# Aide à l'utilisation et à l'exploitation de l'analyse de concepts formels pour des non-spécialistes de l'analyse des données

Ali Jaffal

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

# THESE DE DOCTORAT
# DE L'UNIVERSITE PARIS 1 PANTHEON-SORBONNE

Spécialité : Informatique

Présentée par


## Ali JAFFAL


Pour l'obtention du grade de Docteur de
L'Université Paris 1 Panthéon – Sorbonne

Thèse dirigée par **Bénédicte LE GRAND**

préparée au sein du **Centre de Recherche en Informatique**
à L'École Doctorale de Management Panthéon-Sorbonne

# Aide à l'utilisation et à l'exploitation de l'Analyse de Concepts Formels pour des non spécialistes de l'analyse de données

Soutenue publiquement le jeudi 3 octobre 2019 devant le jury composé de :

| | | |
|---|---|---|
| Mme. Florence Sèdes | Professeur, Université Paul Sabatier, Toulouse | Rapporteur |
| M. Michel Soto | MCF-HDR, Université Paris Descartes | Rapporteur |
| Mme. Christine Largeron | Professeur, Université Jean Monnet Saint-Etienne | Examinatrice |
| Mme. Manuele Kirsch Pinheiro | MCF, Université Paris 1 Panthéon-Sorbonne | Examinatrice |
| Mme. Bénédicte Le Grand | Professeur, Université Paris 1 Panthéon-Sorbonne | Directrice de thèse |

# Remerciements

Je tiens à exprimer ma profonde gratitude et mes sincères remerciements à ma directrice de thèse Madame Bénédicte Le Grand, d'abord pour sa confiance en moi et d'avoir accepté avec enthousiasme de me diriger en thèse, et aussi son support continu ces dernières années. Merci pour les encouragements, les nombreuses remarques, relectures, corrections et suggestions. Je la remercie pour sa disponibilité et ses conseils importants, pour son soutien, sa patience et pour l'encadrement idéal pendant ma thèse. Je lui suis particulièrement reconnaissant.

Je remercie Madame Florence Sèdes et Monsieur Michel Soto de m'avoir fait l'honneur d'être rapporteurs de ce travail. Je les remercie pour leurs recommandations sur ce manuscrit ainsi que pour toutes les questions posées lors de la pré-soutenance. Leurs commentaires m'ont été très utiles pour la version finale de ce manuscrit.

Je remercie Madame Christine Largeron et Madame Manuele Kirsch Pinheiro qui m'ont fait l'honneur d'avoir accepté d'examiner ce travail.

J'ai eu la chance d'enseigner au cours de cette thèse, pour cela je remercie Madame Manuele Kirsch Pinheiro, Madame Carine Souveyet et Madame Bénédicte Le Grand pour la confiance accordée afin des donner des travaux dirigés et des cours depuis ma première année de thèse à l'UFR 06 et l'UFR 27. Je les remercie encore pour leurs nombreux conseils relatifs à la pédagogie.

Je remercie l'École Doctorale de Management Panthéon-Sorbonne (EDMPS) de m'avoir accordé le contrat doctoral ayant financé ma thèse.

Mes remerciements s'adressent aussi à tous les membres du Centre de Recherche en Informatique de l'Université Paris 1 Panthéon-Sorbonne. Je remercie mes collègues et les secrétaires pour l'ambiance amicale et les bons moments que nous avons partagés.

Je remercie mon ami et mon beau-frère le cher Ahmad Farhat pour le temps qu'il m'a accordé depuis et même avant mon arrivée en France, et durant toutes ces années jusqu'à ce jour.

Je remercie mon ami Mohamad Rmayti, pour le temps qu'il m'a accordé, la relecture de ma thèse et pour tous ces conseils.

Je remercie Ángela, Luisa, Elena E., Elena K., Asmaa, Danillo, David, Fabrice, Floriane, Danny, Maximilien, Alice, Sébastien, Khalil, Héla, Sonia, Amina, Sana, Salma, Sabrine, Houssam, Abir, Yosra, Raouia, Nourhene, Juan Carlos, Lamiae, Afef, Abdelkader et Lookman pour tous les jolis moments partagés durant ma thèse ainsi que pour toutes les discussions scientifiques et autres.

Je remercie tous mes amis et mes proches, sans oublier personne, vous étiez toujours là dans mes pensées.

Je remercie affectueusement mes parents, mon frère et mes sœurs, pour leur soutien, je leur envoie ma gratitude pour tout ce qu'ils m'ont donné sans mesure.

Enfin, je remercie de tout mon cœur mon amour et ma princesse Ghadir d'avoir été là et durant les moments magiques, les moments difficiles et pour toujours.

Je dédie cette thèse à ma mère, mon père, ma famille, mon amour, et à tous ceux que j'aime.

# Résumé

**Aide à l'utilisation et à l'exploitation de l'Analyse de Concepts Formels pour des non spécialistes de l'analyse de données**

De nombreuses approches ont été élaborées pour extraire des connaissances à partir des données. On distingue traditionnellement l'analyse descriptive et l'analyse prédictive. Nous nous focalisons dans cette thèse sur l'analyse descriptive des données et plus particulièrement sur l'Analyse de Concepts Formels (ACF), qui permet de construire des clusters recouvrants (appelés *concepts formels*) dont la signification est explicite. Il existe une relation d'ordre partiel entre les concepts formels résultant de l'ACF, qui sont organisés en une structure mathématique appelée *treillis de Galois*. Malgré ses nombreux avantages, l'ACF est peu accessible à des utilisateurs non experts de l'analyse de données. En effet, malgré les représentations graphiques des treillis de Galois, ceux-ci restent difficiles à interpréter, notamment lorsque les données sont volumineuses. De plus, la construction des données d'entrée de l'ACF sous la forme d'un *contexte formel* peut être délicate. Pour cela, nous avons proposé une méthodologie d'interprétation des treillis de Galois reposant sur un ensemble de métriques simples, dont les résultats sont présentés sous une forme visuelle aussi intuitive que possible. Nous avons également développé des stratégies pour construire des contextes formels qui, non seulement, dénaturent le moins possible les données initiales, mais permettent aussi de tenir compte des besoins de l'utilisateur en termes de recherche d'information.

**Mots clés** : Analyse de données, Analyse de Concepts Formels, Treillis de Galois, Exploration de connaissances.

# Abstract

**Assistance in the use and exploitation of Formal Concepts Analysis for non-specialists in data analysis**

Many data analysis techniques have been developed to extract knowledge from the data. The two traditional approaches are descriptive analysis and predictive. We focus in this thesis on descriptive data analysis, and in particular on Formal Concepts Analysis (FCA). This approach builds overlapping clusters (called *formal concepts*) whose meaning is explicit. There is a partial order relationship between the formal concepts resulting from FCA, which are organized in a mathematical structure called a *Galois lattice*. Despite its advantages, FCA is poorly accessible to users who are not experts in data analysis. Although graphical representations of Galois lattices exist, their interpretation remains difficult for large data. Moreover, the construction of FCA input data, called *formal context*, can be tricky. For this, we have proposed a methodology for Galois lattice interpretation based on a set of simple metrics, the results of which are presented in a visual form as intuitive as possible. We have also developed strategies for constructing formal contexts that not only remain as close to the initial data as possible, but also take into consideration the user's needs in terms of information retrieval.

**Keywords**: Data Analysis, Formal Concepts Analysis, Galois Lattice, Knowledge Exploration.

# Table des matières

# Table des figures

# Liste des tableaux

# CHAPITRE I :
# INTRODUCTION GÉNÉRALE

# Sommaire

## I.1. **Contexte**

L'essor du numérique, dans des contextes aussi bien personnels que professionnels, a engendré une augmentation significative du volume des données générées par des sources diverses. Toutes les actions des utilisateurs sur un système informatisé peuvent notamment être enregistrées sous forme de traces, qu'il s'agisse de l'utilisation de smartphones ou d'objets connectés, de l'envoi de messages sur des réseaux sociaux, ou de la recherche d'information sur Internet. Le défi est plus que jamais d'en extraire des informations exploitables, sans que l'on ait forcément d'a priori sur ce que l'on espère y trouver.

De nombreuses techniques d'analyse de données ont été élaborées pour extraire des connaissances à partir des données. On distingue traditionnellement l'analyse descriptive et l'analyse prédictive. Nous nous focalisons dans cette thèse sur l'analyse descriptive des données, qui vise à les « comprendre » sans avoir nécessairement d'idée préconçue sur ce que l'on recherche. Elle peut être utilisée par exemple pour identifier des tendances ou des éléments qui s'éloignent de la norme (les « outliers »). La classification non supervisée, aussi appelée clustering, appartient également à cette catégorie : elle permet de regrouper les données dans des « clusters » qui ne sont pas prédéterminés. Nous nous intéressons ici plus particulièrement à l'Analyse de Concepts Formels (ACF), qui rassemble des éléments (appelés *objets*) dans des clusters (appelés *concepts formels*) qui peuvent se chevaucher, en fonction des caractéristiques (appelées *attributs*) que ces objets ont en commun. Cette approche, que nous décrivons dans le chapitre suivant, permet de construire des clusters dont la signification est explicite, puisque directement liée aux attributs communs aux objets du cluster. Le fait que les objets puissent appartenir à plusieurs clusters est également un avantage par rapport aux techniques de clustering qui effectuent un partitionnement de l'ensemble des données. Enfin, il existe une relation d'ordre partiel entre les concepts formels résultant de l'ACF, qui sont organisés dans un *treillis* dit *de Galois*, comme nous le détaillons dans le chapitre suivant. Un treillis de Galois est une structure mathématique à partir de

laquelle des représentations graphiques peuvent être construites, le diagramme de Hasse[1] étant la plus commune et connue. Le diagramme de Hasse permet de de visualiser l'ensemble des concepts formels et de naviguer dans les données à différents niveaux de spécificité / généralité.

Cependant, l'ACF présente aussi des limites : la complexité de calcul des concepts formels et du diagramme de Hasse associé rend cette approche inadaptée pour des données volumineuses. Dans un contexte de Big Data, cette technique devra donc être appliquée en 2e intention, sur un sous-ensemble de données spécifique.

## I.2. **Problématiques**

L'Analyse de Concepts Formels permet de découvrir des relations non triviales entre des données hétérogènes et complexes. Bien que la théorie sous-jacente ne soit pas récente, nous montrons dans l'état de l'art du chapitre suivant que l'ACF est utilisée dans des domaines très divers et donne des résultats intéressants.

Nous constatons néanmoins qu'**elle est peu accessible à des utilisateurs non experts de l'analyse de données**. En effet, bien que les diagrammes de Hasse construits à partir des treillis de Galois représentent les relations entre les concepts de manière visuelle, leur interprétation devient difficile lorsque leur taille augmente (y compris, d'ailleurs, pour des experts de l'ACF).

D'autre part, **l'ACF requiert en entrée une relation binaire entre un ensemble d'objets et l'ensemble des attributs qui les caractérisent**, sous la forme d'une matrice binaire appelée *contexte formel*. Nous illustrons un exemple de contexte formel à donner en entrée dans le chapitre suivant. Lorsque la relation que l'on étudie n'est pas naturellement sous cette forme, il est nécessaire d'y effectuer des transformations, qui peuvent avoir un

[1] Dont un exemple est donné sur la Figure II-1

impact important sur les résultats qu'on peut obtenir[2]. En effet, les solutions simples que pourraient envisager des utilisateurs novices risquent de fausser significativement les résultats et d'aboutir à des conclusions éloignées de la réalité des données initiales.

## I.3. **Contributions**

Les problématiques soulevées dans la section précédente nous ont amenés à travailler sur deux objectifs durant cette thèse :

- rendre les résultats de l'ACF plus aisément exploitables pour des utilisateurs non experts de l'analyse de données, en facilitant notamment l'interprétation visuelle des treillis de Galois générés. Nous avons proposé pour cela une **méthodologie d'interprétation des treillis de Galois reposant sur un ensemble de métriques simples**, dont les résultats sont présentés sous une forme visuelle aussi intuitive que possible. Ces métriques peuvent être appliquées à des treillis de toute taille, et la présentation visuelle des résultats est plus accessible que la représentation classique des treillis de Galois, inadaptée aux grands treillis.

- faciliter l'utilisation de l'Analyse de Concepts Formels dans le cas où les données d'entrée ne se présentent pas naturellement sous la forme d'une relation binaire. Nous avons pour cela développé **des stratégies pour construire des contextes formels** qui, non seulement, dénaturent le moins possible les données initiales, mais permettent aussi de tenir compte des besoins de l'utilisateur en termes de recherche d'information. Plusieurs des stratégies proposées permettent en outre de tenir compte d'informations sémantiques sur les données d'entrée, lorsque l'on en dispose.

[2] Ce point sera abordé dans le Chapitre V.

## I.4. **Organisation du manuscrit**

Dans le Chapitre II, nous proposons un état de l'art des travaux de recherche récents qui utilisent l'Analyse de Concepts Formels, afin de fournir un aperçu des domaines d'application de l'ACF. Compte-tenu des problématiques sur lesquelles nous nous focalisons dans cette thèse, nous nous sommes concentrés sur la manière dont les résultats de l'ACF sont exploités par les utilisateurs non experts de l'analyse de données. Nous discutons donc plus particulièrement des méthodes de représentation et de navigation dans les treillis de Galois, ainsi que des méthodes de sélection d'information à partir de diverses mesures.

Dans le Chapitre III, nous proposons notre première contribution, qui consiste en de nouvelles mesures pour faciliter l'interprétation des treillis de Galois, pour en extraire des connaissances. Ces mesures ont l'avantage d'être aisées à comprendre et de pouvoir être présentées de manière visuelle intuitive à un utilisateur non spécialiste de l'ACF. Nous illustrons ces mesures sur un ensemble de données relatives au contexte d'usage d'applications mobiles sur des smartphones par un groupe d'étudiants.

Dans le Chapitre IV, nous décrivons les résultats d'une expérimentation que nous avons menée afin d'évaluer la pertinence de l'ACF pour appréhender un ensemble d'articles de recherche, et en particulier pour les organiser en un état de l'art. Nous avons choisi l'ACF comme thème de cet état de l'art à construire, afin de pouvoir comparer les résultats obtenus avec l'état de l'art que nous avons réalisé de manière manuelle dans le Chapitre II de ce manuscrit.

Dans le Chapitre V, nous présentons notre deuxième contribution, sous la forme de **stratégies de construction de contextes formels**, c'est-à-dire les matrices binaires à fournir en entrée de l'ACF. Ces stratégies permettent de tenir compte au mieux des données initiales, et notamment de leur sémantique ainsi que de l'intérêt de l'utilisateur en termes de besoin d'information. L'impact de ces stratégies sur les résultats de l'ACF est illustré grâce aux mesures définies dans le Chapitre III.

Dans le Chapitre VI, nous présentons une autre application de nos travaux, dédiée à l'étude de la communauté des chercheurs dans le domaine de lignes de produits logiciels.

Dans le Chapitre VII, nous concluons ce mémoire et réunissons nos contributions dans une **méthodologie globale d'exploration de connaissances par l'analyse de concepts formels**, visant à en faciliter l'usage et l'exploitation par des non-spécialistes. Nous présentons finalement les perspectives ouvertes par ces travaux de thèse.

## I.5. **Publications**

**A) Chapitre de livre d'audience internationale avec comité de lecture :**

**1.** Rébecca Deneckère, Charlotte Hug, **Ali Jaffal**, Manuele Kirsch Pinheiro, Bénédicte Le Grand, Raúl Mazo, Irina Rychkova. Context Management and Intention Mining for Adaptive Systems in Mobile Environments: from Business Process Management to Video Games?. *In Digital Interfaces in Situations of Mobility: Cognitive, Artistic, and Game Devices*. Common Ground Publishing, 2016, pp. 87-114.

**B) Conférence internationale avec comité de lecture :**

**2.** Abir Gorrab, Ferihane Kboubi, **Ali Jaffal**, Bénédicte Le Grand, Henda Ben Ghezala. Twitter User Profiling Model Based on Temporal Analysis of Hashtags and Social Interactions, *22nd International Conference on Natural Languages and Information Systems, NLDB 2017*, Liège Belgium, pp. 124-130.

**3. Ali Jaffal**, Bénédicte Le Grand. Towards an Automatic Extraction of Smartphone Users' Contextual Behaviors. *IEEE 10th International Conference on Research Challenges in Information Science (RCIS)*, doi: 10.1109/RCIS.2016.7549334, Jun 2016, Grenoble, France. pp. 1-6.

**4. Ali Jaffal**, Manuele Kirsch-Pinheiro, Bénédicte Le Grand. Unified and Conceptual Context Analysis in Ubiquitous Environments, *In: Jaime Lloret Mauri, Christoph Steup, Sönke Knoch (Eds.), UBICOMM 2014 : The Eighth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies*, August 24 - 28, 2014 – Rome, Italy, ISBN: 978-1-61208-353-7, IARIA, pp. 48-55.

**C) Workshop international avec comité de lecture :**

**5. Ali Jaffal**, Bénédicte Le Grand, Manuele Kirsch Pinheiro. Refinement Strategies for Correlating Context and User Behavior in Pervasive Information Systems.

*International Workshop on Big Data and Data Mining Challenges on IoT and Pervasive Systems (BigD2M)*, Jun 2015, London, United Kingdom, pp. 1040-1046.

**D) Conférence nationale avec comité de lecture :**

**6. Ali Jaffal**, Bénédicte Le Grand B., Manuele Kirsh-Pinheiro, Extraction de connaissances dans les Systèmes d'Information Pervasifs par l'Analyse Formelle de Concepts. *Extraction et Gestion des Connaissances (EGC), Revue des Nouvelles Technologies de l'Information.* Jan 2016, Reims France, pp. 291-296.

**E) Revue nationale avec comité de lecture suite à un Forum Jeunes Chercheurs :**

**7.** Guillaume Cabanac, Amira Derradji, **Ali Jaffal**, Jonathan Louëdec, Gloria Elena Jaramillo Rojas, Forum Jeunes Chercheurs à Inforsid 2014. *Revue des Sciences et Technologies de l'Information - Série ISI : Ingénierie des Systèmes d'Information*, Lavoisier, 2015, pp. 125-130.

**F) Forum Jeunes Chercheurs :**

**8. Ali Jaffal**, Analyse formelle de concepts et règles d'association pour la gestion de contexte dans des environnements ubiquitaires. *INFormatique des ORganisation et Systèmes d'Information de Décision*, Forum Jeunes Chercheurs, mai 2014, Lyon, France, pp. 37- 40.

# CHAPITRE II :
# ÉTAT DE L'ART -
# USAGE DE L'ANALYSE DE CONCEPTS FORMELS

# Sommaire

## II.1. **Introduction**

L'Analyse de Concepts Formels (ACF) est utilisée dans de nombreux travaux de recherche pour la représentation des connaissances, l'analyse, la visualisation et l'interprétation de données. Elle a été appliquée à plusieurs domaines de recherche et sur différents types de données (médicales, biologiques, Web, éducation...). Dans cette thèse, nous nous intéressons tout particulièrement à l'utilisation de l'ACF par un public non-spécialiste de l'analyse de données. Nous avons donc orienté notre étude de l'état de l'art en nous intéressant plus spécifiquement aux divers usages qui en sont faits et à la manière dont son utilisation pourrait être étendue à un large public.

Afin de rédiger cet état de l'art, nous avons étudié les articles présentés dans les 4 conférences de référence dans le domaine de l'ACF : CLA[3](2016, 2015, 2014), ICFCA[4](2015, 2014, 2013), FCA4AI[5] (2016, 2015, 2014) et ICCS[6] (2016, 2014, 2013). L'ensemble de ces éditions de conférences nous ont permis de réunir 168 articles. Après une première lecture basée sur les résumés, nous avons sélectionné 76 articles qui nous ont semblé pertinents en termes d'usage de l'ACF[7]. La lecture approfondie de ces articles nous a finalement permis d'en retenir 32, qui constituent la liste des références bibliographiques citées dans ce chapitre. Nous avons étudié les différentes applications de l'ACF présentées dans ces articles et nous avons identifié les points forts et les limites de chaque méthode. Nous nous sommes

[3] CLA : International Conference on Concept Lattices and Their Applications

[4] ICFCA : International Conference on Formal Concept Analysis

[5] FCA4AI : Formal Concept Analysis (FCA) for Artificial Intelligence

[6] ICCS : International Conference on Conceptual Structures

[7] Nous n'avons pas retenu les articles visant à optimiser la construction des treillis (en termes de performances), car nous nous sommes focalisés dans cette thèse sur l'exploitation des treillis, quelle que soit la manière dont ils ont été construits.

tout particulièrement attachés à la manière dont ces travaux pourraient être utilisés par des utilisateurs non nécessairement spécialistes de l'analyse de données.

Ce chapitre est organisé comme suit. Nous rappelons d'abord quelques notions théoriques sur l'ACF en section II.2. Nous décrivons ensuite dans les sections suivantes les diverses manières dont les treillis de Galois sont utilisés actuellement, selon les articles publiés récemment dans les conférences de ce domaine. Nous avons identifié trois principaux types d'exploitation de ces treillis :

- La génération de règles d'association,
- L'interprétation des treillis à partir de visualisation et de navigation,
- La sélection (ou l'élimination) d'information à partir du calcul de mesures sur le treillis.

## II.2. **Analyse de Concepts Formels**

L'ACF constitue un pont entre les mathématiques, en particulier la théorie des ensembles ordonnés, et les applications d'analyse de données (Priss, 2013). Ses origines reposent sur des tentatives de restructuration de la théorie des ensembles ordonnés pour assurer une meilleure communication entre les théoriciens et les utilisateurs potentiels de la théorie. (Wille, 2005) a introduit l'ACF en tant qu'application de la théorie des treillis. Il s'agit d'une méthode de regroupement conceptuel permettant de découvrir et de structurer des connaissances.

L'ACF propose une démarche algébrique d'organisation hiérarchique d'éléments appelés *objets* en fonction de leurs propriétés appelées *attributs*. Cette démarche consiste à structurer la connaissance extraite sous forme d'associations fortes, appelées *concepts formels*, entre des ensembles d'objets et des ensembles d'attributs. L'ACF est centrée sur cette notion de *concept* qui, de manière informelle, peut être vue comme un groupement d'objets et de leurs attributs communs. Un treillis de concepts traduit ensuite la relation d'ordre hiérarchique entre les concepts formels, et peut être utilisé à des fins de clustering, de classification, de prédiction ou d'approximation. Par exemple, (Trabelsi, Meddouri, & Maddouri, 2016) présentent des méthodes de classification basées sur l'ACF. Ces méthodes reposent sur deux étapes principales : une étape d'apprentissage durant laquelle des classes

sont construites pour décrire un ensemble de données, et une deuxième étape qui assigne une classe à chaque nouvel objet.

### II.2.1. **Théorie des treillis**

Nous présentons ici quelques définitions concernant les ensembles ordonnés, sur lesquelles s'appuie la théorie des treillis. Pour cela, nous définissons tout d'abord la notion de relation binaire.

**Définition 1** : Une relation binaire R entre deux ensembles M et N est un ensemble de couples d'éléments (m, n) tels que m ∈ M et n ∈ N, i.e., un sous-ensemble de M × N.

(m, n) ∈ R (noté aussi mRn) signifie que l'élément m est en relation R avec l'élément n. Si M = N, on parle de relation binaire sur M.

L'ACF repose sur l'étude d'une nouvelle branche de l'algèbre moderne qui est la théorie des treillis. Dans ce cadre, l'analyste examine un treillis de données, l'interprète et en extrait la connaissance cachée. On propose, dans ce qui suit, une définition de la notion de treillis. Pour cela, nous présentons d'abord les ensembles partiellement ordonnés, qui sont une structure mathématique plus générale que les treillis. Une relation d'ordre R définie sur un ensemble E est une relation binaire qui permet de comparer les éléments de E entre eux de manière cohérente.

**Définition 2** : Soit ≤ une relation d'ordre sur un ensemble non vide E. On dit que ≤ définit un ordre total sur E si tous ses éléments sont comparables deux à deux par ≤ :

$$\forall x, y \in E^2, x \neq y \Longrightarrow (x \leq y \ \vee y \leq x)$$

Un ordre qui n'est pas total est dit partiel.

**Définition 3** : Un ensemble partiellement ordonné est un couple (E, ≤) où E est un ensemble non vide d'éléments et ≤ une relation d'ordre partiel.

L'intérêt des ensembles partiellement ordonnés est limité non pas par leur généralité mais par la pauvreté des théorèmes que l'on peut démontrer sur de tels ensembles. Cependant, une famille d'ensembles partiellement ordonnés, appelée *treillis*, se distingue par sa richesse et sa complétude (Zenou & Samuelides, 2005).

**Définition 4** : Un treillis est un ensemble partiellement ordonné (E, ≤) dans lequel chaque couple d'éléments admet une borne supérieure et une borne inférieure. Un treillis est

dit complet si et seulement si toute partie non vide S ⊆ E admet une borne supérieure ∨S et une borne inférieure ∧S. En particulier, un treillis complet admet un élément maximal (borne supérieure ou *top*) et un élément minimal (borne inférieure ou *bottom*).

(Wille, 2005) a présenté l'ACF comme un domaine des mathématiques appliquées qui consiste à restructurer la théorie des treillis afin de faciliter son utilisation dans des applications du monde réel et de permettre l'interprétation de ses notions en dehors du cadre théorique.

### II.2.2. **Contexte formel et treillis de Galois**

Un *treillis de Galois* ou *treillis de concepts* permet de construire, sur la base d'une relation binaire entre un ensemble d'objets et un ensemble d'attributs, une hiérarchie de concepts formels traduisant une association entre un sous-ensemble d'objets et un sous-ensemble d'attributs (Guénoche & Mechelen Van, 1993). Ces concepts sont construits à partir de tableaux de données, appelé *contextes formels*, qui constituent le point de départ de l'ACF.

**Définition 5** : Un contexte formel est un triplet $k = (G, M, I)$, où G et M sont respectivement l'ensemble des objets et celui des attributs, et $I \subseteq G \times M$ est une relation binaire reliant G et M et exprimant que $\forall(o, a) \in I$, $a$ est un attribut de l'objet $o$.

Le contexte formel traduit donc la relation binaire entre un ensemble de propriétés et un ensemble d'objets.

Les opérateurs de dérivation $(.)^I$ sont définis comme suit pour $O \subseteq G$ et $A \subseteq M$ :

$$O^I = \{a \in M \mid \forall o \in O : o\ I\ a\}$$

$$A^I = \{o \in G \mid \forall a \in A : o\ I\ a\}$$

$A^I$ est l'ensemble des objets possédant les attributs de $A$ et $O^I$ est l'ensemble des attributs communs à tous les objets de $O$.

La double application de $(.)^I$ est un opérateur de fermeture, c'est-à-dire que l'application $(.)^{II}$ est extensive, idempotente et monotone. Les ensembles $A^I$ et $O^I$ sont par conséquent des ensembles *fermés*.

Un ***concept formel*** du contexte $(G, M, I)$ est une paire $(O, A)$, où $O \subseteq G$ et $A \subseteq M$, $O = A^I$ et $A = O^I$. Dans ce cas, on a aussi $O = O^{II}$ et $A = A^{II}$.

L'ensemble *O* est appelé l'***extension*** du concept (*O, A*) et *A* en est l'***intension***.

Un concept formel peut donc être vu comme un regroupement entre un ensemble d'objets et l'ensemble des attributs communs à cet ensemble d'objets.

**Définition 6** : L'ensemble de tous les concepts formels et de leur relation d'ordre constitue un treillis, appelé ***treillis de Galois*** du contexte *K*. Ce treillis comprend tous les regroupements naturels des éléments des ensembles G et M.

Un concept $(O, A)$ est un sous-concept, ou une spécialisation, du concept $(C, D)$ si $O \subseteq C$, ce qui est équivalent à $D \subseteq A$. Ceci est noté $(O, A) \leq (C, D)$. Réciproquement, le concept $(C, D)$ est une généralisation du concept $(O, A)$.

Nous illustrons ci-dessous les principales notions définies dans cette section sur des données que nous utiliserons également dans le Chapitre III et le Chapitre V. Ces données ont été collectées à partir de questionnaires remplis par 28 étudiants en master de notre université. Dans ces questionnaires, nous leur avons demandé de lister pendant une semaine toutes les applications utilisées à partir de leur smartphone. Selon le vocabulaire de l'ACF, les applications constituent les *objets* et les éléments de contexte (temporels, géographiques, etc.) dans lesquels ils les ont utilisées sont les *attributs*.

Nous donnons dans le Tableau II-1 un exemple de contexte formel, correspondant au contenu du questionnaire rempli par l'un des étudiants. Dans ce contexte formel, les applications exécutées par l'étudiant sont listées dans les lignes du tableau ; elles correspondent aux objets du contexte formel. Les différents éléments de contexte associés à ces applications apparaissent dans les colonnes et correspondent aux attributs. L'ensemble des applications et l'ensemble des éléments de contexte peuvent varier d'un étudiant à un autre, puisque ces données sont spécifiées par les étudiants eux-mêmes, sans contrôle extérieur. Ainsi, l'étudiant associé au contexte formel du Tableau II-1 a utilisé les

applications suivantes : *Gmail*, *SMS*, *telephone*, *VDM*[8], *Flappy bird*[9] et *Youtube*. Dans ce cas précis, on peut considérer que les applications appartiennent à deux grandes catégories : la communication et les loisirs. Les différents éléments de contexte identifiés par cet utilisateur sont liés à sa localisation géographique (*university, Restaurant, Parents' home, home et transportation*[10]), mais aussi à son type de connexion réseau (*3G*) et au moment de la journée (*morning*, *Coffee break, Lunch break, afternoon*, *evening*). Par exemple, la dernière ligne du contexte formel du Tableau II-1 montre que cet étudiant en particulier a utilisé l'application *Youtube* à la maison (*Home)*, à partir d'un réseau *3G* et pendant la matinée (*Morning)*.

**Tableau II-1 : Exemple de contexte formel extrait d'un questionnaire spécifique**

| Applications | University | Restaurant | Parents' home | Home | Transportation | 3G | Morning | Coffee break | Lunch break | Afternoon | Evening |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Gmail | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 |
| SMS | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Telephone | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 |
| VDM | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 |
| Flappy Bird | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| Youtube | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |

À partir de chaque contexte formel fourni en entrée, l'ACF regroupe les objets dans des clusters appelés *concepts formels*, en fonction de leurs attributs communs (Priss, 2006). Dans l'exemple que nous présentons ici, les applications sont par conséquent regroupées en fonction des éléments de contexte qu'elles ont en commun. Un avantage de l'ACF par rapport à la plupart des approches de clustering est que les clusters générés sont recouvrants

[8] *VDM* est site web sur lequel les internautes racontent des anecdotes malheureuses les concernant.

[9] *Flappy Bird* est un jeu.

[10] *Transportation* = moyen de transport.

(i.e., des objets et/ou attributs peuvent apparaître simultanément dans plusieurs clusters). De plus, la sémantique de chaque concept formel est explicite : l'existence de chaque concept formel est expliquée par les attributs communs qui caractérisent les objets regroupés.

Le résultat du processus d'ACF est un treillis de Galois qui représente la relation d'ordre partiel entre les concepts formels. La Figure II-1 est une représentation du treillis généré à partir du contexte formel du Tableau II-1, appelée diagramme de Hasse.

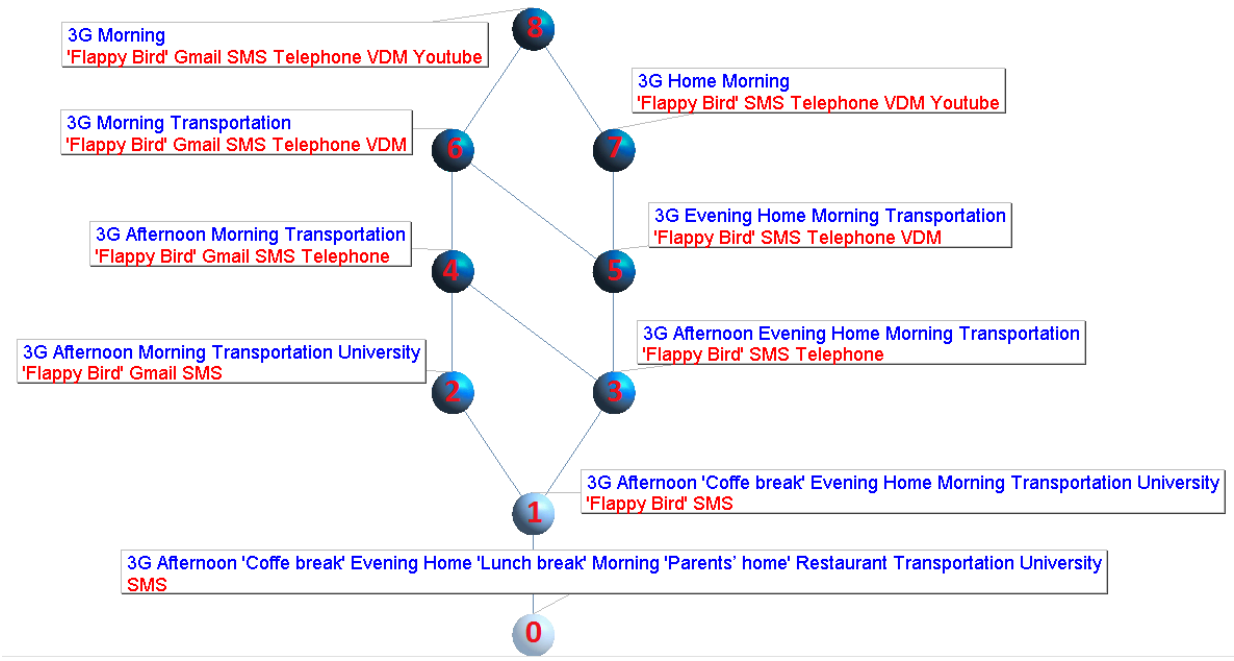



**Figure II-1 : Exemple de treillis de Galois généré à partir du contexte formel du Tableau II-1**

Chaque concept formel du treillis regroupe les applications en fonction des éléments de contexte qu'elles ont en commun. Par exemple, le concept n°2 dans le treillis de la Figure II-1 contient trois applications : *Flappy Bird*, *Gmail* et *SMS*, qui ont toutes été au moins une fois utilisées à partir d'un réseau *3G*, dans les *transports* et à l'*université*, le *matin et* l'*après-midi*. Une interprétation plus détaillée de ce treillis de Galois est fournie dans la section III.1.

L'objectif de nos travaux est de faciliter l'usage et l'interprétation des résultats de l'ACF. C'est pourquoi, dans la suite de ce chapitre, nous présentons un état de l'art établi à partir d'un éventail d'articles acceptés dans des éditions récentes des quatre principales

conférences dans le domaine de l'ACF. De nombreux travaux sont consacrés à l'optimisation de la construction des treillis de Galois, et notamment à la mise à jour incrémentale de treillis existants avec des nouveaux objets (Pattison & Ceglar, 2014). Cet aspect n'est pas le sujet de nos travaux, qui visent à faciliter l'interprétation et l'exploitation d'un treillis, quelle que soit la méthode avec laquelle il a été obtenu.

Dans notre état de l'art, nous avons en particulier retenu les travaux mettant l'accent sur l'utilisation de l'ACF dans différents domaines, par des utilisateurs experts ou non de l'analyse de données. Les trois principaux types d'exploitation des treillis de Galois que nous avons recensés lors de notre étude bibliographique sont :

- La génération de règles d'association, décrite dans la section II.3,
- L'interprétation du treillis à partir de visualisation et de navigation (section II.4),
- La sélection (ou l'élimination) d'information à partir du calcul de mesures sur le treillis (section II.5.

## II.3. **Génération de règles d'association**

Une règle d'association est définie comme une implication entre deux ensembles d'attributs. Elle est de la forme suivante :

$$R: X \Longrightarrow Y \; tel \; que \; X \subseteq I, Y \subseteq I \; et \; X \cap Y = \emptyset$$

Les règles d'association permettent ainsi de faire « parler » les données en identifiant les inférences logiques entre les différents objets et attributs du contexte formel.

Étant donné que le nombre de règles d'association extraites peut être très élevé, plusieurs mesures ont été proposées pour leur vérification. Nous présentons ici les deux mesures les plus fréquemment utilisées :

– Le *support* : le support de la règle R, noté support(R) est égal au support $(X \cup Y)$, c'est-à-dire la fréquence d'apparition simultanée de l'ensemble $X \; et \; Y$.

– La *confiance* : elle exprime le rapport entre support $(X \cup Y)$ et le support $(X)$.

Les mesures de support et de confiance peuvent être utilisées pour sélectionner les règles d'inférence en conservant celles dont les valeurs associées dépassent un certain seuil.

Le Tableau II-2 représente l'ensemble des règles d'association générées à partir du contexte formel du Tableau II-1, ainsi que les valeurs de support et de confiance de chacune.

**Tableau II-2 : Exemples de règles d'association**

| | | | Support | Confiance |
|---|---|---|---|---|
| {3G} | → | {Morning} | 100% | 100% |
| {Morning} | → | {3G} | 100% | 100% |
| **{Transportation}** | **→** | **{3G, Morning}** | **83%** | **100%** |
| {Home} | → | {3G, Morning} | 83% | 100% |
| {Afternoon} | → | {3G, Morning, Transportation} | 67% | 100% |
| {Evening} | → | {3G, Home, Morning, Transportation} | 67% | 100% |
| {Home, Transportation} | → | {3G, Evening, Morning} | 67% | 100% |
| {University} | → | {3G, Afternoon, Morning, Transportation} | 50% | 100% |
| {Afternoon, Evening} | → | {3G, Home, Morning, Transportation} | 50% | 100% |
| {Afternoon, Home} | → | {3G, Evening, Morning, Transportation} | 50% | 100% |
| {Coffe break} | → | {3G, Afternoon, Evening, Home, Morning, Transportation, University} | 33% | 100% |
| {Evening, University} | → | {3G, Afternoon, Coffe break, Home, Morning, Transportation} | 33% | 100% |
| {Home, University} | → | {3G, Afternoon, Coffe break, Evening, Morning, Transportation} | 33% | 100% |
| {Lunch break} | → | {3G, Afternoon, Coffe break, Evening, Home, Morning, Parents' home, Restaurant, Transportation, University} | 17% | 100% |
| {Parents' home} | → | {3G, Afternoon, Coffe break, Evening, Home, Lunch break, Morning, Restaurant, Transportation, University} | 17% | 100% |
| {Restaurant} | → | {3G, Afternoon, Coffe break, Evening, Home, Lunch break, Morning, Parents' home, Transportation, University} | 17% | 100% |
| {3G} | → | {Transportation} | 83% | 83% |
| {3G} | → | {Home} | 83% | 83% |
| {Morning} | → | {Transportation} | 83% | 83% |
| {Morning} | → | {Home} | 83% | 83% |
| {Transportation} | → | {Afternoon} | 67% | 80% |
| {Transportation} | → | {Evening, Home} | 67% | 80% |
| {Home} | → | {Evening, Transportation} | 67% | 80% |
| {Afternoon} | → | {University} | 50% | 75% |
| {Afternoon} | → | {Evening, Home} | 50% | 75% |
| {Evening} | → | {Afternoon} | 50% | 75% |
| {Home, Transportation} | → | {Afternoon} | 50% | 75% |
| {University} | → | {Coffe break, Evening, Home} | 33% | 67% |
| {Afternoon, Evening} | → | {Coffe break, University} | 33% | 67% |
| {Afternoon, Home} | → | {Coffe break, University} | 33% | 67% |

| | | | | |
|---|---|---|---|---|
| {Coffe break} | → | {Lunch break, Parents' home, Restaurant} | 17% | 50% |
| {Evening, University} | → | {Lunch break, Parents' home, Restaurant} | 17% | 50% |
| {Home, University} | → | {Lunch break, Parents' home, Restaurant} | 17% | 50% |

Considérons la règle surlignée dans la le Tableau II-2. La confiance de cette règle est de 100%, ce qui signifie que toutes les applications associées à *Transportation* sont également associées à *3G* et *Morning*. Le support de cette règle est de 83% car, dans le contexte formel du Tableau II-1, 5/6 des applications sont associées à *Transportation*, *3G* et *Morning*.

