Explicit versus Tacit Knowledge in Duquenne-Guigues Basis of Implications: Preliminary Results

Johanna Saoud, Alain Gutierrez, Marianne Huchard, Pascal Marnotte, Pierre Silvie, Pierre Martin

Explicit versus Tacit Knowledge in Duquenne-Guigues Basis of Implications: Preliminary Results

Johanna Saoud[1], Alain Gutierrez[1], Marianne Huchard[1], Pascal Marnotte[2], Pierre Silvie[2,3], and Pierre Martin[2]

[1] LIRMM, Univ Montpellier, CNRS, Montpellier, France
johanna.saoud@etu.umontpellier.fr,
{marianne.huchard,alain.gutierrez}@lirmm.fr
[2] CIRAD, UPR AIDA, F-34398 Montpellier, France
AIDA, Univ Montpellier, CIRAD, Montpellier, France
{pascal.marnotte,pierre.silvie,pierre.martin}@cirad.fr
[3] PHIM Plant Health Institute, Montpellier University,
IRD, CIRAD, INRAE, Institut Agro, Montpellier, France

Abstract. Formal Concept Analysis (FCA) comes with a range of relevant techniques for knowledge analysis, such as conceptual structures or implications. The Duquenne-Guigues basis of implications provides a cardinality minimal set of non-redundant implications. The concern of a domain expert is to discover new knowledge within this implication set. The objective of this prospective paper is to collect and discuss the different patterns of implications extracted from a dataset on plants used in medical care or consumed as food. We identify 16 patterns combining 3 types of knowledge elements (KE). The patterns highlight redundant KEs, or KEs of little interest, in particular, those corresponding to plant taxonomy, as it is familiar knowledge for the experts. Removing these KEs from the implications would make them tacit. We suggest a post-process for cleaning up the implications before reporting them to the experts. In addition, we discuss the different patterns and how an implication classification based on patterns could help the experts.

Keywords: Formal Concept Analysis · Duquenne-Guigues basis · Implication Rules · Life Sciences Knowledge Base · One Health

1 Introduction

Formal Concept Analysis (FCA) is a mathematical framework based on lattice theory which aims to formalize the notion of concept [6]. It gives foundations for a large range of methods for knowledge processing and knowledge discovery [13]. These methods include the construction of formal concepts and their ordering in a concept lattice or in restricted sub-structures of the lattice. For a domain expert (e.g. pathologist, entomologist), navigating through a complex lattice to extract knowledge may remain a challenge. An alternative view on knowledge is

building the Duquenne-Guigues basis (DGB) of implications [8]. An interest of this implication basis is its formulation of pieces of knowledge using a compact and comprehensive formalism. DGB indeed provides a cardinality-minimal set of non-redundant implications.

Through DGB building, expert concern is to discover new knowledge within the implication set. Diverse situations can occur. For instance, if an implication is too obvious for the expert, e.g. because it exclusively describes a domain taxonomy, then it presents a too limited interest to be kept in the implication set. This implication therefore becomes a tacit knowledge, while the others remain explicit. In more complex situations, the implication may contain both obvious knowledge elements (KE) and useful KE. In such cases, the obvious part of the implication may become tacit, when the other part may remain explicit.

The objective of this paper is to observe and discuss different patterns of implications extracted from a dataset on plants used in medical care or consumed as food. With these patterns, we aim to identify obvious KE included in the implications, and how implications can be simplified. The patterns may also provide an opportunity to classify the implications into coherent sets. Section 2 introduces the background and the dataset. Section 3 presents and discusses the preliminary results. Section 4 concludes and draws future work.

2 Background and Dataset

Background FCA elaborates knowledge, including formal concepts or attribute implications, on top of a formal context (FC) $\mathcal{K} = (G, M, I)$ where G is an object set, M is an attribute set and $I \subseteq G \times M$. An implication, denoted by $A \implies B$, is an attribute set pair $(A, B), A, B \subseteq M$ such that all objects owning the attributes of A (premise) also own the ones of B (conclusion): $\{g|\forall m_a \in A, (g, m_a) \in I\} \subseteq \{g|\forall m_b \in B, (g, m_b) \in I\}$. There are several types of implication bases [1]. Here we consider the Duquenne-Guigues basis (DGB) of implications [8], which is a cardinality minimal set of non-redundant implications, from which all implications can be produced. An implication is held (or supported) by a number of objects, that we call the implication scope (S). The support is the proportion of such supporting objects. Let $Imp = A \implies B$, $S(Imp) = |\{g|\forall m \in A, (g, m) \in I\}|$. $Support(Imp) = S(Imp)/|G|$. For this work, the DGB of implications is built on a FC using Cogui software platform[4], which includes a Java implementation of LinCbO [11].

