LGPD: A Formal Concept Analysis and its Evaluation

Antônio Diogo Forte Martins[1], **Patrícia Vieira**[1],
José Maria Monteiro[1], **Javam C. Machado**[1]

[1]Computer Science Department, Federal University of Ceará, Fortaleza-Ceará, Brazil

{diogo.martins, patricia.vieira

jose.monteiro, javam.machado}@lsbd.ufc.br

Abstract. *Nowadays, personal data generated through the use of smart devices are collected and stored by many companies. So, concerns about privacy issues are raised by users. In this context, the Brazilian General Law for the Protection of Personal Data (LGPD) was proposed in 2018. It regulates how any company must store and process personal data. This paper details the results of applying formal concept analysis (FCA) to the LGPD. The goal of applying FCA was to elicit key insights to support software development or re-design in compliance with LGPD. Besides, we propose an automatic manner to apply FCA.*

1. Introduction

Personal data generated through the use of different kinds of smart devices have been collected and stored by companies, which do not always act transparently. So, concerns about privacy issues are raised by users of Information Technology (IT) solutions, who have a keen interest in the processing and storing procedures of personal data by these organizations.

In this context, the Brazilian General Law for the Protection of Personal Data (LGPD) was proposed in 2018. It regulates how any company must store and process Brazilian citizens' personal data. Furthermore, companies that do not comply with the LGPD can be monetarily penalized, which underscores the seriousness of ensuring LGPD compliance. However, achieving LGPD compliance is not trivial, since the organizations must implement controls that track and manage personal data. This requires changes in the existing systems, as well as in the way of developing new soft ware. Besides, LGPD compliance must be proven by analyzing the actions that a system applies in order to gather, process, and safeguard personal data. Therefore, we argue that compliance must be considered in the design phase of a system.

This paper studies the problem of modeling the main concepts found in data protection regulations. This problem is particularly relevant for data holders, compliance auditing systems, and citizens as data owners in general. We focus our investigation on the Brazilian regulation, known as LGPD, to which we apply the formal concept analysis (FCA). FCA is a disciplined method of deriving a concept hierarchy from a collection of objects and their properties expressed in a structured text, such as a regulation. It allows us to formally reasoning about concepts, attributes, and implications hidden in regulation. The goal of applying FCA is to elicit critical insights to support software development in compliance with LGPD. In addition, our approach proposes an automatic manner to apply FCA by pre-processing the law text with a natural language processing tool. Our results, described in Section 4, are promising. We describe our next steps in Conclusion.

2. LGPD and Formal Concept Analysis

The LGPD regulates how any individual (whether natural or legal, under the public or private law) must store and process Brazilian citizens' personal data while respecting citizens' privacy. Furthermore, individuals that do not comply with the LGPD can be monetarily penalized, which underscores the seriousness of ensuring LGPD compliance. With the LGPD, Brazil became part of the countries that have specific legislation to protect data and their citizens' privacy. This legislation was formulated based on the General Data Protection Regulation (GDPR), the European Union's Data Protection Regulation. It is composed by 65 articles, organized in 10 chapters. LGPD defines a set of rights for data holders and establishes rules and limits for companies regarding personal data collection, storage, processing, and sharing, especially in digital media, in order to protect the fundamental rights of freedom, privacy, and the free formation of the personality of each individual. LGPD can be seen as a general guideline for data protection in Brazil.

Formal Concept Analysis (FCA) [Ganter et al. 2005] is a technique of conceptual modeling which can capture how objects (concepts) can be hierarchically clustered based on the attributes they have in common. The input for applying the FCA technique is a cross-table (formal context) that can be represented by a triplet *(G,M,I)* where the concepts *(G)* are the rows and the attributes *(M)* are the columns. *I* denotes a binary relation between G and M, in other words, if concept g, an element of G, has attribute m, an element of M. The FCA technique produces two main outputs. First, the lattice of formal concepts, also known as, concept lattice which is built from hierarchical relation order of the formal concepts by means of subconcept-superconcept [Bogatyrev 2016]. Second, the attribute implications from the concept lattice, an implication describe a dependency between attributes within the data. One of the main advantages of this technique is the high level of abstraction which makes it useful and suitable to work with different data types. In this paper, we work with textual data which is possible to extract concepts and their attributes in order to build the formal context [Carpineto and Romano 2004].

