A Semantics-Based Privacy-Aware Approach for Fragmenting Business Processes

Mohamed-Anis Zemni, Nejib Ben Hadj-Anouane
OASIS Laboratory
National School of Computer Sciences (ENSI)
Manouba, TUNISIA
mohamedaniszemni@gmail.com,
nejib_bha@yahoo.com

Moez Yeddes
CRISTAL Laboratory
National School of Computer Sciences (ENSI)
Manouba, TUNISIA
yeddes@yahoo.fr

Abstract—There is a growing need for the ability to fragment ones business processes effectively, in order to get useful fragments for future reutilization in building business processes. This can prove to increase the productivity and shorten the development time. The decomposition task aims at clustering workflow activities into fragments according to business constraints. Existing approaches lack semantic and privacy concerns. In this paper, we propose a semantic fragment identification approach to assemble activities that are semantically close according a semantic attraction threshold. Moreover, Fragments must be aware of sensitive information preserving. Our fragmentation approach is based on the so-called formal concept analysis approach, while integrating a semantic clustering technique for avoiding the association of sensitive information.

Keywords- Business Process, Fragmentation, Privacy-Aware, Semantic Activity Clustering.

I. INTRODUCTION

With the advent of new Web Technologies, companies tend more and more to improve the quality of their services in order to be competitive in terms of delivery deadlines and productivity. A key strategy to cope with the competition which becomes tough, is to organize their business processes in an agile manner. In [1], a business process (process for short) is made of a structured collection of objects, consisting mainly of activities and data objects respectively linked with control flows and data flows. It models the manner elementary operations should be organized to get a more complex functionality. Providing good designs for this business process, in short periods of time, may improve the quality of the services which can be further improved by taking into account important privacy features. Indeed, organizations have to organize their business effectively while managing privacy risks [2], as the individuals are becoming more and more concerned about the privacy of their personal data that may be seriously violated by malicious third parties.

However, a business process design, solely based on constructing processes completely from scratch, lacks flexibility as business processes should be implemented, optimized and tested in short periods of time. To get over such issues, designers can make profit of past user experience as demonstrated in web service selection [3]. Indeed, it is more efficient to reuse existing processes, rather than build the new process from scratch. This can prove to increase productivity and shorten development time [4]. Even better, process designers are keen to only select interesting fragments from existing processes which are easier to integrate when building new processes. Recent works have shown the relevance of this emerging approach [5], [6].

In [7], authors address two approaches to get business process fragments: developing them from scratch [8] or extracting the fragments from existing process models [9]. The second approach, we focus on in this paper, is more interesting as it better fits companies objectives but is now mainly carried out manually [10]. Existing approaches [11], [12], [13] for decomposing processes into fragments focus either on enhancing the process execution or protecting private information. Their main objective consists of outsourcing a snippet of the main process to third parties to perform the activities. Moreover, even if they handle structural concerns, these approaches lack privacy features and semantic concerns.

In our work, we propose an approach for semantically fragment existing business processes while taking into account privacy for sensitive information preservation.

Our approach is organized as follows:

1) We introduce a privacy-preserving constraint that avoids inferring sensitive information during the fragment reuse task. Indeed, privacy violations of sensitive data, represented via data objects and routed in between process activities by means of data flows, can occur by inferring information, through the construction of so-called sensitive associations of data belonging to multiple published fragments, typically handled within the same original business process boundaries (not intended for sharing or publishing).

2) We propose to annotate business processes with semantic features that the fragmentation technique has to deal with in order to deliver fragments made of activities that are semantically close to each other. Indeed, providing fragments made of semantically close activities improves their reutilization in new business
processes to fulfill a given functionality or part of it. 3) We define a privacy-aware fragmentation technique based on Formal Concept Analysis (FCA) [14] that semantically decomposes processes into privacy-preserving fragments, each involving a given functionality and ensuring the privacy of the sensitive association of data.