The dataset and the taxonomic knowledge To conduct the evaluation, we use an excerpt of the Noctuidae dataset [14], which is itself part of the Knomana dataset [15]. This dataset draws particular attention of experts in the context of One Health initiative [12] for addressing the worrying worldwide invasion of Spodoptera frugiperda (Lepidoptera from the Noctuidae family) which was first detected in Africa in 2016 [7] and is continuously spreading. At the end of 2018, S. frugiperda was first found in Yunnan Province in China [17]. Furthermore, in

[4] http://www.lirmm.fr/cogui/

2018, it was first recorded in South Asia, namely India [9]. In January 2020, it was trapped in Australia's special biosecurity zone in the Torres Strait islands of Saibai and Erub, and confirmed on 3 February 2020, and on mainland Australia in Bamaga on 18 February 2020 [16].

Table 1. On the top, excerpt of [14] that describes organisms uses, and, on the bottom, the associated formal context (FC) after the nominal scaling of *Species*, *Genus*, *Family*, *food*, and *medical*. *S*, *G*, and *F* are short notation for Species, Genus, and Family.

Organism	Species	Genus	Family	food	medical
p1	AcorusCalamus	Acorus	Acoraceae	no	yes
p2	CychoriumIntybus	Cychorium	Asteraceae	yes	yes
p3	AchilleaCollina	Achillea	Asteraceae	no	no

FC	S_Acorus Calamus	S_Cychorium Intybus	S_Achillea Collina	G_Ac orus	G_Cy chorium	G_Ach illea	F_Aco raceae	F_Aste raceae	food	no-food	medical	no-medical
p1	X			X			X			X	X	
p2		X			X			X	X		X	
p3			X			X		X		X		X

This dataset indicates for each plant organism, out of the 600 in the dataset, its species, its genus, its family, and whether it is consumed as food and used in medical care. There is one-to-one mapping between organisms (objects) and *Species* values (cf. Tab. 1). Species, genera, and families respect a taxonomy which is a 3-level tree structure with this general shape: Species $\prec$ Genus $\prec$ Family; E.g. Species Acorus calamus $\prec$ Genus Acorus $\prec$ Family Acoraceae (see Fig. 1). This taxonomy is a familiar knowledge for the experts. The dataset comprises 600 species from 376 genera and from 98 families.

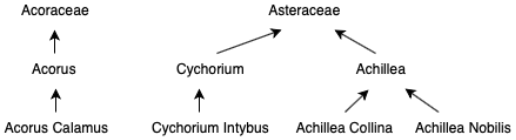

Fig. 1. Example of taxonomy for plants. Family, genus, and species names are respectively presented on the top, in the middle, and at the bottom of the figure. An arrow represents a generalization relation.

Formal context built from the dataset A FC describes a set of objects using a set of Boolean attributes. When the dataset contains a multi-valued attribute, conceptual scaling can be used to obtain Boolean attributes [6]. Various conversion methods are adopted, among which the nominal scaling for categorical attributes, such as the species name (e.g. Achillea collina or Acorus calamus), where each value is converted into a Boolean attribute [10]. Applied to this

work, the conversion of the dataset as a FC consisted in the nominal scaling of the attributes *Species*, *Genus*, *Family*, *food*, and *medical*. The taxonomy KE is expressed, in the FC (G, M, I), by the fact that for a plant organism $p \in G$ and a given species s from genus g and family f, with s, g *and* $f \in M$, if $(p, s) \in I$, then $(p, g) \in I$. Similarly, if $(p, g) \in I$ then $(p, f) \in I$. In addition, the dual of the *food* (i.e. *no-food*) and *medical* (i.e. *no-medical*) attributes are added in the FC to explicit respectively the fact to be not consumed (*no-food*) or not used in medical care (*no-medical*). Note that, in [14], the attribute *medical* is encoded by *Medical_X*, *no-medical* by *Medical_*, *food* by *Food_X*, and *no-food* by *Food_*.

3 Results and Discussion

As noticed in [4], the implications from the DGB are not redundant one with the others. But they may contain redundant attributes in the premise and in the conclusion, due to the fact that pseudo-intents are used instead of minimal generators. Besides, we can expect that implications respect a limited number of patterns, and that some of these patterns have different meanings. A long-term objective of this work is to provide the experts with a minimal set of implications filtered and classified to assist them in the analysis. In this section, we observe patterns in implications of DGB. Then, we discuss how implications may be post-processed and classified making them more appropriate to the domain expert.