3. Related Works

The state of the art in data privacy regulations is upon a growing progress. In [Tamburri 2020], the authors describe the results of applying formal concept analysis (FCA) to the General Data Protection Regulation (GDPR). The FCA analysis revealed many key insights about the regulation. Practitioners can use these insights to prioritize their redesign and refactoring for compliance campaigns.

[Basin et al. 2018] proposes an approach that identifies an objective with a business process showing how formal models of communication between processes can be used to audit or even derive privacy policies. Based on this insight, a methodology has been proposed to audit GDPR compliance, in addition to showing how GDPR compliance processes can be determined algorithmically. In [Gjermundrød et al. 2016] a privacy tracker was proposed. It consists of a structure based on the principles of GDPR, in which users can control their personal data. In [Pandit and Lewis 2017], a linked data ontology for expressing provenance of consent and data lifecycles with a view towards documenting compliance. In [Allegue et al. 2019], the authors propose a security model for data privacy and an original solution where a GDPR consent manager is integrated using Complex Event Processing (CEP) system and following the edge computing.

4. Applying FCA to LGPD

In order to apply FCA to the LGPD regulation, we build the LGPD formal context, as a cross-table, using an automatic approach to identify concepts, attributes and the relationship between them. We use this cross-table as input for the FCA tools to retrieve the formal concepts, build the concepts lattice, find association rules, and identify Voronoi partitions in the concepts lattice. Figure 1 shows the general workflow of our research.

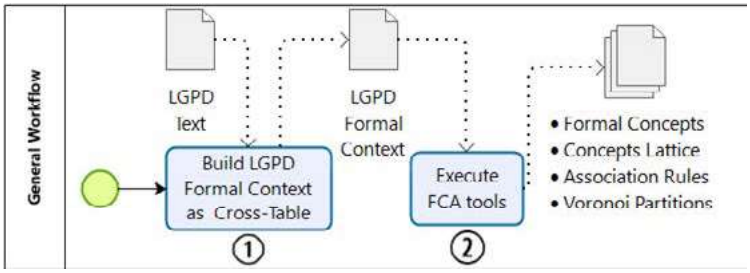

Figure 1. General workflow.

4.1. Building the Cross-Table

The input of an FCA analysis is a formal context, as a cross-table, including the relationship between concepts and attributes. Building a formal context for textual data is an exhausting and error pruning activity, since it relies on human interpretation to find what is considered to be a concept or an attribute. We propose an algorithm to perform the first process of the workflow that automatically builds the formal context. The algorithm uses a natural language processing (NLP) technique known as Part-Of-Speech (POS) tagging which is able to return the grammatical class of a word. In our technique nouns are considered as *concepts*; adjectives, adverbs and verbs are considered as *attributes*. We used spaCy [Honnibal and Montani 2017], a python module, with the *pt_core_news_lg* as language model to perform all the NLP tasks we need in order to build the cross table. Alongside POS tagging, we also use the dependency parser available in the module to find the relationship between the words in the LGPD text. Algorithm 1 describes the building process of the cross-table for the LGPD textual data. We also use Pandas[McKinney et al. 2010] python module to handle the cross-table data. We found 337 concepts and 408 attributes within the LGPD text.

Figure 2 contains an example of an article of the LGPD with the words classified as attributes or concepts. Figure 3 shows an example of a formal concept extracted from the LGPD.

4.2. Identifying Formal Concepts

We apply FCA to the formal context of the LGPD textual data to identify the formal concepts in order to build the concept lattice. A formal concept is a set of concepts that share the same attributes in the formal context.