Nous citons dans la suite de cette section quelques exemples d'exploitation des résultats de l'ACF via la génération de règles d'association.

### II.3.1. **Filtrage de règles d'association**

L'un des défis concernant l'extraction de règles d'association est la sélection des règles les plus « pertinentes ». L'ACF est utilisée dans (Alam, Buzmakov, Codocedo, & Napoli, 2015) pour améliorer la cohérence et la complétude de la base DBPedia[11] contenant de gros volumes de données liées. Les auteurs conservent ici les règles d'association qui ont une mesure de confiance égale à 100%.

Néanmoins, les mesures de confiance et de support ne sont pas toujours suffisantes pour permettre une sélection fine des règles d'association « significatives ».

Les règles d'association sont utilisées dans (Alam & Napoli, 2014) pour classifier les résultats de requêtes SPARQL[12] dans le contexte du web sémantique. La méthode proposée est appliquée sur des données complexes et les règles d'association sont filtrées de trois façons :

[11] Le projet DBpedia est une initiative collaborative pour extraire des informations structurées à partir de Wikipédia et rendre cette information accessible sur le web.

[12] SPARQL est un langage de requête pour l'interrogation des données RDF (Resource Description Framework).

- tout d'abord en fonction du support,
- deuxièmement en fonction du nombre d'éléments dans le corps de la règle,
- troisièmement en sélectionnant les règles ayant des attributs différents en tête de l'implication (pour obtenir des implications entre attribut-attribut).

### II.3.2. **Règles d'association et chemins de concepts**

Les auteurs de (Antoni, Gunis, Krajci, Krıdlo, & Snajder, 2014) ont utilisé l'ACF pour analyser les tâches et les objectifs éducatifs de cinq enseignants en informatique. Les auteurs ont tout d'abord construit un treillis pour chaque enseignant et les ont exploités grâce à des règles d'association. Ces travaux ont permis aux auteurs de recommander une méthodologie pour identifier les éléments à enseigner au début d'un enseignement, en phase d'ancrage et de diagnostic. Ces travaux ont également permis d'identifier les éléments les moins appropriés dans le processus éducatif. Cet article nous semble très intéressant dans le contexte de nos travaux, puisque les auteurs proposent une méthodologie d'utilisation d'un treillis de Galois. Néanmoins cette approche est très spécifique au domaine de l'enseignement, dans lequel la notion de « chemin d'apprentissage » est très pertinente.

Dans (Tang, Buzmakov, Toussaint, & Napoli, 2016), les auteurs ont proposé une méthode pour intégrer la connaissance des experts du domaine dans un treillis de Galois, tout en préservant sa structure. Les contraintes indiquées par les experts sont exprimées sous forme de dépendances d'attributs. Les auteurs font la différence entre la représentation des données basées sur le treillis et la représentation des données imaginées par un expert du domaine, c'est-à-dire que l'expert peut contredire certaines implications entre les attributs. Les implications sont construites à partir du contexte formel ou du treillis ; la méthode fournit aux experts les concepts formels qui satisfont leurs contraintes ; les autres ne sont pas pris en compte. Le treillis initial est donc modifié en fonction des concepts sélectionnés ; ce travail réduit la taille du treillis en ne conservant que les concepts qui vérifient les contraintes d'un expert, au prix bien entendu d'une potentielle perte d'information. On peut noter qu'au-delà de la réduction du treillis, les auteurs ne fournissent pas de méthodologie pour interpréter le treillis résiduel.

L'ACF et les treillis de Galois peuvent être utilisés pour la recommandation d'idées ou de personnes dans le contexte des plateformes collaboratives (Ignatov, Kaminskaya,

Bezzubtseva, Konstantinov, & Poelmans, 2013; Ignatov, Kaminskaya, Konstantinova, Malyukov, & Poelmans, 2014). Le système développé dans ces articles est basé sur plusieurs modèles et méthodes pour l'analyse des données : l'ACF pour le regroupement des informations, ainsi que l'extraction de règles d'association pour l'extraction de mots clés à partir de textes. Les auteurs ont également proposé d'appliquer des méthodes d'analyse de réseaux sociaux (Social Network Analysis). Les auteurs ont défini deux types de recommandation utilisés dans différentes phases d'un projet de crowdsourcing[13]: la recommandation de personnes partageant les mêmes idées et la détection d'antagonistes, i.e., des utilisateurs qui ont discuté des mêmes sujets mais avec des opinions opposées.

Les règles d'association peuvent également être utilisées avec l'extension de l'ACF que sont les *pattern structures* (Kuznetsov, 2013), qui s'appliquent à des données complexes telles que les formules, les graphes, intervalles numériques, etc.

L'avantage de l'extraction des règles d'association est que toutes les implications existantes entre les données sont identifiées (à la différence de certaines des techniques « manuelles » décrites dans la section suivante et reposant sur l'interprétation visuelle du treillis). Néanmoins le nombre de règles générées est si élevé que les utilisateurs non experts de l'analyse de données peuvent éprouver des difficultés à identifier les plus significatives. Cette manière d'interpréter les résultats de l'ACF peut sembler trop technique à un utilisateur non spécialiste.

## II.4. **Visualisation et navigation dans le treillis**

Une autre façon d'exploiter les résultats de l'ACF consiste à interpréter le treillis à partir de techniques de visualisation et de navigation. Nous présentons ici une sélection de

[13] Le mot "crowdsourcing", proposé par Jeff Howe en 2006 (Howe, 2006), est une combinaison de "crowd" et "outsourcing" et peut se traduire par « production participative », consistant à faire appel à la créativité et au savoir-faire de la collectivité.

travaux récents qui utilisent la représentation visuelle du treillis pour en déduire des informations.

### II.4.1. **Utilisateurs experts**

Les travaux de recherche de (Priss, 2013) utilisent l'ACF pour comprendre le processus d'apprentissage des étudiants qui découvrent la programmation. Cette application de l'ACF au monde de l'éducation est un exemple intéressant du potentiel de cette technique pour une utilisation « grand public ». Nous notons cependant deux limites : tout d'abord l'interprétation est ici effectuée par des experts de l'ACF – serait-elle aussi intuitive pour des non-spécialistes ? D'autre part, les treillis générés dans ce contexte sont de petite taille et leur analyse manuelle est possible ; que se passe-t-il, y compris pour des spécialistes de l'ACF, lorsque l'on a affaire à des treillis de grande taille ?

Les travaux de (Carbonnel, Bertet, Huchard, & Nebut, 2016) sont un autre exemple d'utilisation de l'ACF destiné à des utilisateurs non spécialistes de l'analyse de données. Le treillis de Galois est utilisé ici pour l'ingénierie des lignes de produits, et plus particulièrement pour trouver les paramètres adéquats pour leur configuration. Les auteurs montrent tout d'abord comment les configurations et les caractéristiques peuvent être réunies dans un contexte formel. Ils proposent aussi des interprétations du treillis, où les caractéristiques partagées par les mêmes configurations sont regroupées dans les mêmes concepts.

Voici un exemple d'interprétation : si des caractéristiques sont partagées par tous les concepts, alors cela signifie qu'elles sont liées à toutes les configurations. Les caractéristiques qui sont liées à des configurations spécifiques se trouvent quant à elles isolées dans les concepts les plus bas dans le treillis. Les auteurs déduisent des informations par inférence (sans pour cela extraire les règles d'association de manière systématique comme nous l'avons vu précédemment avec d'autres utilisations de l'ACF). Par exemple : si une caractéristique (*feature*) F1 se trouve dans un sous-concept d'une autre fonctionnalité F2, alors F1 implique F2, autrement dit la caractéristique F1 est liée au même ensemble de configurations que celles de F2, avec des configurations en plus.

Ces travaux sont intéressants dans le cadre de notre recherche, dans la mesure où ils fournissent à un utilisateur non expert de l'ACF des pistes pour l'interprétation du treillis de

Galois obtenu. Nous souhaitons aller encore plus loin en proposant une méthodologie d'interprétation complète qu'un utilisateur non expert pourrait suivre pas à pas.

Avec l'aide d'experts dans le domaine d'ACF, (Stanley & Astudillo, 2015) ont proposé une méthode dont les étapes sont bien détaillées, pour construire une ontologie concernant le treillis. Cette ontologie est utilisée ensuite pour instancier un wiki sémantique accessible à des utilisateurs non experts de l'analyse de données.

Les auteurs de (Coulet, Domenach, Kaytoue, & Napoli, 2013) fournissent un exemple d'utilisation de l'ACF pour l'extraction de connaissances dans des données biomédicales. Ils abordent tout d'abord le problème du prétraitement des données, nécessaire du fait que les données biomédicales ne sont pas binaires et qu'elles sont volumineuses. Les auteurs résolvent ce problème en utilisant une ontologie. Après la construction du contexte formel et la génération du treillis, les auteurs proposent une interprétation graphique de ce dernier pour trouver des corrélations entre les données. Cependant, la question du passage à l'échelle se pose : le treillis généré contient 204 801 concepts et l'interprétation graphique manuelle n'est plus possible. La solution proposée par les auteurs est d'avoir recours à un expert du domaine. Nos travaux visent à proposer une méthodologie pour l'interprétation des treillis de Galois qui soit utilisable y compris sur des treillis de grande taille.

Une autre forme possible d'aide à l'utilisateur non expert pour l'exploitation des résultats de l'ACF est proposée dans (Mimouni, Nazarenko, & Salotti, 2015), sous la forme de requêtes appliquées au treillis. Les auteurs utilisent l'ACF pour structurer, interroger et parcourir des collections de documents juridiques. Dans ces travaux, l'ACF et la Relational Concept Analysis (RCA), son adaptation aux données relationnelles, permettent de prendre en compte non seulement le contenu sémantique des documents, mais aussi les liens intertextuels qui les relient. Par exemple, dans le domaine juridique, les données sources sont liées par plusieurs types de relations, ce qui permet de construire un réseau de documents. La méthodologie proposée est la suivante : après avoir collecté les données afin de construire les treillis de Galois, ceux-ci sont enrichis par les informations relationnelles pour obtenir un ensemble de treillis dits relationnels ; ces derniers sont ensuite exploités grâce à des requêtes des utilisateurs. Une recherche des concepts formels les plus pertinents est alors effectuée à partir des attributs qui apparaissent dans la requête de l'utilisateur.

Dans le même esprit, nous avons aussi identifié les travaux de (Dolques, Mondal, Braud, Huchard, & Le Ber, 2014), qui visent également à faciliter l'utilisation de l'ACF, mais qui se situent plus en amont : en effet les auteurs proposent une approche pour transformer des données afin de pouvoir utiliser la RCA. Par contre, ils ne fournissent pas d'aide pour l'interprétation des treillis obtenus.

### II.4.2. **Outils de navigation**

Les auteurs de (Alam, Napoli, & Osmuk, 2015) proposent un outil basé sur l'ACF, appelé RV-Xplorer, permettant à des experts d'analyser des données Web structurées au format Resource Description Framework (RDF), l'un des standards de représentation de données sémantiques (Alam & Napoli, 2014). L'objectif est de trouver des liens intéressants et de répondre à des questions telles que : Quels sont les principaux sujets de recherche et les principaux chercheurs dans l'équipe ? Quel est le domaine de recherche principal du chef de l'équipe et des diverses personnes clés ? Les auteurs proposent de répondre à ce genre de questions avec RV-Xplorer en navigant dans le treillis. Pour chaque concept sélectionné dans le treillis, RV-Xplorer permet de visualiser l'extension et l'intension (i.e., les objets et leurs attributs communs), ainsi que les liens vers les concepts supérieurs et les concepts inférieurs. L'outil donne la possibilité de naviguer selon plusieurs points de vue, changer d'espace de navigation, détendre une zone ou encore de cacher des parties non intéressantes de la vue. Pour chaque intension ou extension donnée, il est possible de visualiser les autres concepts du treillis qui les possèdent. Une mesure est également proposée pour choisir les concepts d'intérêt dans la partie inférieure du treillis ; il s'agit de la mesure de *summarization*, qui permet de décrire la répartition des objets contenus dans l'extension du concept courant dans les sous-concepts qui lui sont reliés. Cette mesure, égale au rapport du nombre d'objets de l'extension entre un concept et un concept inférieur (sous-concept), fournit des indications sur le niveau inférieur dans le treillis, à partir de la position courante de l'utilisateur durant sa navigation. Les auteurs précisent que cet utilisateur est un expert. Notre objectif est d'aller encore plus loin en fournissant une méthodologie pour l'interprétation des treillis de Galois, accessible à des non spécialistes de l'ACF. Comme nous le décrivons dans le Chapitre III, notre approche repose sur des mesures très simples ; nous pourrons tout à fait l'étendre en incluant cette mesure de *summarization*.

D'autres outils ont été développés pour la construction et la visualisation de treillis, comme dans (Kis, Sacarea, & Troanca, 2016) où les auteurs présentent une collection d'outils pour l'ACF (exemple http://latviz.loria.fr/). L'outil LatViz est également présenté dans (Alam, Le, & Napoli, 2016b, 2016a). Plusieurs fonctionnalités sont proposées pour permettre l'interaction avec les experts, comme afficher les sous-concepts et super-concepts d'un concept donné dans le treillis.

Une autre approche consiste à proposer des chemins entre les regroupements d'objets dans le treillis pour permettre aux visiteurs de naviguer de l'un à l'autre en fonction de leurs points d'intérêt. Les auteurs de (Wray & Eklund, 2014) s'appuient pour cela sur une mesure de similarité, ainsi que sur des règles d'association pour représenter les relations entre les objets d'un même cluster. Ils utilisent l'ACF pour une application dans les musées : il s'agit de trouver les relations entre les objets des collections et de proposer une aide au visiteur en lui proposant un itinéraire pour sa visite du musée.

### II.4.3. **Outils de visualisation et d'analyse de grands treillis**

Les auteurs de (Neznanov, Ilvovsky, & Kuznetsov, 2013; Neznanov & Parinov, 2014) proposent un outil pour la construction de treillis, après avoir dressé un panorama des différentes approches d'ACF. Les auteurs indiquent que la plupart des outils existants permettent de construire des treillis à partir de contextes formels de petite taille. Selon eux, il manque les prétraitements de données nécessaires pour permettre aux chercheurs d'utiliser ces outils pour analyser des données volumineuses et complexes. Afin de pallier ce problème, ils proposent un outil appelé FCART, utilisant des méthodes de clustering pour construire des treillis à partir de données volumineuses. FCART fournit une interface utilisateur qui lui permet de visualiser et de naviguer dans des grands treillis. Cet outil est très intéressant et les auteurs proposent une méthodologie pour son utilisation, en quatre étapes (Neznanov & Parinov, 2016), allant du stockage de données jusqu'à la construction du treillis. Par contre, l'analyse du treillis obtenu reste manuelle.

Dans (Grissa, Comte, Pujos-Guillot, & Napoli, 2016b, 2016a), l'ACF est utilisée pour l'analyse de données biologiques complexes, possédant un nombre élevé d'attributs. Les auteurs proposent de combiner des approches numériques pour sélectionner les attributs les plus pertinents, et l'approche symbolique que constitue l'ACF pour l'aspect visualisation.

Les résultats obtenus consistent en l'identification des 10 biomarqueurs prédictifs potentiels pour les maladies métaboliques. Ces travaux montrent bien l'intérêt de l'ACF dans des domaines très variés et pour des utilisateurs non spécialistes de l'analyse de données. L'accent est ici porté sur la sélection d'attributs en amont de la construction du treillis de Galois. Les méthodes que nous présentons dans la section suivante s'intéressent quant à elles à l'interprétation (et parfois la réduction) du treillis, une fois que celui-ci est construit.

## II.5. **Sélection (ou élimination) d'information à partir de mesures**

### II.5.1. **Réduction du contexte formel**

(Ganter & Glodeanu, 2014) utilisent le treillis de Galois pour réaliser une factorisation du contexte formel : un treillis de Galois est construit à partir d'un contexte formel initial. Ce treillis est ensuite utilisé pour construire deux contextes formels, Objets×Concepts et Concepts×Attributs, afin d'effectuer une factorisation. Par contre, les auteurs ne fournissent pas de méthode pour interpréter le treillis.

### II.5.2. **Sélection de concepts pertinents**

Les auteurs de (Kuznetsov & Makhalova, 2015) abordent la question de la sélection des concepts pertinents dans les treillis ; ceci est essentiel lorsque ces derniers contiennent un grand nombre de concepts. Quand la taille et la densité du contexte formel augmentent, le treillis peut croître de façon exponentielle et son interprétation devient très délicate et impossible par la seule visualisation de sa structure. (Kuznetsov & Makhalova, 2015) effectuent une comparaison des méthodes de réduction de treillis, reposant sur des mesures conçues pour évaluer des ensembles d'*itemsets* (concepts formels), des *itemsets* arbitraires et des mesures pour évaluer l'appartenance à un niveau de base (approche fondée sur la psychologie) :

- Indices pour les concepts formels: stabilité *Stabi(A,B)*, échelle de stabilité logarithmique *LStab(c)*, stabilité de probabilité *PB*, Robustesse $r(c,\alpha)$, Séparation *s(A, B)*, métriques de niveau de base *BLs*, diverses mesures de similarité $sim_{SMC}(B1,B2)$ $sim_{J}(B1,B2)$, approche de prévisibilité *P(A,B)*, validité de *Cue(CV)*, regroupement d'entités de catégorie *(CFC)*, utilitaire de catégorie *(CU)*.

- Métriques pour *itemsets* arbitraires: Fréquence (support) *sup(A,B)*, *lift(B)*, force collective *cs(B)*.

Ces mesures permettent de diminuer la taille du contexte formel et/ou de sélectionner les concepts les plus intéressants. L'objectif de nos travaux est de proposer une méthodologie pour aider un utilisateur non expert de l'ACF à interpréter les résultats qu'il obtient. Notre méthodologie peut tout à fait intégrer certaines de ces mesures existantes.

(Buzmakov, Kuznetsov, & Napoli, 2014) utilisent un contexte formel pondéré. Ils présentent également des mesures telles que la stabilité et la pertinence pour identifier les concepts les plus significatifs du treillis. L'approche présentée est intéressante mais nous paraît difficile d'accès pour un utilisateur non spécialiste.

L'idée d'identifier des clusters (regroupements) de données avec l'ACF est utilisée dans (Andrews, Brewster, & Day, 2016) pour regrouper dans des concepts formels les ressources relatives au crime organisé. Les données analysées sont des tweets, et la détection des concepts importants dans le treillis (grâce à la mesure de support) permet d'alerter la police. Un autre exemple d'utilisation de mesures est donné dans (Ciobanu, Horne, & Vuaideanu, 2014), où il s'agit d'extraire de l'information à partir d'une collection de documents Web.

Les travaux de (Coste, Garet, Groisillier, Nicolas, & Tonon, 2014) sont une application de l'ACF à la bio-informatique : l'objectif est de réaliser une classification automatique d'enzymes. Une fois le treillis construit, les « meilleurs » concepts sont sélectionnés à partir de mesures telles que le support.

### II.5.3. **Réduction de treillis**

Les travaux de (Makhalova, Ilvovsky, & Galitsky, 2015) montrent que la sélection de concepts pertinents est également une préoccupation lorsque l'on utilise les pattern structures. Celles-ci sont mises en œuvre ici pour construire des clusters à partir d'un texte. Les auteurs présentent plusieurs mesures pour faciliter l'interprétation des concepts et éliminer les clusters redondants.

(Ikeda, Otaki, & Yamamoto, 2014) appliquent l'ACF au domaine de la fouille de processus (*process mining*). Dans ce travail, les auteurs visent à améliorer les processus

d'entreprise, en identifiant des séquences d'événements dont les interruptions sont fatales à l'exécution des processus ; l'extension de chaque concept est un ensemble de types d'événement et l'intention est l'ensemble des personnes qui ont été à l'origine de ces évènements. Afin d'obtenir un contexte formel, une matrice est construite, contenant la fréquence d'implication des personnes dans les divers événements. Le contexte formel binaire est ensuite obtenu en remplaçant les fréquences non nulles par 1 et le reste par 0. Une fois le treillis construit, comme dans les travaux mentionnés précédemment, les concepts importants sont identifiés grâce à des mesures définies sur le treillis. Les auteurs présentent notamment une mesure d'importance pour chaque concept, calculée en fonction de l'intention et de l'extension de chaque concept, ainsi qu'une mesure de charge (*Load(c)*), calculée à partir de la fréquence de chaque concept. Enfin, la mesure de faiblesse est obtenue en multipliant la mesure d'importance et la mesure de charge. Cette mesure de faiblesse permet de supprimer certains concepts pour améliorer les processus, en supprimant le point le plus faible ou en réduisant la faiblesse totale. L'approche proposée ici est très intéressante pour nos travaux. Néanmoins les auteurs n'indiquent pas comment l'utiliser automatiquement à grande échelle : par exemple, combien de concepts faut-il éliminer dans le treillis ? L'exemple présenté montre l'élimination d'un seul concept.

## II.6. **Conclusion**

Dans ce chapitre nous avons présenté un éventail de travaux de recherche appliquant l'analyse de concepts formels à differents domaines et sur des types de données variés.

Nous avons étudié l'utilisation de l'ACF décrite dans les 4 conférences suivantes : CLA (2016, 2015, 2014), ICFCA (2015, 2014, 2013), FCA4AI (2016, 2015, 2014) et ICCS (2016, 2014, 2013).

**Tableau II-3 : Synthèse de l'état de l'art sur l'exploitation des résultats de l'ACF**

| Méthodes d'exploitation des treillis de Galois | Points positifs | Points négatifs |
|---|---|---|
| **Règles d'association** | - Implications simples à comprendre.<br>- Identification de toutes les implications entre les données. | - Nombre de règles générées souvent très élevé.<br>- Difficulté pour les utilisateurs non experts à identifier les règles les plus pertinentes. |
| **Visualisation et navigation dans le treillis** | - Interprétation intuitive pour des utilisateurs non spécialistes<br>- Possibilité de naviguer à différents niveaux de détail. | - Que se passe-t-il, lorsque l'on a affaire à des treillis de grande taille ?<br>- Comment interpréter l'intégralité du treillis ? |
| **Mesures pour la sélection d'information et la réduction de treillis** | - Sélection de concepts pertinents selon des métriques données. | - Expertise requise pour comprendre les mesures et sélectionner la plus adaptée dans un contexte donné. |

Le Tableau II-3 montre les points positifs et les points négatifs de chaque famille de méthodes présentées dans les trois sections précédentes pour exploiter les résultats de

l'ACF : la génération de règles d'association, l'interprétation des treillis à partir de visualisation et de navigation, et la sélection ou l'élimination d'information à partir de calcul de mesures sur les treillis.

L'avantage de l'extraction des règles d'association est que toutes les implications existantes entre les données sont identifiées, par contre le nombre de règles générées est souvent si élevé que les utilisateurs non experts de l'analyse de données peuvent éprouver des difficultés à identifier les plus significatives.

Deuxièmement pour les travaux liés à la visualisation et la navigation dans le treillis permettent aux utilisateurs non spécialistes de faire des interprétations aisées, et de naviguer à différents niveaux de détail dans le treillis de Galois. Par contre la représentation et la manipulation de treillis de grande taille demeurent des défis.

Enfin, la plupart des mesures que nous avons identifiées dans la littérature visent à séléctionner ou éliminer des concepts particuliers pour réduire les treillis, ce qui n'est pas notre objectif. De plus, elles sont souvent complexes et difficiles à appréhender pour un utilisateur non expert. La sélection de la mesure la mieux adaptée à un cas d'étude précis peut être délicate.

Dans notre thèse nous souhaitons aller au-delà de cet état de l'art en proposant une méthodologie d'interprétation de treillis de galois[14], pour qu'un utilisateur non expert puisse les exploiter. Nous présentons dans le chapitre suivant les mesures que nous proposons, qui sont au cœur de cette méthodologie.

[14] La méthodologie complète est présentée dans la section VII.2

# CHAPITRE III :
# MESURES POUR L'INTERPRÉTATION DES TREILLIS DE GALOIS

# Sommaire

## III.1. **Introduction**

L'objectif de nos travaux de thèse est de proposer une méthodologie afin de permettre à des utilisateurs non experts de L'Analyse de Concepts Formels d'interpréter eux-mêmes les treillis de Galois qu'ils génèrent. Cette méthodologie, décrite dans le Chapitre VII, requiert le calcul de diverses mesures. Nous avons listé dans le Chapitre II un éventail d'approches existantes, que nous proposons d'enrichir dans ce chapitre avec des mesures originales que nous illustrons sur des données réelles.

Comme nous l'avons vu dans le Chapitre II, l'ACF permet, à partir d'un ensemble d'objets et d'attributs organisés dans un contexte formel, de construire un treillis de concepts formels, i.e., de regroupements d'objets et des attributs qu'ils ont en commun.

Tout au long de ce chapitre, nous illustrons les mesures proposées sur les données présentées dans le Chapitre II, décrivant le contexte d'utilisation des applications utilisées par 28 étudiants de notre université.

L'utilisation que font les étudiants de leur smartphone est très variée du fait du grand nombre d'applications disponibles et des contextes très différents où elles sont utilisées. Comprendre l'usage que ces étudiants font de leur smartphone ainsi que l'influence de leur contexte permet d'obtenir des informations intéressantes, par exemple pour optimiser certaines applications ou en concevoir de nouvelles. Nous montrons ici comment l'ACF peut être utilisée pour étudier les interactions entre les utilisateurs et leur smartphone et d'évaluer l'impact du contexte sur les applications qu'ils utilisent.

Nous reprenons ici l'exemple du treillis représenté par le digramme de Hasse de la Figure II-1, afin d'en détailler l'interprétation. Nous reproduisons cette figure ici (Figure III-1) afin de faciliter la lecture de ce chapitre.

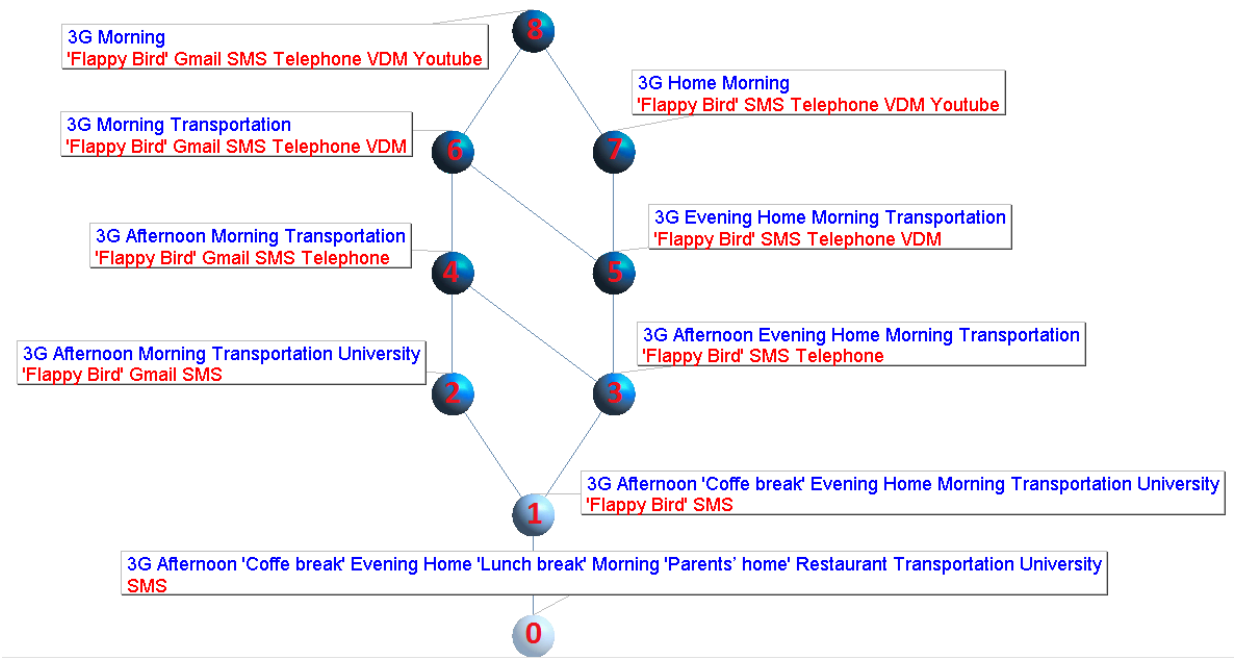



**Figure III-1 : Exemple de treillis de Galois généré à partir du contexte formel du Tableau II-1**

Dans le treillis de la Figure II-1 (et la Figure III-1), les concepts qui sont proches du haut du diagramme de Hasse sont les plus génériques, c'est-à-dire qu'ils contiennent beaucoup d'applications, mais qui ont peu d'éléments de contexte en commun. Par conséquent, nous considérons que ces applications ont une faible *similarité conceptuelle* (nous définissons cette notion de manière formelle dans la section III.3).

A l'inverse, les concepts qui sont proches du bas du treillis sont spécifiques, c'est-à-dire qu'ils contiennent peu d'applications, mais ayant beaucoup d'éléments de contexte en commun, et sont par conséquent très *similaires*.

Donnons à présent quelques exemples d'interprétations possibles de ce treillis (Figure III-1) : le concept n°8 - qui correspond à la borne inférieure du treillis - contient les éléments de contexte *3G et Morning* qui sont communs à toutes les applications : ces éléments de contexte peuvent par conséquent être considérés comme « universels » pour cet utilisateur. De la même manière, l'application *SMS* est « universelle » puisqu'elle apparaît dans le concept n°0 qui est la borne supérieure du treillis, contenant tous les éléments de contexte.

En complément de l'interprétation des bornes supérieure et inférieure du treillis, nous pouvons identifier les applications qui apparaissent fréquemment simultanément dans les concepts formels, c'est ce que nous qualifions de *similarité conceptuelle* : c'est le cas des applications *Gmail* et *SMS*, qui sont présentes ensemble dans quatre concepts. Il est intéressant de noter que cette similarité que nous qualifions de conceptuelle n'est pas liée à une quelconque information sémantique sur la nature de ces applications, bien que toutes les deux soient dédiées à la communication. On apprend également que ces applications ont été associées à l'après-midi (*afternoon*) et au matin (*Morning*), aux *transports* et à l'*université*.

Il est également possible de se concentrer sur les éléments de contexte et d'identifier ceux qui sont conceptuellement similaires, tels que *Morning* et *3G*, car associés à un grand nombre d'applications communes : *Flappy Bird, Gmail, SMS, telephone, VDM et Youtube*. De telles conclusions peuvent être stratégiques, par exemple pour un fournisseur de services de téléphonie mobile, à qui elles permettent de mieux comprendre les préférences et les habitudes de ses clients. L'interprétation manuelle de ce treillis nous fournit donc des informations utiles à propos de l'utilisation du smartphone faite par l'étudiant associé, ainsi qu'à propos de l'impact des éléments de contexte sur les applications qu'il a lancées. Cependant questions se posent : tout d'abord, comment sélectionner les conclusions les plus significatives ? De plus, parmi les observations faites pour certains individus, lesquelles peuvent être généralisées (ou non) à l'ensemble des étudiants ? Cela nécessite de comparer et synthétiser les résultats obtenus à partir de l'analyse des treillis individuels, afin de pouvoir distinguer les comportements fréquents de ceux qui sont plus marginaux.

L'objectif du travail présenté dans ce chapitre est 1) de concevoir et d'implémenter des mesures afin de faciliter l'interprétation de treillis individuels pour comprendre le comportement de chaque utilisateur, et 2) d'analyser différents treillis pour comparer les comportements des utilisateurs les uns par rapport aux autres. Comme nous l'avons expliqué dans le Chapitre II, nos travaux s'inscrivent de manière orthogonale aux travaux dédiés à la réduction de treillis ; en effet, nous souhaitons développer des outils pour permettre une interprétation systématique et pertinente des treillis, quelle que soit leur taille.

Les mesures que nous proposons permettent d'analyser les objets, attributs et concepts des treillis.

## III.2. **Poids conceptuel des objets et des attributs**

Le poids conceptuel d'un objet correspond à la proportion de concepts contenant cet objet. De la même manière, le poids conceptuel d'un attribut correspond à la proportion de concepts contenant cet attribut.

Étant donné un objet *Oi* :

$$Poids\ conceptuel(Oi) = \frac{Nombre\ de\ concepts\ contenant\ (Oi)}{Nombre\ total\ de\ concepts\ dans\ le\ treillis}$$

De la même manière, étant donné un attribut *Ai* :

$$Poids\ conceptuel(Ai) = \frac{Nombre\ de\ concepts\ contenant\ (Ai)}{Nombre\ total\ de\ concepts\ dans\ le\ treillis}$$

La Figure III-2 illustre le calcul de ce poids conceptuel sur les mêmes données que dans la section précédente. On s'intéresse ici au poids conceptuel de l'objet (application) *Telephone*, dans le treillis de la Figure III-1. Cette application apparaît dans 6 concepts du treillis, qui contient 9 concepts au total : son poids conceptuel est donc égal à 66%, comme le montre la Figure III-4, qui indique aussi les poids conceptuels de toutes les autres applications.

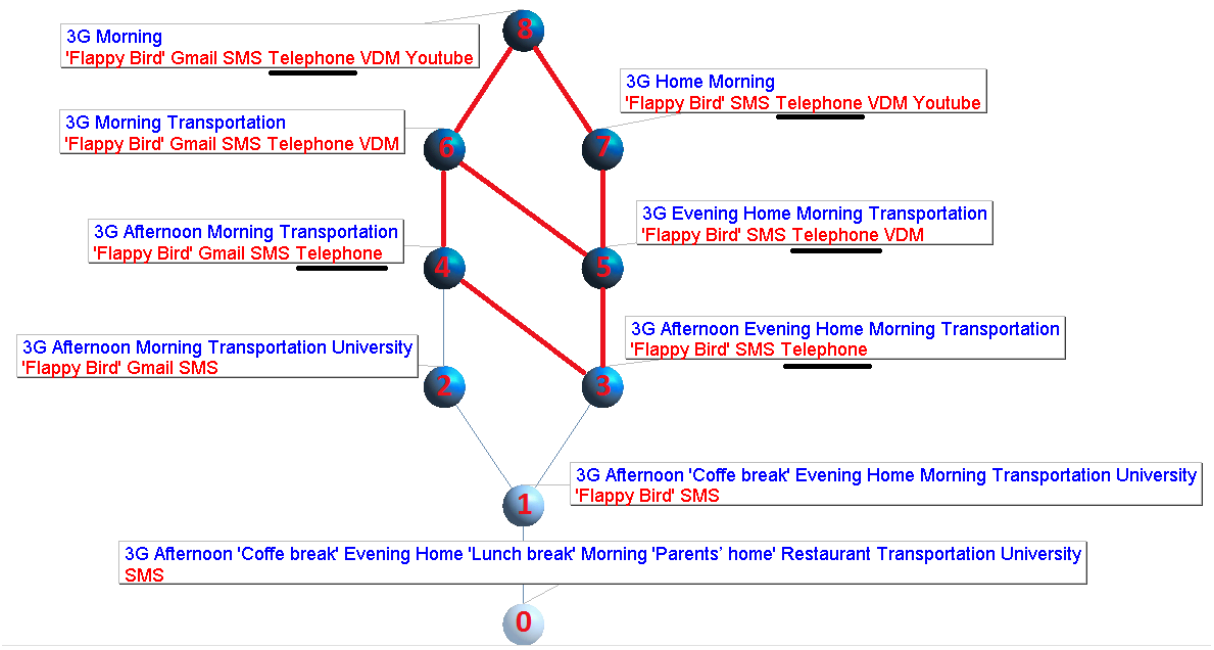



**Figure III-2 : Exemple de calcul du poids conceptuel de l'application *Telephone***

Ce poids que nous qualifions de conceptuel repose sur le clustering réalisé par l'ACF, et fournit des résultats différents de ce que donnerait la simple fréquence d'apparition des objets/attributs dans le contexte formel initial. Par exemple, la fréquence de l'application *Telephone* dans le contexte formel de Tableau II-1 est égale à 55% (Figure III-3), puisque cette application est associée à 6 éléments de contexte sur 11 au total.

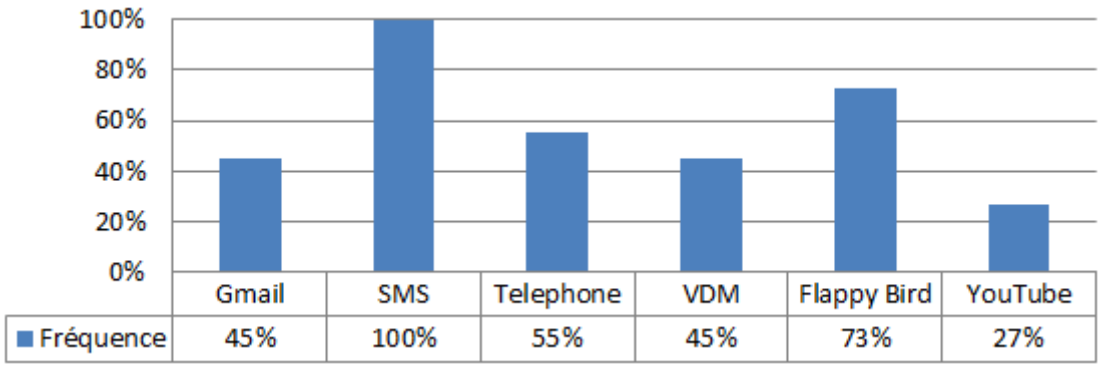



**Figure III-3 : Fréquence des applications dans le contexte formel du Tableau II-1**

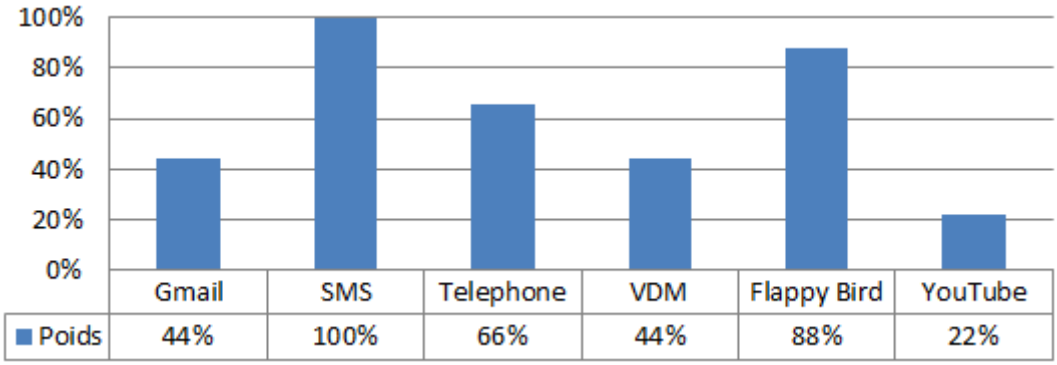



**Figure III-4 : Poids conceptuel des applications dans le treillis de la Tableau II-1**

Les valeurs de poids conceptuel (Figure III-4) permettent de saisir l'importance relative des objets et/ou attributs dans un treillis donné. Ici, elles illustrent donc l'importance relative des applications et/ou éléments de contexte pour un utilisateur de smartphone donné. Cette mesure permet d'identifier les applications qui apparaissent fréquemment dans le treillis, i.e., qui sont souvent regroupées avec d'autres applications du fait d'éléments de contexte communs. Appliquée aux attributs, cette mesure permet de mettre en évidence les éléments de contexte qui sont fréquemment regroupés avec d'autres du fait d'applications communes.