3.1 The implications from DGB

Due to the scaling of the attributes *Species*, *Genus*, and *Family*, the FC associated to the dataset has 1078 Boolean attributes, i.e. 600 attributes to inform on the species, 376 on the genus, 98 on the family, and 4 on the medical and food use. For the 600 plant organisms, Table 2 shows that 1168 implications were extracted from this FC. Most of the implications, i.e 1007, are held by one object, and thus are specific to a plant species. Among the 161 remaining ones, 9 implications are supported by more than 9 objects. The maximum scope, i.e. 35, corresponds to the implication informing that none of the 35 species from the Meliaceae Family, present in the dataset, is consumed (cf. ID 1 in Table 3).

Table 2. Number of implications per scope.

Scope	1	2	3	4	5	6	7	8	10	11	16	18	29	35
#Implications	1007	76	37	18	12	3	1	5	3	2	1	1	1	1

These 1168 implications are formulated using 16 patterns, where a pattern corresponds to the pair (Premise, Conclusion) in which each of the declarative sentence is designed using the SGFp schema (Table 3). This schema is the ordered list of presence of the attributes *Species* (S), *Genus* (G), *Family* (F), *food* or *no-food*, *medical* or *no-medical* (p), in the declarative sentence. By grouping *food*,

no-food, *medical* and *no-medical*, our intent is to focus our analysis on the types of KEs, and not the KEs themselves.

Various combinations of S, G, F, and p can be observed in the premise or the conclusion. For instance, pattern 1 informs on the use of all plants from a family. Pattern 2 provides the taxonomic relation of a genus with a family. Some patterns are more extended, such as pattern 5 that states on the genus of plants from a family with a given use.

Table 3. Implication patterns from DGB. The premise and the conclusion are designed using the SGFp schema. KU, KT, and KD are respectively short notations for Knowledge on plant Use, Knowledge on plant Taxonomy, and Knowledge on the Dataset.

ID	Premise	Conclusion	knowledge elements	Example of implication	#implications	Max scope
1	F	p	KU	F_Meliaceae ⇒ no-food	35	35
2	G	F	KT	G_Salvia ⇒ F_Lamiaceae	12	18
3	Fp	p	KU	no-food,F_Annonaceae ⇒ no-medical	10	16
4	G	Fp	KU, KT	G_Trichilia ⇒ no-food,F_Meliaceae	84	10
5	Fp	G	KU, KD	food,medical,F_Rutaceae ⇒ G_Citrus	6	5
6	F	G	KD	F_Piperaceae ⇒ G_Piper	1	5
7	GFp	p	KU, KT	medical,F_Asteraceae,G_Artemisia ⇒ no-food	7	4
8	Fp	Gp	KU, KD	medical,F_Annonaceae ⇒ food,G_Annona	1	3
9	F	Gp	KU, KD	F_Lythraceae ⇒ no-food,no-medical,G_Lythrum	5	2
10	S	GFp	KU, KT	S_ZygophyllumAlbum ⇒ no-food,no-medical,G_Zygophyllum,F_Zygophyllaceae	600	1
11	G	SFp	KU, KT, KD	G_Zygophyllum ⇒ no-food,no-medical,S_ZygophyllumAlbum,F_Zygophyllaceae	280	1
12	Fp	SG	KU, KT, KD	medical,no-food,F_Zingiberaceae ⇒ S_HedychiumCoronarium,G_Hedychium	29	1
13	GFp	S	KU, KT, KD	medical,no-food,G_Cinnamomum,F_Lauraceae ⇒ S_CinnamomumCassia	38	1
14	F	SGp	KU, KT, KD	F_Zygophyllaceae ⇒ no-food,no-medical,S_ZygophyllumAlbum,G_Zygophyllum	42	1
15	Fp	SGp	KU, KT, KD	food,F_Lauraceae ⇒ medical,G_Cinnamomum,S_CinnamomumVerum	3	1
16	GFp	Sp	KU, KT, KD	medical,F_Solanaceae,G_Solanum ⇒ food,S_SolanumLycopersicum	15	1

3.2 Observing patterns and implications

Three types of KEs were identified in the implications. KU type informs on the relationship of a plant, at any taxonomic level, with a use as food or medical care. The second KE type is the Taxonomic relationship type (KT), such as giving the family in the conclusion when the species is indicated in the premise. The third type (KD) corresponds to a KE resulting from the content of the Dataset, which represents a limit in this work. For instance, taxonomic referential web sites list 5 genera from the Piperaceae family [5]. As only one is present in the dataset (i.e. Piper), inferring that "*a plant from the Piperaceae family is of the genus Piper*" is wrong in the real life, but is true in this work as it results from a side effect of the dataset.