The present paper is organized as follows. In section II, we state related work on process decomposition and privacy concerns in the process management field. In section III, we propose our FCA-based extended approach that copes with privacy-constraints satisfaction while dealing with semantic decomposition. In section IV, implementation details are discussed. Section V concludes the paper and outlines some perspectives.

II. RELATED WORK

The decomposition of business processes has been addressed in [11], [12], [15]. In [11], [12], the decomposition task results into several coordinated partitions outsourced to different process engines. The partitions are defined at the design stage of the decomposition task where the activities are assigned to the execution sites manually. The aim of such approach is to enhance the execution of the original process. The communication infrastructure is then added among the newly created processes to exchange data and propagate the execution flow. The idea of distributed processes execution has also been tackled in [15]. The authors dynamically separate one integrated workflow model into small partitions at process runtime and allot them to different servers to be executed. Although these papers address business process fragmentation, reusing the resulted process fragments with privacy guarantee during fragmentation are not considered.

Writing processes or process fragments that can be reused in different places within a process or multiple processes has been considered in [9]. In this aim, a resolution of extending BPEL to support modularization and reuse has been proposed by using two different approaches: macros in WS-BPEL and the WS-BPEL on Event structure. In [8], process fragments are modeled and annotated with local information from scratch by people having fragmentary knowledge about the desired process model. However, modeling process fragments, in [8], from scratch is as critical as building new process models. Moreover, even if related works, analyzed above, relies on reusing process fragments, none of them addresses the privacy concerns while decomposing and reusing process fragments.

Privacy has been studied in various contexts: data mining, social networks, statistical databases, etc. However, there has been little investigation on privacy in business process area. The works in [13] and [16], propose a fragment identification approach for outsourcing, while dealing with some privacy issues; so-called predefined policies are used to restrict access to certain information, by the external parties, involved in fragment outsourcing business. However, policies are fixed following the external parties to which fragments will be assigned. [17], deals with privacy-aware workflows where sensitive data of subjects, who demand that their data should be protected are safe. This work, however, does not rely on fragmentation nor reutilization.

The focus of these works has been set on either ensuring privacy in workflows processing or process fragmentation to enhance the execution of the overall process. To the best of our knowledge, our work is the first to address the problem of semantic fragmentation while maintaining the privacy of sensitive information.

III. BUSINESS PROCESS FRAGMENTATION

In this section, we propose to semantically fragment a given business process model while preserving sensitive information from leakage. A business process fragment (or fragment for short), as defined in the literature [18], [7], [8], is a portion of a process destined for reuse purposes. It is a connected graph with relaxed completeness and consistency criteria, compared to a complete processes. A fragment depicts incomplete knowledge, and can be composed with other fragments for the purpose of building a complete and useful process. It is composed of at least one activity, and of several edges, defining control and data flows. As for the case of a complete process, a fragment should contain no cycles, and a single control flow linking up two distinct activities. A fragment can contain dangling control and data flows, i.e., arcs with either no source or target activities specified. For this, we propose privacy-preserving constraints to avoid sensitive information inference while rooting data objects between process model activities. We, then, propose to annotate activities with semantic features to enable computing activities’ attraction to enable semantically assembling them. Finally, we propose a privacy-aware fragmentation technique that provides semantically coherent fragments that may be useful for future reutilization.

A. Privacy-preserving Constraints

In this part, we detail some principles from an initial approach [19]. We seek to improve on our privacy enforcing mechanism. For this, we propose a privacy-preserving constraint to avoid disclosing sensitive associations between data objects (information) that are rooted between activities. A sensitive association is the association between non-sensitive data objects that are safe when they are considered independently, but turn out to be critical when they are associated together (e.g. name/illness). Indeed, reusing a fragment, that contains the data objects involved in a sensitive association while building new processes that contains some malicious activities, may infer the association. In our work, data objects are either activity inputs or activity outputs and are denoted by $E$. $E^* \subseteq E$ is the set of data objects involved in the sensitive association called, critical
data object set and \( E \setminus E^* \subseteq E \) is the set of neutral data objects. Activities that output some critical data objects in \( E^* \) are called critical activities. The rest of activities are called neutral. Activities that input some critical data objects in \( E^* \) are not critical but neutral because, while reused in a new process, these activities may receive other similar data objects that substitute the critical data objects.