Si l'on considère le poids conceptuel des éléments de contexte *Home* et *Transportation*, illustré sur la Figure III-6, on peut remarquer que bien que leur fréquence dans le contexte formel de Tableau II-1 soit identique (83%) (Figure III-5), leur poids conceptuel est différent (respectivement 55% et 77%) (Figure III-6). En effet, *Transportation* est associé à un plus grand nombre de regroupements d'applications dans le treillis que *Home*.

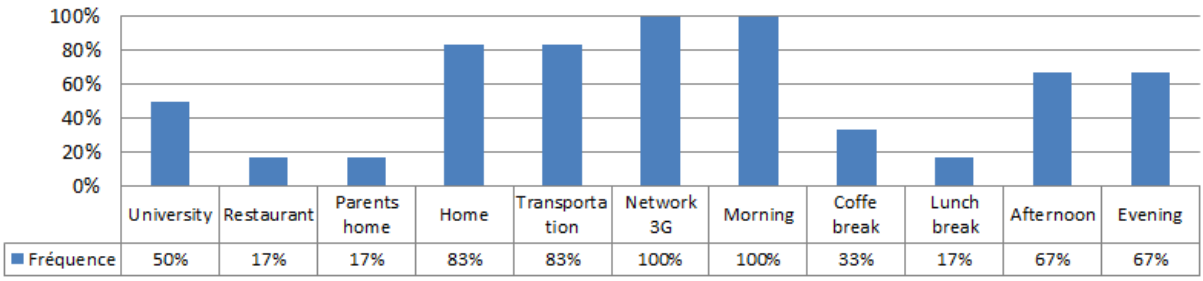
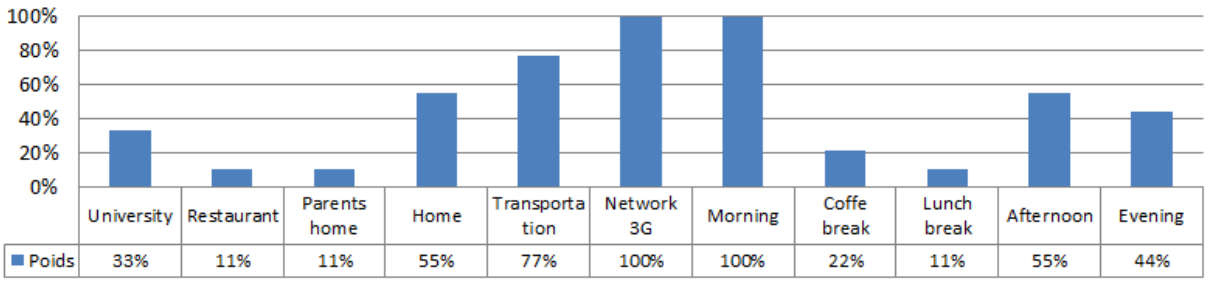



**Figure III-5 : Fréquence des éléments de contexte dans le contexte formel du Tableau II-1**




| | University | Restaurant | Parents home | Home | Transportation | Network 3G | Morning | Coffe break | Lunch break | Afternoon | Evening |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Poids | 33% | 11% | 11% | 55% | 77% | 100% | 100% | 22% | 11% | 55% | 44% |

**Figure III-6 : Poids conceptuel des éléments de contexte du treillis de la Figure III-1**

Cette mesure de poids conceptuel se concentre sur un objet (ou un attribut) spécifique. Dans la section suivante, nous proposons une mesure de similarité conceptuelle, qui permet de comparer deux objets ou deux attributs. Dans notre exemple, il s'agira de comparer deux applications ou deux éléments de contexte.

## III.3. **Similarité conceptuelle entre deux objets ou deux attributs**

Nous définissons la similarité conceptuelle entre deux objets *Oi* et *Oj* comme suit :

$$Similarité\ conceptuelle\ (Oi, Oj) = \frac{Nombre\ de\ concepts\ contenant\ Oi\ et\ Oj}{Nombre\ de\ concepts\ contenant\ Oi\ ou\ Oj}$$

De la même manière, nous définissons la similarité conceptuelle entre deux attributs *Ai* et *Aj* comme suit :

$$Similarité\ conceptuelle\ (Ai, Aj) = \frac{Nombre\ de\ concepts\ contenant\ Ai\ et\ Aj}{Nombre\ de\ concepts\ contenant\ Ai\ ou\ Aj}$$

Cette similarité conceptuelle permet de comparer deux objets, ou deux attributs ; elle est donc complémentaire à la mesure de poids conceptuel présentée dans la section précédente.

Le Tableau III-1 représente la similarité conceptuelle entre tous les attributs (éléments de contexte) du treillis de la Figure III-1. Il s'agit donc ici d'une similarité conceptuelle liée à l'usage de smartphone. Nous avons choisi de représenter la matrice complète (bien qu'elle soit symétrique) pour des raisons de lisibilité.

**Tableau III-1 : Similarité conceptuelle entre les éléments de contexte (attributs) du treillis de la Figure III-1**

| | University | Restaurant | Parents home | Home | Transportation | Network 3G | Morning | Coffe break | Lunch break | Afternoon | Evening |
|---|---|---|---|---|---|---|---|---|---|---|---|
| University | | 33% | 33% | 33% | 43% | 33% | 33% | 67% | 33% | 60% | 40% |
| Restaurant | 33% | | 100% | 20% | 14% | 11% | 11% | 50% | 100% | 20% | 25% |
| Parents home | 33% | 100% | | 20% | 14% | 11% | 11% | 50% | 100% | 20% | 25% |
| Home | 33% | 20% | 20% | | 50% | 56% | 56% | 40% | 20% | 43% | 80% |
| Transportation | 43% | 14% | 14% | 50% | | 78% | 78% | 29% | 14% | 71% | 57% |
| Network 3G | 33% | 11% | 11% | 56% | 78% | | 100% | 22% | 11% | 56% | 44% |
| Morning | 33% | 11% | 11% | 56% | 78% | 100% | | 22% | 11% | 56% | 44% |
| Coffe break | 67% | 50% | 50% | 40% | 29% | 22% | 22% | | 50% | 40% | 50% |
| Lunch break | 33% | 100% | 100% | 20% | 14% | 11% | 11% | 50% | | 20% | 25% |
| Afternoon | 60% | 20% | 20% | 43% | 71% | 56% | 56% | 40% | 20% | | 50% |
| Evening | 40% | 25% | 25% | 80% | 57% | 44% | 44% | 50% | 25% | 50% | |

Une valeur élevée de la mesure de similarité conceptuelle entre deux attributs *Ai* et *Aj*, i.e., proche de 100%, indique que tous les objets associés à l'attribut *Ai* sont également associés à l'attribut *Aj*. Dans notre exemple, si la valeur de similarité conceptuelle entre deux attributs est élevée, cela signifie que la plupart des applications utilisées avec l'élément de contexte *Ai* sont aussi utilisées avec l'élément de contexte *Aj*. Par exemple la plupart des applications utilisées par cet utilisateur à l'université sont aussi utilisées durant la pause-café, avec une connexion *3G* et l'après-midi. On constate également que les éléments de contexte *Restaurant* et *Parents-home* ont une similarité conceptuelle égale à 100%, alors que *Restaurant* et *Network 3G* n'ont une similarité conceptuelle que de 11%.

Si l'on s'intéresse aux éléments de contexte temporel dans nos données, on remarque que *morning* et *afternoon* sont plus fortement similaires à *3G network* que *Lunch break* ne l'est, ce qui est compréhensible puisque d'autres types d'accès au réseau sont disponibles en soirée (probablement depuis le domicile - *home*). Une autre observation est que *morning* est plus similaire à *afternoon* qu'il ne l'est à *Lunch break*, en termes d'applications associées.

La Figure III-7 et la Figure III-8 proposent deux autres types de représentation de la similarité conceptuelle entre les éléments de contexte (attributs) du treillis de la Figure III-1 :

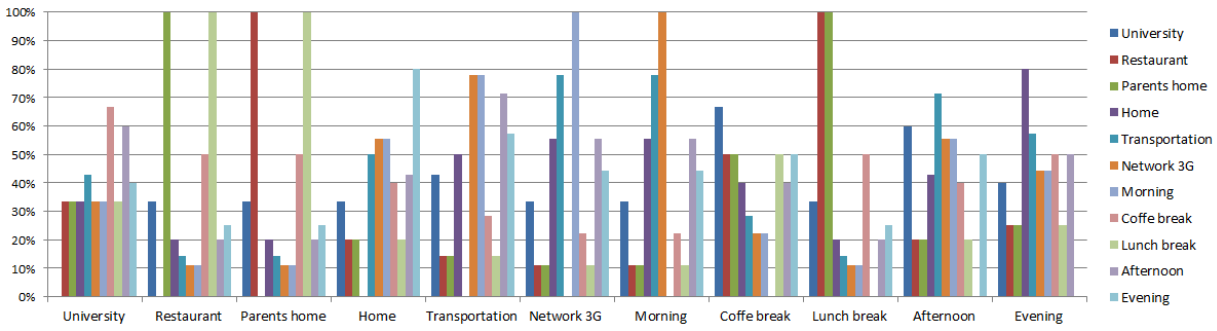



**Figure III-7 : 1er type de représentation de la similarité conceptuelle entre les éléments de contexte (attributs) du treillis de la Figure III-1**

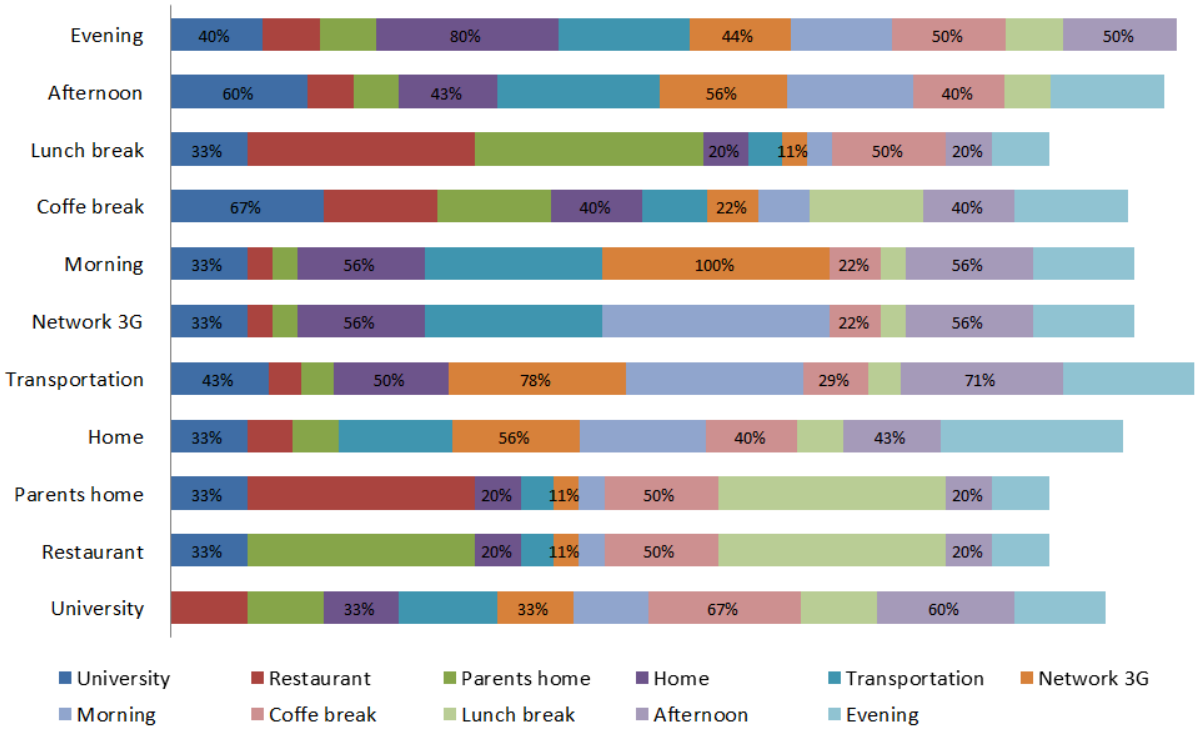



**Figure III-8 : 2e type de représentation de la similarité conceptuelle entre les éléments de contexte (attributs) du treillis de la Figure III-1**

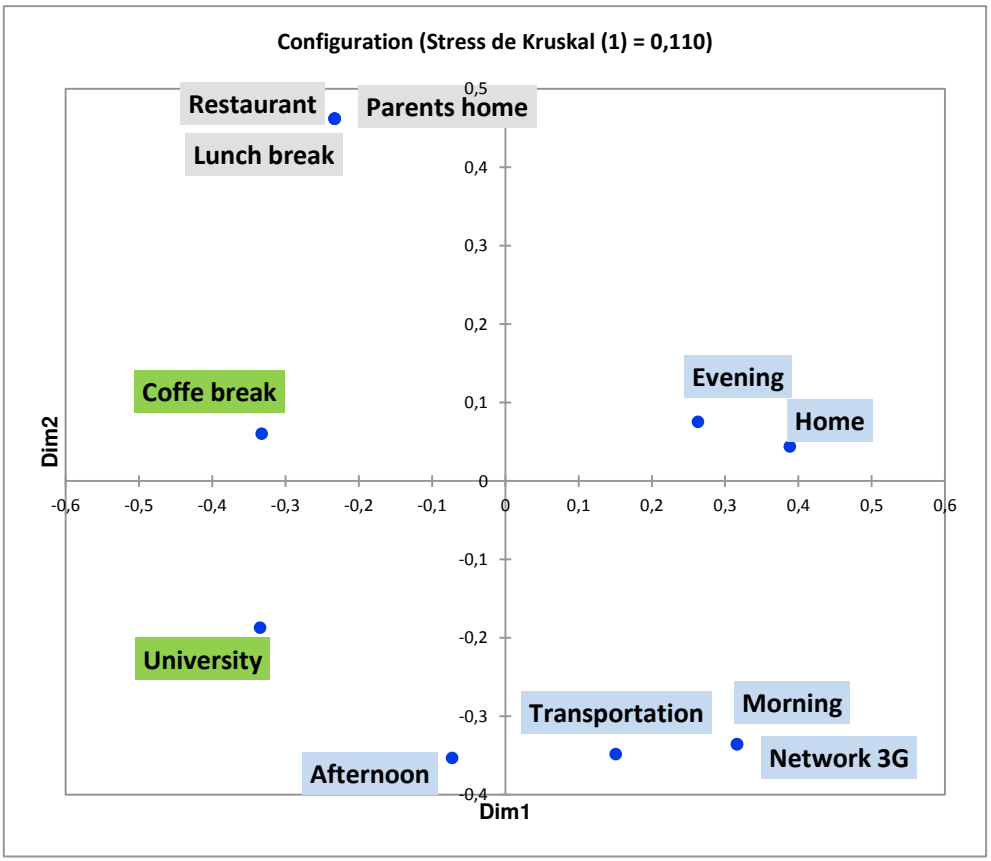



**Figure III-9 : 3e type de représentation de la similarité conceptuelle entre les éléments de contexte (attributs) du treillis de la Figure III-1**

La Figure III-9 fournis une représentation de la similarité conceptuelle entre les éléments de contexte sous forme de carte en deux dimensions. Cette carte est obtenu grâce à la technique de *Multidimensional Scaling* (MDS) (Cox & Cox, 2000) qui permet de passer d'une matrice de proximité (similarité ou dissimilarité) entre une série de *N* objets aux coordonnées de ces mêmes objets dans un espace à *p* dimensions.

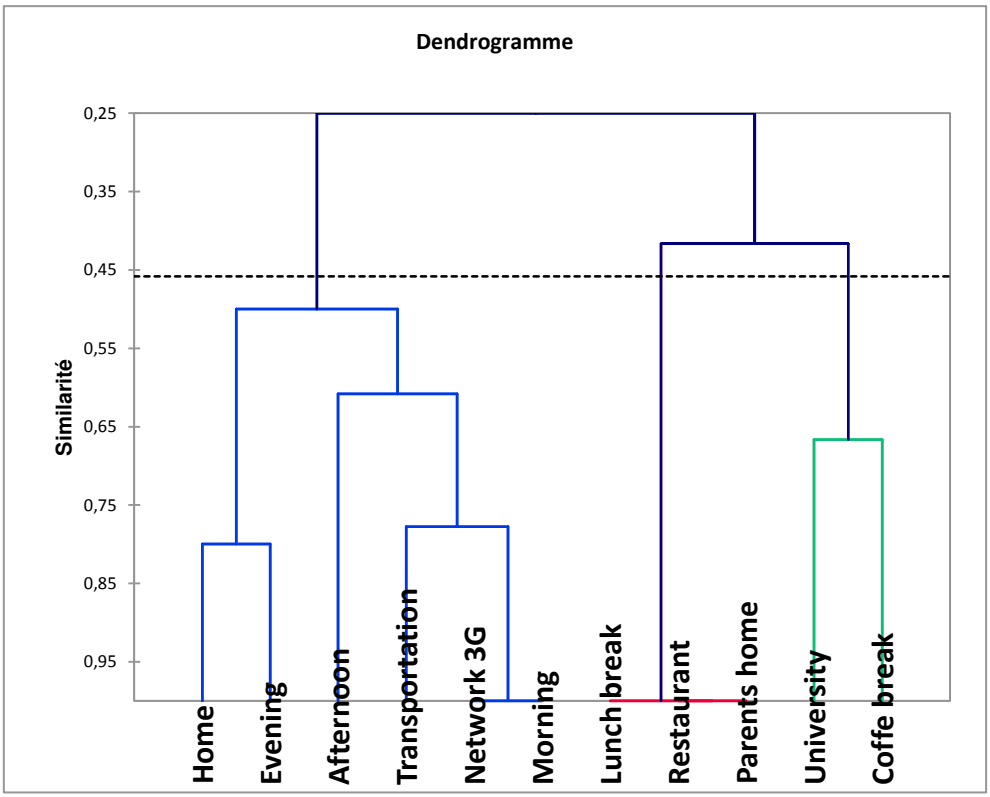



**Figure III-10 : CAH appliquée aux éléments de contexte (attributs) du treillis de la Figure III-1**

La représentation en deux dimensions n'est pas parfaite lorsque le nombre de dimensions initial est supérieur à deux. La distorsion est mesurée par un paramètre appelé stress. Malgré cette imperfection, la représentation cartographique est intéressante car très intuitive : Nous avons enrichi cette carte MDS en reflétant par des couleurs les groupes d'éléments de contexte identifiés par l'algorithme de Classification Ascendante Hiérarchique (CAH) (Johnson, 1967). Cet algorithme place au départ chaque élément dans un cluster diffèrent, et regroupe les deux clusters les plus proches jusqu'à ce qu'il n'y en ait plus qu'un.

La Figure III-10 montre le résultat de la CAH sur les éléments de contexte. La ligne horizontale en pointillés indique le nombre de clusters optimal (ici, 3). C'est pourquoi nous avons utilisé trois couleurs sur la carte MDS de la Figure III-9.

La Figure III-11 représente la valeur de la similarité conceptuelle entre *3G* et *transportation* pour l'ensemble des étudiants qui ont répondu à notre questionnaire. Ce diagramme montre que la similarité conceptuelle entre ces deux éléments de contexte est plutôt faible dans la plupart des treillis, malgré quelques exceptions, ce qui indique que l'observation qui a été faite dans le Tableau III-1 (Similarité conceptuelle = 78%) sur un treillis

spécifique étudié ne peut pas être généralisée. Ce treillis spécifique appartient à l'étudiant E6 qui est présenté dans le Chapitre II par le Tableau II-1 et la Figure II-1 (Figure III-1). La similarité conceptuelle entre *3G* et *transportation* pour l'étudiant E6 est présenté dans la Figure III-11 par la couleur verte. Certains étudiants (E8, E16, E18, E24) n'ont pas utilisé d'applications à la fois avec les deux éléments de contexte *3G* et *transportation*, c'est pour quoi ils ont disparu du diagramme de la Figure III-11.

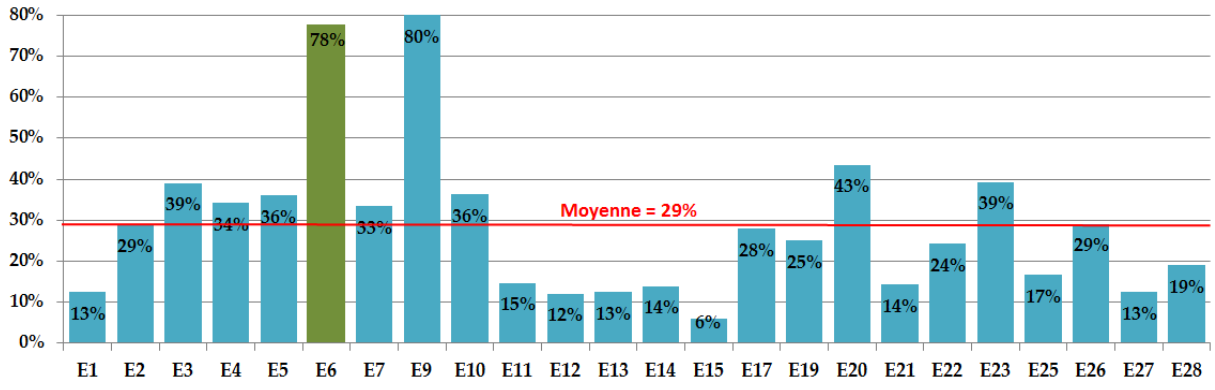



**Figure III-11 : Comparaison des valeurs de similarité conceptuelle entre les éléments de contexte *3G* et *transportation* pour tous les étudiants**

D'autres observations sont possibles si l'on compare les valeurs de la similarité conceptuelle entre d'autres éléments de contexte. Par exemple, la Figure III-12 et la Figure III-13 montrent respectivement la similarité conceptuelle entre *university* et *afternoon* ainsi qu'entre *university* et *morning*.

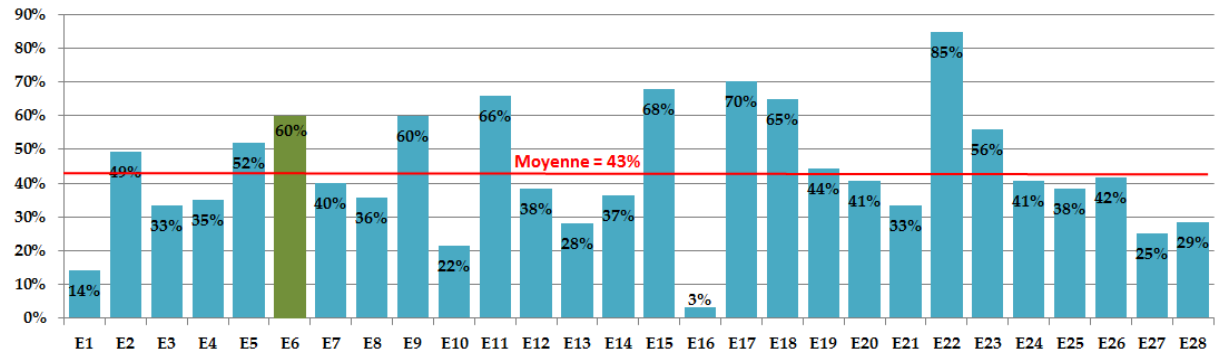



**Figure III-12 : Comparaison des valeurs de similarité conceptuelle entre les éléments de contexte *university* et *afternoon* pour tous les étudiants**

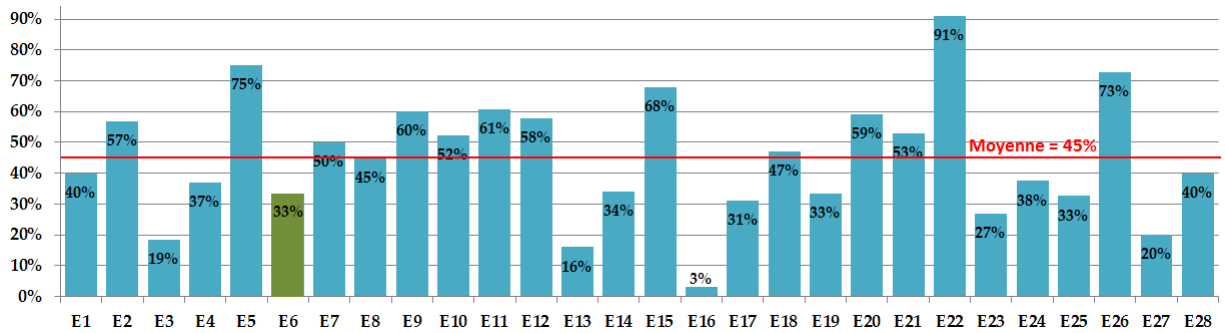



**Figure III-13 : Comparaison des valeurs de similarité conceptuelle entre les éléments de contexte *university* et *morning* pour tous les étudiants**

Après avoir étudié les attributs (ici, les éléments de contexte), nous nous intéressons à présent aux objets (applications). Le Tableau III-2 représente la similarité conceptuelle « d'usage » entre toutes les applications du treillis de la Figure III-1 (étudiant E6).

Certaines applications apparaissent comme similaires, de manière attendue. Par exemple, *Telephone* et *SMS* ont une similarité supérieure à 66%, ou encore *Youtube* et *VDM* avec une similarité proche de 50% ; cependant, il est important de comprendre que la similarité conceptuelle reflète le fait que les applications sont utilisées dans des contextes similaires : c'est pourquoi des applications de nature très différente peuvent être conceptuellement similaire en termes d'usage. C'est le cas pour les applications *Flappy bird* et *SMS*, qui ont une similarité conceptuelle élevée de 88%, alors qu'elles ont des finalités très différentes : loisir et communication. On peut aussi observer que, en termes de similarité (d'usage) et pour l'utilisateur (E6) dont nous avons détaillé les résultats, *Flappy bird* est plus similaire à *SMS* qu'à *Youtube*, ce qui pourrait sembler surprenant.

**Tableau III-2 : Similarité conceptuelle (d'usage) entre tous les objets (applications) du treillis de la Figure III-1**

| | *Gmail* | *SMS* | *Telephone* | *VDM* | *Flappy Bird* | *YouTube* |
|---|---|---|---|---|---|---|
| *Gmail* | | 44% | 43% | 33% | 50% | 20% |
| *SMS* | 44% | | 67% | 44% | 89% | 22% |
| *Telephone* | 43% | 67% | | 67% | 75% | 33% |
| *VDM* | 33% | 44% | 67% | | 50% | 50% |
| *Flappy Bird* | 50% | 89% | 75% | 50% | | 25% |
| *YouTube* | 20% | 22% | 33% | 50% | 25% | |

La Figure III-14, la Figure III-15, la Figure III-16 et la Figure III-17 proposent d'autres formes de représentation de la similarité conceptuelle (d'usage) entre les applications (objets) du treillis de la Figure III-1 :

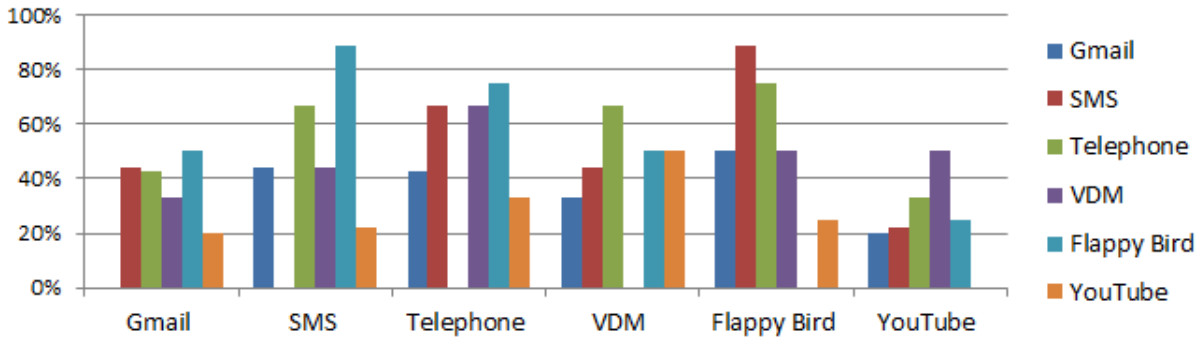



**Figure III-14 : 1ere représentation de la similarité conceptuelle entre toutes les applications (objets) du treillis de la Figure III-1**

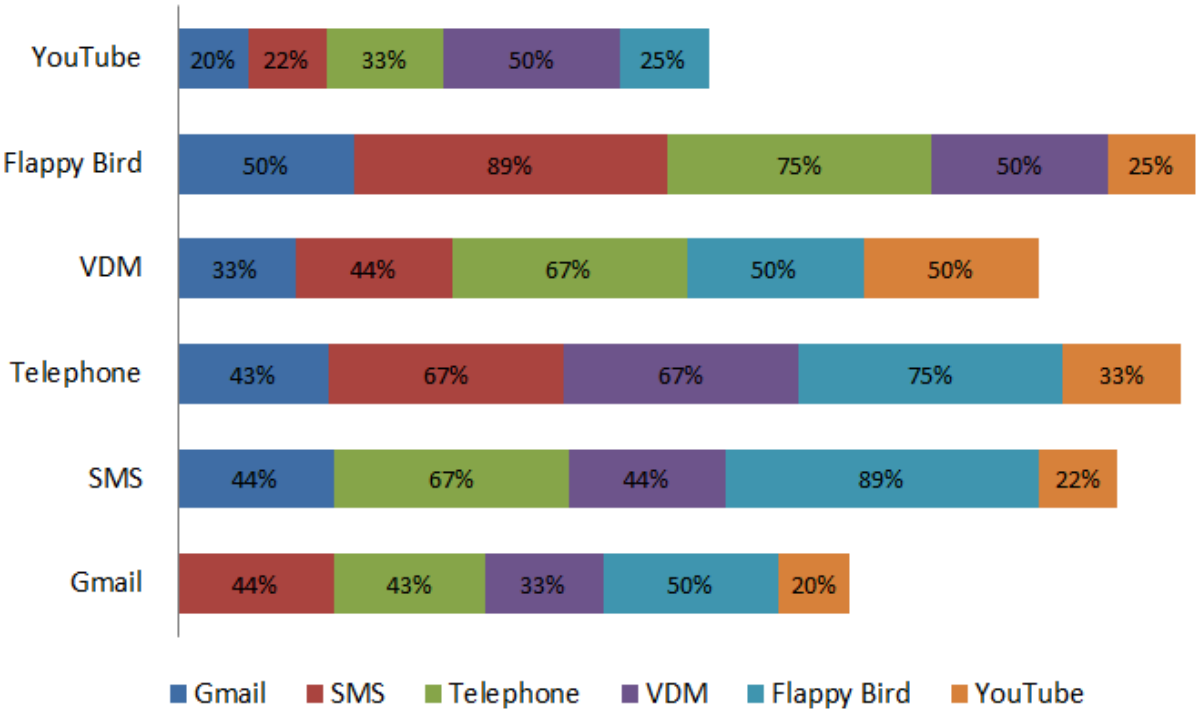



**Figure III-15 : 2e représentation de la similarité conceptuelle entre toutes les applications (objets) du treillis de la Figure III-1**

La Figure III-15 permet notamment d'identifier les applications qui sont globalement le plus (ou le moins) similaires aux autres.

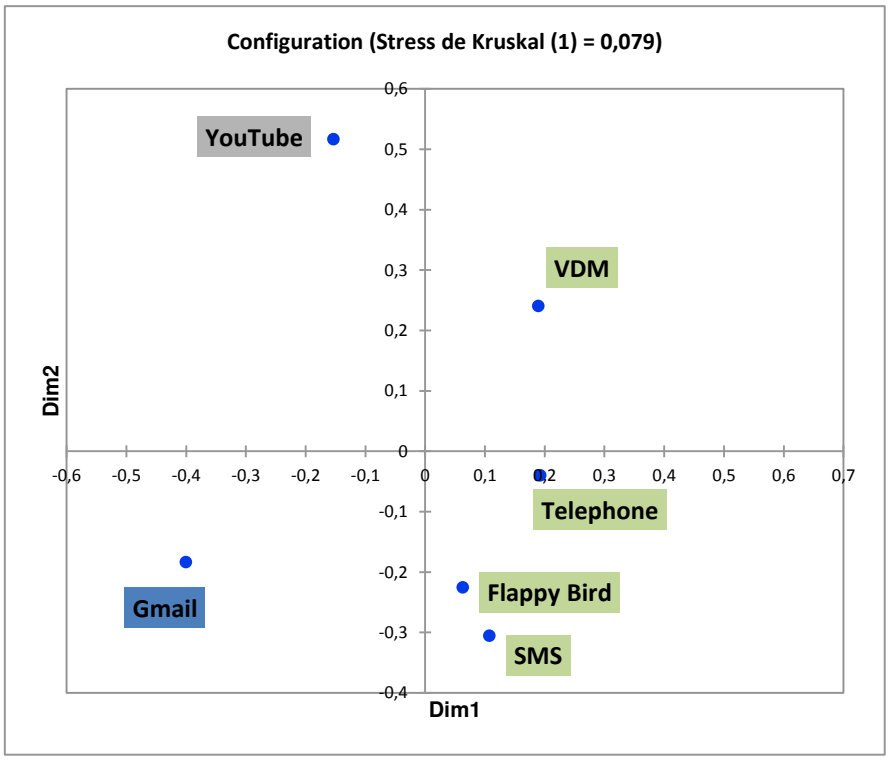



**Figure III-16 : Carte 2D enrichie des clusters (Figure III-17) pour les applications (objets) du treillis de la Figure III-1**

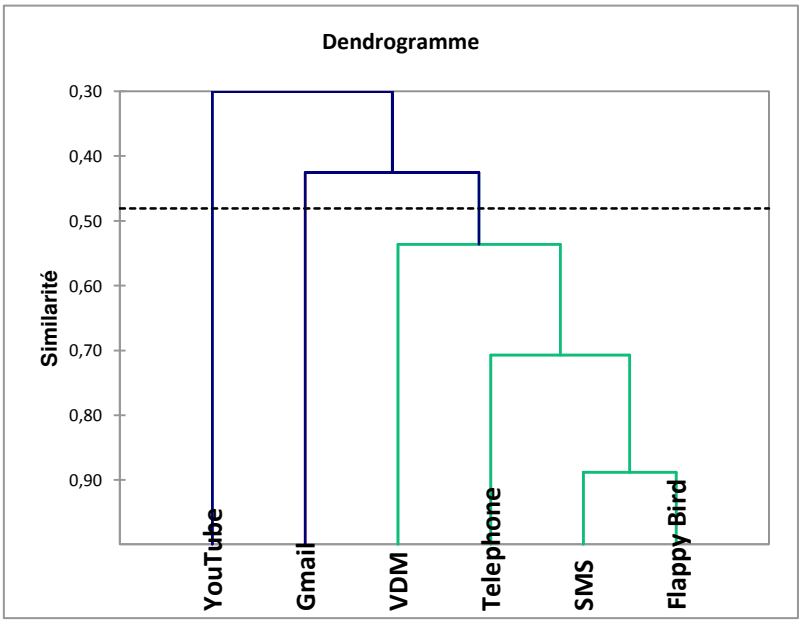



**Figure III-17 : CAH appliquée aux applications (objets) du treillis de la Figure III-1**

La Figure III-18 représente les valeurs de la similarité conceptuelle entre les applications *Flappy bird* et *SMS* pour les 28 étudiants. Cette figure montre une valeur élevée de la similarité conceptuelle de ces deux applications pour plusieurs étudiants.

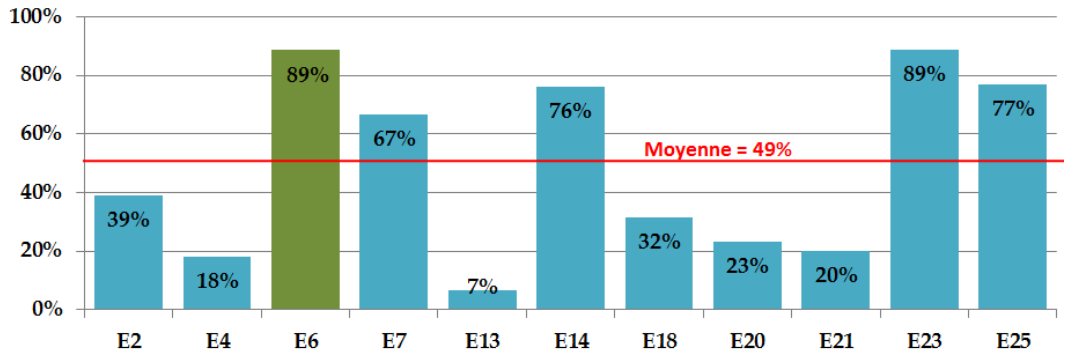



**Figure III-18 : Comparaison des valeurs de la similarité conceptuelle entre les applications *Flappy bird* et *SMS* pour tous les étudiants**

La Figure III-19 montre que la similarité conceptuelle entre *Flappy bird* et *Youtube* est globalement significativement plus faible que la similarité conceptuelle entre *Flappy bird* et *SMS*.

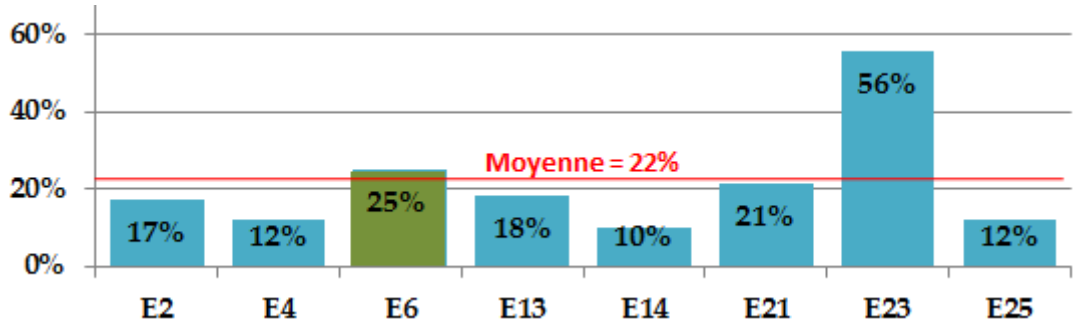



**Figure III-19 : Comparaison des valeurs de la similarité conceptuelle entre les applications *Flappy bird* et *Youtube* pour tous les étudiants**

La Figure III-20 fournit un autre exemple de conclusion, relative à deux applications de types différents : *SMS* et *Youtube*. Pour de nombreux utilisateurs, ces deux applications n'ont rien de commun (similarité conceptuelle faible), mais pour d'autres la similarité conceptuelle entre les deux est étonnamment élevée (50%-53%). Par conséquent, on peut constater que ces étudiants utilisent ces applications dans des conditions comparables, malgré leur nature très différente.

