ISIMA

Institut Supérieur d'Informatique, de Modélisation et de leurs Applications

1 rue de la Chebarde
Aubières, 63178, France

NATIONAL RESEARCH UNIVERSITY

Higher School of Economics

Kochnovskiy Proyezd, 3
Moscow, 125319, Russia

Master Thesis report:
Data science and 3rd year of computer science engineering

# Horn Minimization: an overview of existing algorithms

***Author*** : Simon Vilmin

*Academic Supervisor :* Sergei A. Obiedkov

*Held* : June, 26, 2018

# Acknowledgement

# List of Figures

# List of Algorithms

# Abstract

# Contents

# Introduction

# Chapter 1

# Introduction to implications through closure systems

In this first chapter, we will be involved in presenting our topic of minimization. For this ground to be understandable by as much readers as possible, we will heavily rely on toy examples to illustrate and provide intuition on the various notions we will introduce. To be more precise on the path we are about to follow in this chapter, we are first to expose an informal small example of the task we want to achieve. Then, we shall investigate the history of research on our topic, to act as an exposition of the actual knowledge on the question and to give a context to our study. For the rest of this chapter we will get familiar with mathematical objects called *closure operators* and *closure systems* modelling our problem. As we shall observe, the topic of minimization can be described in several mathematical frameworks. However, even if we describe briefly other objects in next chapters, we will stick to our closure framework in all the report in order to have a leading light among various different terminologies.

## 1.1 Implications and minimization: first meeting

Let us imagine we are some specialist of flowers and plants in general. As such, we are interested in studying *correlations* between plant characteristics. Some possible traits are: *colourful, bloom, wither, aquatic, seasonal, climbing, scented, flower, perennial* and so forth. Having observed countless plants during our studies, we are able to draw relations among all those *attributes*. For instance, we know that a plant having the attribute *flower* is likely to have traits *scent, bloom, wither* while a plant being *perennial* (i.e: does not need a lot of water to survive, like a cactus) is not likely to be *aquatic*.

Those relations *"if we have some attributes, we get those ones too"* depict correlation between attributes (not cause/consequence!). It is important to stress on the knowledge those relations bring. They just indicate that whenever we have say *flower*, we have also *colourful*. This is very different from saying that *because* some plant is a flower, it will be colourful. We call those correlation relations *implication* and use $flower \longrightarrow colourful$ to denote *"if we have the attribute flower, then we have colourful"*. Now let us give

some implications:

$$(colourful,\ bloom \longrightarrow seasonal),\ (colourful,\ wither \longrightarrow seasonal),\ (bloom \longrightarrow wither)$$

All those implications represent a certain amount of knowledge. While in our example they are not numerous we could imagine having tons of them. Hence we would wonder whether there is a way to reduce the number of implications while keeping all the knowledge they represent. This question is *minimization*. Actually, in our small example we can reduce the number of implications. Take *(colourful, bloom* $\longrightarrow$ *seasonal)*. We can derive this implication relation only with the two other ones. Indeed, because a plant *blooming* is likely to *wither* (3rd implication), we have *(colourful, bloom* $\longrightarrow$ *wither)*, but since we now have *wither* and *colourful* we also have *seasonal* (2nd implication). That is, the implication *(colourful, bloom* $\longrightarrow$ *seasonal)* is useless (or *redundant*) in our context and can be removed. Our set of implications will then be smaller, but pointing out the same relations as before.

To summarize, we have seen that out of a set of *attributes* we can draw several relations called *implications* providing some knowledge. We also realized that sometimes, some implications are not necessary. Consequently, the set of implications we are given can be *minimized* without altering the information it contains. This is the topic we were interested during this master thesis. In the next section, we will trace back the overhaul knowledge on this question.

## 1.2 Research on implications theories minimization

This section is intended to supply the reader with a general overview of the minimization topic. After a short contextual information, we focus on some relevant results on the question by providing references to algorithms and properties dedicated to our problem. Eventually, we situate our work within this context.