These principles should be integrated in the fragmentation technique to ensure that distinct critical activities that output different critical data objects involved in the same sensitive association should never figure in the same fragment. For this, we propose the following privacy-preserving constraint

\[
C_N : E^* \times E^*, \text{ that protects the sensitive association between two critical data object sets.}
\]

**Definition 1 (Privacy-preserving Constraints).** \( C_N \) is a privacy-preserving constraint such as \( C_N = \{ (E_i^p \setminus E_j^p) | E_i^p \cap E_j^p \subseteq E^*, E_i^p \cap E_j^p = \phi \} \)

Notice that we use privacy-preserving constraint and privacy-constraint interchangeably.

Given a privacy-constraint set, \( C_N \), we classify activities into groups, that we call privacy-groups and denote \( G_{E_i} = (E_i^*) \) involved in \( C_N \) where \( G_{E_i} \) would enclose the activities that are not in conflict with \( E_i^* \), i.e. critical activities outputting some elements in \( E_i^* \), and neutral activities. Let us denote \( G = \bigcup G_{E_i} \) the set of privacy-groups.

Now, suppose that we have many constraints. When we generate the privacy-groups w.r.t. critical data objects in each privacy-constraint, some sensitive association discover are noticed. Indeed, a privacy-group, for a given critical data object, may enclose critical activities that are not in conflict with the group, but, handling other critical data objects of a single privacy-constraint. Thus, in case we have many constraints, and in order to avoid generating groups enclosing such activities, we propose the combination operator, \( \otimes \), combining the privacy-constraints before performing the extended FCA technique. Therefore, given two privacy-preserving constraints, \( (E_i^p \setminus E_j^p) \), and, \( (E_k^p \setminus E_l^p) \),

\[
\otimes C_N = (E_i^p \setminus E_j^p) \otimes (E_k^p \setminus E_l^p) = (E_i^p \setminus E_j^p) \otimes (E_k^p \setminus E_l^p) =
\]

\[
\left\{
\begin{align*}
((E_i^* \cup E_k^*) \setminus (E_j^* \cup E_l^*)) \land
((E_j^* \cup E_l^*) \setminus (E_i^* \cup E_k^*))
& \quad \text{if } E_i^* \neq E_k^*, E_j^* \neq E_l^*
\quad \text{and } E_i^* \neq E_j^*, E_k^* \neq E_l^*
\end{align*}
\right.
\]

Thus, we rewrite the privacy-constraint as:

\[
\otimes C_N = \{(E_i^p \setminus E_j^p) | E_i^p \subseteq E^*, E_i^p \cap E_j^p = \phi, E_i^p \cup E_j^p = E^* \}
\]

**B. Annotating Activities with Weighted Words**

Our work proposes a novel fragmentation approach that is based on semantic features where activities, that are semantically close to each others, are assembled into clusters. Indeed, it is easier to integrate a cluster of activities that fulfills the desired functionality in the new process. Thus, activities that belong to the same fragment have semantic attraction with the involved fragment functionality. The semantic attraction of an activity to a given functionality is basically computed from the description of the activity business (i.e. operation) and depicts the relevance of an activity for a functionality. For this, we propose to annotate the activities in the business process model with semantic features related to the description of the activity business. In [20], similar business processes, each defined as a mixture of topics (in our work, topics correspond to functionalities), are grouped into clusters using textual and structural features. Indeed, during the first level clustering, i.e. textual clustering, each process is described with a set of words retrieved from the process textual description (delivered with the process). Then, each process is assigned a given topic, from a topic database, that corresponds to the process description words. Similarly, we propose to annotate process activities with relevant words so that to enable computing their attraction with a given functionality. Relevant words may be manually defined by human or automatically retrieved, e.g. from activity textual description as demonstrated in keywords retrieval [21]. Let \( W \) be the set of relevant words that are used to describe the process activities. Ideally, in order to improve the computation of the activities attraction, words must follow a set of unifying features in a single format. Indeed, words should be semantically and syntactically different. For example, “medicine”, “drug” and “medication” are words having the same meaning, so only one of them should be used in the activity description. Also, the word “gift” should be used instead of the word “present” when “present”, i.e. meaning “existing”, is already used in the description. Likewise, words should be formed of alphanumeric unigrams but not fully numeric. The use of stems also ameliorates the attraction of activities. For example, “manag” is used instead of “management” and “managing”.