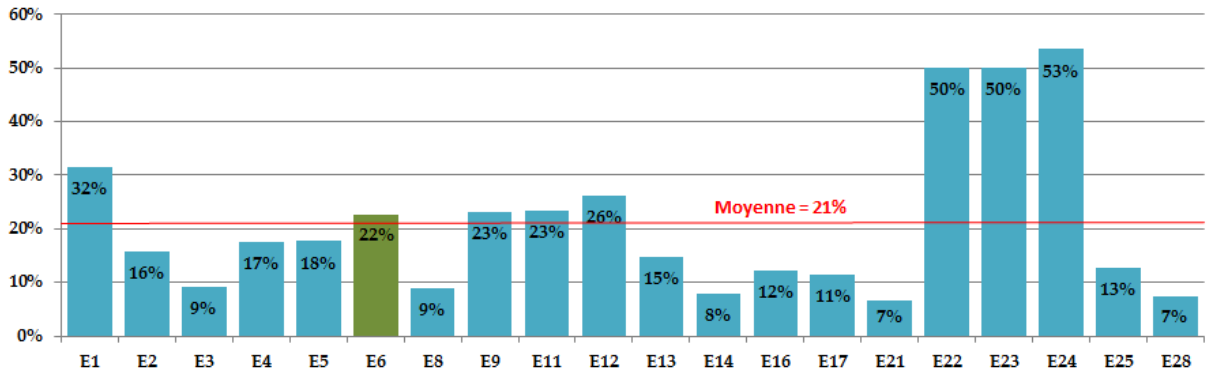



**Figure III-20 : Comparaison des valeurs de la similarité conceptuelle entre les applications *SMS* et *Youtube* pour tous les étudiants**

La mesure de similarité conceptuelle présentée dans cette section considère les objets et les attributs de manière séparée. Dans la suite de ce chapitre, nous proposons une mesure qui reflète l'impact mutuel entre les objets et les attributs. Dans l'exemple que nous utilisons, il s'agit donc de l'impact mutuel entre les applications et les éléments de contexte qui leur sont associés.

## III.4. **Impact mutuel entre un objet et un attribut**

L'objectif de l'impact mutuel est d'étudier la force de la relation entre un un objet *Oi* et un attribut *Aj* en fonction des concepts formels qui les associent.

Nous définissons donc l'impact mutuel entre un objet *Oi* et un attribut *Aj* comme le rapport entre le nombre de concepts partagés par *Oi* et *Aj* et le nombre de concepts dans le treillis contenant *Oi* ou *Aj*.

$$Impact\ mutuel\ (Oi, Aj) = \frac{Nombre\ de\ concepts\ contenant\ Oi\ et\ Aj}{Nombre\ de\ concepts\ contenant\ Oi\ ou\ Aj}$$

Le Tableau III-3 représente l'impact mutuel entre les objets (applications) et les attributs (éléments de contexte) du treillis de la Figure III-1. Ce tableau peut être interprété comme suit : on remarque tout d'abord que les valeurs d'impact mutuel peuvent varier de manière significative selon les objets et les attributs considérés. Par exemple, *Lunch break* est associé uniquement à l'application *SMS*, et *University* uniquement à *Gmail*, *SMS* et *Flappy Bird*. Si l'on s'intéresse à la localisation *Transportation*, on constate que la valeur maximale de

l'impact mutuel est obtenue avec *SMS* (avec une valeur de 78%), et que l'impact mutuel est nul avec *Youtube*. Nous rappelons que ces observations concernent un utilisateur spécifique.

Notons que, tout comme la mesure de poids, le contexte formel du Tableau II-1 ne montre pas cet impact, calculable uniquement à partir du treillis.

Dans l'exemple que nous avons choisi ici pour illustrer nos mesures, l''impact mutuel permet de savoir quels éléments de contexte caractérisent l'usage d'une application donnée. Pour cet utilisateur, on voit par exemple que *Gmail* est majoritairement associée à *Morning* et *Network 3G*.

**Tableau III-3 : Impact mutuel calculé à partir du treillis de la Figure III-1**

| | *University* | *Restaurant* | *Parents home* | *Home* | *Transportation* | *Network 3G* | *Morning* | *Coffe break* | *Lunch break* | *Afternoon* | *Evening* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Gmail* | 17% | 0% | 0% | 0% | 38% | 44% | 44% | 0% | 0% | 29% | 0% |
| *SMS* | 33% | 11% | 11% | 56% | 78% | 100% | 100% | 22% | 11% | 56% | 44% |
| *Telephone* | 0% | 0% | 0% | 38% | 44% | 67% | 67% | 0% | 0% | 22% | 25% |
| *VDM* | 0% | 0% | 0% | 29% | 22% | 44% | 44% | 0% | 0% | 0% | 14% |
| *Flappy Bird* | 22% | 0% | 0% | 44% | 67% | 89% | 89% | 11% | 0% | 44% | 33% |
| *YouTube* | 0% | 0% | 0% | 17% | 0% | 22% | 22% | 0% | 0% | 0% | 0% |

Les Figure III-21 à Figure III-24 proposent d'autres formes de représentation de l'impact mutuel entre les objets (applications) et les attributs (éléments de contexte) du treillis de la Figure III-1 :

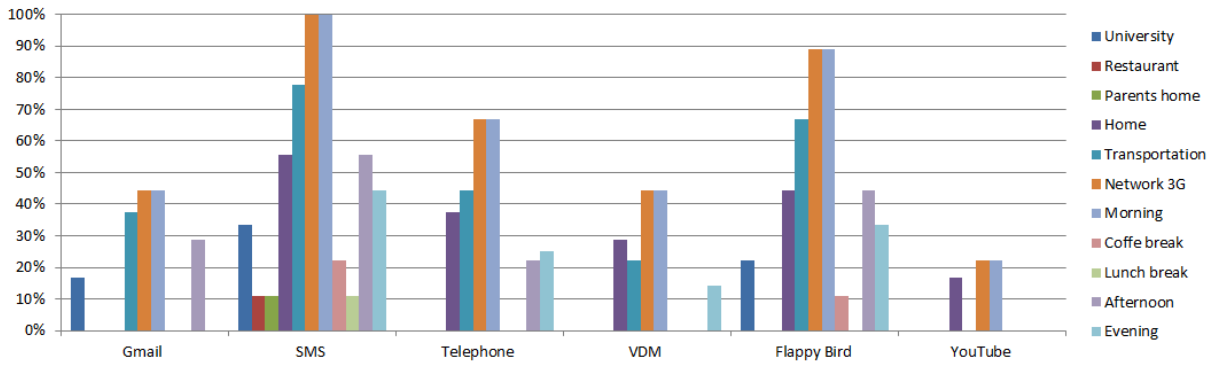



**Figure III-21 : 1ere type de représentation de l'impact mutuel entre les objets (applications) et les attributs (éléments de contexte) du treillis de la Figure III-1**

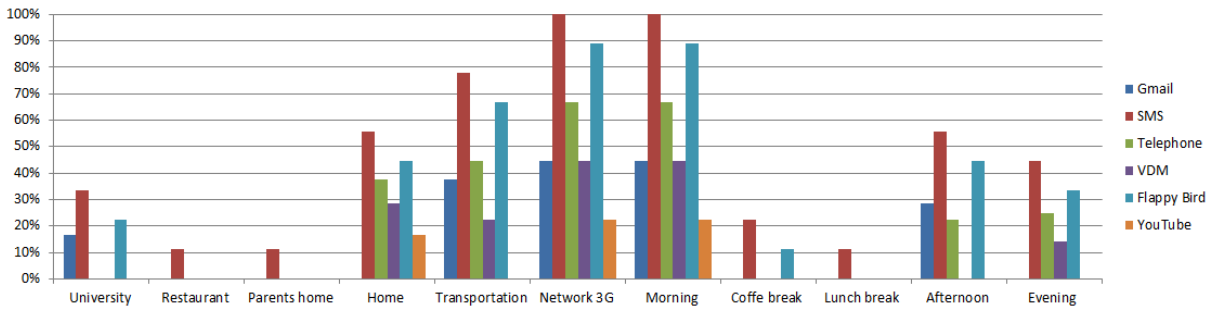



**Figure III-22 : 1ere type de représentation de l'impact mutuel avec inversion des objets et des attributs**

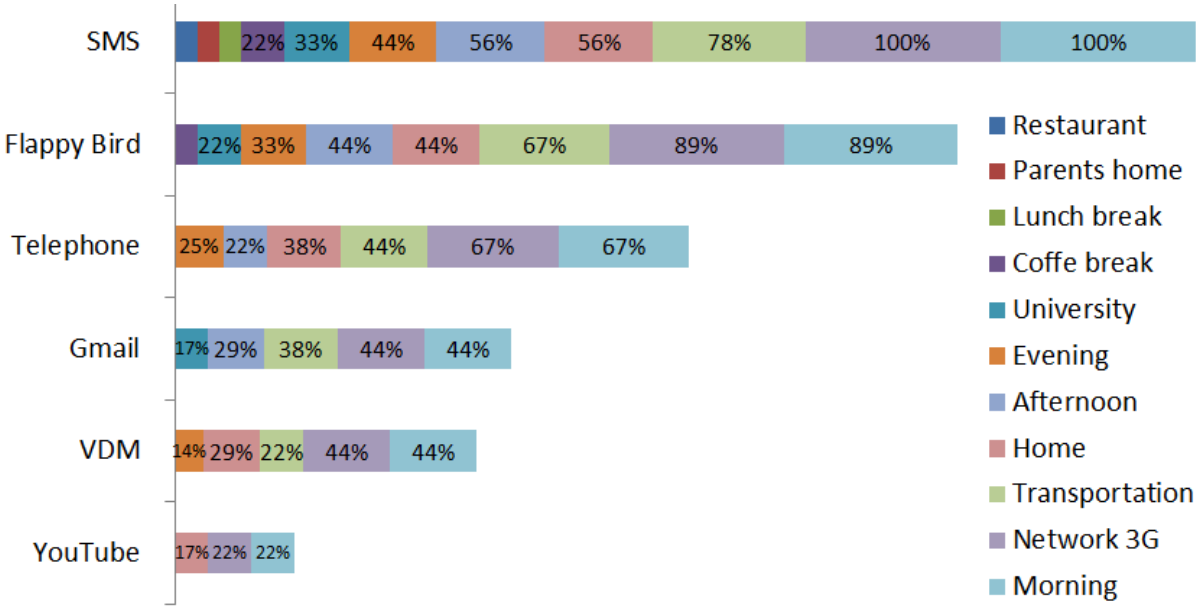



**Figure III-23 : 2e type de représentation de l'impact mutuel entre les objets (applications) et les attributs (éléments de contexte) du treillis de la Figure III-1**

La Figure III-23 montre que l'application *SMS* est celle qui a globalement l'impact le plus élevé sur les éléments de contexte. Inversement, l'impact de l'application *Youtube* est faible et limité à trois éléments de contexte (*Home*, *Network 3G* et *Morning*).

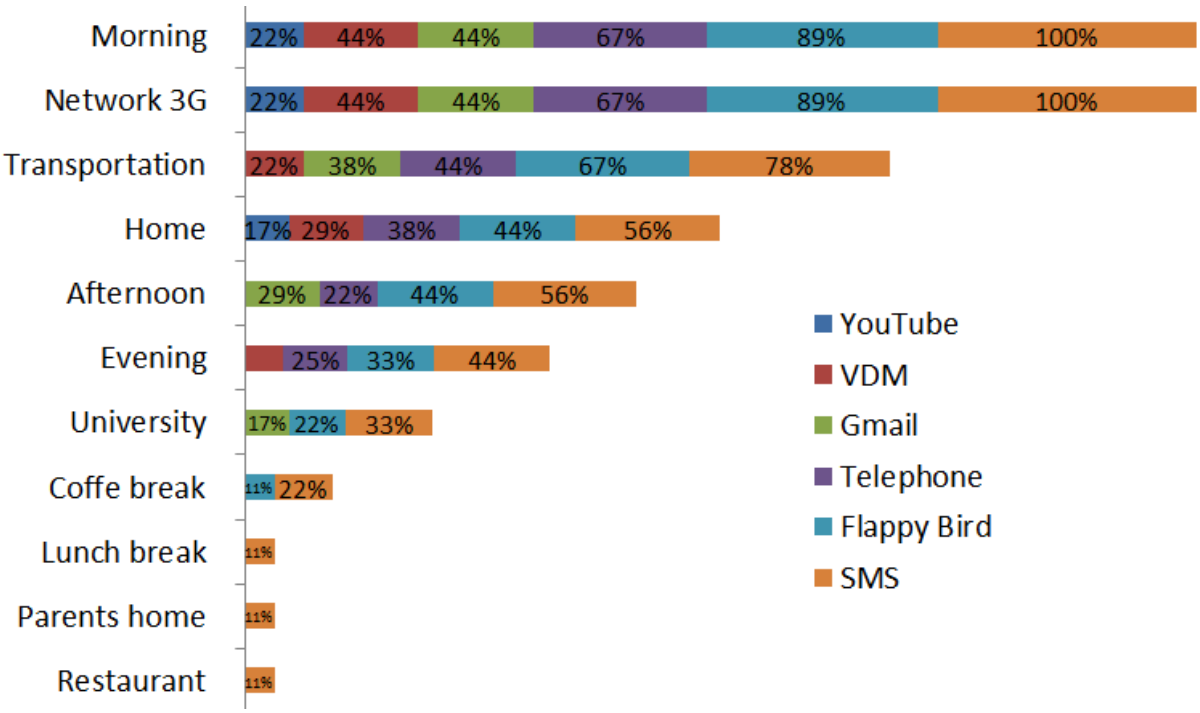



**Figure III-24 : 2e type de représentation de l'impact mutuel avec inversion des objets et des attributs**

De la même manière, on voit sur la Figure III-24 que les éléments de contexte *Lunch break, Parents home* et *Restaurant* ont un impact très faible, et uniquement avec l'application *SMS*. Au contraire, les éléments de contexte *Morning* et *Network 3G* ont un impact moyen à très élevé avec toutes les applications.

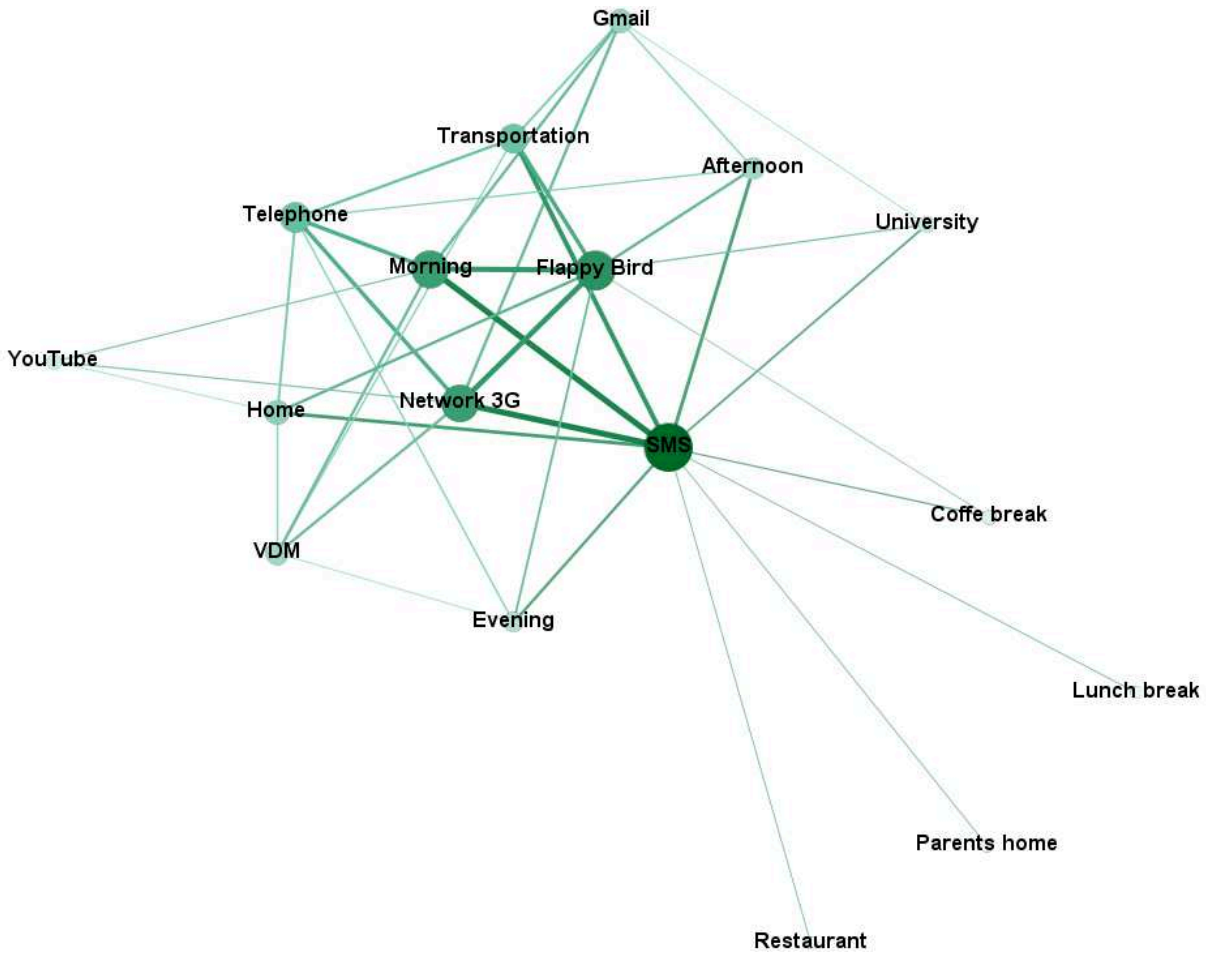



**Figure III-25 : 3e type de représentation de l'impact mutuel entre les objets (applications) et les attributs (éléments de contexte) du treillis de la Figure III-1**

La Figure III-25 est une représentation de l'impact mutuel sous la forme d'un graphe réalisé avec le logiciel Gephi[15]. Gephi est une plate-forme d'exploration qui permet une visualisation interactive des données pour l'analyse de tous types de réseaux et systèmes complexes, dynamiques et graphiques hiérarchiques (Bastian, Heymann, & Jacomy, 2009). La taille des nœuds sur ce graphe reflète le nombre de leurs voisins (appelé degré) pondéré par la valeur de l'impact mutuel avec ces voisins. L'impact mutuel entre les nœuds du graphe est également reflété par l'épaisseur des arcs du graphe. On voit bien sur ce graphe quels sont

[15] https://gephi.org/

les nœuds les plus importants en termes d'impact mutuel : *SMS*, *Flappy Bird*, *Network 3G* et *Morning*, et ceux qui sont plus marginaux et moins « connectés » aux autres nœuds, comme *Restaurant*, *Parents home*, *Lunch break* et *Coffe break*.

De la même manière que nous l'avons fait avec la mesure précédente, il est possible de comparer l'impact mutuel entre *Gmail* et *university* pour tous les étudiants (i.e., treillis). Nous avons fait les calculs de cette mesure sur l'ensemble des treillis pour tous les étudiants, dont les résultats sont présentés dans la Figure III-26.

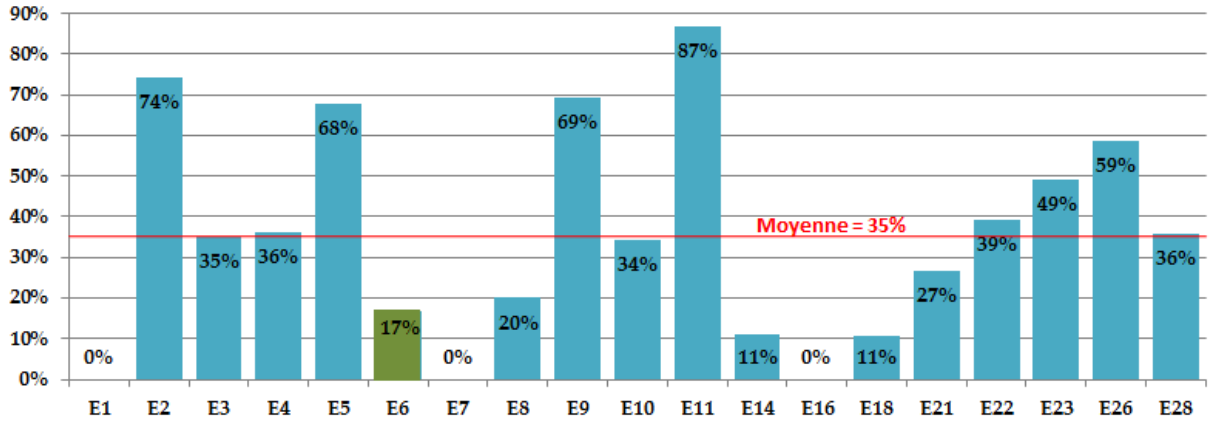



**Figure III-26 : Impact mutuel entre *Gmail* et *university* pour tous les étudiants (treillis)**

La Figure III-26 montre que la valeur d'impact mutuel de 17% entre *Gmail* et *university* pour l'étudiant E6 correspondant au treillis de la Figure III-1 est plus basse que la valeur moyenne sur tous les treillis. Le comportement de cet étudiant en termes d'utilisation de *Gmail* au sein de l'université n'est donc pas observé fréquemment à l'échelle de l'effectif global des étudiants. L'observation au niveau individuel ne peut donc dans ce cas pas être généralisée. L'impact mutuel entre *Gmail* et *university* pour l'étudiant E7 est égale à 0%, cela signifie que cet étudiant n'a jamais utilisé *Gmail* à l'*university*, par contre les étudiants (E12, E13, E15 …) qui n'appartiennent pas au diagramme de la Figure III-26, soit n'ont pas utilisé l'application *Gmail*, soit n'ont pas utilisé aucune application à l'*university*.

Des conclusions additionnelles peuvent être tirées si on compare l'ensemble des valeurs d'impact mutuel pour d'autres paires (objet, attribut). La Figure III-27 montre par exemple que bien que *Gmail* et *Morning* aient un impact mutuel de 44% pour l'étudiant (E6) du treillis de la Figure III-1, cet impact est plus élevé que la valeur moyenne sur tous les treillis.

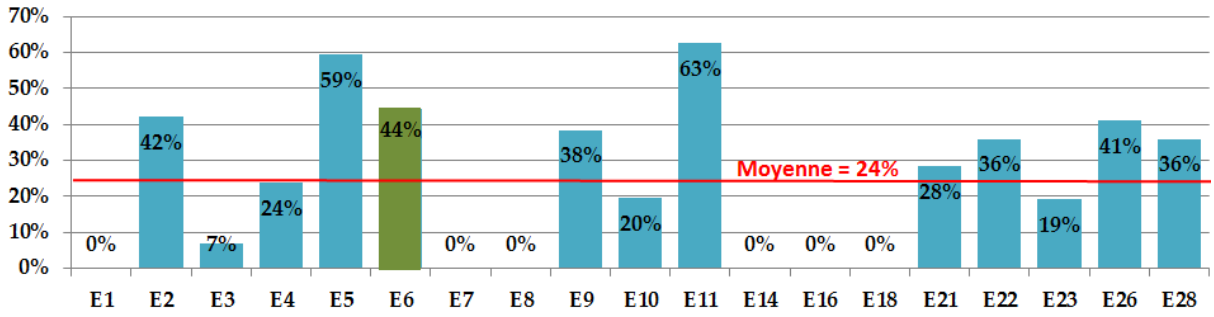



**Figure III-27 : Impact mutuel entre *Gmail* et *Morning* pour tous les étudiants (treillis)**

D'un autre côté, la Figure III-28 montre quant à elle que l'impact faible (17%) constaté à l'échelle individuelle entre *Youtube* et *Home* est également bas à l'échelle globale (moyenne = 19%).

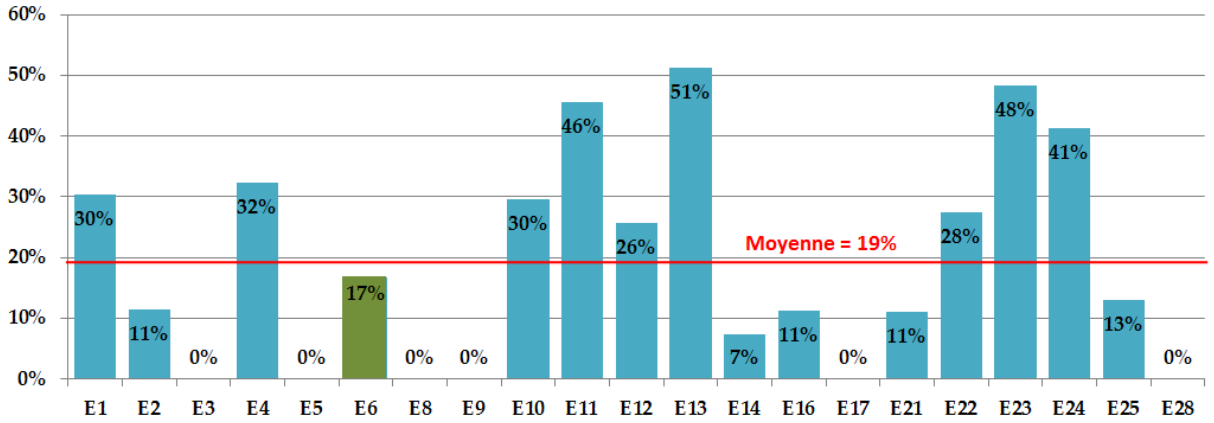



**Figure III-28 : Impact mutuel entre *Youtube* et *Home* pour tous les étudiants (treillis)**

## III.5. **Conclusion**

Nous avons proposé dans ce chapitre trois métriques pour l'interprétation de treillis de Galois dans le cadre de l'Analyse de Concepts Formels : le poids conceptuel, la similarité conceptuelle et l'impact mutuel. Les mesures proposées permettent aux utilisateurs non spécialistes de l'ACF d'interpréter les résultats obtenus. Dans notre étude de cas dédiée à l'usage d'applications sur des smartphones, les treillis de Galois regroupent, pour chaque utilisateur, les applications utilisées en fonction des éléments de contexte associés. Nous avons donné des exemples d'interprétations rendues possibles par ces mesures. Du point de vue d'un fournisseur de services mobiles, les conclusions que l'on peut tirer de ces analyses peuvent permettre de recommander des applications selon le contexte où se trouve l'utilisateur ; de nouvelles applications peuvent également être développées pour des situations contextuelles spécifiques, en tenant compte de l'impact mutuel entre les applications et les éléments de contexte, ainsi que de la similarité conceptuelle (d'usage) entre applications et entre éléments de contexte.

Chacun des treillis de Galois générés à partir des données d'entrée représente le comportement d'un utilisateur particulier ; nous avons aussi montré comment les mesures que nous avons proposées, conçues au départ pour l'analyse de treillis individuels, peuvent également être utilisées pour comparer plusieurs treillis, et donc plusieurs utilisateurs. Cette comparaison peut être mise à profit pour classifier les utilisateurs en fonction de leur comportement contextuel, et peut aussi aider à déceler des tendances générales. Pour cela nous avons développé une méthode pour la comparaison des treillis en fonction des mesures proposées, ainsi que des objets et des attributs communs.

Pour l'instant, les combinaisons les plus significatives d'objets et/ou d'attributs (parmi l'ensemble des paires existantes) ont été choisies manuellement. L'une des perspectives de nos travaux sera d'automatiser le processus de sélection de ces paires « notables ».

# CHAPITRE IV :
# APPLICATION : AIDE À L'ÉLABORATION D'UN ÉTAT DE L'ART SUR L'ACF

# Sommaire

## IV.1. **Introduction**

Nous avons présenté dans le chapitre précédent des mesures pour faciliter l'interprétation des treillis de Galois issus de l'ACF. Nous décrivons ici les résultats d'une expérimentation que nous avons menée afin d'évaluer la pertinence de l'ACF pour élaborer un état de l'art, c'est-à-dire une revue de la littérature sur un sujet donné. L'objectif est de proposer un panorama de ce sujet en proposant de regrouper des articles de recherche dans des sections pertinentes, liées au contenu de ces articles. Selon le vocabulaire de l'ACF, les *objets* que nous considérons dans cette étude sont des articles de recherche, et leurs *attributs* sont les mots-clés qui leur sont associés.

Nous avons choisi l'ACF comme thème de cet état de l'art à construire, afin de pouvoir comparer les résultats obtenus avec l'état de l'art que nous avons réalisé de manière manuelle dans le Chapitre II de ce manuscrit.

Les objets que nous avons utilisés pour cette étude consistent en 44 articles de recherche, issus des éditions récentes (2013 à 2016) des conférences ICFCA, CLA, FCA4AI et ICCS, dédiées à l'ACF ou aux structures conceptuelles. Nous avons retenu tous les articles de recherche qui abordaient la question de l'usage de l'ACF, par opposition à des contributions théoriques dans ce domaine. Les articles que nous avons sélectionnés sont listés dans le Tableau IV-1. Chaque papier possède un identifiant unique, indiqué dans la dernière colonne du tableau (P1, P2, etc.).

**Tableau IV-1 : Articles retenus pour la revue de la littérature**

| **ICFCA** | | |
|---|---|---|
| 2015 | Formal Concept Analysis and Information Retrieval – A Survey | P1 |
| 2015 | A Note on Pattern Structures and Their Projections | P2 |
| 2015 | Exploring Pattern Structures of Syntactic Trees for Relation Extraction | P3 |
| 2015 | Revisiting Pattern Structure Projections | P4 |
| 2014 | Factors and Skills | P5 |
| 2014 | Automated Enzyme Classification by Formal Concept Analysis | P6 |
| 2013 | Using Pattern Structures for Analyzing Ontology-Based Annotations of Biomedical Data | P7 |
| 2013 | Applications of Ordinal Factor Analysis | P8 |
| 2013 | Tri-ordinal Factor Analysis | P9 |

| | | |
|---|---|---|
| 2013 | User-Friendly Fuzzy FCA | P10 |
| 2013 | Using FCA to Analyse How Students Learn to Program | P11 |
| 2013 | Soundness and Completeness of Relational Concept Analysis | P12 |
| 2013 | Fitting Pattern Structures to Knowledge Discovery in Big Data | P13 |
| **FCA4AI** | | |
| 2016 | A Reachability-based Navigation Paradigm for Triadic Concepts | P14 |
| 2016 | Steps Towards Interactive Formal Concept Analysis with LatViz | P15 |
| 2016 | Contribution to the Classification of Web of Data based on Formal Concept Analysis | P16 |
| 2016 | How Fuzzy FCA and Pattern Structures are connected? | P17 |
| 2016 | A Tool for Classification of Sequential Data | P18 |
| 2016 | New taxonomy of classification methods based on Formal Concepts Analysis | P19 |
| 2016 | Contributions to the Formalization of Order-like Dependencies using FCA | P20 |
| 2016 | A Hybrid Approach for Mining Metabolomic Data | P21 |
| 2015 | Bridging DBpedia Categories and DL-Concept Definitions using Formal Concept Analysis | P22 |
| 2015 | A Conceptual-KDD tool for ontology construction from a database schema | P23 |
| 2015 | Pattern structures for news clustering | P24 |
| 2015 | Lazy Classification with Interval Pattern Structures : Application to Credit Scoring | P25 |
| 2015 | RAPS: A Recommender Algorithm Based on Pattern Structures | P26 |
| 2014 | Using Formal Concept Analysis to Create Pathways through Museum Collections | P27 |
| 2014 | Towards an FCA-based Recommender System for Black-Box Optimization | P28 |
| 2014 | Concept Stability as a Tool for Pattern Selection | P29 |
| **ICCS** | | |
| 2016 | A Semiotic-Conceptual Analysis of Conceptual Learning | P30 |
| 2016 | Organised Crime and Social Media: Detecting and Corroborating Weak Signals of Human Trafficking Online | P31 |
| 2016 | Distilling Conceptual Structures from Weblog Data Using Polyadic FCA | P32 |
| 2014 | Extracting Threshold Conceptual Structures from Web Documents | P33 |
| 2014 | FCA-Based Recommender Models and Data Analysis for Crowdsourcing Platform Witology | P34 |
| 2014 | Conceptual Structures in LEADing and Best Enterprise Practices | P35 |
| 2014 | Investigating Oncological Databases Using Conceptual Landscapes | P36 |
| 2013 | Using Conceptual Structures in the Design of Computer-Based Assessment Software | P37 |
| **CLA** | | |
| 2016 | FCA for Software Product Lines Representation: Mixing Configuration and Feature Relationships in a Unique Canonical Representation | P38 |
| 2016 | Exploring Temporal Data Using Relational Concept Analysis: An Application to Hydroecology | P39 |

| | | |
|---|---|---|
| 2016 | Building a Domain Knowledge Model Based on a Concept Lattice Integrating Expert Constraints | P40 |
| 2015 | RV-Xplorer: A Way to Navigate Lattice-Based Views over RDF Graphs | P41 |
| 2015 | Concept interestingness measures: a comparative | P42 |
| 2014 | The educational tasks and objectives system within a formal context | P43 |
| 2014 | Formal Concept Analysis for Process Enhancement Based on a Pair of Perspectives | P44 |

Les attributs associés à ces articles sont les mots-clés, tels que proposés explicitement par les auteurs ou bien extraits manuellement à partir des résumés. Cela nous a donné un ensemble de 32 mots-clés, certains plus spécifiques que d'autres. Afin d'avoir une certaine homogénéité, nous avons choisi de conserver les 15 mots-clés les plus génériques, listés dans le Tableau IV-2.

**Tableau IV-2 : Mots-clés associés aux articles retenus**

| |
|---|
| 1 (Biomedical) |
| 2 (Social Media) |
| 3 (Ontology) |
| 4 (Cultural Heritage) |
| 5 (Classification) |
| 6 (Clustering) |
| 7 (Interpretation) |
| 8 (Measure) |
| 9 (Data Analysis) |
| 10 (Data Mining) |
| 11 (Web) |
| 12 (FCA variant) |
| 13 (Information Retrieval) |
| 14 (Knowledge) |
| 15 (Patterns) |

L'objectif étant ici de proposer une structure pour la revue de la littérature, c'est-à-dire une manière de fournir un panorama en regroupant les articles de recherche dans différentes sections en fonction de leurs mots-clés, nous avons tout particulièrement exploité la mesure de similarité conceptuelle que nous avons définie dans la section III.3. Nous avons choisi cette mesure car elle peut donner lieu à une représentation sous forme de carte, aisée à interpréter pour l'utilisateur et adaptée à ce besoin de « panorama ».

Notre mesure de similarité conceptuelle pouvant s'appliquer aussi bien aux objets qu'aux attributs, nous proposons une cartographie des mots-clés associés à l'ACF en plus de la cartographie des articles eux-mêmes.

## IV.2. **Résultats obtenus avec les 15 mots-clés retenus**

Nous présentons tout d'abord les résultats que nous avons obtenus en appliquant notre mesure de similarité conceptuelle au treillis généré à partir des 44 articles (objets) caractérisés par l'ensemble des 15 mots-clés que nous avons retenus (attributs).

### IV.2.1. **Contexte formel et treillis de Galois**

Le Tableau IV-3 représente le contexte formel fourni en entrée. Ce contexte formel contient 44 lignes, correspondant aux 44 articles, et 15 colonnes, correspondant aux 15 mots-clés. La valeur de chaque case est de 1 si le mot-clé est associé à l'article et de 0 sinon.

**Tableau IV-3 : Contexte formel – 15 mots-clés**

| | Biomedical | Social Media | Ontology | Cultural Heritage | Classification | Clustering | Interpretation | Measure | Data analysis | Data mining | Web | FCA variant | Information Retrieval | Knowledge | Patterns |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 |
| P2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| P3 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| P4 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| P5 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| P6 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| P7 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| P8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| P9 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| P10 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| P11 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 |
| P12 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 |
| P13 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| P14 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 |
| P15 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 |
| P16 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| P17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 |
| P18 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 |
| P19 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 |
| P20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| P21 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 |
| P22 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 |
| P23 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| P24 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 |
| P25 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| P26 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| P27 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| P28 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| P29 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| P30 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| P31 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| P32 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 |
| P33 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 |
| P34 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| P35 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| P36 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 |
| P37 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| P38 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 |
| P39 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| P40 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| P41 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 |
| P42 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| P43 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| P44 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |

Le treillis de Galois obtenu en sortie de l'algorithme est représenté sur la Figure IV-1. La complexité de ce treillis ne permet pas son interprétation à partir de sa représentation visuelle. Nous étudions dans la section suivante les conclusions que l'analyse de la similarité conceptuelle entre les mots-clés peut nous fournir.

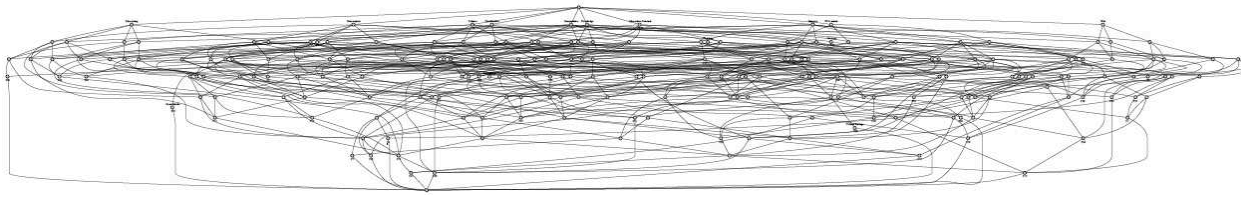


**Figure IV-1 : Treillis de Galois obtenu avec les 15 mots-clés retenus**

### IV.2.2. Similarité conceptuelle entre les 15 mots-clés retenus

Nous nous sommes tout d'abord intéressés à la similarité conceptuelle entre les mots-clés. La similarité conceptuelle entre chaque paire de mots-clés est représentée sur la Figure IV-2, de la même manière que nous l'avons fait dans le Chapitre III.

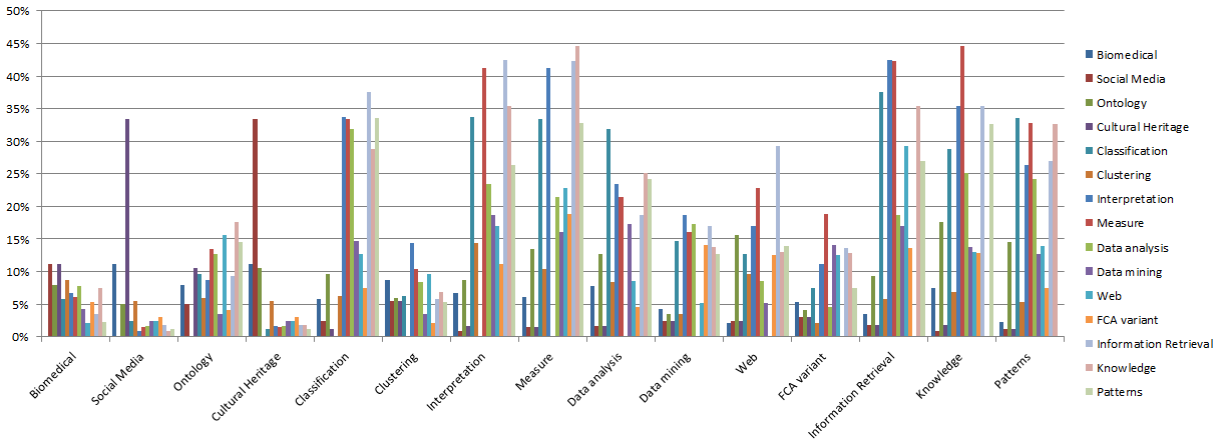



**Figure IV-2 : Similarité conceptuelle entre les 15 mots-clés**

Cette représentation permet de distinguer aisément les paires d'attributs fortement similaires ou au contraire très dissimilaires. Il est plus difficile en revanche de tirer des conclusions pour les paires d'attributs dont la similarité conceptuelle est moins tranchée. C'est pourquoi, tout comme dans le Chapitre III, nous proposons une représentation sous forme de carte en 2D, sur la Figure IV-3.