Except for patterns 2 and 6, all the implication patterns include KU (Table 3), suggesting at first sight that the latter are useful implications for the expert. But attention should be paid on implications combining KU, KT, or KD, and respected by implications with a scope value of 1, meaning that each one is associated to a single plant organism. Pattern 10 (scope 600) only reports information from the context as these rules only describe an organism by its

[5] E.g. `http://www.plantsoftheworldonline.org/` lists 5 plant genera.

attributes. Similarly, patterns 11 and 14 indicate that a given genus or family has only one species in the dataset, the other attributes being those of the species, and these patterns could be considered as containing only KD. Most of the patterns include KT. Including the taxonomy was crucial in this work for FC processing in order to discover knowledge at a higher generic level, but corresponds to a redundancy in the implications.

This preliminary analysis gives directions for pattern and implication post-processing and classification. Some may be specific to some characteristics of Knomana, such as the fact that attribute species is an object identifier, and some could be generalized to all datasets.

The post-processing may have different forms for redundant or evident information: either removing it, or simply separating it from the rest, so that it remains written but not distracting. KU, KD, KT can be highlighted in different ways to distinguish them. Highlighting the different reasons under redundancy or evident information (e.g. KT information versus logical redundancy due to the fact that minimal generators are not used) would be useful. We could consider removing redundancy only in premise [5] or only in conclusion, or in both.

Patterns also have to be analyzed. For example, pattern 4 (G -> Fp) could be simplified as G -> p, as F is tacit given G. This would change the pattern classification (initially KU KT), as it is now reduced to KU.

As regard with implications, a post-process could remove KT from the implications, corresponding to a tacit knowledge as experts are familiar with the taxonomy. As the redundancy due to KT may appear in the premise (e.g. *the species and its genus*), in the conclusion, or in both (e.g. *a species implies a family*), the post-process has to consider the implication in its entirety. This approach differs with [5] where authors consider exclusively the left-minimal premises for technical purpose (i.e. a fast computation of attribute closure and a minimal left hand side in the implication). In addition, a filtering could lead to remove pattern 2 implications that only express tacit knowledge.

KUs are the explicit KEs investigated by the expert and thus have to be put forward. KDs present a particular situation in the implications as they result from a lack of knowledge in the dataset. Pattern 6 has to be considered carefully as it contains only KD. Thus, the experts have to be alerted of KD presence to consider this aspect in the dataset analysis.

A classification of implications based on patterns seems to us relevant for presenting them to the expert by groups having a coherent meaning: E.g. implications providing information on the diversity of some plant families in the dataset or pure information on the One Health Approach. Depending of the number of rules in some categories, a classification may have a significant impact for the expert, e.g. discarding or at least separating rules from pattern 10 distinguishes 600 rules that only recall initial data. We also guess that if a post-processing is made, the way it is made has to be notified to the expert and it should be indicated if this is reversible operation.

Finally, literature on association rules also faced the issues of extension of non-redundant rules [18] and redundancy removal introduced by using a taxon-

omy in concept-based rules building [3]. Classifying association rules in a lattice has been addressed in [2] in the context of fault localization, where rules are described by elements of their premises, that can be inspiring in our case, using patterns as an implication description.

4 Conclusion

This paper identifies 3 types of knowledge elements and 16 patterns that constitute the implications from the Duquenne-Guigues basis on a formal context. Each implication needs a specific consideration before being presented to the expert. For instance, a post-process can be conducted to remove tacit knowledge elements from implications, which may drive to delete some of them.

In a future work, we will study how the different patterns can be used to display implications to experts by categories, that may help them to focus on different aspects of the dataset. For instance, the user may focus on knowledge elements related to some plant families in the dataset, or pure knowledge elements on the plant uses at the family level.

This work is a preliminary study to the analysis of the Duquenne-Guigues basis of implications resulting from a Relational Context Family of the Knomana knowledge base. The general objective is to contribute in the decision support process to identify plants that could be used by farmers to control pest. These pesticidal plants will be an alternative to pesticide and antibiotics, considering the One Health approach. Using this approach, one must be aware of the multi-uses of these plants to prevent the intentional effects on the animals, the humans, and their environment.

In a more general perspective, it would be relevant to examine how the forms of post-processing, filtering and classification of patterns and rules can be generalized in order to be able to apply these approaches to other datasets, more particularly when several taxonomic relations are involved.

Acknowledgments. We warmly thank the reviewers for their comments. Part of the discussion has been notably improved thanks to their relevant remarks. This work was supported by the French National Research Agency under the Investments for the Future Program, referred as ANR-16-CONV-0004.