After performing the second process of the workflow, we found 920 formal concepts in the LGPD formal context with the assistance of *Con-Exp NG*[1], a tool to support tasks of FCA (workflow process 2). The highlighted area in Figure 3 shows an example of one of the formal concepts identified by the tool.

[1]https://github.com/fcatools/conexp-ng

Algorithm 1 Cross-table automatic build.

Input: Text processed as one paragraph
Output: LGPD cross-table

Apply spaCy NLP to text input.
for each word **in** processed text **do**
  if word POS tag in concepts POS tag list **then**
    Add word to concepts list.
  else if word POS tag in attributes POS tag list **then**
    Add word to attributes list.
  end if
end for
Create a pandas DataFrame with concepts as rows and attributes as columns, and initialize all values with 0.
for each concept **in** concepts list **do**
  Get concept's dependency list.
  if attribute in dependency list **then**
    Set relation between concept and attribute (DataFrame[concept, attribute] = 1).
  end if
end for
Merge all concepts with same lemma in a single row.
Merge all attributes with same lemma in a single column.

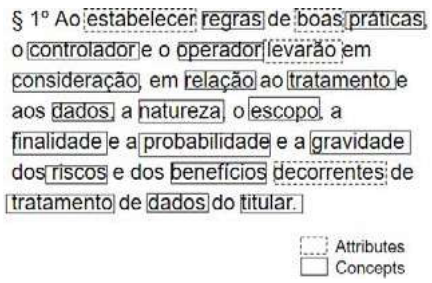

Figure 2. Text analysis.

Concepts/Attributes	observadas	trata	tem
informação	X	X	X
medida	X	X	
multa	X	X	
finalidade	X	X	X
dado	X	X	X
titular	X	X	X
acordo	X	X	
casos	X	X	
termos	X		X
interesse	X		X
tratamento	X		X
direito	X		X

Figure 3. Formal concept example found in the LGPD formal context.

4.3. Building the Concept Lattice

The FCA concept lattice express the hierarchical relation of the formal concepts ordered by the subconcept and superconcept relations. We used *Con-Exp NG* and *Concept Explorer FX*[2] to assist the creation and visualization of the concept lattice. Figure 4 shows the concept lattice for the formal context depicted in Figure 3. The blue and black boxes represents attributes and concepts, respectively, of a formal concept represented by a circle. Fully colored circles represent a fully connected formal concept, top-colored are the intents and bottom-colored the extents of the formal concept.

4.4. Identifying the Voronoi Partitions

We conduct Voronoi partitioning [Lee and Yang 2005] on the concept lattice to identify the number of distinguishable design rules and to investigate LGPD's complexity. After analysis, we found 21 Voronoi partitions. From the software and system design point of view, this means that at least 21 groups of formal concepts need specific and explicit requirements [Cai et al. 2019] to be implemented to guarantee LGPD compliance.

[2]https://github.com/francesco-kriegel/conexp-fx

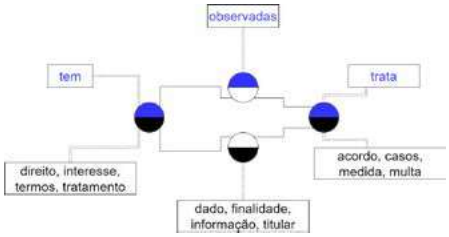

Figure 4. Concept lattice.

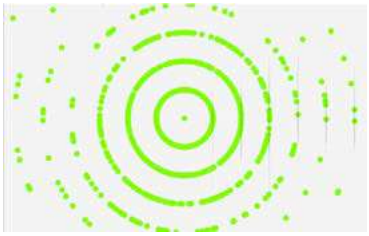

Figure 5. Polar layout.

4.5. Analyzing the Relationship Between Concepts and Attributes

Figure 5 shows the formal concepts in a polar layout. This diagram suggests 7 levels of implication within the LGPD. The formal concepts relation get weaker at each level. This result reveals that LGPD has different levels of compliance that can be addressed incrementally. For instance, software developers can use these levels of compliance to refactor their systems following the formal concepts descriptions from the innermost circle, supremum, to the most external level of compliance.