After defining the words format, we present how activities can be assembled intuitively using partial attraction. The partial attraction computes the attraction of an activity to a functionality. If the quotient of the number of common words between an activity and a functionality is over a small but relevant set of words, the computed attraction is smaller than the threshold \( \alpha_p \), then there is a partial attraction.

Notice that, activity attraction and partial attraction are used interchangeably. The partial attraction of an activity to a functionality is used to assemble activities into clusters. Therefore, activities of a given cluster are semantically close to each other. However, when an activity is richly described with words and the semantic attraction is over a small but relevant set of words, the computed attraction is smaller than the threshold \( \alpha_p \) (relatively, \( \alpha_c \)). Consequently, the activity is not important for the functionality processing and is not kept in the cluster corresponding to that functionality. This is the reason why we propose to assign a weight to words
used to describe a given activity. The weight corresponds to the relevance of the word for an activity. Therefore, the partial attraction of an activity to a functionality is computed over shared weighted words, indicating the relevance degree of the activity to the functionality.

C. Privacy-aware Semantic Fragmentation

In the following, we propose a technique to fragment a given business process into clusters made of semantically close activities while ensuring the privacy of sensitive associations.

The Formal Concept Analysis (FCA) [14], [22] is a data analysis technique, used for classifying objects within object collections, w.r.t., their common attributes. Due to the space limitations, we invite the readers to take a look to the cited references. To adapt the FCA technique to our needs, we need to fix the features that it is based on to generate fragments involving a given functionality. In our work, he proposed technique combines the concepts that have been previously defined in section (III-A) and section (III-B). Indeed, the privacy-constraints \( C_N \) and the words \( W \) used for activity annotations will respectively ensure the preservation of sensitive associations, and the semantic logic. Therefore, the FCA technique would be used to group activities and catch the main functionalities involved in the process while ensuring the preservation of sensitive associations. This technique has been used as a generic approach in [23] to describe the data objects for data compliance checking, structure analysis and fragmentation. Nevertheless, it is not as interesting in fragmentation as in data compliance or structure analysis because describing the data objects does not influence the sharing analysis processing, and consequently, the obtained fragments are the same as when data objects are taken without description.

A start point could be the translation of the process model, e.g. modeled in BPMN, annotated with description words. The formal context, defined in the FCA technique, depicts the fragmentation features definition. In the present work, objects correspond to activities \( A \). Activities are organized into privacy-groups \( G \) within which the combination of activities does not disclose sensitive associations. The activities’ properties can be mapped onto the Context Attributes. These latter are represented by the words that are used to describe the activities. Therefore, the activity clustering is performed within each privacy-group to avoid sensitive associations as each group contains non conflicting activities. A function is also defined to return the weight of a word for an activity within a privacy-group. This is formally defined as follows.

Definition 2 (Privacy-Aware Formal Context). A Privacy-Aware Formal Context is a tuple, \( C = (W, A, G, f) \), where:

- \( W \) is a word set as defined in subsection(III-B),
- \( A \) is an activity set,
- \( G \) is a privacy-group set as defined in subsection(III-A),
- \( f : A \times W \times G \rightarrow \mathbb{N} \) is a function that returns the weight of a word \( w \in W \), for a given activity \( a \in A \) within the group \( G_{E_k} \in G \).