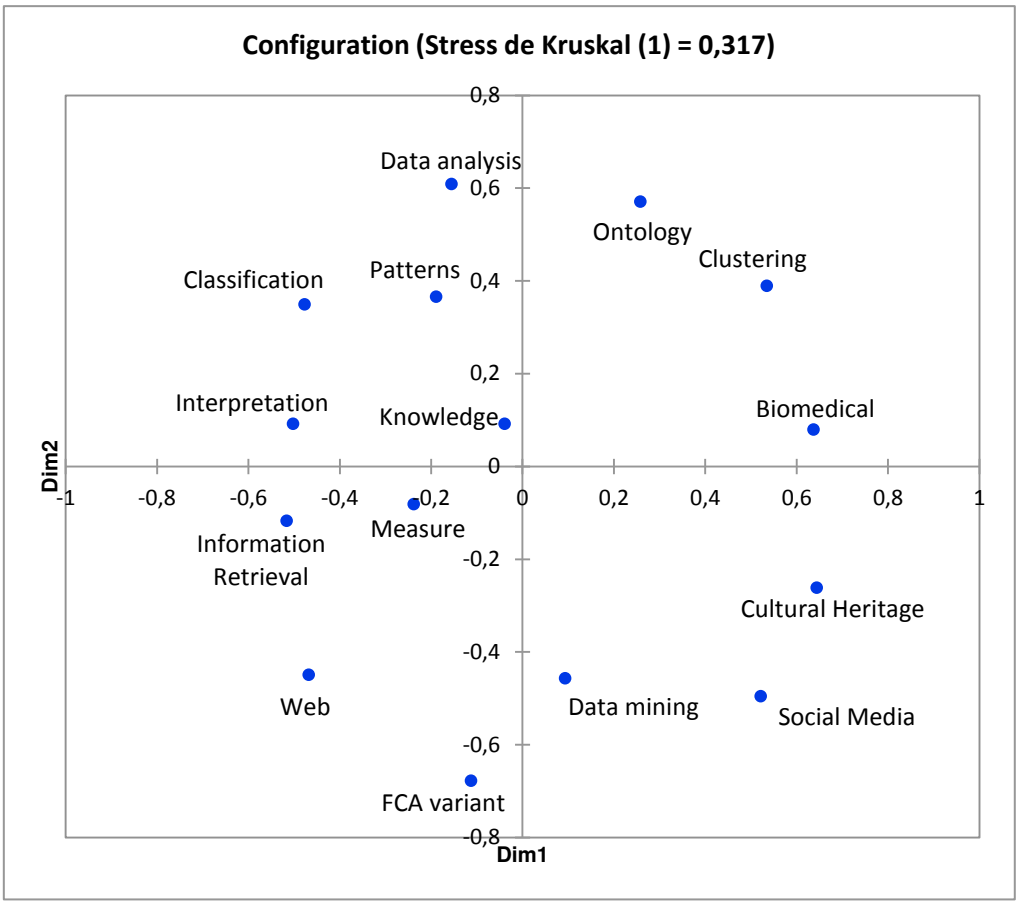



**Figure IV-3 : Carte 2D obtenue à partir de la matrice de similarité conceptuelle entre les 15 mots-clés**

Cette carte reflète ainsi la proximité, ou au contraire la distance, entre les mots-clés. Néanmoins, la valeur de stress associée à cette représentation est élevée, ce qui signifie qu'elle est imparfaite. Nous avons alors appliqué l'algorithme de CAH afin de compléter cette représentation en indiquant la répartition de ces mots-clés dans différents clusters. Le résultat de cet algorithme de clustering est représenté sur la Figure IV-4.

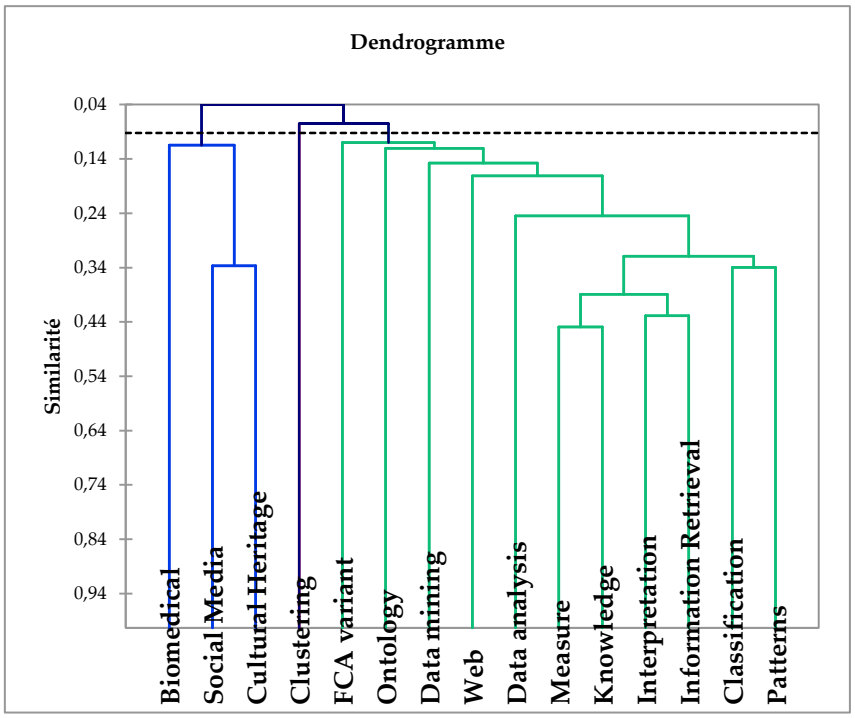



**Figure IV-4 : Résultat du clustering ascendant hiérarchique appliqué à la matrice de similarité conceptuelle entre les 15 mots-clés**

Le résultat de la CAH n'est malheureusement pas satisfaisant car les trois clusters obtenus sont de taille très différente, ce qui ne nous donne pas une répartition harmonieuse des différents mots-clés. C'est pourtant un objectif important dans le contexte d'une revue de la littérature, où il s'agit de construire des sections aussi équilibrées que possible en termes de thèmes abordés et d'articles référencés. Nous ne creusons pas l'analyse (en étudiant par exemple la similarité entre les articles), car l'absence d'une structure satisfaisante au niveau des mots-clés est bloquante.

Nous concluons de cette étude que la description des articles à l'aide des 15 mots clés listés dans le Tableau IV-2 ne nous permet pas d'atteindre notre objectif.

Notre hypothèse pour expliquer cela est que les mots-clés identifiés sont trop génériques ; nous avons donc réintégré les mots-clés plus spécifiques qui avaient été éliminés en première intention, pour obtenir un ensemble de 32 mots-clés.

Nous avons alors procédé à une analyse similaire à celle de cette section, en travaillant cette fois avec la nouvelle liste de mots-clés. Nous présentons les résultats obtenus dans la section suivante.

## IV.3. **Résultats obtenus avec les 32 mots-clés plus spécifiques**

Nous présentons dans le Tableau IV-4 les 32 mots-clés dérivés des 15 mots-clés initialement retenus. Nous avons notamment choisi d'ajouter des mots-clés plus spécifiques pour préciser les notions qui sont centrales dans cette thèse : l'interprétation, les mesures et les ontologies.

**Tableau IV-4 : Liste des 32 mots-clés « spécifiques » associés aux articles**

| Niveau 1 | Niveau 2 | |
|---|---|---|
| 1 (Biomedical) | 1.1 | Biomedical |
| 2 (Social Media) | 2.1 | Social Media |
| 3 (Ontology) | 3.1 | DBpedia |
| | 3.2 | RDF |
| | 3.3 | Ontology |
| 4 (Cultural Heritage) | 4.1 | Cultural Heritage |
| 5 (Classification) | 5.1 | Classification |
| 6 (Clustering) | 6.1 | Clustering |
| 7 (Interpretation) | 7.1 | Interpretation |
| | 7.2 | Navigation |
| | 7.3 | Recommendation |
| | 7.4 | Visualization |
| | 7.5 | Association rules |
| | 7.6 | Interaction |
| 8 (Measure) | 8.1 | Measure |
| | 8.2 | Similarity |
| | 8.3 | Stability |
| | 8.4 | Dependencies |
| | 8.5 | Relation extraction |
| | 8.6 | Boolean factors |
| | 8.7 | Factor Analysis |
| | 8.8 | Random Forest |
| | 8.9 | Scaling |
| 9 (Data Analysis) | 9.1 | Data analysis |
| 10 (Data Mining) | 10.1 | Data mining |
| 11 (Web) | 11.1 | Web |
| 12 (FCA variant) | 12.1 | Fuzzy FCA |
| | 12.2 | Relational Concept Analysis |
| | 12.3 | Triadic Concept Analysis |
| 13 (Information Retrieval) | 13.1 | Information Retrieval |

| | | |
|---|---|---|
| 14 (Knowledge) | 14.1 | Knowledge |
| 15 (Patterns) | 15.1 | Patterns |

Nous reprenons dans la suite de cette section la même méthodologie que celle que nous avons suivie avec les 15 mots-clés initiaux : nous construisons le treillis de Galois et l'interprétons à l'aide de la mesure de similarité conceptuelle.

## IV.3.1. **Contexte formel et treillis de Galois**

Le Tableau IV-5 représente le contexte formel fourni en entrée. Ce contexte formel contient toujours 44 lignes, correspondant aux 44 articles (*objets*), mais cette fois 32 colonnes (au lieu de 15), correspondant aux 32 mots-clés qui correspondent à leurs attributs.

**Tableau IV-5 : Contexte formel – 32 mots-clés plus spécifiques**

| | Biomedical | Social Media | DBpedia | RDF | Ontology | Cultural Heritage | Classification | Clustering | Interpretation | Navigation | Recommendation | Visualization | Association rules | Interaction | Measure | Similarity | Stability | Dependencies | Relation extraction | Boolean factors | Factor Analysis | Random Forest | Scaling | Data analysis | Data mining | Web | Fuzzy FCA | Relational Concept Analysis | Triadic Concept Analysis | Information Retrieval | Knowledge | Patterns |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 |
| P2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| P3 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| P4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| P5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| P6 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| P7 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| P8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| P9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| P10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| P11 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| P12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 |
| P13 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| P14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 |
| P15 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| P16 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 |
| P17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 |
| P18 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 |
| P19 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| P20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| P21 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| P22 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| P23 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| P24 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| P25 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| P26 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| P27 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| P28 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| P29 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| P30 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| P31 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| P32 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| P33 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 |
| P34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| P35 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| P36 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 |
| P37 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| P38 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 |
| P39 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 |
| P40 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| P41 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| P42 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| P43 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| P44 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

Le treillis de Galois obtenu en sortie de l'algorithme est représenté sur la Figure IV-5. Tout comme le treillis obtenu avec les 15 mots-clés initiaux, sa représentation visuelle ne permet pas de tirer des conclusions. Nous utilisons donc la mesure de similarité conceptuelle afin d'en faciliter l'interprétation.

**Figure IV-5 : Treillis de Galois obtenu avec les 32 mots-clés plus spécifiques**

### IV.3.2. **Similarité conceptuelle entre les 32 mots-clés**

Nous présentons sur la Figure IV-6 et Figure IV-7 respectivement le résultat du clustering et la carte 2D reflétant les clusters de mots-clés obtenus.

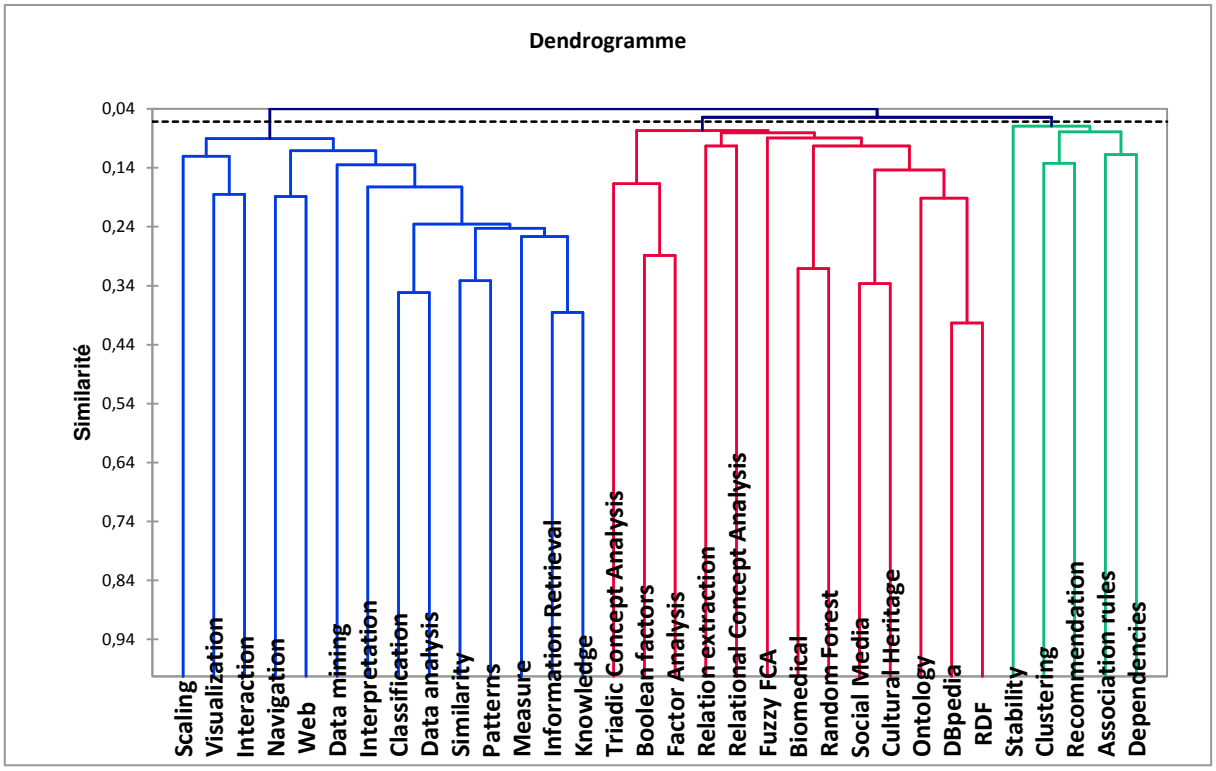



**Figure IV-6 : Résultat du clustering ascendant hiérarchique appliqué à la matrice de similarité conceptuelle entre les 32 mots-clés plus spécifiques**

Le résultat de la CAH est beaucoup plus satisfaisant que celui que nous avions obtenu en décrivant les articles avec les 15 mots-clés initiaux : les trois clusters identifiés sont bien plus homogènes.

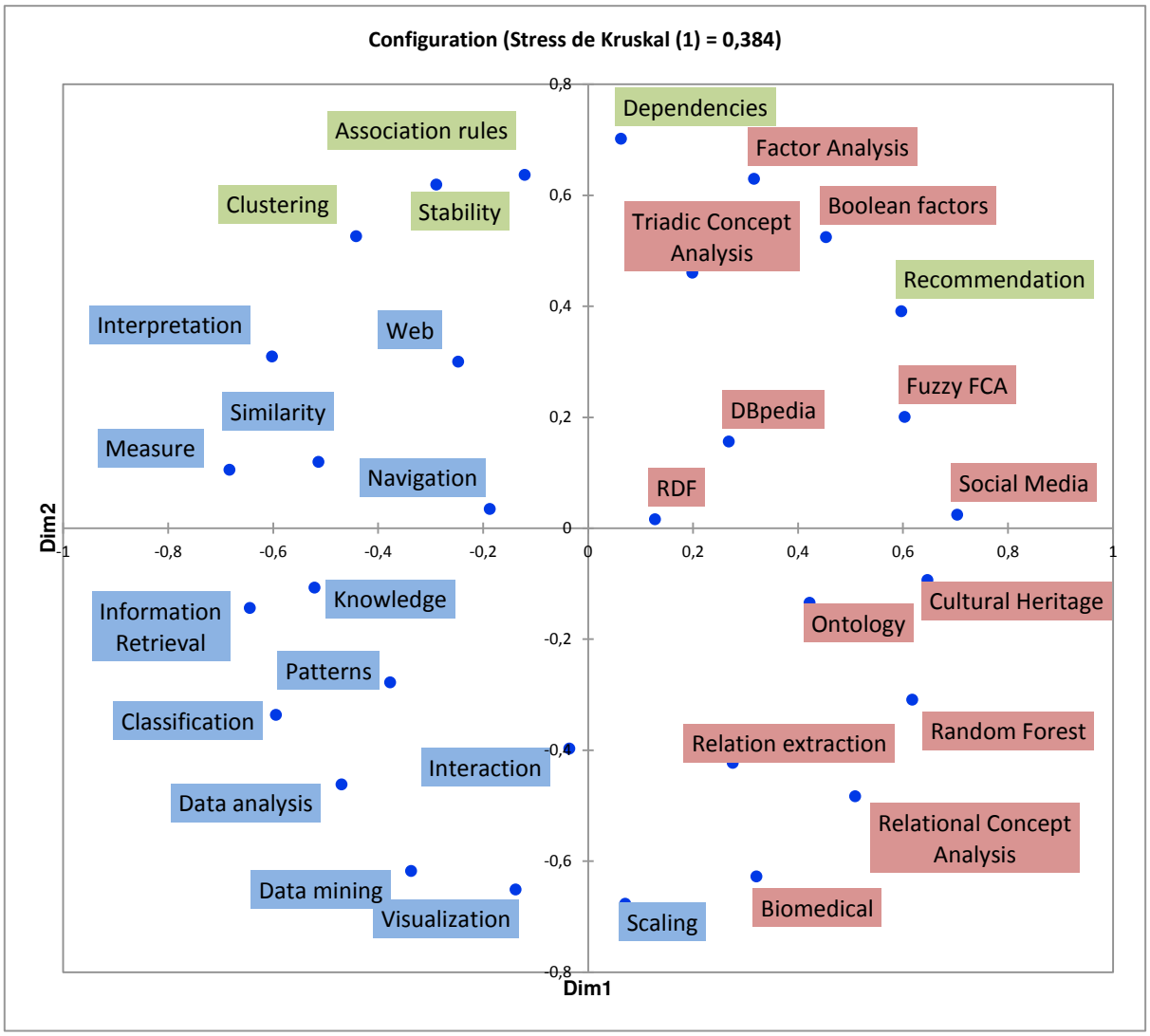



**Figure IV-7 : Carte 2D des 32 mots-clés plus spécifiques, enrichie des clusters obtenus par CAH**

La carte MDS permet de proposer à un utilisateur une « vue d'ensemble » des thèmes abordés dans l'ensemble des articles de la littérature dédiée à l'ACF. Le découpage en trois clusters pourrait correspondre à une structure possible de l'état de l'art.

Nous poursuivons cette analyse dans la section suivante, en étudiant cette fois, à l'aide de la similarité conceptuelle, la répartition des articles dans les différentes sections selon cette structure.

### IV.3.3. **Similarité conceptuelle entre les articles décrits par les 32 mots-clés**

Nous avons calculé la similarité conceptuelle entre l'ensemble des articles (décrits par les 32 mots-clés plus spécifiques). Nous présentons sur la Figure IV-8 et la Figure IV-9 respectivement le résultat du clustering et la carte 2D reflétant les clusters obtenus.

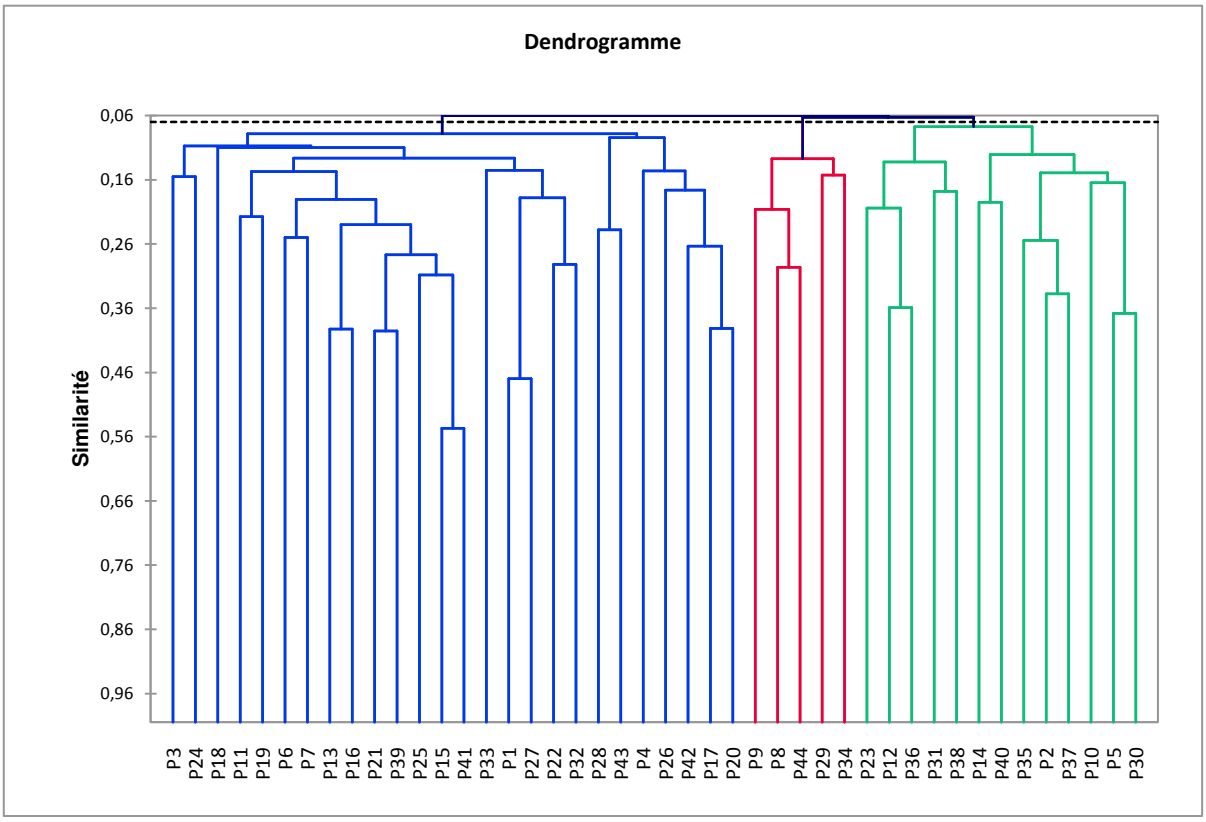



**Figure IV-8 : Résultat du clustering ascendant hiérarchique appliqué à la matrice de similarité conceptuelle entre les articles décrits par les mots-clés plus spécifiques**

Les résultats de la CAH font apparaître trois clusters, ce qui est cohérent avec le nombre de clusters de mots-clés. On note néanmoins que l'un des clusters d'articles est plus grand que les deux autres, ce qui n'est pas optimal en termes d'équilibre du nombre d'articles par section d'un état de l'art. Ces résultats sont toutefois nettement plus exploitables que ceux que l'on avait obtenus en utilisant les 15 mots-clés retenus initialement pour décrire les mots-clés.

Le panorama auquel on aboutirait selon les résultats de notre étude apparaît sur la carte MDS de la Figure IV-9, où chaque article est associé à la couleur de son cluster.

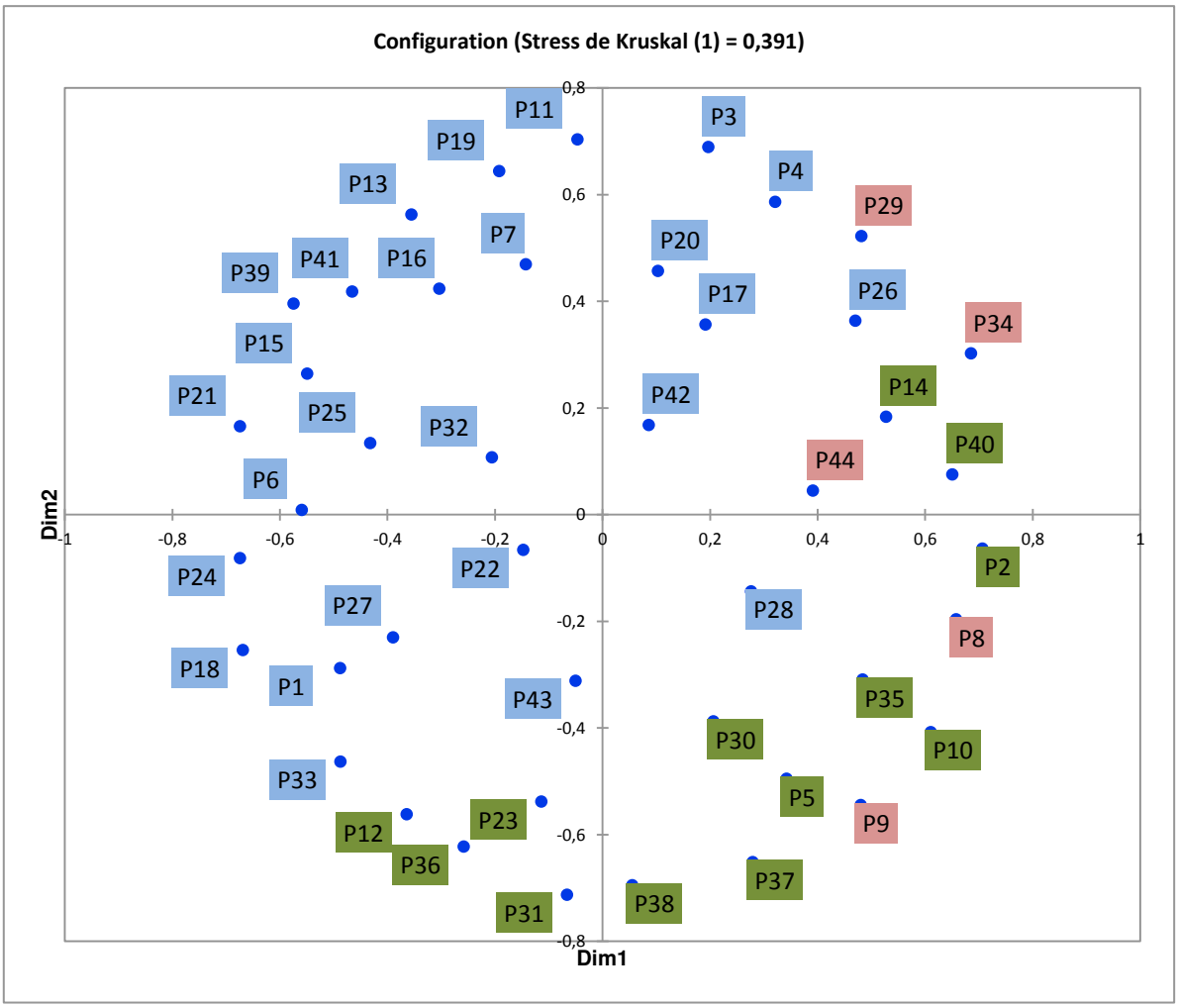



**Figure IV-9 : Carte 2D des articles décrits par les 32 mots-clés plus spécifiques, enrichie des clusters obtenus par CAH**

Puisque notre méthodologie nous a permis d'obtenir une proposition de structure pour un état de l'art dédié à l'ACF, nous avons comparé la répartition des articles déduits automatiquement et la répartition que nous avons effectuée manuellement lorsque nous avons rédigé le chapitre d'état de l'art de ce manuscrit dans le Chapitre II.

### IV.3.4. **Comparaison avec la structure de l'état de l'art manuel**

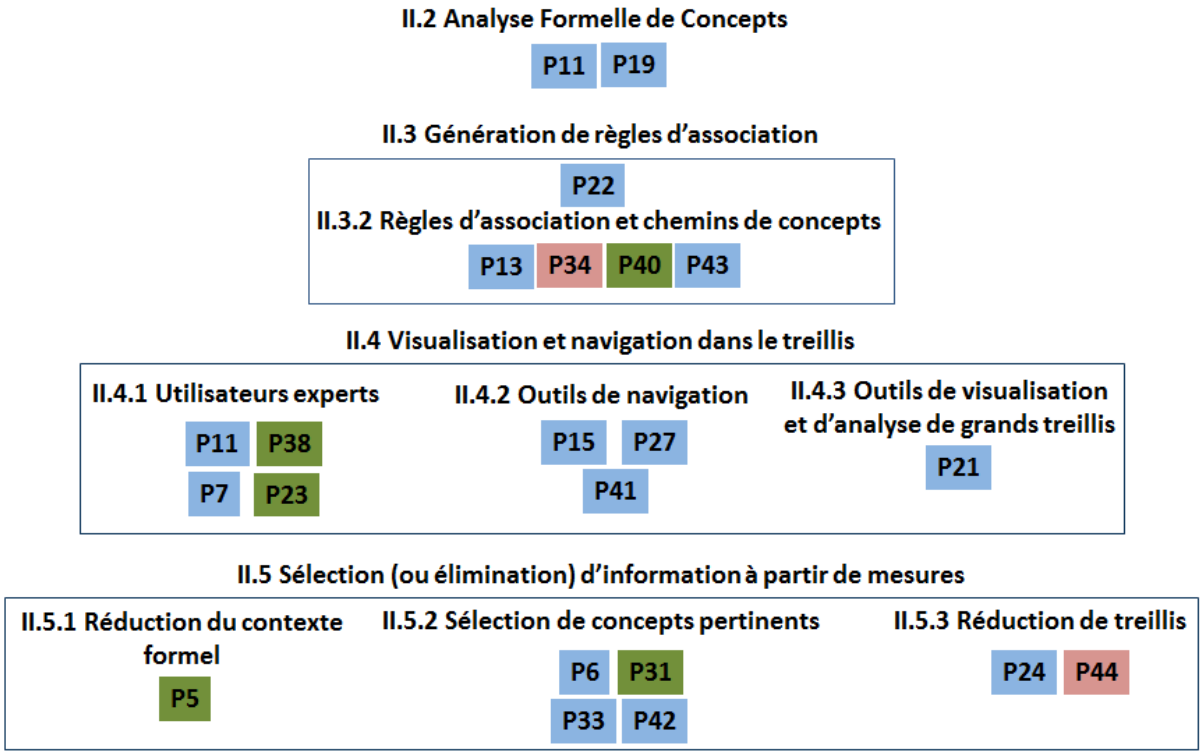



**Figure IV-10 : Répartition des articles dans l'état de l'art réalisé manuellement pour ce manuscrit**

La Figure IV-10 représente la répartition des articles dans les différentes sections (ou sous-sections) de notre chapitre d'état de l'art. Il est à noter que certains articles (parmi les 44 utilisés dans l'étude) n'apparaissent pas, car ils ont soit été cités dans d'autres chapitres du manuscrit, soit ils n'ont pas été retenus dans notre état de l'art manuel. Il est intéressant de voir que cette expérimentation aurait peut-être pu nous permettre de citer plus d'articles que ce que nous avons fait. Nous rappelons néanmoins que l'expérimentation ne tient pas compte que des mots clés et que tous les articles n'apparaissaient plus forcément pertinents, une fois lus en détail. Inversement, nous avons aussi cité dans notre état de l'art des articles n'appartenant pas à la l'ensemble des 44 articles de l'expérimentation.

Nous pouvons constater que les sections que nous avons définies ne correspondent pas aux mêmes clusters que ceux qui ont été identifiés de manière automatique à partir de notre analyse. Cela ne signifie pas pour autant que la structure déduite par l'ACF ne soit pas pertinente ; cela signifie seulement que nous n'avons pas utilisé les mêmes critères pour construire notre état de l'art manuel.

Lorsque nous avons réalisé notre revue de la littérature, nous nous sommes tout particulièrement concentrés sur la manière dont les articles abordaient la question de l'interprétation des résultats de l'ACF. C'est pourquoi nous avons réalisé une 3e analyse, en ciblant cette fois les mots-clés dédiés à l'interprétation, c'est à dire aux mots-clés listés dans la Tableau IV-6.

**Tableau IV-6 : Mots-clés dédiés à l'interprétation des treillis de Galois**

| | | |
|---|---|---|
| 7 (Interpretation) | 7.1 | Interpretation |
| | 7.2 | Navigation |
| | 7.3 | Recommendation |
| | 7.4 | Visualization |
| | 7.5 | Association rules |
| | 7.6 | Interaction |

## IV.4. **Analyse ciblée sur les mots-clés liés à l'interprétation des résultats de l'ACF**

### IV.4.1. **Contexte formel et treillis de Galois**

Le Tableau IV-7 représente le contexte formel fourni en entrée. Ce contexte formel contient contient 6 colonnes, correspondant aux 6 mots-clés dédiés à l'interprétation des résultats de l'ACF et trouvés dans les articles issus des conférences récentes du domaine. Les 6 attributs retenus ici sont donc : *Interpretation, Navigation, Recommendation, Visualization, Association rules* et *Interaction.* Le mot-clé « interprétation » peut sembler un peu vague, mais il a été rencontré dans des articles qui ne précisaient pas plus avant la technique utilisée pour interpréter le treillis. On peut noter que le contexte formel contient moins de 44 lignes ; ceci est dû au fait que les articles qui ne sont associés à aucun de ces 6 mots-clés ne sont pas présents.

**Tableau IV-7 : Contexte formel – 6 mots-clés ciblés sur l'interprétation des treillis de Galois**

| | Interpretation | Navigation | Recommendation | Visualization | Association rules | Interaction |
|---|---|---|---|---|---|---|
| P1 | 1 | 1 | 1 | 0 | 1 | 1 |
| P3 | 0 | 0 | 0 | 0 | 0 | 1 |
| P4 | 0 | 0 | 0 | 1 | 0 | 0 |
| P5 | 1 | 0 | 0 | 0 | 0 | 0 |
| P6 | 1 | 0 | 0 | 0 | 0 | 0 |
| P7 | 1 | 0 | 0 | 1 | 0 | 0 |
| P9 | 1 | 0 | 0 | 0 | 0 | 0 |
| P10 | 1 | 0 | 0 | 0 | 0 | 0 |
| P11 | 0 | 0 | 0 | 1 | 0 | 1 |
| P12 | 1 | 0 | 0 | 1 | 0 | 1 |
| P13 | 0 | 0 | 0 | 0 | 1 | 0 |
| P14 | 0 | 1 | 0 | 1 | 0 | 0 |
| P15 | 1 | 1 | 0 | 1 | 0 | 1 |
| P18 | 0 | 0 | 0 | 0 | 1 | 0 |
| P19 | 0 | 1 | 0 | 0 | 0 | 0 |
| P21 | 1 | 0 | 0 | 1 | 0 | 0 |
| P22 | 0 | 0 | 0 | 0 | 1 | 0 |
| P23 | 1 | 0 | 0 | 0 | 0 | 0 |
| P24 | 1 | 0 | 0 | 0 | 0 | 0 |
| P25 | 1 | 0 | 0 | 0 | 0 | 0 |
| P26 | 0 | 0 | 1 | 0 | 0 | 0 |
| P27 | 1 | 1 | 0 | 0 | 0 | 1 |
| P28 | 0 | 0 | 1 | 0 | 0 | 0 |
| P30 | 1 | 0 | 0 | 0 | 0 | 0 |
| P32 | 0 | 1 | 0 | 1 | 0 | 0 |
| P33 | 1 | 1 | 0 | 0 | 0 | 0 |
| P34 | 0 | 0 | 1 | 0 | 0 | 0 |
| P36 | 0 | 0 | 0 | 1 | 0 | 1 |
| P38 | 0 | 0 | 0 | 1 | 0 | 0 |
| P39 | 1 | 1 | 0 | 0 | 0 | 0 |
| P41 | 0 | 1 | 0 | 1 | 0 | 1 |
| P42 | 0 | 0 | 0 | 0 | 1 | 0 |
| P43 | 0 | 0 | 1 | 1 | 1 | 0 |

Le treillis de Galois obtenu en sortie de l'algorithme est représenté sur la Figure IV-11. Bien qu'il soit de taille nettement plus raisonnable que les treillis obtenus avec les deux analyses précédentes, son interprétation reste délicate pour un non spécialiste de l'ACF. Nous poursuivons donc notre étude en utilisant la mesure de similarité conceptuelle, comme nous l'avons fait pour les cas précédents.

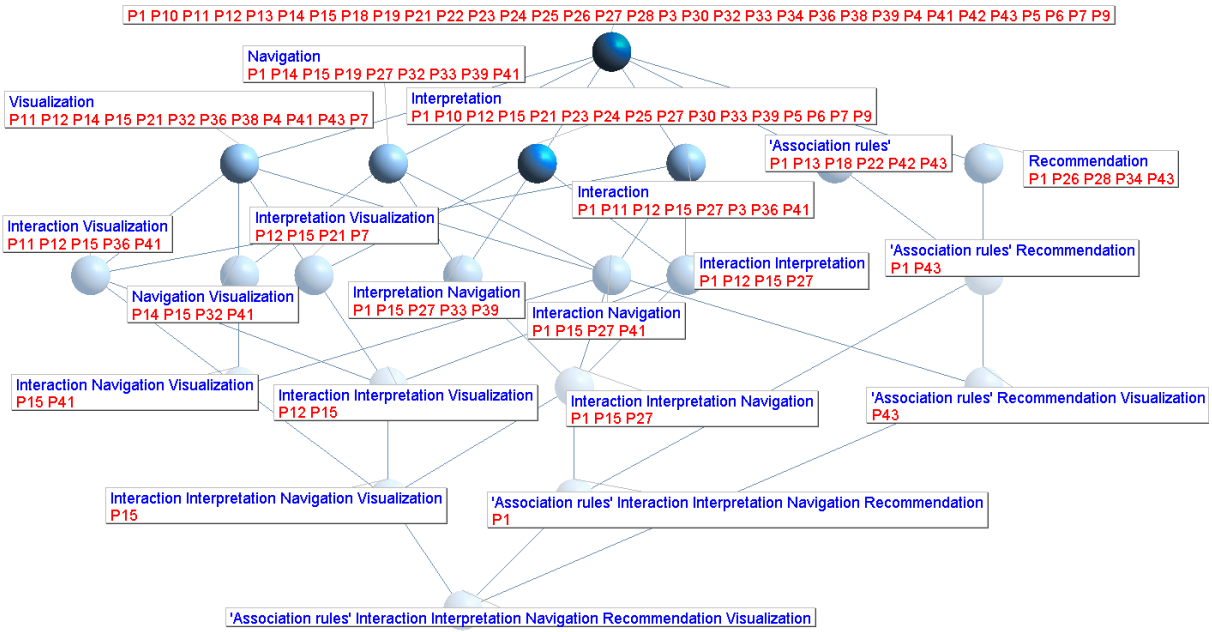



**Figure IV-11 : Treillis de Galois obtenu avec les 6 mots-clés dédiés à l'interprétation des treillis de Galois**

### IV.4.2. Similarité conceptuelle entre les 6 mots-clés ciblés sur l'interprétation des treillis de Galois

La Figure IV-12 représente la similarité conceptuelle entre les 6 mots-clés ciblés sur l'interprétation des treillis de Galois, de la même manière que nous l'avons fait dans le Chapitre III. Cette représentation permet d'identifier facilement les mots-clés très similaires (par exemple *Recommandation* et *Association rules*), mais la représentation sous forme de carte MDS de la Figure IV-15 facilitera l'interprétation globale.