References

1. Bertet, K., Demko, C., Viaud, J.F., Guérin, C.: Lattices, closures systems and implication bases: A survey of structural aspects and algorithms. Theor. Comp. Sci. **743**, 93–109 (2018)
2. Cellier, P., Ducassé, M., Ferré, S., Ridoux, O.: Dellis: A data mining process for fault localization. In: Proceedings of the 21st International Conference on Software Engineering & Knowledge Engineering (SEKE). pp. 432–437 (2009)
3. Cellier, P., Ferré, S., Ridoux, O., Ducassé, M.: A parameterized algorithm to explore formal contexts with a taxonomy. Int. J. Found. Comput. Sci. **19**(2), 319–343 (2008)

4. Cordero, P., Enciso, M., Mora, A., Ojeda-Aciego, M.: Computing minimal generators from implications: a logic-guided approach. In: Szathmary, L., Priss, U. (eds.) Proceedings of The Ninth International Conference on Concept Lattices and Their Applications (CLA). CEUR Workshop Proceedings, vol. 972, pp. 187–198 (2012)
5. Cordero, P., Enciso, M., Mora, A., Ojeda-Aciego, M.: Computing left-minimal direct basis of implications. In: Ojeda-Aciego, M., Outrata, J. (eds.) Proceedings of the Tenth International Conference on Concept Lattices and Their Applications (CLA). CEUR Workshop Proceedings, vol. 1062, pp. 293–298 (2013)
6. Ganter, B., Wille, R.: Formal Concept Analysis - Mathematical Foundations. Springer (1999)
7. Goergen, G., Kumar, P., Sankung, S., Togola, A., Tamò, M.: First report of outbreaks of the fall armyworm Spodoptera frugiperda (J. E. Smith) (Lepidoptera, Noctuidae), a new alien invasive pest in West and Central Africa. PLoS ONE **11** (2016)
8. Guigues, J.L., Duquenne, V.: Famille minimale d'implications informatives résultant d'un tableau de données binaires. Math. et Sci. Hum. **24**(95), 5–18 (1986)
9. Guo, J., He, K., Wang, Z.: Biological characteristics, trend of fall armyworm Spodoptera frugiperda, and the strategy for management of the pest. Chin. J. Appl. Entomol. **56**(3) (2019)
10. Ignatov, D.I.: Introduction to Formal Concept Analysis and Its Applications in Information Retrieval and Related Fields. CoRR **abs/1703.02819** (2017)
11. Janostik, R., Konecny, J., Krajča, P.: Pruning techniques in LinCbO for computation of the Duquenne-Guigues basis. In: Hanika, T., Buzmakov, A. (eds.) The 16th International Conference on Formal Concept Analysis (ICFCA) 2021, Strasbourg, France, Jun 29-Jul 2 (to appear). LNCS/LNAI, Springer (2021)
12. Kahn, L.H., Kaplan, B., Monath, T.P., Conti, L.A., Yuill, T.M., Chapman, H.J., Carter, C.N., Barrentine, B.: One-health initiative. `http://www.onehealthinitiative.com` (2020), accessed: 2020-04-01
13. Kuznetsov, S.O., Poelmans, J.: Knowledge representation and processing with formal concept analysis. Wiley Interd. Rev. Data Min. Knowl. Disc. **3**(3), 200–215 (2013)
14. Martin, P., Gutierrez, A., Marnotte, P., Huchard, M., Keip, P., Mahrach, L., Silvie, P.: Dataset on Noctuidae species used to evaluate the separate concerns in conceptual analysis: Application to a life sciences knowledge base (2021). https://doi.org/10.18167/DVN1/HTFE8T
15. Martin, P., Silvie, P., Sarter, S.: Knomana - usage des plantes à effet pesticide, antimicrobien, antiparasitaire et antibiotique (patent APP IDDN.FR.001.130024.000.S.P.2019.000.31235) (2019)
16. Queensland-Government: Department of Agriculture and Fisheries. First mainland detection of fall armyworm (2020)
17. Shylesha, A., Jalali, S., Gupta, A., Varshney, R., Venkatesan, T., Shetty, P., Ojha, R., Ganiger, P., Navik, O., Subaharan, K., Bakthavatsalam, N., Ballal, C., Raghavendra, A.: Studies on new invasive pest Spodoptera frugiperda (J. E. Smith) (Lepidoptera: Noctuidae) and its natural enemies. J. Biol. Cont. **32**(3), 145–151 (2018)
18. Zaki, M.J.: Closed itemset mining and non-redundant association rule mining. In: Liu, L., Özsu, M.T. (eds.) Encyclopedia of Database Systems, pp. 365–368. Springer US (2009)