When a privacy-aware formal context is represented with a 3D-table (arrows are activities, columns are words and the height correspond to privacy-groups), the weight is indicated in the cells. ‘0’ indicates that the activity does not use the corresponding word.

The FCA technique exposes two functions to check the attraction between activities according to their description words and their weights within privacy-groups. These functions represent the privacy-aware Galois Correspondence.

Definition 3 (Privacy-Aware Galois Correspondence). A Privacy-Aware Galois Correspondence involves two functions, \( \Theta \) and \( \Delta \), for a Privacy-Aware Formal Context \( C = (W, A, G, f) \). \( \Theta \) is defined over the subsets of activities \( A \) within a privacy-groups in \( G \), and returns the words that are shared by all the activities, so that, for a given \( A_i \subseteq A \), \( \Theta(A_i, G_{E_k}) = \{ w \in W | f(w, a, G_{E_k}) \neq 0, \forall a \in A_i \} \). \( \Delta \) is defined over the subsets of words \( W \), and returns relevant-enough activities, according to a fixed threshold \( \alpha_P \), within a privacy-groups in \( G \), that share all the words, so that, for a given \( W_j \subseteq W \), \( \Delta(W_j, G_{E_k}) = \{ a \in A | \sum_{w \in W_j} \frac{f(a, w, G_{E_k})}{\sum_{w \in \Theta(a, G_{E_k})} f(a, w, G_{E_k})} > \alpha_P, \forall w \in W_j \} \).

The function \( \Theta \) is used to define the functionality that the set of activities participate to, and the function \( \Delta \) fixes the boundaries, in terms of activities, of a given functionality.

Using these functions, we generate Privacy-Aware Formal Concepts where, for a given Privacy-Aware Formal Concept, all activities share the same word set and words are all shared by those activities.

Definition 4 (Privacy-Aware Formal Concept/Cluster). Given a privacy-aware formal context \( C = (W, A, G, f) \), a privacy-aware formal concept is a tuple \( Con = (W_j, A_i, G_{E_k}) \), where \( W_j \subseteq W \) is a set of words shared by each \( a \in A_i \subseteq A \) within \( G_{E_k} \) with \( \Theta(A_i, G_{E_k}) = W_j \), and, \( \Delta(W_j, G_{E_k}) = A_i \). \( Cl = A_i \) is a privacy-aware cluster derived from the privacy-aware formal concept \( Con = (W_j, A_i, G_{E_k}) \).

In our work, privacy-aware clusters correspond to final fragments where fragments boundaries are fixed by function \( \Delta \), and the functionality involved by the fragment is fixed by function \( \Theta \). Activities in a privacy-aware formal concept are relevant enough for the functionality described by the set of words. When the Privacy-Aware Formal Context is represented as a table, each privacy-aware formal concept refers to a maximal portion where all cells are not ‘0’ and all activities have partial attraction greater than \( \alpha_P \).
Example. Consider a process $P$ made of a set of activities $A = \{a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_9, a_{10}\}$, and a set of data objects $E$. We are interested by the data objects $e_1$ and $e_2$ that are respectively $a_2$ and $a_{10}$, and $a_6$ outputs. $e_1$ and $e_2$ are involved in a privacy-preserving constraint $C_N = \{(P \rightarrow C)\}$. This results into the privacy-group set $G = \{G_{\{e_1\}}, G_{\{e_2\}}\}$ where $G_{\{e_1\}}$ would enclose activities that are not in conflict with $\{e_1\}$ and $G_{\{e_2\}}$ would enclose activities that are not in conflict with $\{e_2\}$. Let $W = \{o, d, n, q, c, i, b\}$ be a set of words used to describe the activities. $C = (W, A, G, f)$ (illustrated in Figure 1) is the formal context corresponding to the process. Words are columns, activities are rows, table 1a and table 1b respectively correspond to the privacy-groups ($G_{\{e_1\}}$ and $G_{\{e_2\}}$). The numeric values correspond to the words' weight for the activities. The weights are given by function $f$. An activity having all words' weight 0 in a given table, means that the activity is in conflict with that group. Let us consider an attraction threshold $\alpha_P = 0.5$