The question of minimization has been discussed and developed through various frameworks, and several computer scientists communities. Notice that in order not to make this synthesis too long, we will stay within the context of minimization and will not trace the field of implication theories in general. For a survey of this domain anyway, the reader should refer to [28]. Also, note that minimality in general terms is not unique. Indeed, one can define several type of minimality among implication systems. For instance, not only we can define minimality with respect to the number of implication within a system (which is our interest) but also with respect to the number of attributes in each implications. The former one is called *canonical* in relational database field, and *hyperarc minimum* within the graph context. Especially in the graph-theoretic and boolean logic settings, one can derive more types of minimality. For general introduction to boolean logic notations, we invite the reader to see [14]. In terms of propositional logic, implications are represented through Horn formulae. Interestingly, the minimization problem we are going to consider is the only one being polynomial time solvable. Other problems are proved to be NP-Complete or NP-Hard. For more discussion on other minimality definitions and their computational complexity, the reader should refer to [12, 6, 7, 5, 28, 10]. In particular for NP-Completeness in the canonical case, one can see [22]. In subsequent explanations, we will refer to minimization with respect to the number of implications.

To the best of our knowledge, the two first fields in which algorithms and properties of minimality arose are Formal Concept Analysis (FCA) (see [20, 19] for an introduction) and Database Theory (DB) (see [23]). Both sides were developed independently in the early 80's. For the first domain, characterization of minimality goes to Duquenne and Guigues [21], in which they describe the so-called *canonical basis* (also called *Duquenne-Guigues basis* after its authors) relying on the notion of pseudo-closed sets. For the database part, study of implications is made by Maier through FD's ([23, 16]). The polynomial time algorithm he gives for minimization heavily relies on a fast subroutine discovered by Beeri and Bernstein in [8], 1979.

From then on, knowledge increased over years and spread out over domains. Another algorithm based on a minimality theorem is given by Shock in 1986 ([24]). Unfortunately, as we shall see and as already discussed by Wild in [27] the algorithm may not be correct in general, even though the underlying theorem is. During the same period, Ausiello and al. brought the problem to graph-theoretic ground, and provided new structure known as *FD-Graph* and algorithm to represent and work on implication systems in [6, 4, 5]. This approach has been seen in graph theory as an extension of the transitive closure in graphs ([1]), but no consideration equivalent to minimization task seems to have been taken beforehand, as far as we know. Still in the 1980 decade, Ganter expressed the canonical basis formalized by Duquenne and Guigues in his paper related to algorithms in FCA, [19] through closure systems, pseudo-closed and quasi-closed sets. Next, Wild ([25, 26, 27]) linked within this set-theoretic framework both the relational databases, formal concept analysis and lattice-theoretic approach. In relating those fields, he describes an algorithm for minimizing a basis, similar to algorithms of Day and, somehow, Shock (resp. [17], [24]). This framework is the one we will use for our study, and can be found in more recent work by Ganter & Obiedkov in [7]. Also, the works of Maier and Duquenne-Guigues have been used in the lattice-theoretic context by Day in [17] to derive an algorithm based on congruence relations. For in-depth knowledge of implication system within lattice terminology, we can see [15] as an introduction and [9] for a survey. Later, Duquenne proposed some variations in Day's work with another algorithm in [18]. More recently, Boròs and al. by working in a boolean logic framework, exhibited a theorem on the size of canonical basis [11, 12]. They also gave a general theoretic approach that algorithm should do one way or another on reduction purpose. Out of these papers, Berczi & al. derived a new minimization procedure based on hypergraphs in [13]. Furthermore, an algorithm for computing the canonical basis starting from any system is given in [7].

Even though the work we are going to cite is not designed to answer this question of minimization, it must also be exposed as the algorithm is intimately related to DG basis and can be used for base reduction. The paper of Angluin and al. in query learning, see [2], provides an algorithm for learning a Horn representation of an unknown initial formula. It has been shown later by Ariàs and Alcazar ([3]) that the output of Angluin algorithm was always the Duquennes-Guigues basis.