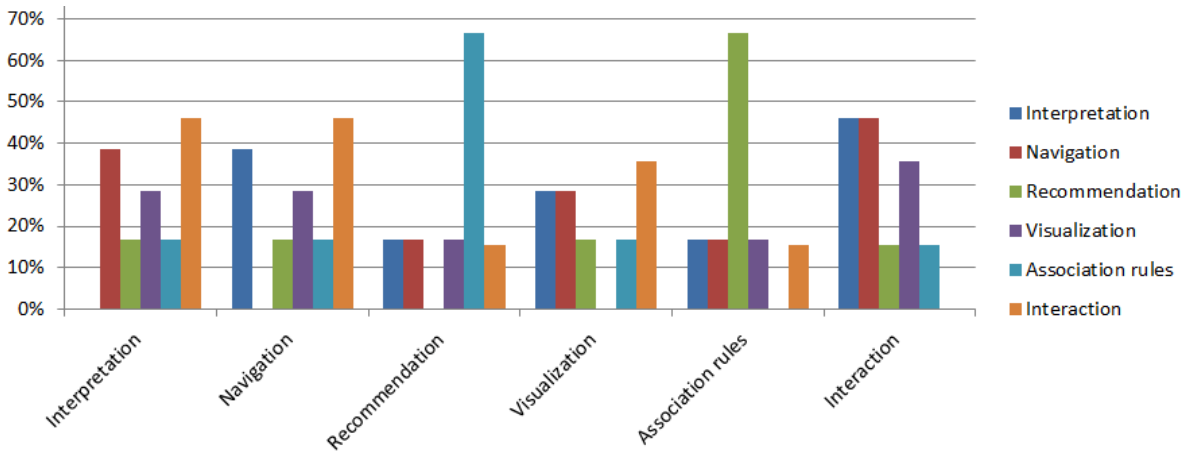



**Figure IV-12 : Similarité conceptuelle entre les 6 mots-clés dédiés à l'interprétation du treillis de Galois**

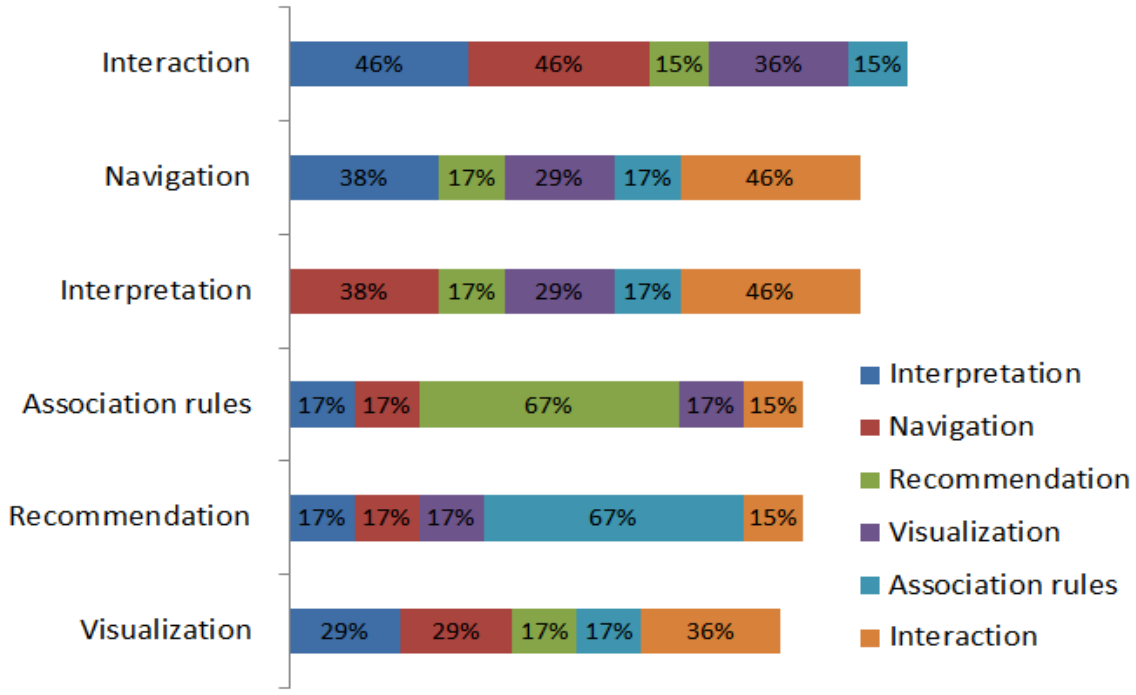



**Figure IV-13 : Similarité conceptuelle entre les 6 mots-clés dédiés à l'interprétation du treillis de Galois**

La Figure IV-13 montre que les 6 mots-clés retenus ici sont très comparables en termes de similarité « globale » avec les 5 autres attributs. Ceci est un résultat positif dans le contexte d'une revue de la littérature, dans la mesure où cela montre que ces mots-clés sont tous autant les uns que les autres au cœur du sujet.

Nous présentons sur la Figure IV-14 et la Figure IV-15 respectivement le résultat du clustering et la carte 2D reflétant les clusters de mots-clés obtenus.

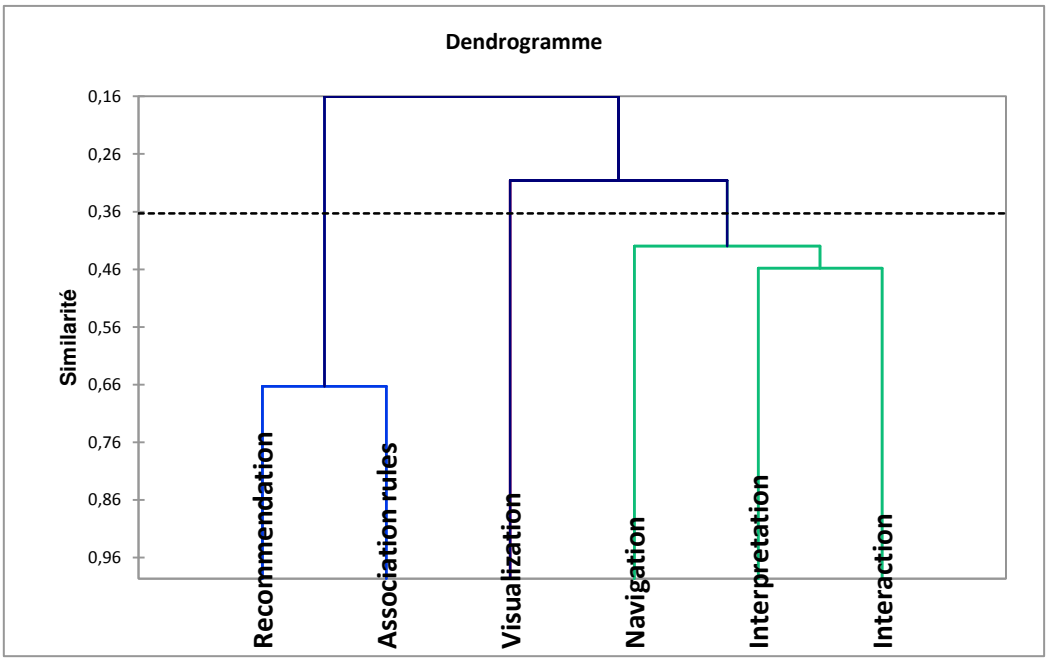



**Figure IV-14 : Résultat du clustering ascendant hiérarchique appliqué à la matrice de similarité conceptuelle entre les 6 mots-clés ciblés sur l'interprétation des résultats d'ACF**

La CAH nous propose de séparer les mots-clés en trois clusters, réunissant *recommendation* et *association rules* d'une part, *navigation, interprétation et interaction* d'autre part, et enfin *visualization*.

La représentation de ces clusters sur la carte MDS est fournie sur la Figure IV-15.

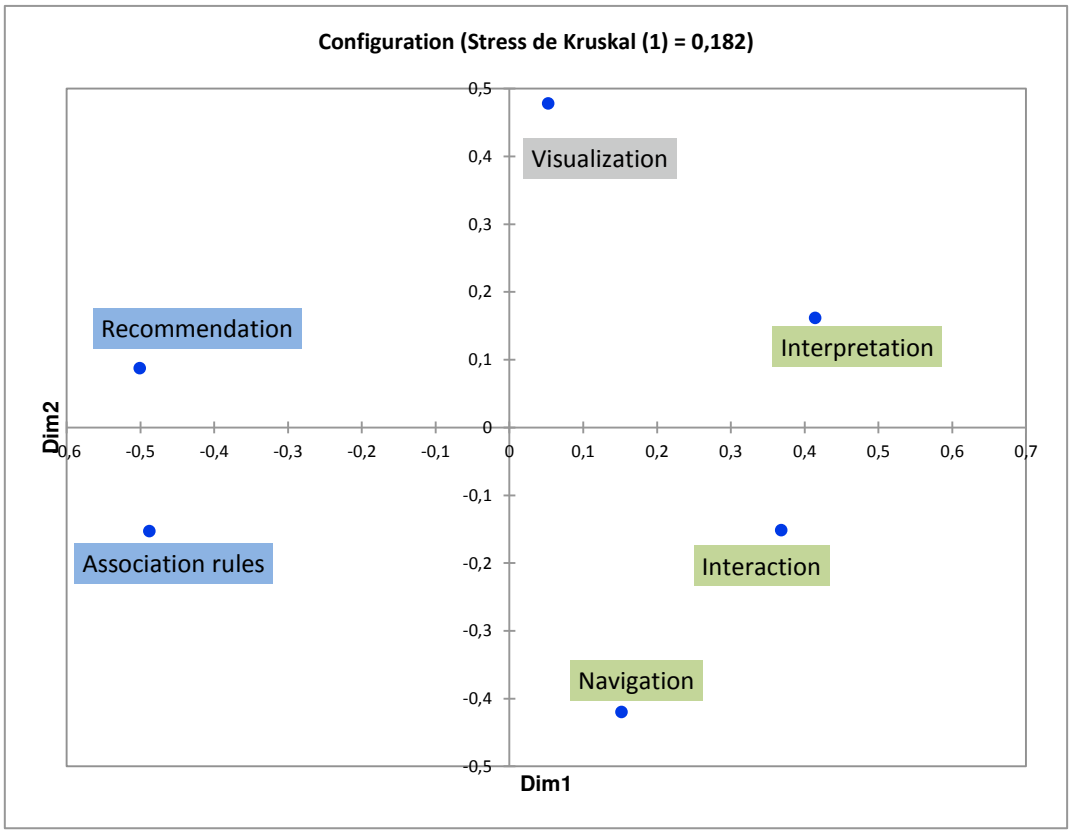



**Figure IV-15 : Carte 2D des 6 mots-clés ciblés sur l'interprétation des résultats de l'ACF, enrichie des clusters obtenus par CAH**

Nous avons poursuivi l'analyse en étudiant la répartition des articles décrits par ces 6 mots-clés. Nous en donnons les résultats dans la section suivante.

### IV.4.3. **Similarité conceptuelle entre les articles décrits par les 6 mots-clés dédiés à l'interprétation des résultats de l'ACF**

Les résultats de notre analyse sur la similarité conceptuelle des articles correspondant sont présentés sur la Figure IV-16 et la Figure IV-17, avec respectivement le résultat du clustering et la carte 2D reflétant les clusters obtenus.

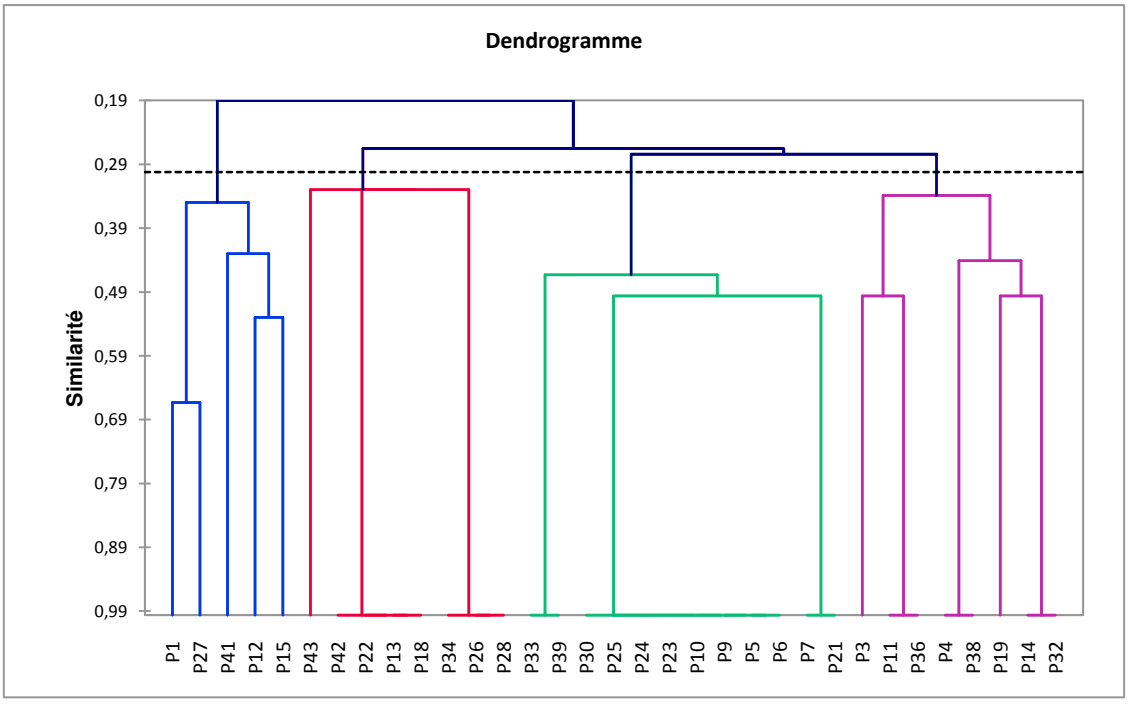



**Figure IV-16 : Résultat du clustering ascendant hiérarchique appliqué à la matrice de similarité conceptuelle entre les articles décrits par les 6 mots-clés ciblés sur l'interprétation des résultats de l'ACF**

Les résultats sont satisfaisants dans la mesure où les articles sont répartis de manière assez homogène entre les 4 clusters identifiés par la CAH sur la Figure IV-16, ce qui est confirmé par la représentation sous forme de carte de la Figure IV-17.

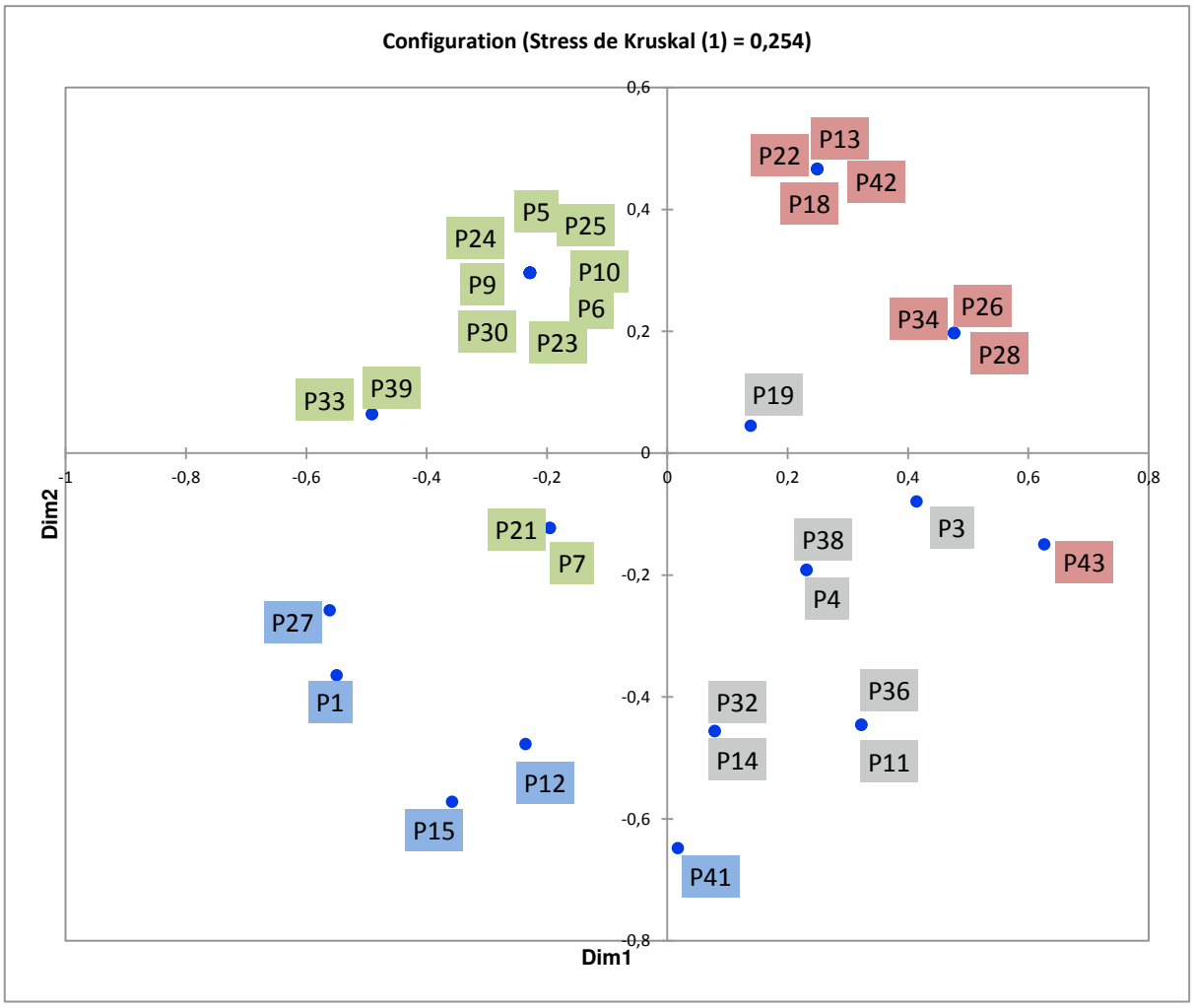



**Figure IV-17 : Carte 2D des articles décrits par les 6 mots-clés de niveau 2 dédiés à l'interprétation, enrichie des clusters obtenus par CAH**

Dans la section suivante, nous comparons la répartition des articles dans les sections de notre état de l'art manuel avec les clusters identifiés de manière automatique.

### IV.4.4. **Comparaison avec la structure de l'état de l'art manuel**

Sur la Figure IV-18, la couleur des articles reflète le cluster (identifié par la CAH) auquel ils appartiennent.

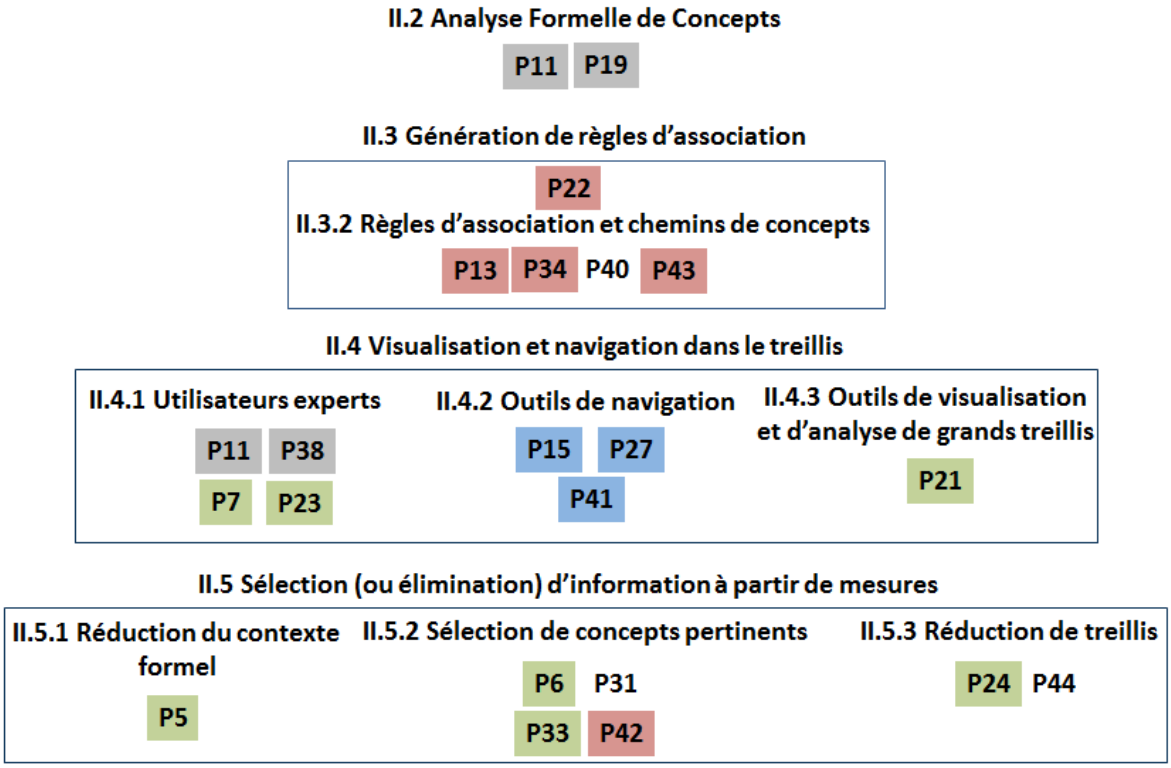



**Figure IV-18 : Répartition des articles dans l'état de l'art réalisé manuellement pour ce manuscrit**

Nous constatons une meilleure cohérence entre la structure de l'état de l'art manuelle et la structure suggérée de manière automatique avec notre méthodologie. On identifie pour quasiment chaque section une couleur « majoritaire » (et même unique pour certaines).

Cela confirme que notre état de l'art manuel a bien été construit en étudiant tout particulièrement, dans les articles que nous avons lus, la manière dont l'interprétation des treillis de Galois était effectuée.

## IV.5. **Conclusion**

Nous avons proposé dans ce chapitre une étude de cas dans le but d'expérimenter notre méthodologie pour faciliter l'interprétation des treillis de Galois. Afin d'avoir un élément de comparaison, nous avons choisi d'utiliser notre approche pour guider la construction d'un état de l'art, en prenant comme objets des articles de recherche dédiés à l'ACF et comme attributs les mots-clés décrivant le contenu de ces articles. L'objectif est, d'une part, d'identifier le thème de chaque section de l'état de l'art et, d'autre part, de répartir les articles dans chacune de ces sections. Parmi les mesures que nous avons définies dans le Chapitre III, c'est la similarité conceptuelle qui est la plus adaptée, notamment lorsqu'elle est complétée par l'identification de clusters à l'aide de la classification ascendante hiérarchique. C'est pourquoi nous avons exclusivement exploité cette mesure dans ce chapitre.

Cette étude de cas s'est déroulée en trois temps :

- Nous avons tout d'abord identifié une première liste de 15 mots-clés pour décrire les articles de recherche. Nous avons rapidement conclu que ces mots-clés ne permettaient pas de proposer une structure exploitable pour un état de l'art.
- Nous avons alors étendu la liste de mots-clés initiale, en proposant des mots-clés plus spécifiques. Nous avons obtenu des résultats satisfaisants avec cette 2e liste de mots-clés, dans la mesure où elle a permis de proposer des sections pertinentes et une répartition des articles dans ces sections.
- Nous avons ensuite appliqué notre approche au sous-ensemble des mots-clés dédiés à l'interprétation des treillis de Galois, puisque c'est le focus que nous avons choisi pour le chapitre d'état de l'art construit manuellement pour ce manuscrit. Les résultats obtenus ont confirmé la cohérence entre la structure manuelle et la structure automatique issue de notre approche.

Cette étude de cas a permis d'illustrer un cas d'usage de notre approche, en nous focalisant tout particulièrement sur l'une des trois mesures que nous avons proposées pour l'interprétation automatique des treillis de Galois.

Dans le chapitre suivant, nous étendrons notre approche en proposant une manière de construire des contextes formels pertinents lorsque la relation entre les données fournies en entrée n'est pas naturellement sous forme binaire.

# CHAPITRE V :
# STRATÉGIES DE CONSTRUCTION DU CONTEXTE FORMEL

# Sommaire

## V.1. **Introduction**

Nous avons vu dans les chapitres précédents que les treillis de Galois étaient générés à partir d'une matrice binaire appelée contexte formel, qui représente la relation entre des objets et des attributs. Nous nous intéressons ici à la manière dont cette matrice binaire peut être construite lorsque les données de départ peuvent prendre des valeurs autres que 0 ou 1.

Nous présentons tout d'abord une stratégie simpliste, dont les limites nous ont amenés à proposer des stratégies alternatives qui reflètent mieux l'intensité des relations entre les objets et les attributs des données initiales.

## V.2. **Stratégie simpliste pour la construction d'un contexte formel**

Considérons l'exemple du Tableau V-1, qui contient, pour un utilisateur donné, le nombre de fois que chaque application a été utilisée dans chaque élément de contexte. Ce tableau montre par exemple que l'application *SMS* a été utilisée 80 fois à l'université et une seule fois au restaurant (en une semaine). Une stratégie simpliste pour construire un contexte formel à partir d'un tel tableau est de remplacer par 1 toute valeur non nulle, comme le montre le Tableau V-2. Cela reviendrait alors à ne pas considérer le nombre de fois que les applications ont été utilisées. Par exemple, l'application SMS qui est utilisée 80 fois à l'université et une seule fois au restaurant est représentée comme ayant indifféremment une relation avec ces deux éléments (valeur binaire égale à 1).

**Tableau V-1 : Matrice d'utilisation des applications en fonction des éléments de contexte**

| Applications | *University* | *Restaurant* | *Parents' home* | *Home* | *Transportation* | *3G* | *Morning* | *Coffee break* | *Lunch break* | *Afternoon* | *Evening* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Gmail* | 2 | 0 | 0 | 0 | 1 | 2 | 1 | 0 | 0 | 1 | 0 |
| *SMS* | 80 | 1 | 5 | 30 | 50 | 170 | 40 | 20 | 40 | 60 | 10 |
| *Telephone* | 0 | 0 | 0 | 3 | 2 | 5 | 1 | 0 | 0 | 2 | 1 |
| *VDM* | 0 | 0 | 0 | 4 | 2 | 5 | 4 | 0 | 0 | 0 | 2 |
| *Flappy Bird* | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| *Youtube* | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |

**Tableau V-2 : Représentation binaire de la matrice d'utilisation des applications selon la stratégie simpliste**

| Applications | *University* | *Restaurant* | *Parents' home* | *Home* | *Transportation* | *3G* | *Morning* | *Coffee break* | *Lunch break* | *Afternoon* | *Evening* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Gmail* | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 |
| *SMS* | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| *Telephone* | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 |
| *VDM* | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 |
| *Flappy Bird* | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| *Youtube* | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |

Nous avons utilisé le Tableau V-2 dans le Chapitre III pour illustrer les mesures d'interprétation de treillis.

L'avantage de la stratégie simpliste de construction du contexte formel, comme son nom l'indique, est que l'obtention de la matrice binaire à partir de données initiales ne nécessite pas un calcul complexe. Il suffit de repérer les valeurs positives et celles qui sont nulles pour obtenir une matrice binaire. Par contre, l'utilisation de cette méthode peut produire une représentation éloignée des données initiales et biaiser l'ensemble de connaissances qui peuvent extraites du treillis de Galois résultant, car le contexte formel obtenu ne reflète pas les valeurs relatives lorsque les données initiales ne sont pas binaires, comme c'est le cas dans le Tableau V-1.

## V.3. **Stratégies alternatives de construction de contexte formel**

Pour obtenir un contexte formel plus représentatif de la force des relations entre les données initiales, nous proposons dans ce chapitre d'autres stratégies de construction d'une matrice binaire, adaptées à la réalité des données et aux besoins d'information de l'utilisateur. Nous proposons de tenir compte de l'intensité de la relation entre chaque objet et chaque attribut, via le calcul d'une valeur de fréquence. Cette fréquence est conçue pour pouvoir exploiter les éventuelles informations sémantiques relatives aux attributs, comme nous le décrivons dans la section V.3.1. Une fois cette fréquence calculée pour chaque objet, nous définissons un seuil-bas et un seuil-haut, qui permettent de définir trois ensembles :

- l'ensemble des fréquences strictement supérieures au seuil-haut ; ces fréquences élevées correspondent à des relations fortes entre des objets et des attributs.

- l'ensemble des fréquences strictement inférieures au seuil-bas, appelées fréquences basses ; ces fréquences faibles correspondent à des relations faibles entre des objets et des attributs.
- l'ensemble des fréquences comprises entre le seuil-bas et le seuil-haut, appelées fréquences moyennes ; ces fréquences moyennes correspondent à des relations d'intensité modérée entre des objets et des attributs.

La stratégie simpliste de construction de contexte formel évoquée, utilisée dans les exemples du Chapitre III, considère comme pertinentes toutes les relations entre des objets et des attributs, dès lors qu'une relation existe, même de faible intensité. A l'inverse, nous proposons de définir trois stratégies de construction de contexte formel, selon que l'on souhaite mettre l'accent sur les relations fortes, moyennes ou faibles, en utilisant les trois ensembles définis ci-dessus :

- la **stratégie à haute dépendance**, où les fréquences élevées sont transformées en 1 et toutes les autres en 0 pour obtenir une matrice binaire,
- la **stratégie à faible dépendance**, où les fréquences faibles sont transformées en 1 et toutes les autres en 0 pour obtenir une matrice binaire,
- la **stratégie à dépendance moyenne**, où les fréquences moyennes sont transformées en 1 et toutes les autres en 0 pour obtenir une matrice binaire.

La Figure V-1 illustre les trois intervalles de fréquences délimités par les deux seuils, haut et bas, ainsi que les trois stratégies correspondantes.

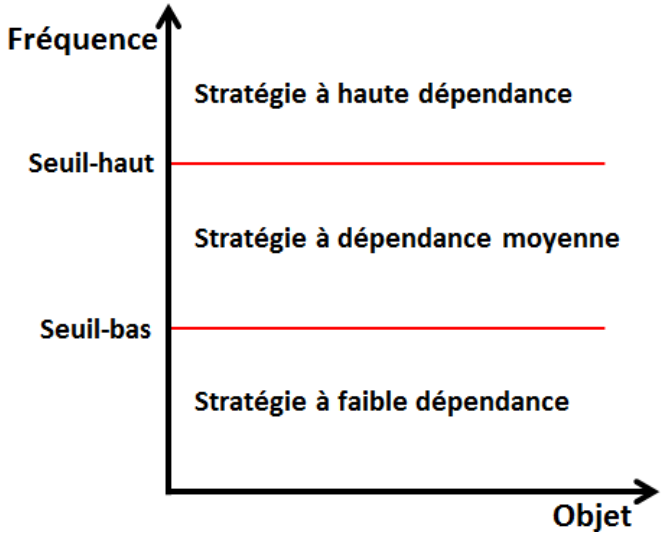



**Figure V-1 : Stratégies de construction de contexte formel dépendantes de la fréquence**

Dans la suite de ce chapitre, nous proposons tout d'abord deux manières calculer la fréquence, exploitant ou non des informations sémantiques relatives aux attributs, avant de détailler les trois stratégies de construction de contexte formel et d'illustrer leur impact en termes d'interprétation du treillis.

### V.3.1. **Calcul de fréquence pour évaluer l'intensité des relations entre objets et attributs**

#### Fréquence sémantique

Nous définissions pour un objet et un élément de contexte donnés, une fréquence dite sémantique $f_s(o,a)$ qui tient compte de la signification (*i.e.* type) des différents éléments de contexte :

$$f_s(o,a) = \frac{\text{Nb d'occurrences de l'objet o dans l'élément de contexte a}}{\text{Nb d'occurrences de l'objet o dans les éléments de contexte connexes}}$$

Les éléments de contexte sont considérés comme connexes s'ils sont liés sémantiquement entre eux, par exemple s'ils sont de la même catégorie comme le montre la Figure V-2.

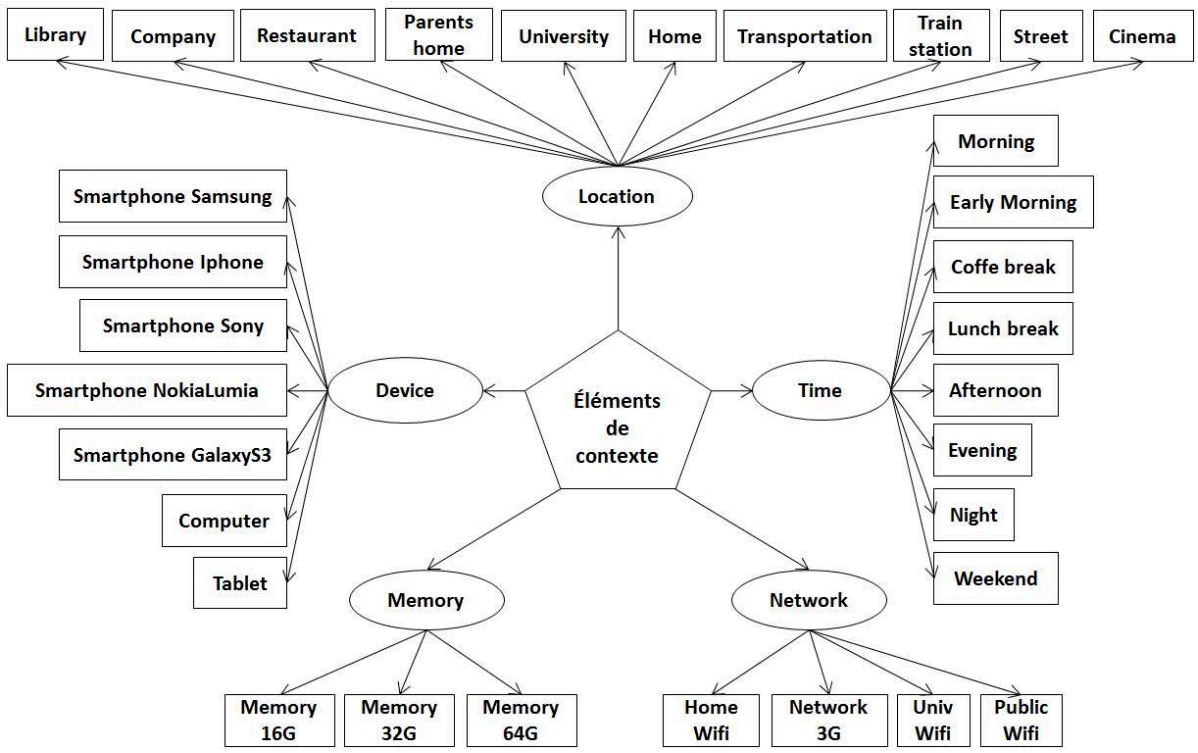



**Figure V-2 : Taxonomie des éléments de contexte**

Nous présentons dans la Figure V-2 une taxonomie des éléments de contexte, qui sont divisés en cinq catégories :

*Location* : les localisations géographiques dans lesquelles les applications sont utilisées.

*Time* : les instants durant lesquels les applications sont utilisées.

*Network* : les réseaux utilisés sur chaque appareil.

*Memory* : la taille de la mémoire pour chaque appareil utilisé.

*Device* : le type d'appareil utilisé pour les applications.

**Fréquence non sémantique**

Nous avons également défini une fréquence non-sémantique $f_{ns}(o,a)$ pour un objet et un attribut donnés, qui est indépendante de toute relation sémantique entre les éléments de contexte. Ceci permet de refléter l'intensité de la relation entre cet objet et cet attribut même si on ne dispose d'aucune information sur le type des attributs :

$$f_{ns}(o,a) = \frac{\text{Nb d'occurrences de l'objet o dans l'élément de contexte a}}{\text{Nb d'occurrence de l'objet o dans tous les éléments de contexte}}$$

Le Tableau V-3 indique par exemple que les éléments de contexte *university, restaurant, parent's home, home* et *transportation* relèvent de la catégorie *localisation*, alors que *morning, coffee break, lunch break, afternoon* et *evening* appartiennent à la *catégorie* temps. Ce même tableau illustre le calcul des deux types de fréquence définis ci-dessus, pour l'application *SMS*. La fréquence sémantique de cette application pour un élément de contexte donné, par exemple *University*, est calculée à partir du nombre d'occurrences de l'application dans ce contexte, divisé par le nombre d'occurrences dans tous les éléments de contexte relevant de la même catégorie (i.e., de localisation), soit 80/166. La fréquence non sémantique, quant à elle, ne distingue aucune catégorie d'éléments de contexte. C'est pourquoi le nombre d'occurrences de l'application à l'université est divisé par le nombre d'occurrences total de cette application, tous éléments de contexte confondus, soit 80/506.

**Tableau V-3 : Fréquence de l'application *SMS* dans les différents contextes**

| | Localisation | | | | | Temps | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | University | Restaurant | Parents' home | Home | Transportation | Morning | Coffee break | Lunch break | Afternoon | Evening |
| $f_s$ | 80/166 | 1/166 | 5/166 | 30/166 | 50/166 | 40/170 | 20/170 | 40/170 | 60/170 | 10/170 |
| $f_{ns}$ | 80/506 | 1/506 | 5/506 | 30/506 | 50/506 | 40/506 | 20/506 | 40/506 | 60/506 | 10/506 |

### V.3.2. **Stratégies de construction de contexte formel dépendantes de la fréquence**

Comme nous l'avons expliqué plus haut, les stratégies alternatives de construction de contexte formel que nous proposons considèrent le nombre d'occurrences objet-attribut, en calculant une fréquence. Nous partitionnons ensuite les valeurs de fréquences obtenues en trois intervalles délimités par deux valeurs seuils : seuil-bas ($S_b$) et seuil-haut ($S_h$) (voir Figure V-1). Nous définissons ces deux seuils comme suit :

$$Seuil\text{-}haut = \bar{X} + \beta \times \sigma$$

$$Seuil\text{-}bas = \bar{X} - \beta \times \sigma$$

où $\bar{X}$ et $\sigma$ représentent respectivement la moyenne et l'écart-type des valeurs de fréquences d'utilisation d'une application dans les différents éléments de contexte. $\beta$ est une variable réelle $\in [0,1]$ qui sert à changer la valeur des seuils dans un intervalle centré en $\bar{X}$ et ayant $\bar{X} - \beta \times \sigma$ comme borne inférieure, et $\bar{X} + \beta \times \sigma$ comme borne supérieure[16]. Les fréquences sont ainsi partitionnées en trois intervalles[17] :$]-\infty, \bar{X} - \sigma[$ , $[\bar{X} - \sigma, \bar{X} + \sigma]$ *et* $]\bar{X} + \sigma, +\infty[$.

[16] Dans toute la suite du manuscrit, nous avons choisi la valeur 0,25 pour le paramètre $\beta$. L'optimisation du choix de ce paramètre n'a pas été abordée durant cette thèse et fera l'objet de travaux futurs.

[17] Une valeur de $\beta$=0 produit une seule valeur seuil $S_b = S_h = \bar{X}$, qui divise l'ensemble de fréquences en deux intervalles : $]-\infty, \bar{X}[ \cup ]\bar{X}, +\infty[$.

Ces trois intervalles de fréquences sont respectivement associés à nos trois stratégies de construction de contexte formel :

- Stratégie à haute dépendance : elle remplace par la valeur 1 les fréquences appartenant à l'intervalle]$S_h$, +∞[. En se basant sur la valeur de $S_h$ déjà définie, la stratégie qui s'intéresse aux relations à forte dépendance peut être représentée par l'équation suivante :

$$Stratégie\_haute(x) = \begin{cases} 1 \ si \ x > S_h \\ 0 \ si \ x \leq S_h \end{cases}$$

où $x$ est la fréquence des occurrences de la relation entre un objet et un attribut, et *Stratégie_haute(x)* la valeur binaire associée à $x$ selon cette stratégie dite *haute*.