With a classical FCA the activities $\{a_1, a_2, a_3, a_4, a_5\}$ would have been in a single fragment, as they all share the word 'o' in their description. However, activities $a_1$ and $a_2$ are not relevant enough for the fragment functionality described by 'o'. Indeed, their attraction is smaller than the threshold $\alpha_P = 0.5$. Consequently, they are not selected as part of that fragment. Moreover, the fragment $Cl = \{a_6, a_7, a_8, a_{10}\}$, generated with the classical FCA, is decomposed into several fragments to avoid disclosing the sensitive association, e.g. $Cl_1 = \{a_7, a_{10}\}$ and $Cl_2 = \{a_6, a_7\}$. Indeed, activity $a_6$ and activity $a_7$ were responsible for the association discloser between the data objects $e_1$ and $e_2$.

We have generated privacy-aware clusters that correspond to useful fragments. Indeed, each resulted cluster contains semantically close activities participating to a given functionality. The functionality is described by the word set used for activity description involved in the privacy-aware formal concept. We have derived clusters, with no sensitive association disclosure, depicted by the set of activities. However, some clusters are disconnected where some activities are not reachable from other activities. Such clusters cannot be reused in new processes and can be decomposed into several connected clusters. For this, we navigate within the clusters as in oriented graphs and check whether they are strongly connected or not. As processes are seen as graphs, we can make use of what is proposed in graph theory [24]. It is straightforward to check whether a cluster is connected or not and some existing algorithms can be adopted [25], [26].

Connected clusters are considered as final fragments. They are structurally consistent. They contain no circles and only one flow linking an activity to another one (i.e. ensured in the original process). Some flows are still dangling. Indeed, some target or source activities are missing. They will be fulfilled with activities when reusing the fragment to build a new process. Some techniques can be used to glue fragments into processes [8].

IV. IMPLEMENTATION

To implement our approach, a Colibri-java library 1, for classical Formal Concept Analysis, has been extended with a prototypical coding of the semantic and privacy extensions. To form the formal context, activities are retrieved from the business process model. Relevant words are generated using the activities textual description delivered with the process model. Numerical words and stopwords are removed based on a set of 425 stopword set [27]. Words are then unified using a 5 step stemmer [28]. In the present implementation, we have not considered semantic and syntactic distinctions between similar words and keep them for future work. The words weight for a given activity correspond to the occurrence of the words in the activities description. To ensure the preservation of the sensitive associations, the activities are classified into privacy-groups, obtained from the fixed privacy-constraints, as explained in Section III-A. Then, the fragmentation technique is performed on activities, with their corresponding weighted words, within each privacy-group, as many times as resulted privacy-groups. To generate formal concepts that correspond to fragments and the set of activities participating to the fragments functionality, we refine formal concepts to contain only activities that have partial attraction to the functionality represented with the set of words, figuring in the formal concept, over a given partial attraction threshold $\alpha_P$. This is performed with new additional modules. Notice that accommodating available FCA implementations to our needs is straightforward.

V. CONCLUSION

In this paper, we have provided safe fragments that may be useful when building new processes. Fragments may be useful as they contain semantically close activities participating for a given functionality. Fragments are safe as sensitive associations cannot be inferred by malicious third parties. Structurally speaking, the fragments are consistent

1http://code.google.com/p/colibri-java/source/browse/#svn%2Ftrunk%2Fcolibri
as they respect the fragments properties. The proposed FCA based technique can be furthermore improved to catch all the activities participating to even a part of a functionality by relaxing the formal concept definition. The results still be interesting and to the best of our knowledge, this is the first research on meaningful fragmentation while respecting privacy-concerns.

REFERENCES