Our purpose with this master thesis is to review and implement as much as possible the algorithms we exposed to provide a comparison. This comparison shall act as both theoretical and experimental statement of algorithm efficiency. As we already mentioned we will focus on closure theory framework. The reason for this choice is our starting point. Because we start from the algorithms provided by Wild and

because the closure framework is the one we are the most familiar with, we focus on clearly explain this terminology with examples. However, once we will be comfortable with those definitions, we will relate other frameworks to our main approach in the next chapter, to explain and draw parallels with other algorithms. In the next section we will focus on theoretical definitions we shall need to understand the algorithms we have implemented.

## 1.3 Implications and minimization: theoretic approach

Here we will dive into mathematical representation of the task we gave in the first section of this chapter. For the recall, our aim here is to get familiar with the representation being closest from closure systems. Most of the notions initially come from [21, 19, 26, 20] but the reader can also find more than sufficient explanations in [7, 28]. Readers with knowledge in relational databases will recognize most of functional dependency notations. The reason is close vicinity between implications and functional dependencies. Talking about our needs, we can consider them as equivalent notations. Actually, the real-life application our set up will be the closest from is FCA ([20]) as we shall see in the last chapter.

### 1.3.1 Implications and closure systems

The easiest object to project onto mathematical definitions is our attribute set. For all the report, we fix $\Sigma$ to be a set of *attributes*. Usually, we will denote attributes by small letters: *a, b, c, ...* and subsets of $\Sigma$ (groups of attributes) will be denoted by capital letters: *A, B, C, ...* We assume the reader to have few background in elementary set-theoretic notations.

**Definition 1** (*Implication, implication system*)**.** *An implication over $\Sigma$ is a pair $(A, B)$ with $A, B \subseteq \Sigma$. It is usually denoted by $A \longrightarrow B$. A set $\mathcal{L}$ of implications is called an implication system, implication theory or implication(al) base(is).*

Note that given as is, this definition seems to lose the semantic relation we depicted earlier. But we should keep in mind that in our set up, we will be given implications more than an attribute set. Hence, implications will make sense on their own, independently from the attribute set they are drawn from. Quickly, remark that implications in logical terms are expressed as *Horn formulae* giving another of its names to implication theories. Also, in $A \longrightarrow B$, $A$ is said to be the *premise* (or *body*) and $B$ the *conclusion* (*head*).

**Definition 2** (*Model*)**.** *Let $\mathcal{L}$ be an implication system over $\Sigma$, and $M \subseteq \Sigma$. Then:*

*(i) $M$ is a model of an implication $A \longrightarrow B$, written $M \models A \longrightarrow B$, if $B \subseteq M$ or $A \nsubseteq M$,*

*(ii) $M$ is a model of $\mathcal{L}$ if $M \models A \longrightarrow B$ for all $A \longrightarrow B \in \mathcal{L}$.*

The notion of model may seem disarming at first sight. But $M$ being a model of $A \longrightarrow B$ simply means that, if $A$ is included in $M$, then for the implication $A \longrightarrow B$ to hold in $M$, we must have $B$ in $M$ too. This still suits the intuitive notion of premise/conclusion. Placed in the context of $M$, $A \longrightarrow B$ says *"whenever we have A, we must also have B"*. Reader with some background in mathematical logic should be familiar with the notation $\models$, denoting semantic entailment, as opposed to $\vdash$ for syntactic deduction (see [14]). By a fortunate twist of fate, semantic entailment is our next step:

**Definition 3** (*Semantic entailment*)**.** *We say that an implication $A \longrightarrow B$ semantically follows from $\mathcal{L}$, denoted $\mathcal{L} \models A \longrightarrow B$, if all models $M$ of $\mathcal{L}$ are models of $A \longrightarrow B$.*

Because next definitions are going to be on a slightly different structure, even though closely related to implication systems of course, let us rest for a while and illustrate our definitions with an example.