In order to find association rules between concepts, we applied the Apriori algorithm using two different metrics, Support, and Confidence, to select the most strong ones. Support metric calculates the proportion of transactions that contain the set of items, showing their importance and significance. Confidence is an indication of how often the rule has been found to be true. Table 4.5 shows the set of selected rules.

Table 1. Association rules found for concepts.

Rule	Antecedent Support	Consequent Support	Rule Support	Confidence
titular ⇒ dado	0.06	0.13	0.03	0.42
informação ⇒ dado	0.06	0.13	0.02	0.39
titular ⇒ controlador	0.06	0.08	0.02	0.35
direito ⇒ dado	0.07	0.13	0.02	0.32
controlador ⇒ dado	0.08	0.13	0.02	0.28
controlador ⇒ titular	0.08	0.06	0.02	0.28
tratamento ⇒ dado	0.10	0.13	0.02	0.25
dado ⇒ titular	0.13	0.06	0.03	0.20
dado ⇒ tratamento	0.13	0.10	0.02	0.18
dado ⇒ controlador	0.13	0.08	0.02	0.16
dado ⇒ direito	0.13	0.07	0.02	0.16
dado ⇒ informação	0.13	0.06	0.02	0.16

5. Conclusion

This paper described the results of applying formal concept analysis (FCA) to the Brazilian Data Protection Regulation, LGPD. Unlike other studies over GDPR, we searched to automatically generate the formal concepts to our analysis method. This approach is less costly and can assist a specialist in building the cross-table to FCA analysis. We also applied an Apriori technique to reason about the association rules between the concepts found in the protection regulation. That opens several new interesting problems we plan to tackle. For instance, what can one learn from the association rules? Is there any conflicting regulation in the law? Can we list every single responsibility for data holders? What about the data owner's rights? What type of data can an organization publish or

share with its partner without breaking the law? We hope our future investigations will help citizens, system developers as data holders or processors, and auditing organizations to better understand the Brazilian data protection regulation.

References

Allegue, S., Rhahla, M., and Abdellatif, T. (2019). Toward gdpr compliance in iot systems. In *International Conference on Service-Oriented Computing*, pages 130–141. Springer.

Basin, D., Debois, S., and Hildebrandt, T. (2018). On purpose and by necessity: compliance under the gdpr. In *International Conference on Financial Cryptography and Data Security*, pages 20–37. Springer.

Bogatyrev, M. (2016). Conceptual modeling with formal concept analysis on natural language texts. In *DAMDID/RCDL*.

Cai, Y., Xiao, L., Kazman, R., Mo, R., and Feng, Q. (2019). Design rule spaces: A new model for representing and analyzing software architecture. *IEEE Transactions on Software Engineering*, 45(7):657–682.

Carpineto, C. and Romano, G. (2004). Exploiting the potential of concept lattices for information retrieval with credo. *J. UCS*, 10:985–1013.

Ganter, B., Stumme, G., and Wille, R. (2005). *Formal concept analysis: foundations and applications*, volume 3626. springer.

Gjermundrød, H., Dionysiou, I., and Costa, K. (2016). privacytracker: a privacy-by-design gdpr-compliant framework with verifiable data traceability controls. In *International Conference on Web Engineering*, pages 3–15. Springer.

Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Lee, I. and Yang, J. (2005). Voronoi-based topological information for combining partitioning and hierarchical clustering. In *2005 International Conference on Computational Intelligence for Modelling Control and Automation (CIMCA 2005), International Conference on Intelligent Agents, Web Technologies and Internet Commerce (IAWTIC 2005), 28-30 November 2005, Vienna, Austria*, pages 484–489. IEEE Computer Society.

McKinney, W. et al. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX.

Pandit, H. J. and Lewis, D. (2017). Modelling provenance for gdpr compliance using linked open data vocabularies. In *PrivOn@ ISWC*.

Tamburri, D. A. (2020). Design principles for the general data protection regulation (gdpr): A formal concept analysis and its evaluation. *Information Systems*, 91:101469.