(R_i.att=m_j.G_i.Att) then
4:        R^P_k =GV_i_Authorization(R_i.att):=G_i R_i.C_i
5:        if R^P_k∉ Policy^G then add R^P_k to Policy^G
6:    endif
7:  endfor
8: endfor
```

Algorithm 6. CBAC Policy Translation Algorithm

In this paper, we propose an algorithm that translates a global policy into VBAC policy at the mediator level. This model offers very fine grained access constraints. It is the most suitable model at relational data integration and conjunctive query as mapping. This is a global authorization view example: **GV1_Authorization(SSN, Department):=RG1(SSN, Department, AdmissionDate), $Role= "Doctor".**

5 Conclusion

Based on the major advantages of our FCA-based approach (an access control policy-centric and a mapping language-independent solution), we intend to deal with semantic data by considering other issues, such as heterogeneities, data dependencies, and semantic constraints. We will also address the problem of consistency and compliance between global and local policies. Although, we will tackle the policy reconfiguration and query revocation to defeat inference problem that may appear at mediator level playing on the data dependencies.

References

1. Backes, M., Dürmuth, M., Steinwandt, R.: An algebra for composing enterprise privacy policies. In: Samarati, P., Ryan, P.Y.A., Gollmann, D., Molva, R. (eds.) ESORICS 2004. LNCS, vol. 3193, pp. 33–52. Springer, Heidelberg (2004)
2. Bellahsene, Z., Bonifati, A., Rahm, E. (eds.): Schema Matching and Mapping. Data-Centric Systems and Applications. Springer, Heidelberg (2011)
3. Belohlavek, R., Vychodil, V.: Closure-based constraints in formal concept analysis. Discrete Appl. Math. 161(13-14), 1894–1911 (2013)
4. Bertino, E., Sandhu, R.: Database security-concepts, approaches, and challenges. IEEE Trans. Dependable Secur. Comput. 2(1), 2–19 (2005)
5. Bertino, E., Jajodia, S., Samarati, P.: Supporting multiple access control policies in database systems. In: Proceedings IEEE Symposium on Security and Privacy, pp. 94–107 (May 1996)
6. Chen, B.C., LeFevre, K., Ramakrishnan, R.: Privacy skyline: Privacy with multidimensional adversarial knowledge. In: Proceedings of the 33rd International Conference on Very Large Data Bases, VLDB 2007, pp. 770–781 (2007)
7. Doan, A., Halevy, A.Y., Ives, Z.G.: Principles of Data Integration. M. Kaufmann (2012)
8. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations, 1st edn. Springer- Verlag New York, Inc., Secaucus (1997)
9. Haddad, M., Hacid, M.S., Laurini, R.: Data integration in presence of authorization policies. In: Min, G., Wu, Y., Liu, L.C., Jin, X., Jarvis, S.A., Al-Dubai, A.Y. (eds.) TrustCom, pp. 92–99. IEEE Computer Society (2012)
10. Pincus, J., Wing, J.M.: Towards an algebra for security policies. In: Ciardo, G., Darondeau, P. (eds.) ICATPN 2005. LNCS, vol. 3536, pp. 17–25. Springer, Heidelberg (2005)
11. Poelmans, J., Kuznetsov, S.O., Ignatov, D.I., Dedene, G.: Formal concept analysis in knowledge processing: A survey on models and techniques. Expert Systems with Applications 40(16), 6601–6623 (2013)
12. Rao, P., Lin, D., Bertino, E., Li, N., Lobo, J.: Fine-grained integration of access control policies. Computers & Security 30(2-3), 91–107 (2011)