**Example** Consider again our plant properties. Let $\Sigma = \{$*colourful, bloom, wither, seasonal, aquatic, perennial, flower, scented*$\}$. An implication could be *flower* $\longrightarrow$ *scented*, or even *(bloom, aquatic)* $\longrightarrow$ *colourful* if we get rid off semantic interpretations. An implication basis $\mathcal{L}$ is for instance:

*(colourful, bloom* $\longrightarrow$ *seasonal), (colourful, wither* $\longrightarrow$ *seasonal), (bloom* $\longrightarrow$ *wither)*

and $M =$*(colourful, bloom, seasonal)* is a model of *colourful, bloom* $\longrightarrow$ *seasonal* because both the head and the body of the implication belong to $M$. Also, $M$ is not a model of $\mathcal{L}$ because it is not a model of *bloom* $\longrightarrow$ *wither*. A model of $\mathcal{L}$ could be *(bloom, wither)* or even the empty set $\emptyset$.

Next definitions are about closure operators, and closure systems. We need to ground ourselves in those definitions before returning to implications. $2^\Sigma$ is the set of all subsets of $\Sigma$, also named the *power set* of $\Sigma$.

**Definition 4** (*Closure operator*)**.** *Let $\Sigma$ be a set and $\phi : 2^\Sigma \longrightarrow 2^\Sigma$ an application on the power set of $\Sigma$. $\phi$ is a closure operator if $\forall X, Y \subseteq \Sigma$:*

*(i)* $X \subseteq \phi(X)$ *(extensive),*

*(ii)* $X \subseteq Y \longrightarrow \phi(X) \subseteq \phi(Y)$ *(monotone),*

*(iii)* $\phi(X) = \phi(\phi(X))$ *(idempotent).*

*$X \subseteq \Sigma$ is called closed if $X = \phi(X)$.*

**Definition 5** (*Closure system*)**.** *Let $\Sigma$ be a set, and $\Sigma^\phi \subseteq 2^\Sigma$. $\Sigma^\phi$ is called a closure system if:*

*(i)* $\Sigma \in \Sigma^\phi$,

*(ii) if* $\mathcal{S} \subseteq \Sigma^\phi$, *then* $\bigcap \mathcal{S} \in \Sigma^\phi$ *(closed under intersection).*

In the second definition, it is worth stressing on the fact that $\Sigma^\phi$ is a set of sets. Also, the notation $\Sigma^\phi$ may seem surprising, but it has been chosen purposefully. Indeed, to each closure system $\Sigma^\phi$ over $\Sigma$, we can associate a closure operator $\phi$ and vice-versa:

- from $\phi$ to $\Sigma^\phi$: compute all closed sets of $\phi$ to obtain $\Sigma^\phi$,
- from $\Sigma^\phi$ to $\phi$: define $\phi(X)$ as the smallest element of $\Sigma^\phi$ (inclusion-wise) containing $X$. Observe that such a set always exists in $\Sigma^\phi$ because $\Sigma \in \Sigma^\phi$.

In any case, this notation used for clear exposition of the link between closure systems and closure operators will be adapted to our context of implication systems as we shall see later on. Notice that one can encounter another object, *closure space*, being a pair $(\Sigma, \phi)$ where $\Sigma$ is a set and $\phi$ a closure operator over $\Sigma$. We are likely to find this notation notably in [25, 26] where a general theory of closure spaces is addressed.

**Example** Let us imagine we have four people: *Jezabel, Neige, Seraphin* and *Narcisse*. Let us assume they all know each other and then define a relation *"like"* between them. For instance, say *Séraphin likes Jezabel*. this relation is a *binary relation*: it relates pairs of elements. We can represent this relation by a graph where nodes are people and edges represent relations:

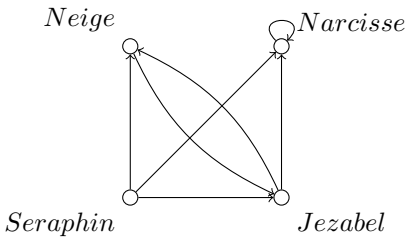


Figure 1.1: Graph of "like" relation

The arrow from *Seraphin* to *Jezabel* stands for *"Seraphin likes Jezabel"* and the arrow from *Narcisse* to itself means equivalently *"Narcisse likes Narcisse"*. With this clear, let us introduce an operation of gathering people. Starting from any group $A$ of persons presented here, let's add to $A$ every person liked by at least one element of $A$, until we can no more add people. For instance:

- if we start from *Neige*, because *Neige* likes *Jezabel* and *Jezabel* likes *Narcisse* we will add both of them to the group of *Neige*,
- because *Narcisse* only likes himself, we have no people to add in his group.