Cette stratégie se concentre ainsi sur les objets et les attributs qui sont fortement liés. Par exemple l'application *Gmail* qui est souvent utilisée dans l'élément de contexte *university* sera incluse dans le treillis de Galois à construire (Tableau V-4).

- Stratégie à faible dépendance : à l'inverse de ce que nous avons vu dans la stratégie à haute dépendance, nous traitons dans cette stratégie les objets et les attributs qui sont faiblement liés. La construction du contexte formel est exprimée par l'équation suivante :

$$Stratégie\_basse(x) = \begin{cases} 1 \ si \ x < S_b \\ 0 \ si \ x \geq S_b \end{cases}$$

où $x$ est la fréquence des occurrences de la relation entre un objet et un attribut, et *Stratégie_basse(x)* la valeur binaire associée à $x$ selon cette stratégie dite *basse*. Par exemple, dans le Tableau V-6 , l'application *Gmail* est faiblement utilisée dans les *Transports* car sa fréquence d'utilisation est inférieure à la valeur $S_b$.

- Stratégie à dépendance moyenne : elle considère les valeurs de fréquences appartenant à l'intervalle [$S_b$, $S_h$]. Cette stratégie se focalise sur les relations de fréquence moyenne entre les objets et les attributs. La construction du contexte formel est exprimée par l'équation suivante :

$$Stratégie\_moyenne(x) = \begin{cases} 1 \ si \ S_b \leq x \leq S_h \\ 0 \qquad\quad sinon \end{cases}$$

où $x$ est la fréquence des occurrences de la relation entre un objet et un attribut, et *Stratégie_moyenne(x)* la valeur binaire associée à $x$ selon cette stratégie dite *moyenne*.

### V.3.3. **Stratégie inverse**

Cette stratégie est similaire à la stratégie simpliste du point de vue de son indépendance vis-à-vis des valeurs de fréquence. Par contre, cette stratégie sert à identifier les applications et les éléments de contexte qui n'ont aucune relation de dépendance entre eux. Cette stratégie peut se traduire par une fonction de transformation exprimée par l'équation suivante :

$$Stratégie\_inverse(x) = \begin{cases} 1 \ si & x = 0 \\ 0 \ si & x \in \mathbb{N}^* \end{cases}$$

Cette stratégie permet de se focaliser sur les relations qui n'existent pas entre les données. Pour cela le contexte formel est une matrice binaire où toutes les fréquences non nulles sont remplacées par des 0 et les fréquences nulles par des 1. Le treillis obtenu par cette stratégie peut être utilisé pour mettre en évidence les relations qui ne sont pas présentes entre les données initiales.

Toutes les stratégies alternatives que nous avons proposées ici pour construire un contexte formel sont illustrées dans la section suivante. Cela nous permettra également de donner des pistes en termes de critères de prescription d'une stratégie plutôt qu'une autre ; ce point nécessitera néanmoins d'être approfondi dans l'avenir.

## V.4. **Illustration des stratégies alternatives de construction d'un contexte formel**

L'objectif de cette section est d'illustrer sur l'exemple des données du Chapitre III, dédié à l'étudiant E6, la construction du contexte formel selon les stratégies alternatives que nous avons proposées. Une étude détaillée de l'interprétation des treillis obtenus sera fournis dans la section V.5.1.

### V.4.1. **Stratégie à haute dépendance**

Dans le Tableau V-4, nous calculons tout d'abord la valeur $S_h$ associée à chaque application en utilisant une valeur de $\beta$ = 0,25. Ensuite, nous comparons chaque entrée du tableau à la valeur seuil associée en utilisant la fonction *Stratégie_haute(x)*. Le résultat de cette transformation est présenté dans le Tableau V-5, qui représente une matrice binaire

permettant de construire un treillis de Galois. Les valeurs supérieures au seuil $S_h$ associé (cases bleues) seront remplacées par 1 dans le tableau de sortie (ex. *Gmail-University*), et celles qui sont inférieures seront remplacées par 0 (ex. *SMS-Restaurant*). Pour des raisons de lisibilité nous avons choisi de supprimer les lignes et les colonnes qui ne contiennent que des zéros. Nous avons également laissé vides les cases contenant un zéro.

**Tableau V-4 : Valeurs de fréquence supérieures au seuil $S_h$**

| | *Univ* | *Restaurant* | *Parents' home* | *Home* | *Transportation* | *3G* | *Morning* | *Coffe break* | *Lunch break* | *Afternoon* | *Evening* | $S_h$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***Gmail*** | 0,666 | | | | 0,333 | 1 | 0,5 | | | 0,5 | | **0,656** |
| ***SMS*** | 0,481 | 0,006 | 0,03 | 0,18 | 0,301 | 1 | 0,235 | 0,117 | 0,235 | 0,352 | 0,058 | **0,339** |
| ***Telephon*** | | | | 0,6 | 0,4 | 1 | 0,25 | | | 0,5 | 0,25 | **0,564** |
| ***VDM*** | | | | 0,666 | 0,333 | 1 | 0,666 | | | | 0,333 | **0,662** |
| ***Flappy Bird*** | 0,333 | | | 0,333 | 0,333 | 1 | 0,25 | 0,25 | | 0,25 | 0,25 | **0,434** |
| ***Youtube*** | | | | 1 | | 1 | 1 | | | | | **1** |

**Tableau V-5 : Représentation binaire des valeurs de fréquences**

| | *Univ* | *Home* | *3G* | *Morning* | *Afternoon* |
|---|---|---|---|---|---|
| *Gmail* | 1 | 0 | 1 | 0 | 0 |
| *SMS* | 1 | 0 | 1 | 0 | 1 |
| *Telephone* | 0 | 1 | 1 | 0 | 0 |
| *VDM* | 0 | 1 | 1 | 1 | 0 |
| *Flappy Bird* | 0 | 0 | 1 | 0 | 0 |

En nous basant sur les valeurs obtenues dans le Tableau V-5, nous construisons le treillis de Galois associé qui est présenté dans la Figure V-3. Ce treillis contient six concepts pour les applications qui ont une relation de dépendance forte avec les éléments de contexte. L'interprétation de ce treillis sera fournie dans la section V.5.1.

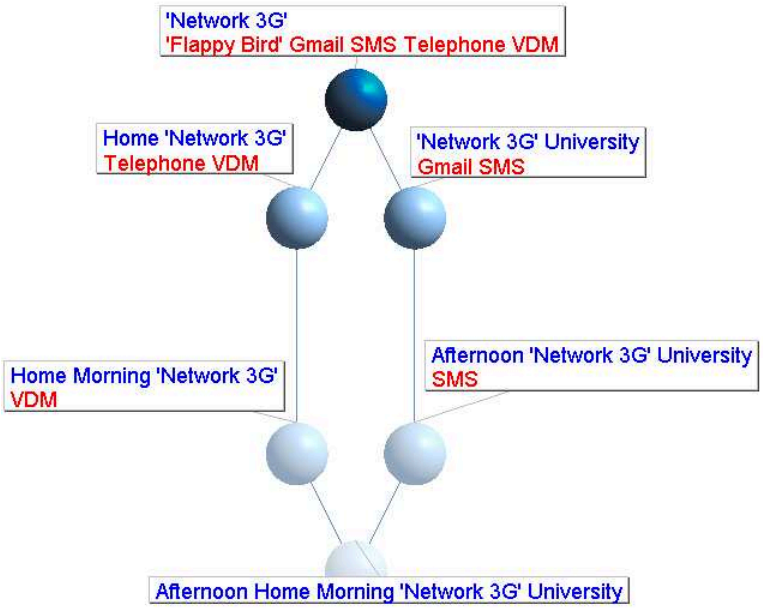



**Figure V-3 : Treillis de Galois des relations à haute dépendance**

## V.4.2. **Stratégie à faible dépendance**

Dans le Tableau V-6, nous calculons tout d'abord la valeur seuil-bas ($S_b$) associée à chaque application en utilisant une valeur de $\beta$ = 0,25. Ensuite, nous comparons chaque entrée du tableau à la valeur seuil associée en utilisant la fonction *Stratégie_basse(x)*. Le résultat de cette transformation est présenté dans le Tableau V-7, qui représente une matrice binaire permettant de construire un treillis de Galois. Selon cette stratégie, les valeurs inférieures au seuil $S_b$ seront remplacées par 1 dans le tableau de sortie (ex. *Gmail-Transport*), et celles qui sont supérieures à $S_b$ seront remplacées par 0 (ex. *SMS-University*).

**Tableau V-6 : Exemple de relations à faible dépendance**

| | *Univ* | *Restaurant* | *Parents' home* | *Home* | *Transportation* | *3G* | *Morning* | *Coffe break* | *Lunch break* | *Afternoon* | *Evening* | $S_b$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Gmail* | 0,666 | | | | 0,333 | 1 | 0,5 | | | 0,5 | | **0,543** |
| *SMS* | 0,481 | 0,006 | 0,0301 | 0,18 | 0,301 | 1 | 0,235 | 0,117 | 0,235 | 0,352 | 0,0588 | **0,205** |
| *Telephone* | | | | 0,6 | 0,4 | 1 | 0,25 | | | 0,5 | 0,25 | **0,435** |
| *VDM* | | | | 0,666 | 0,333 | 1 | 0,666 | | | | 0,333 | **0,537** |
| *Flappy Bird* | 0,333 | | | 0,333 | 0,333 | 1 | 0,25 | 0,25 | | 0,25 | 0,25 | **0,315** |
| *Youtube* | | | | 1 | | 1 | 1 | | | | | **1** |

**Tableau V-7 : Contexte formel résultant**

| | *Restaurant* | *Parents' home* | *Home* | *Transportation* | *Morning* | *Coffe break* | *Afternoon* | *Evening* |
|---|---|---|---|---|---|---|---|---|
| *Gmail* | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| *SMS* | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 |
| *Telephone* | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| *VDM* | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| *Flappy Bird* | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |

En nous basant sur les valeurs obtenues dans le Tableau V-7, nous construisons le treillis de Galois associé qui est présenté dans la Figure V-4. Ce treillis contient plus de concepts que ceux qui sont obtenus selon la stratégie à haute dépendance ; il est composé de 14 concepts. L'interprétation du treillis à l'aide des mesures sera fournie dans la section V.5.2.

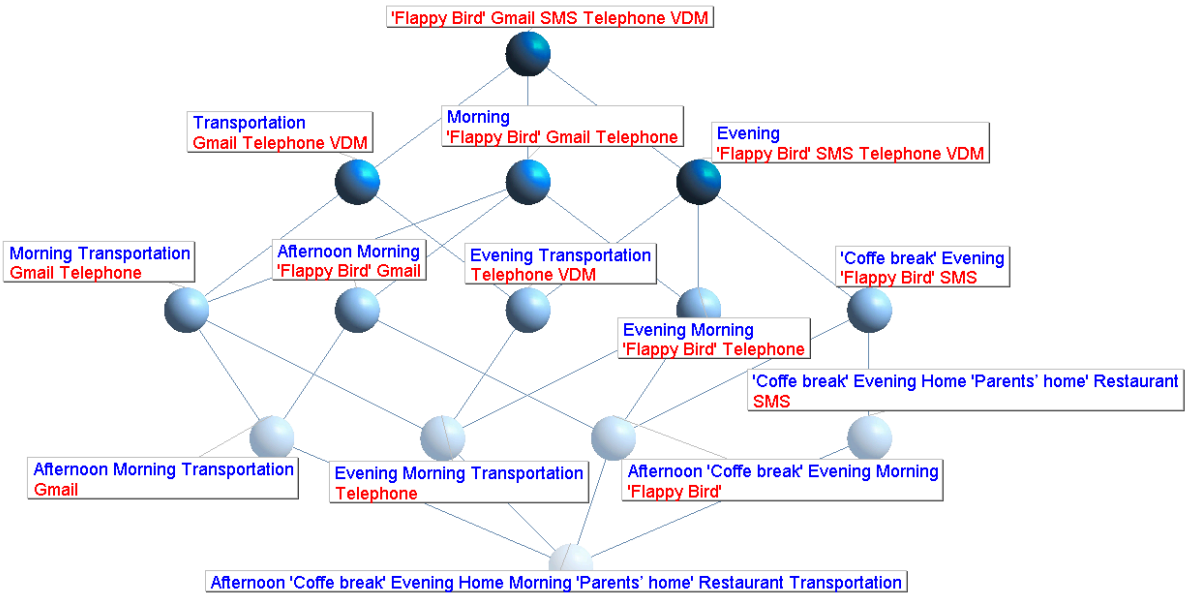



**Figure V-4 : Treillis de Galois des relations à faible dépendance**

### V.4.3. **Stratégie à dépendance moyenne**

Dans le Tableau V-8, nous appliquons la stratégie à dépendance moyenne en utilisant une valeur de $\beta = 0{,}25$ avec la fonction *Stratégie_moyenne(x)*.

**Tableau V-8 : Exemple de relations à dépendance moyenne**

| | *Univ* | *Restaurant* | *Parents' home* | *Home* | *Transportation* | *3G* | *Morning* | *Coffe break* | *Lunch break* | *Afternoon* | *Evening* | $S_b$ | $S_h$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Gmail* | 0,666 | | | | 0,333 | 1 | 0,5 | | | 0,5 | | **0,543** | **0,656** |
| *SMS* | 0,481 | 0,006 | 0,03 | 0,18 | 0,301 | 1 | 0,235 | 0,117 | 0,235 | 0,352 | 0,058 | **0,205** | **0,339** |
| *Telephone* | | | | 0,6 | 0,4 | 1 | 0,25 | | | 0,5 | 0,25 | **0,435** | **0,564** |
| *VDM* | | | | 0,666 | 0,333 | 1 | 0,666 | | | | 0,333 | **0,537** | **0,662** |
| *Flappy Bird* | 0,333 | | | 0,333 | 0,333 | 1 | 0,25 | 0,25 | | 0,25 | 0,25 | **0,315** | **0,434** |
| *Youtube* | | | | 1 | | 1 | 1 | | | | | **1** | **1** |

**Tableau V-9 : Contexte formel résultant**

| | *Univ* | *Home* | *Transportation* | *3G* | *Morning* | *Lunch break* | *Afternoon* |
|---|---|---|---|---|---|---|---|
| *SMS* | 0 | 0 | 1 | 0 | 1 | 1 | 0 |
| *Telephone* | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| *Flappy Bird* | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| *Youtube* | 0 | 1 | 0 | 1 | 1 | 0 | 0 |

Le Tableau V-9 montre la matrice binaire associée aux résultats obtenus en appliquant la stratégie à dépendance moyenne. Cette matrice est générée par le treillis de Galois dans la Figure V-5.

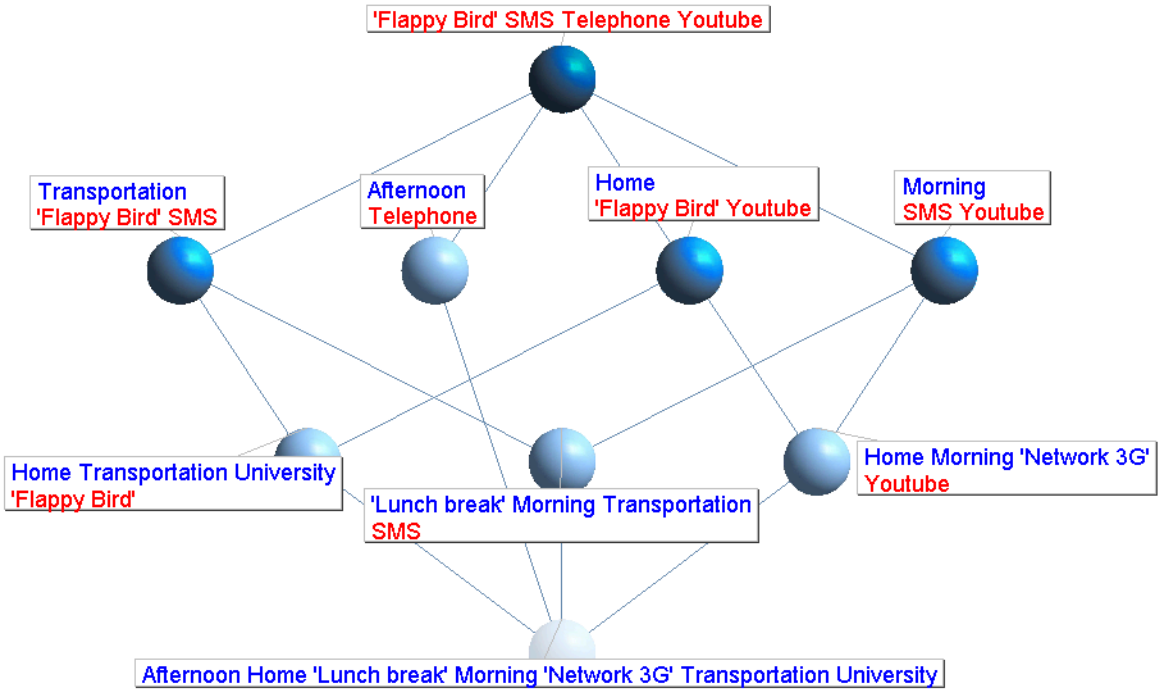



**Figure V-5 : Treillis de Galois des relations à dépendance moyenne**

Pour ne pas alourdir le manuscrit, nous n'avons pas détaillé l'interprétation des résultats obtenus avec cette stratégie.

### V.4.4. **Stratégie inverse**

Le Tableau V-10 montre la matrice binaire associée aux résultats obtenus en appliquant la stratégie inverse. Cette matrice donne lieu au treillis de Galois de la Figure V-6.

**Tableau V-10 : Contexte formel résultant**

| | *University* | *Restaurant* | *Parents home* | *Home* | *Transportation* | *Coffe break* | *Lunch break* | *Afternoon* | *Evening* |
|---|---|---|---|---|---|---|---|---|---|
| *Gmail* | | 1 | 1 | 1 | | 1 | 1 | | 1 |
| *Telephone* | 1 | 1 | 1 | | | 1 | 1 | | |
| *VDM* | 1 | 1 | 1 | | | 1 | 1 | 1 | |
| *Flappy Bird* | | 1 | 1 | | | | 1 | | |
| *YouTube* | 1 | 1 | 1 | | 1 | 1 | 1 | 1 | 1 |

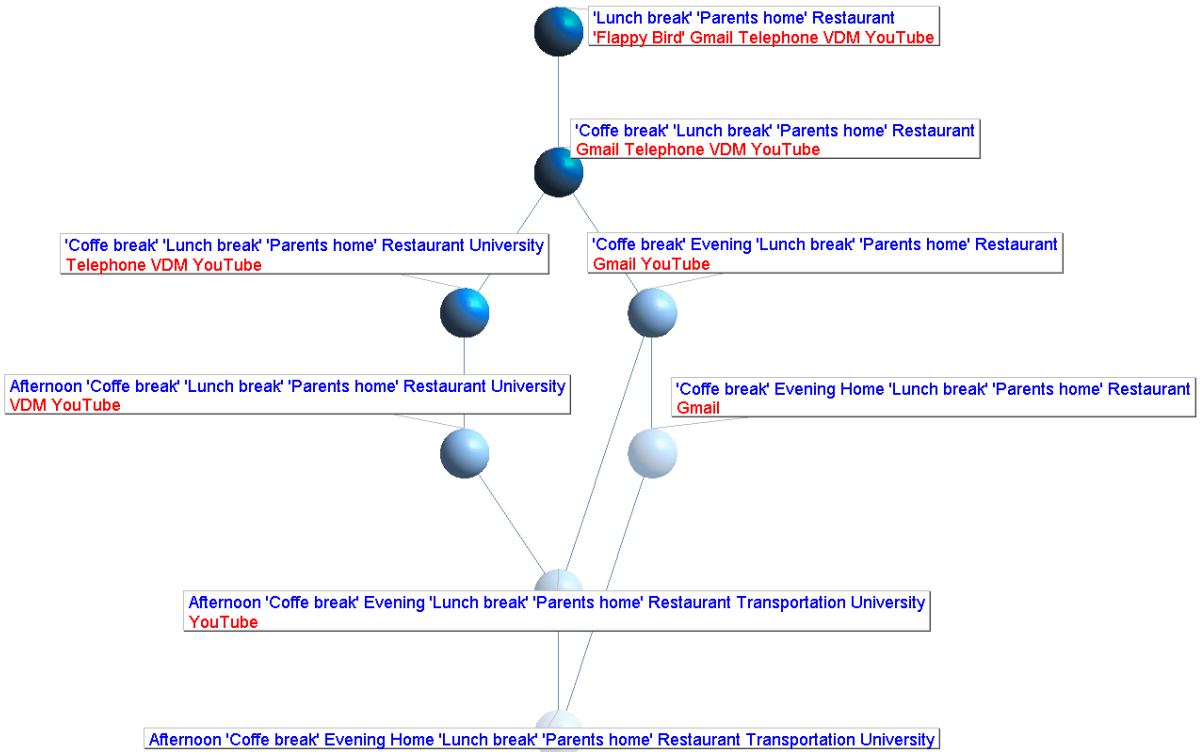



**Figure V-6 : Treillis de Galois pour la stratégie inverse**

L'interprétation du treillis à l'aide des mesures sera fournie dans la section V.5.3.

## V.5. **Discussion sur les stratégies proposées**

Dans ce chapitre, nous avons proposé de nouvelles stratégies pour la construction d'un contexte formel, visant à être à la fois plus représentatives des données initiales et mieux adaptées aux informations recherchées par l'utilisateur en termes d'interprétation des résultats. L'ensemble des stratégies auxquelles nous nous sommes intéressés est représenté sur la Figure V-7. La stratégie simpliste est celle que nous avons illustrée dans le Chapitre III et utilisée pour l'étude de cas du Chapitre IV. Les quatre autres sont les stratégies que nous avons qualifiées d'alternatives. Les stratégies dites « indépendantes de la fréquence » sont adaptées quand la relations entre les données initiales est naturellement binaire. Lorsque ce n'est pas le cas, nous recommandons l'utilisation de l'une des trois stratégies dépendante de la fréquence. Il est idéal de tenir compte de la sémantique des attributs pour calculer cette fréquence, mais il est également possible d'utiliser la fréquence non sémantique.

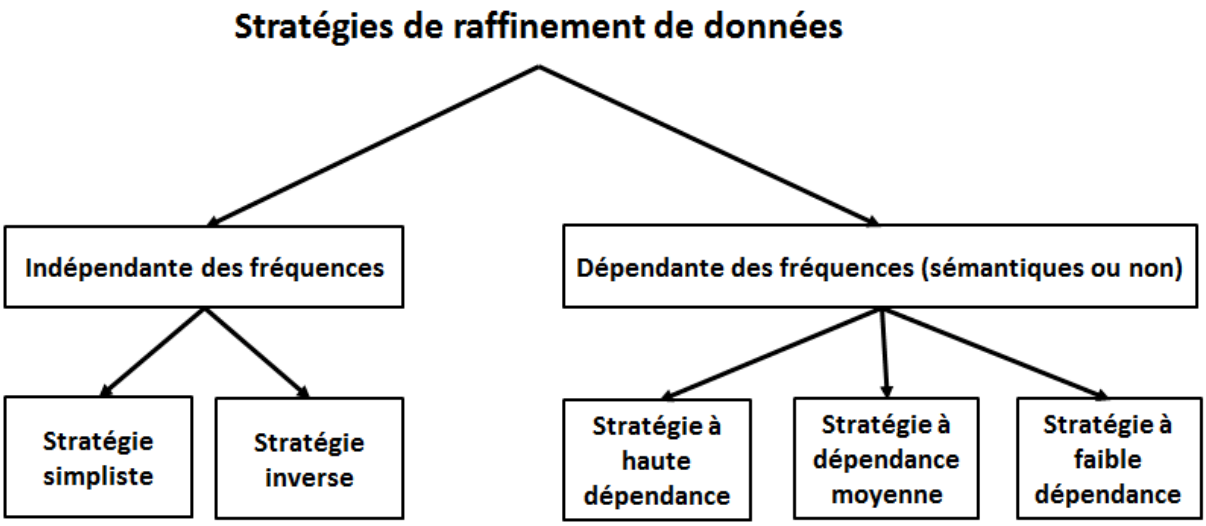



**Figure V-7 : Synthèse des stratégies de construction du contexte formel**

Dans le Tableau V-11, nous présentons les avantages et les inconvénients de chaque stratégie.

**Tableau V-11 : Avantages et inconvénients de chaque stratégie.**

| | Stratégies | Avantages | Inconvénients |
|---|---|---|---|
| **Stratégie indépendante de la fréquence** | **Simpliste** | - Pas de calcul | - Si la relation entre les données initiales n'est pas naturellement binaire, perte d'information sur l'intensité de la relation entre objet et attribut. |
| | **Inverse** | - Pas de calcul<br>- Proposition de nouvelles relations | |
| **Stratégie dépendante d'une fréquence sémantique** | **Haute dépendance** | - Conservation de la valeur relative des informations<br>- Identification de relations fortes, faibles et moyennes | - Paramétrage nécessaire pour fixer les seuils haut et bas |
| | **Faible dépendance** | | |
| | **Dépendance moyenne** | | |

Dans la suite de cette section, nous illustrons les résultats obtenus avec les différentes stratégies[18] :

- la stratégie à haute dépendance, dont nous comparons les résultats avec ceux du Chapitre III, obtenus avec la stratégie simpliste,
- la stratégie à faible dépendance,
- la stratégie inverse.

Pour la stratégie à haute et à faible dépendance, nous avons utilisé dans cette section la fréquence sémantique présentée dans la section V.3.1. Afin de compléter notre expérimentation, nous terminons cette section avec quelques éléments de comparaison entre les résultats obtenus selon le type de fréquence utilisée (sémantique ou non), pour le cas particulier de l'application *Gmail*.

## V.5.1. **Comparaison de la stratégie à haute dépendance selon la stratégie simpliste**

### V.5.1.a. **Comparaison des poids conceptuels**

Nous rappelons que la fréquence qui a été utilisée est la fréquence sémantique.

La première observation que nous pouvons faire en étudiant le Tableau V-12 et la Figure V-8 est que l'application *Youtube* n'apparaît pas dans la stratégie à haute dépendance. Cela signifie que sa fréquence d'utilisation n'est pas assez élevée pour apparaître ici, ce que la stratégie simpliste ne permettait pas de voir.

[18] En dehors de la stratégie moyenne, que nous avons définie mais pas détaillée dans ce mémoire.

**Tableau V-12 : Comparaisons des poids conceptuels**

| | Strategie Simpliste | Strategie à haute dépendance |
|---|---|---|
| *Gmail* | 44% | 33% |
| *SMS* | 100% | 50% |
| *Telephone* | 66% | 33% |
| *VDM* | 44% | 50% |
| *Flappy Bird* | 88% | 16% |
| *Youtube* | 22% | |

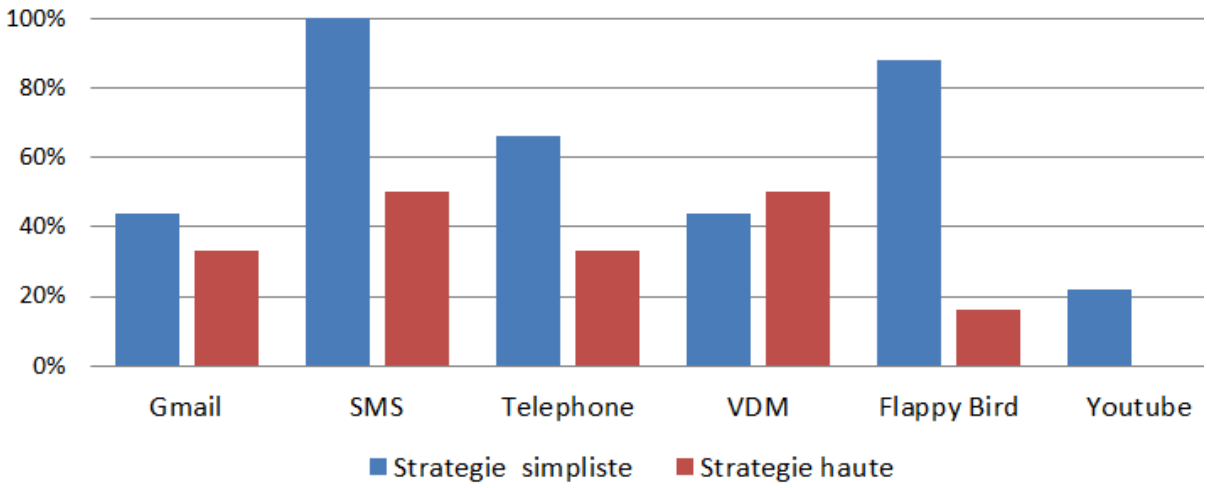



**Figure V-8 : Comparaisons des valeurs de poids conceptuel**

On constate également que la prise en compte de la fréquence d'utilisation peut modifier significativement le poids relatif des applications. C'est le cas notamment pour l'application *Flappy Bird*, dont le poids conceptuel est bien plus faible lorsque l'on tient compte de la fréquence. La Figure V-8 nous montre que, selon la stratégie à haute dépendance, les applications de plus fort poids conceptuel pour cet étudiant sont *SMS* et *VDM*, au lieu de *SMS* et *Flappy Bird* selon la stratégie simpliste. Les résultats les plus proches de la réalité sont ceux obtenus selon la stratégie à haute dépendance, car celle-ci tient compte de l'intensité des relations entre les applications et les éléments de contexte.

### V.5.1.b. **Comparaison de la similarité conceptuelle des applications**

La Figure V-9 et la Figure V-10 permettent de comparer la similarité conceptuelle des applications, selon la stratégie simpliste et selon la stratégie à haute dépendance.

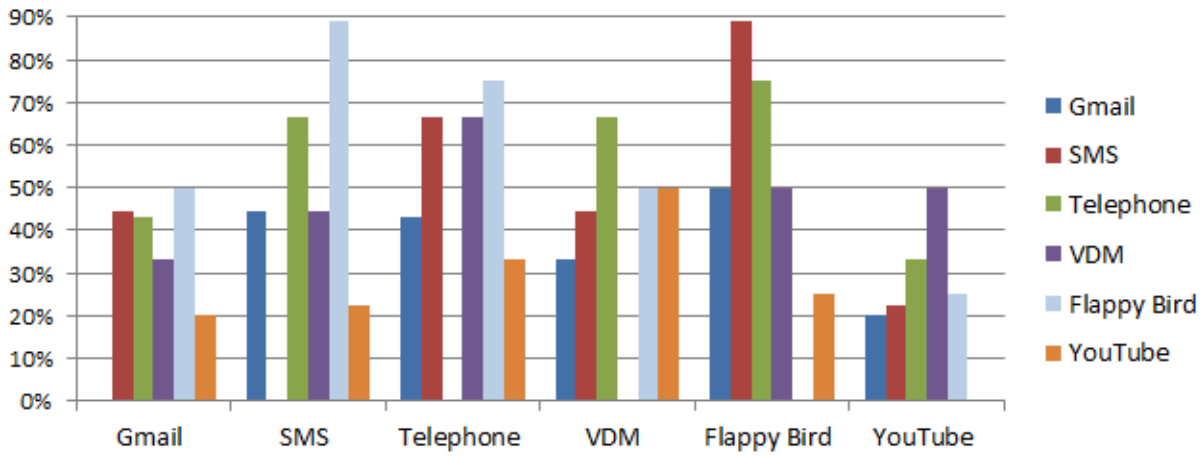



**Figure V-9 : Similarité conceptuelle des applications selon la stratégie simpliste (indépendante de la fréquence)**

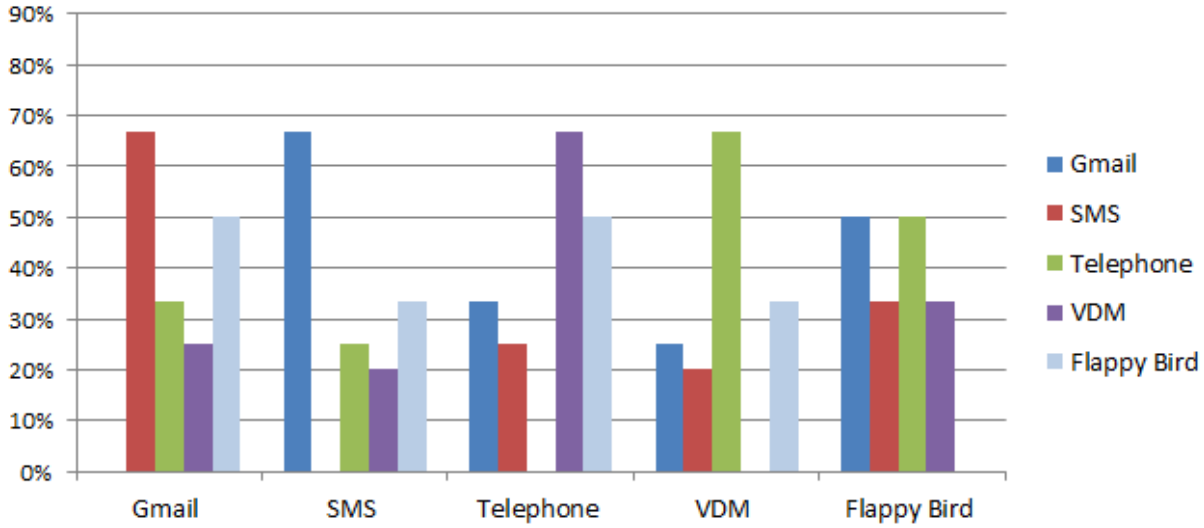



**Figure V-10 : Similarité conceptuelle des applications selon la stratégie à haute dépendance (dépendante de la fréquence)**

Dans la Figure V-9 et la Figure V-10, nous avons remarqué que, selon la stratégie simpliste, l'application *SMS* est très similaire à *Telephone* et à *Flappy Bird*. Ceci est contredit

par la stratégie à haute dépendance ; en effet, *SMS* et *Telephone* n'apparaissent pas aussi similaires lorsque l'on tient compte de la fréquence. Le même phénomène se produit pour *SMS* et *Flappy Bird*.

La Figure V-11 et la Figure V-12 fournissent une autre représentation de la similarité conceptuelle entre les applications, selon la stratégie simpliste et selon la stratégie à haute dépendance. Cette représentation a l'avantage de faciliter l'identification des applications les plus similaires aux autres, selon les deux stratégies.

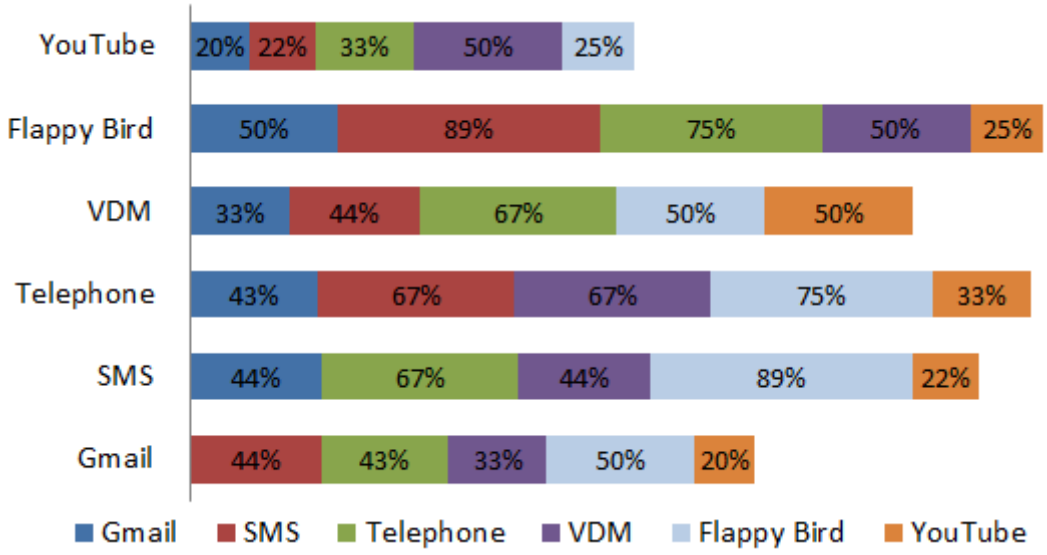



**Figure V-11 : Similarité conceptuelle des applications selon la stratégie simpliste (indépendante de la fréquence)**

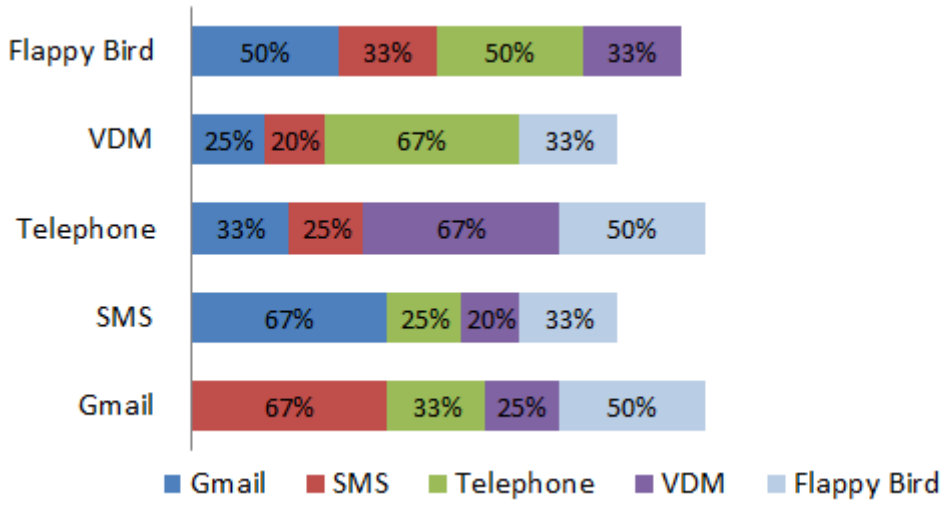


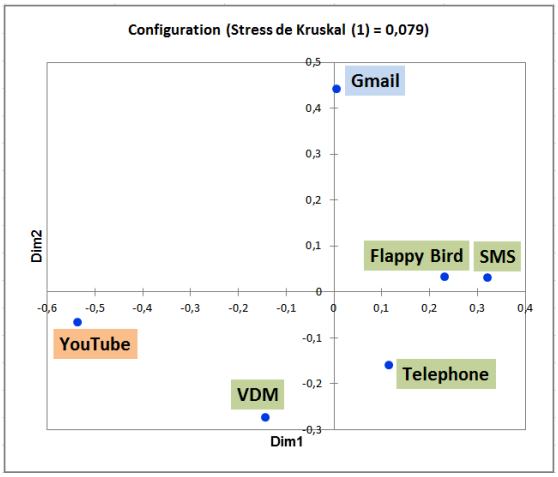
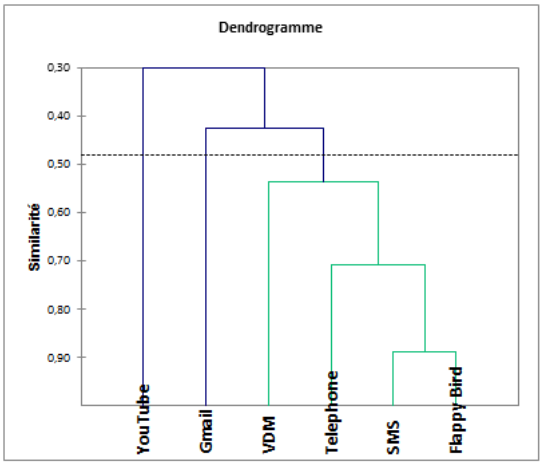
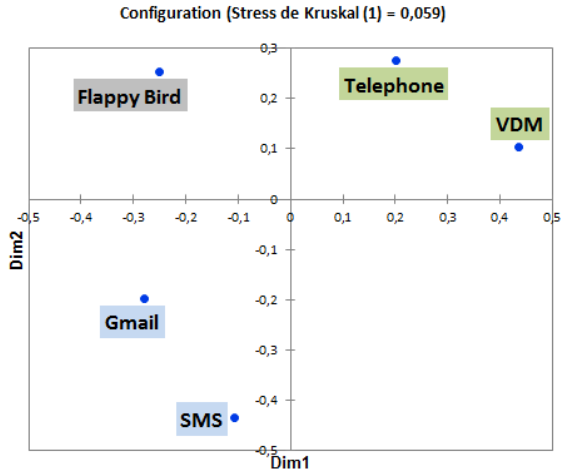
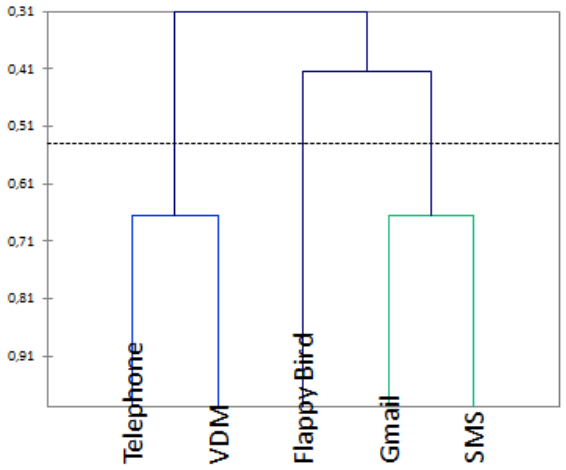

**Figure V-12 : Similarité conceptuelle des applications selon la stratégie à haute dépendance (dépendante de la fréquence)**

Les applications les plus similaires dans le contexte de forte utilisation sont *SMS* et *Gmail*, ainsi que *Telephone* et *VDM*.