Now observe that this operation of gathering people is in fact a closure operator:

(i) it is *extensive*: starting from any group of people, we can only add new ones, hence either the group does not change (e.g: *Narcisse*) or it grows,

(ii) it is *monotone*: if we start from a group $A$ containing a group $B$, it is clear that we will at least gather in $A$ all the people we would add with $B$,

(iii) *idempotency*: once we added all the people we had to reach, then trying to find new people is useless by definition. Hence the group will remain the same if we apply our operation once more.

We are going to get back to our main implication purpose to illustrate the notion of closure in our context. It turns out that given a basis $\mathcal{L}$ over some set $\Sigma$, the set of models of $\mathcal{L}$, $\Sigma^{\mathcal{L}}$, is a closure system. Moreover, the operator $\mathcal{L} : 2^{\Sigma} \longrightarrow 2^{\Sigma}$ associating to a subset $X$ of $\Sigma$ the smallest model (inclusion wise) containing $X$ is a closure operator. Furthermore, the closure system it defines is $\Sigma^{\mathcal{L}}$. An interesting point is the mathematical computation of $\mathcal{L}(X)$ given $\mathcal{L}$ as a set of implications. We rely on [25, 7] to this end. Let us define a temporary operation $\circ : 2^{\Sigma} \longrightarrow 2^{\Sigma}$ as follows:

$$X^{\circ} = X \cup \bigcup \{B \mid A \longrightarrow B \in \mathcal{L},\ A \subseteq X\}$$

Applying this operator up to stability provides $\mathcal{L}(X)$. In other words $\mathcal{L}(X) = X^{\circ\circ\cdots}$. It is clear that we have a finite amount of iterations since $X$ cannot grow more than $\Sigma$. Readers with background in logic (see [12]) or graph theory ([13]) might see this operation as the marking or forward chaining procedure.

**Example** Let's stick to our vegetable example, but reducing $\Sigma$ to {*bloom, flower, colourful* } (abbreviated *b, f, c*) for the sake of simplicity. Furthermore, let $\mathcal{L} = \{$*((colourful, bloom)* $\longrightarrow$ *flower), (flower* $\longrightarrow$ *bloom)*$\}$, abbreviated then $cb \longrightarrow f$, $f \longrightarrow b$. For instance, because $f \longrightarrow b \in \mathcal{L}$, the smallest model of $\mathcal{L}$ containing $f$ is $bf$, and $bf$ is closed. More precisely, the set of closed sets is the following:

$$\Sigma^{\mathcal{L}} = \{\emptyset,\ b,\ c,\ bf,\ bcf\}$$

Pouet.

Having presented the main definitions we shall need, we are to investigate practical computation of closures and more elaborated structures like the canonical basis (or Duquenne-Guigues basis) in the next section.

# Conclusion

# Bibliography


[1] Aho, A. V., Garey, M. R., and Ullman, J. D. The Transitive Reduction of a Directed Graph. *SIAM Journal on Computing* (July 2006).

[2] Angluin, D., Frazier, M., and Pitt, L. Learning conjunctions of Horn clauses. *Machine Learning 9*, 2 (July 1992), 147–164.

[3] Arias, M., and Balcázar, J. L. Canonical Horn Representations and Query Learning. In *Algorithmic Learning Theory* (Berlin, Heidelberg, 2009), R. Gavaldà, G. Lugosi, T. Zeugmann, and S. Zilles, Eds., Springer Berlin Heidelberg, pp. 156–170.

[4] Ausiello, G., D'Atri, A., and Saccà, D. Graph Algorithms for Functional Dependency Manipulation. *J. ACM 30*, 4 (Oct. 1983), 752–766.

[5] Ausiello, G., D'Atri, A., and Saccá, D. Minimal Representation of Directed Hypergraphs. *SIAM J. Comput. 15*, 2 (May 1986), 418–431.

[6] Ausiello, G., and Laura, L. Directed hypergraphs: Introduction and fundamental algorithms—A survey. *Theoretical Computer Science 658*, Part B (2017), 293 – 306.

[7] B. Ganter, S. O. *Conceptual Exploration*. Springer, 2016.

[8] Beeri, C., and Bernstein, P. A. Computational Problems Related to the Design of Normal Form Relational Schemas. *ACM Trans. Database Syst. 4*, 1 (Mar. 1979), 30–59.

[9] Bertet, K., Demko, C., Viaud, J.-F., and Guérin, C. Lattices, closures systems and implication bases: A survey of structural aspects and algorithms. *Theoretical Computer Science* (Nov. 2016).

[10] Boros, E., Čepek, O., and Kogan, A. Horn minimization by iterative decomposition. *Annals of Mathematics and Artificial Intelligence 23*, 3-4 (Nov. 1998), 321–343.

[11] Boros, E., Čepek, O., Kogan, A., and Kučera, P. Exclusive and essential sets of implicates of Boolean functions. *Discrete Applied Mathematics 158*, 2 (2010), 81 – 96.

[12] Boros, E., Čepek, O., and Makino, K. Strong Duality in Horn Minimization. In *Fundamentals of Computation Theory* (Berlin, Heidelberg, 2017), R. Klasing and M. Zeitoun, Eds., Springer Berlin Heidelberg, pp. 123–135.

[13] Bérczi, K., and Bérczi-Kovács, E. R. Directed hypergraphs and Horn minimization. *Information Processing Letters 128* (2017), 32 – 37.

[14] Cori, R., and Lascar, D. *Mathematical Logic: Part 1: Propositional Calculus, Boolean Algebras, Predicate Calculus, Completeness Theorems*. OUP Oxford, Sept. 2000. Google-Books-ID: Cle6_dOLt2IC.

[15] Davey, B. A., and Priestley, H. A. *Introduction to Lattices and Order*. Cambridge University Press, 2002.

[16] David, M. Minimum Covers in Relational Database Model. *J. ACM 27*, 4 (1980), 664 – 674.

[17] Day, A. The Lattice Theory of Functional Dependencies and Normal Decompositions. *International Journal of Algebra and Computation* (1992).

[18] Duquenne, V. Some variations on Alan Day's algorithm for calculating canonical basis of implications. In *Concept Lattices and their Applications (CLA)* (Montpellier, France, 2007), pp. 17–25.

[19] Ganter, B. Two Basic Algorithms in Concept Analysis. In *Formal Concept Analysis* (Mar. 2010), Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, pp. 312–340.

[20] Ganter, B., and Wille, R. *Formal Concept Analysis: Mathematical Foundations*. Springer-Verlag, Berlin Heidelberg, 1999.

[21] Guigues J.L, D. V. Familles minimales d'implications informatives résultant d'un tableau de données binaires. *Mathématiques et Sciences Humaines 95* (1986), 5–18.

[22] Hammer, P. L., and Kogan, A. Optimal compression of propositional Horn knowledge bases: complexity and approximation. *Artificial Intelligence 64*, 1 (Nov. 1993), 131–145.

[23] Maier, D. *Theory of Relational Databases*. Computer Science Pr, 1983.

[24] Shock, R. C. Computing the minimum cover of functional dependencies. *Information Processing Letters 22*, 3 (Mar. 1986), 157–159.

[25] Wild, M. Implicational bases for finite closure systems. *Informatik-Bericht 89/3, Institut fuer Informatik* (Jan. 1989).

[26] Wild, M. A Theory of Finite Closure Spaces Based on Implications. *Advances in Mathematics 108*, 1 (Sept. 1994), 118–139.

[27] Wild, M. Computations with finite closure systems and implications. In *Computing and Combinatorics* (Aug. 1995), Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, pp. 111–120.

[28] Wild, M. The joy of implications, aka pure Horn formulas: Mainly a survey. *Theoretical Computer Science 658* (2017), 264 – 292.