On voit que *Youtube*, qui était l'application globalement la moins similaire aux autres dans la stratégie simpliste, a disparu dans la stratégie à haute dépendance. Par contre, *Gmail*, qui était aussi globalement moins similaire aux autres applications, apparait un peu plus similaire dans la deuxième approche tenant compte des fréquences.

Le Tableau V-13 montre la troisième représentation permettant de comparer la similarité conceptuelle des applications selon les deux stratégies, sous forme de carte 2D enrichie des clusters découverts par la Classification Ascendante Hiérarchique.

**Tableau V-13 : Comparaison des cartes MDS et des clusters déduits de la similarité conceptuelle des applications pour la stratégie simpliste et la stratégie à haute dépendance**

| **Strategie simpliste( indépendante de fréquence)** | |
|---|---|
| **MDS** | **CAH** |
|  |  |
| **Strategie à haute dépendance (dépendante de fréquence)** | |
| **MDS** | **CAH** |
|  |  |

Même si cette représentation est moins fiable, puisqu'elle introduit un facteur d'erreur, elle est très intuitive et permet de confirmer les observations obtenues à partir des précédentes représentations.

Nous rappelons que les couleurs qui apparaissent sur la carte MDS correspondent aux clusters identifiés avec l'algorithme de CAH.

Nous avons remarqué que, selon la stratégie simpliste, l'application *SMS* appartient à un cluster qui contient aussi *Telephone* et *Flappy Bird*, alors que ces applications appartiennent à trois clusters différents selon la stratégie à haute dépendance.

### V.5.1.c. **Comparaison de la similarité conceptuelle des applications**

La Figure V-13 et la Figure V-14 permettent de comparer la similarité conceptuelle de tous les éléments de contexte (attributs) du treillis, selon la stratégie simpliste et selon la stratégie à haute dépendance.

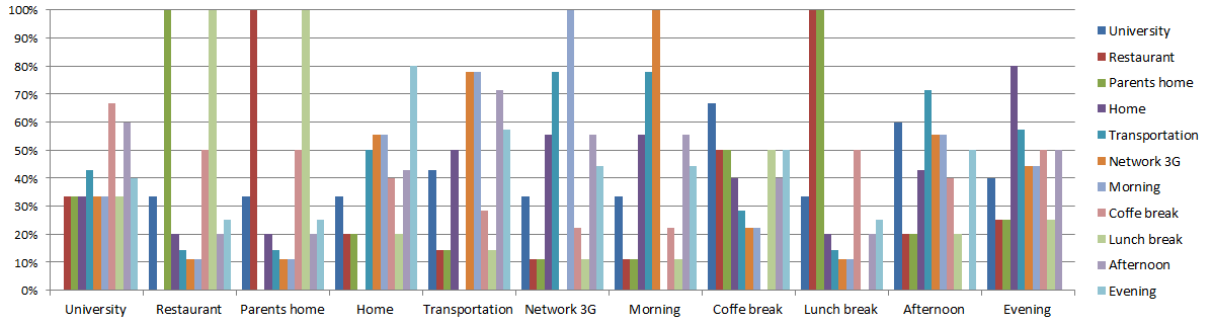



**Figure V-13 : Similarité conceptuelle des éléments de contexte selon la stratégie simpliste (indépendante de la fréquence)**

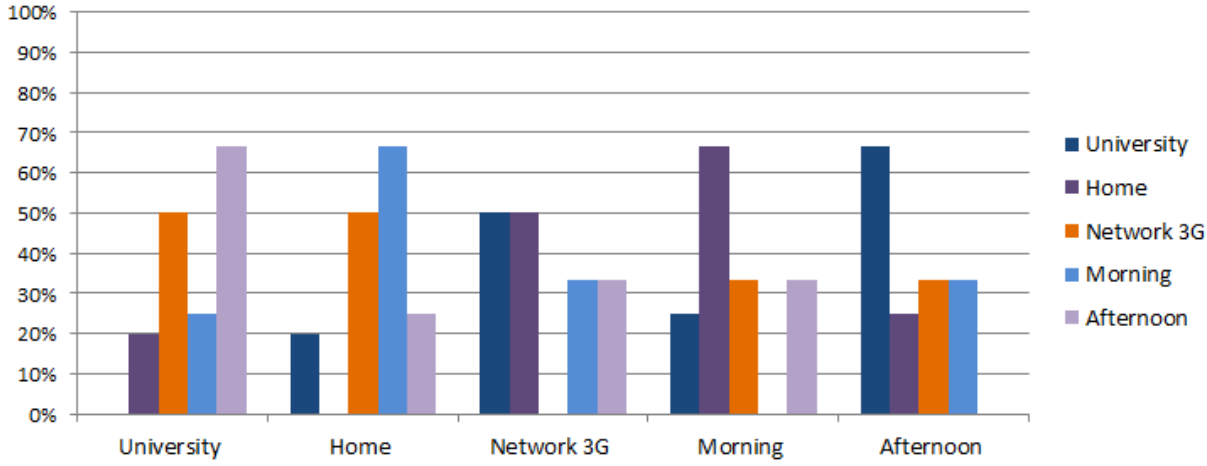



**Figure V-14 : Similarité conceptuelle des éléments de contexte selon la stratégie à haute dépendance (dépendante de la fréquence)**

L'interprétation de la représentation des résultats pour la stratégie simpliste, fournie par la Figure V-13, est difficile car il y a beaucoup d'éléments de contexte. La représentation sous forme de carte 2D, enrichie des clusters, est beaucoup plus accessible dans ce cas.

On peut néanmoins déjà constater que plusieurs éléments de contexte ont disparu dans la stratégie à haute dépendance : *restaurant, parents' home, transportation, coffee break, lunch break* et *evening*.

La Figure V-15 et la Figure V-16 fournissent une autre représentation de la similarité conceptuelle entre les éléments de contexte, selon les deux stratégies : simpliste et à haute dépendance.

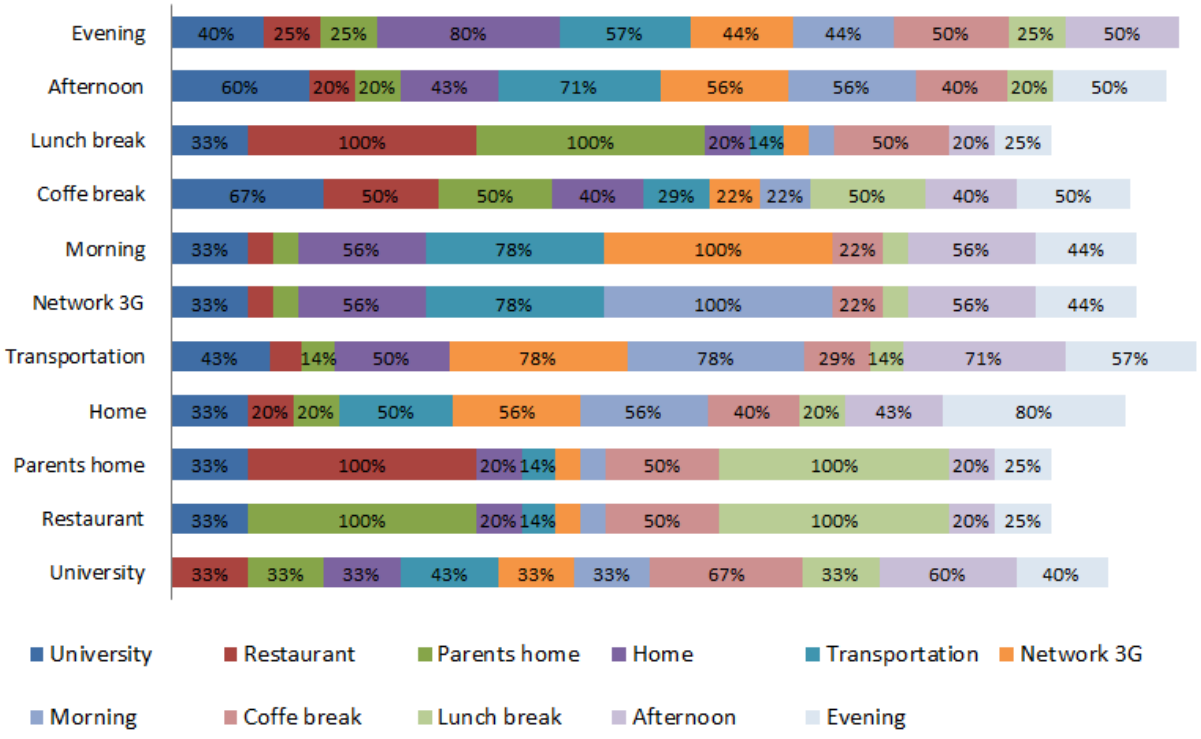



**Figure V-15 : Similarité conceptuelle des éléments de contexte selon la stratégie simpliste (indépendante de la fréquence)**

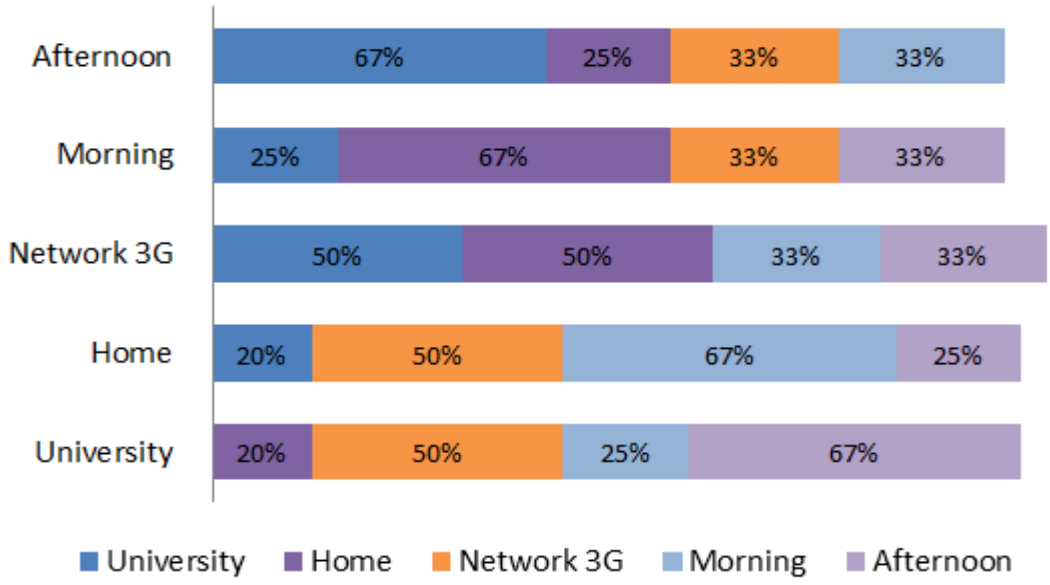



**Figure V-16 : Similarité conceptuelle des éléments de contexte selon la stratégie à haute dépendance (dépendante de la fréquence)**

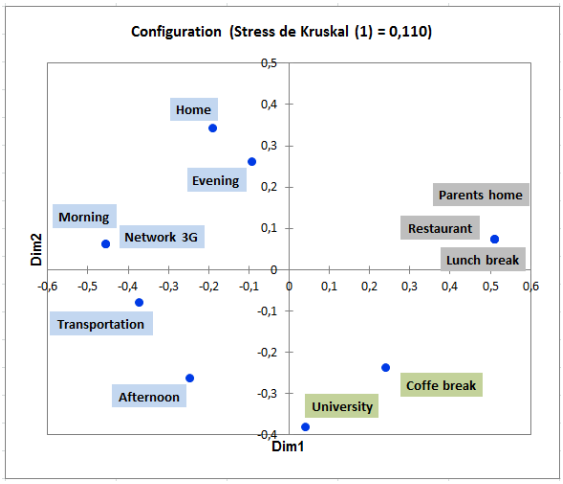
Peu de nouvelles informations sont tirées de cette représentation, du fait du grand nombre d'éléments de contexte selon la stratégie simpliste.

Le Tableau V-14 montre la troisième représentation, sous forme de carte 2D, que nous espérons plus facile à interpréter.

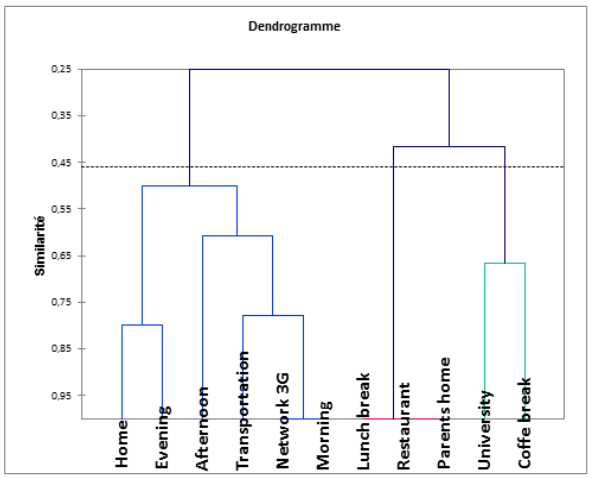
**Tableau V-14 : Comparaison des cartes MDS et des clusters déduits de la similarité conceptuelle des élements de contexte pour la stratégie simpliste et la stratégie à haute dépendance**

**Strategie simpliste( indépendante de fréquence)**

**MDS**

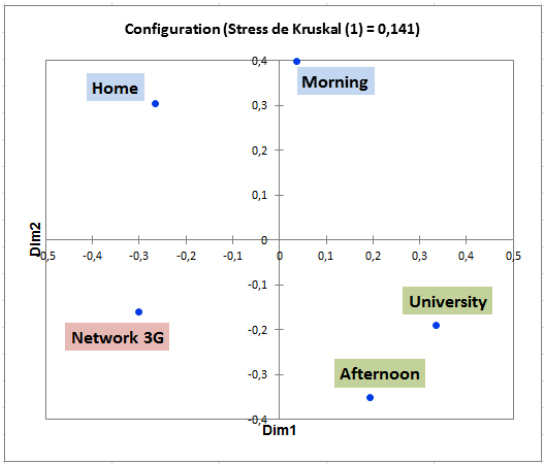



**CAH**

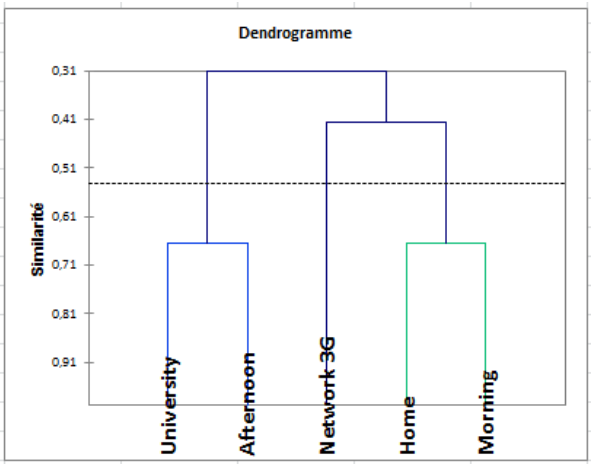



**Strategie à haute dépendance (dépendante de fréquence)**

**MDS**




**CAH**




D'après le Tableau V-14, selon la stratégie simpliste les éléments de contexte *Morning* et *Network* appartenaient au même cluster, alors qu'ils sont dans des clusters différents selon la stratégie à haute dépendance. Le phénomène inverse se produit pour *University* et *Afternoon*, qui n'étaient pas dans le même cluster selon la stratégie simpliste et qui sont réunis selon la stratégie à haute dépendance.

Cela confirme ce que nous avions déjà observé en étudiant les applications : il est important de tenir compte de la fréquence d'utilisation des applications selon le contexte, car des différences importantes apparaissent par rapport à la stratégie simpliste.

Dans la section suivante, nous comparons les résultats obtenus en termes d'impact mutuel selon les deux stratégies (simpliste et à haute dépendance).

### V.5.1.d. **Comparaison de l'impact mutuel**

La Figure V-17 et la Figure V-18 représentent l'impact mutuel entre les applications et les éléments de contexte selon les deux stratégies.

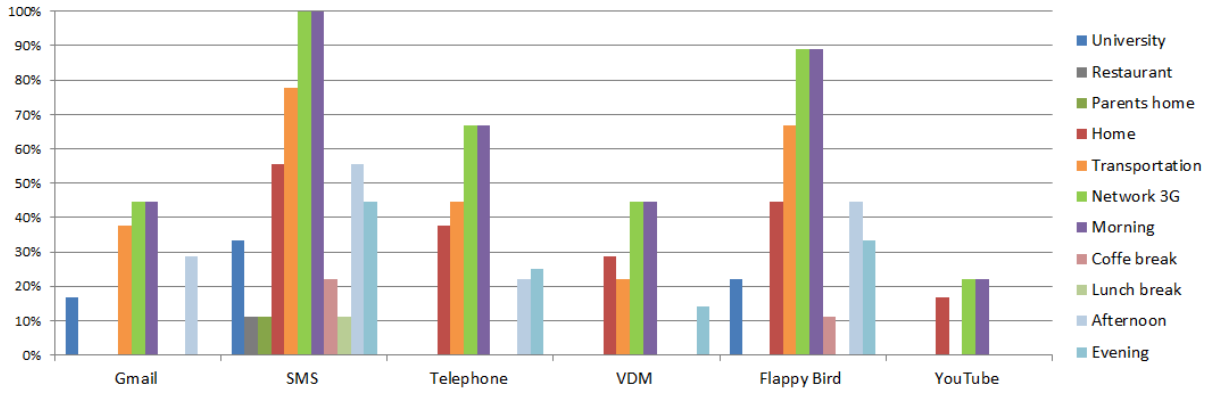



**Figure V-17 : Impact mutuel selon la stratégie simpliste (indépendante de la fréquence)**

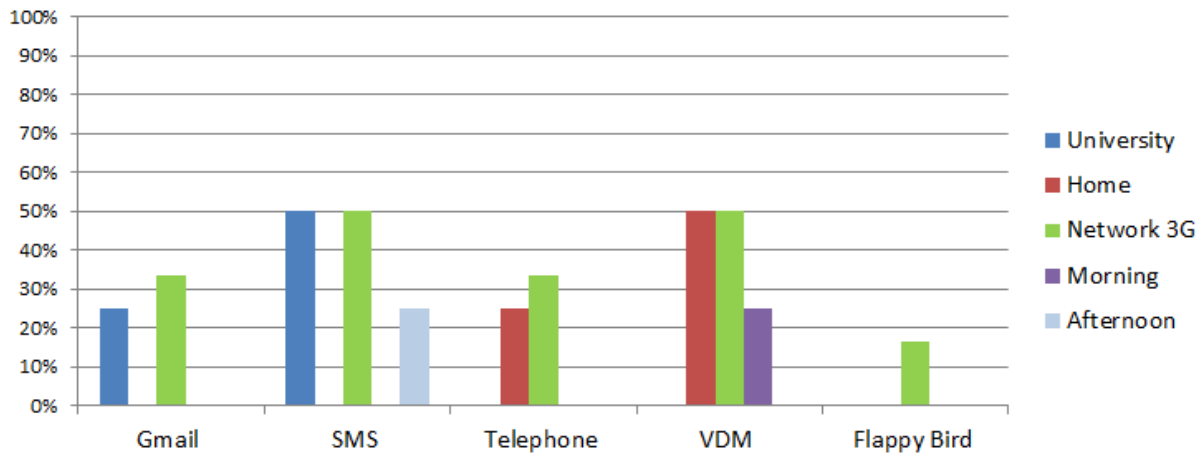



**Figure V-18 : Impact mutuel selon la stratégie à haute dépendance (dépendante de la fréquence)**

Ces figures montrent bien la disparition d'une application et de 6 éléments de contexte quand on passe de la stratégie simpliste à la stratégie à haute dépendance. On constate également que l'impact mutuel entre *Network 3G* et *SMS* est fortement réduit, alors que celui entre *SMS* et *university* augmente, tout comme l'impact mutuel entre *VDM* et *home*.

La Figure V-19 et la Figure V-20 fournissent une représentation de l'impact mutuel entre les applications et les éléments de contexte sous la forme d'un graphe, généré avec l'application Gephi. Sur ces graphes, la taille des nœuds est liée au degré pondéré par la valeur de l'impact mutuel. L'épaisseur des arcs représente également la valeur de l'impact mutuel entre les applications et les éléments de contexte. L'algorithme de spatialisation choisi est Force Atlas.

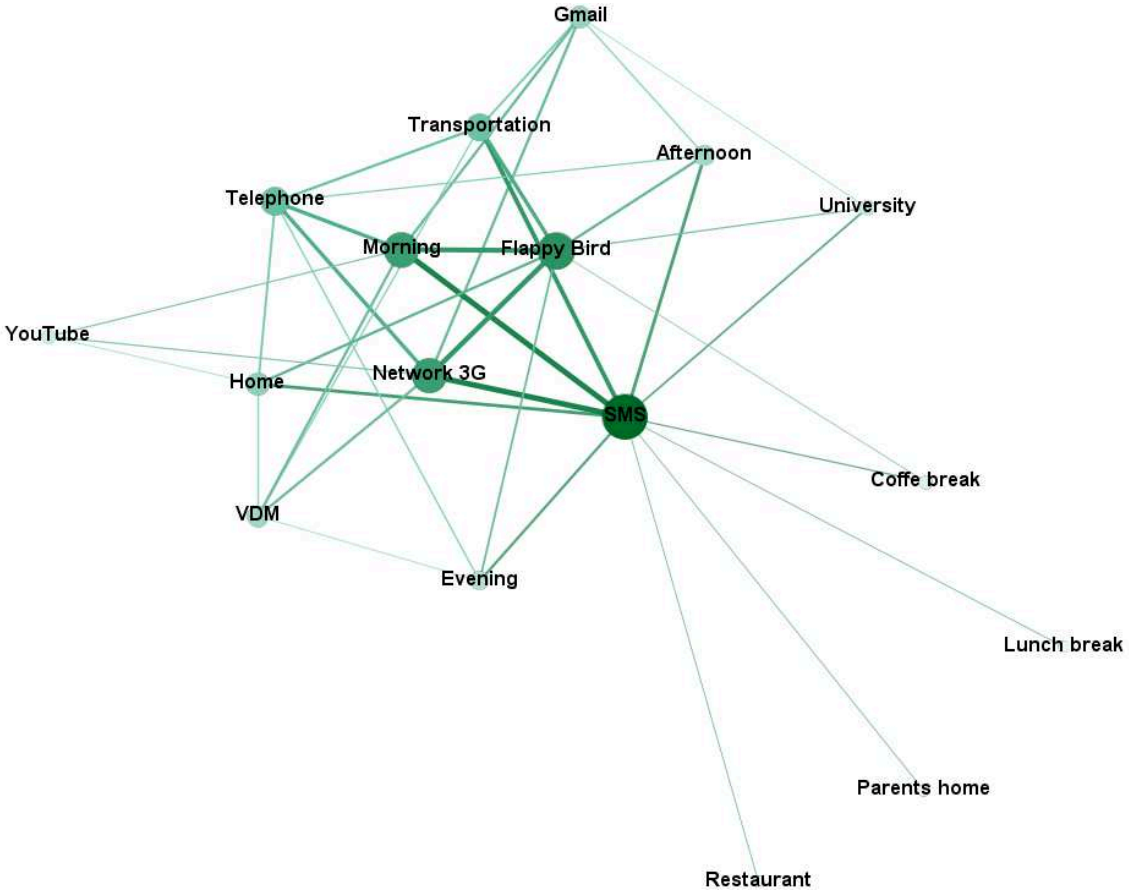



**Figure V-19 : Impact mutuel selon la stratégie simpliste (indépendante de la fréquence)**

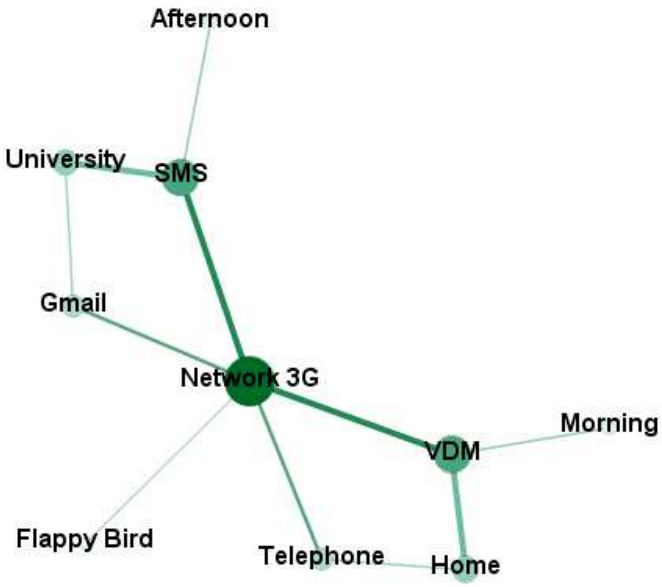



**Figure V-20 : Impact mutuel selon la stratégie à haute dépendance (dépendante de la fréquence)**

Ces graphes permettent de confirmer les observations faites à partir des précédentes représentations : on constate par exemple que *SMS* et *University* sont plus proches selon la stratégie à haute dépendance.

On voit également que la plupart des applications et des éléments de contexte qui disparaissent selon la stratégie haute n'étaient pas « centraux », à l'exception de *transportation* et *Morning*, qui avaient une importance faussement élevée selon la stratégie simpliste. Dans la stratégie à haute dépendance, *transportation* a disparu, et l'impact de *Morning* a fortement diminué.

Inversement, *university* était bien plus éloigné qu'il n'aurait dû l'être, selon la stratégie simpliste. On voit aussi que l'application *Flappy Bird* devient un peu moins centrale, car elle n'a plus d'impact que sur *Network 3G* dans la stratégie à haute dépendance.

Nous avons comparé dans cette section les résultats des trois mesures que nous avons proposées, selon la stratégie à haute dépendance et selon la stratégie simpliste. Nous constatons dans tous les cas que la prise en compte de la fréquence (ici, sémantique) d'utilisation des applications selon le contexte a un impact non négligeable sur les résultats. Par conséquent, lorsque la relation entre les données n'est pas naturellement binaire, il est important de tenir compte de cette fréquence afin de ne pas tirer de conclusions erronées.

Dans le Chapitre III, nous avions comparé plusieurs treillis de Galois à l'aide de nos mesures, selon la stratégie simpliste. Nous avons refait ce travail selon la stratégie à haute dépendance, mais nous présentons les résultats en annexe A de ce manuscrit, car ils n'apportent pas d'information supplémentaire en termes de comparaison des deux stratégies.

### V.5.2. **Stratégie à faible dépendance**

Dans cette section, nous réalisons l'interprétation des résultats obtenus selon la stratégie à faible dépendance. L'intérêt de cette stratégie est de voir par exemple si les éléments de contexte similaires en termes d'applications fortement utilisées le sont aussi lorsqu'il s'agit des applications peu utilisées.

Ainsi, l'analyse de la similarité conceptuelle entre les applications sur la Figure V-21 montre que les applications *Telephone* et *VDM* sont présentes simultanément dans 50% des

concepts. Cela signifie que ces deux applications sont fortement similaires en termes de contexte dans lesquels elles sont faiblement utilisées. Ces conclusions peuvent aider un concepteur d'applications à faire évoluer les applications selon leur similarité d'usage.

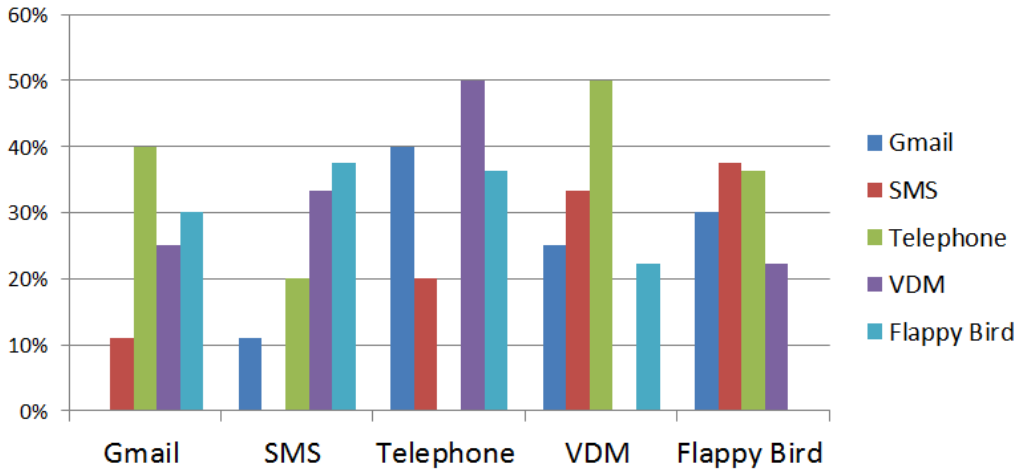



**Figure V-21 : 1ere représentation de la similarité conceptuelle des applications selon la stratégie à faible dépendance**

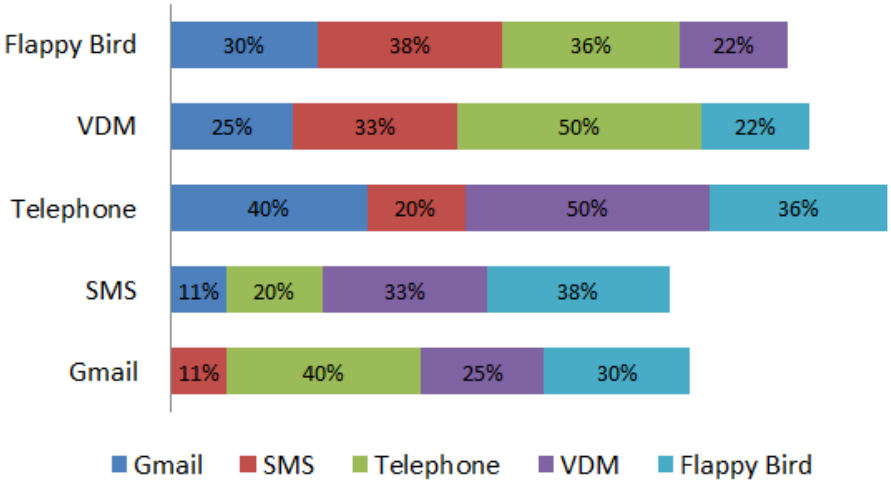
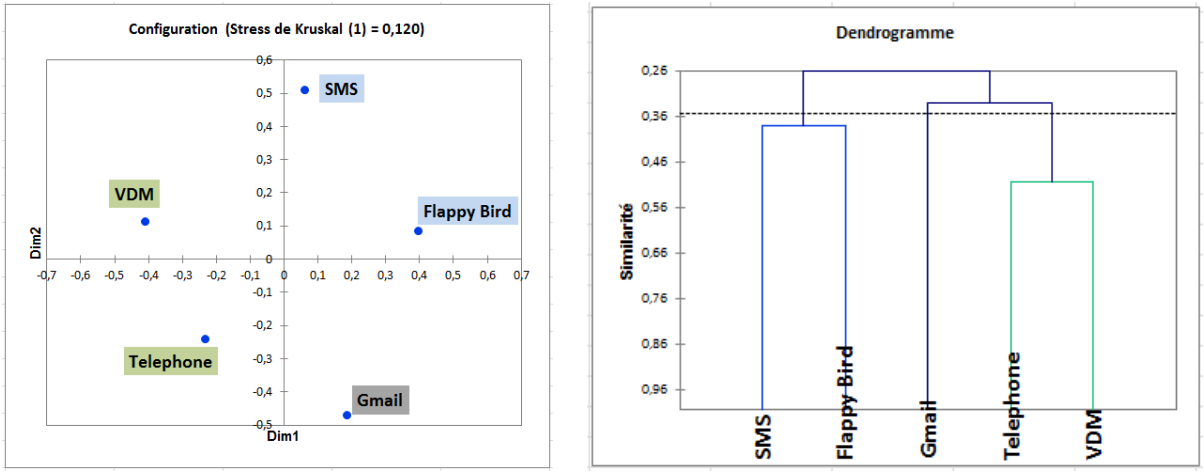



**Figure V-22 : 2e représentation de la similarité conceptuelle des applications selon la stratégie à faible dépendance**

On voit, en comparant les résultats de la stratégie à faible dépendance à ceux de la stratégie à haute dépendance de la section V.5.1.b, que *Telephone* et *VDM* restent très similaires dans leur contexte d'utilisation, que ce soit dans les cas où elles sont beaucoup

utilisées ou dans les cas où elles le sont peu. Par contre *Gmail* et *SMS*, très similaires dans leur utilisation fréquente, le sont beaucoup moins dans les contextes de faible utilisation. Ceci est confirmé par la carte MDS et les clusters du Tableau V-15, car *Telephone* et *VDM* restent dans le même cluster selon la stratégie à faible dépendance, alors que *Gmail* et *SMS* sont a présent séparés.

**Tableau V-15 : Carte MDS et clusters déduits de la similarité conceptuelle des applications selon la stratégie à faible dépendance**




La Figure V-23 et la Figure V-24 montrent la similarité conceptuelle des éléments de contexte du treillis de la Figure V-4 qui représente les relations à faible dépendance.

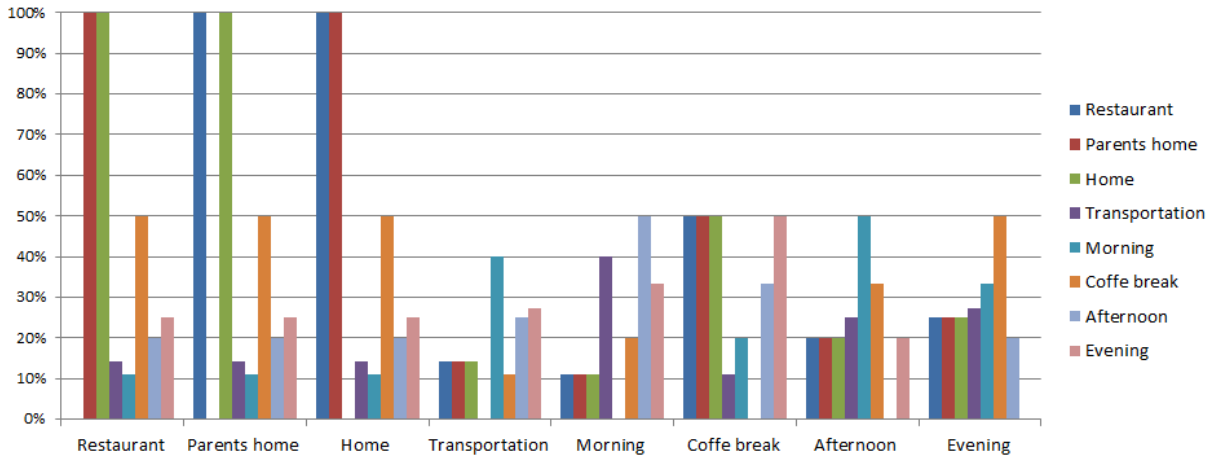



**Figure V-23 : 1ere représentation de la similarité conceptuelle des éléments de contexte selon la stratégie à faible dépendance**

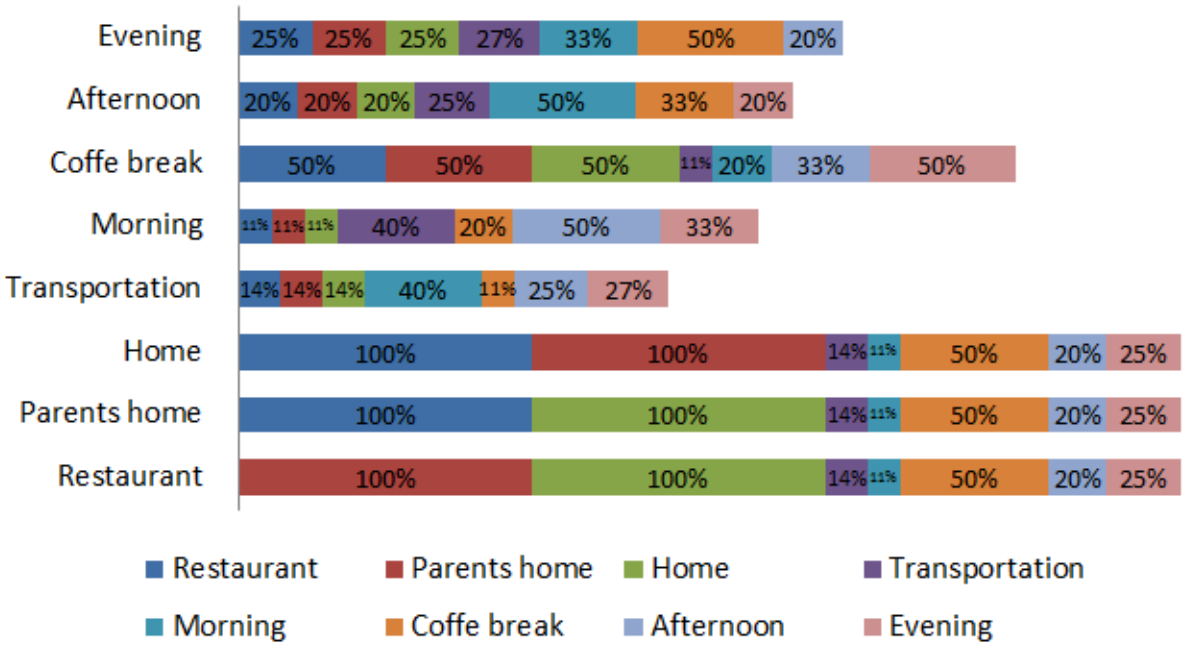


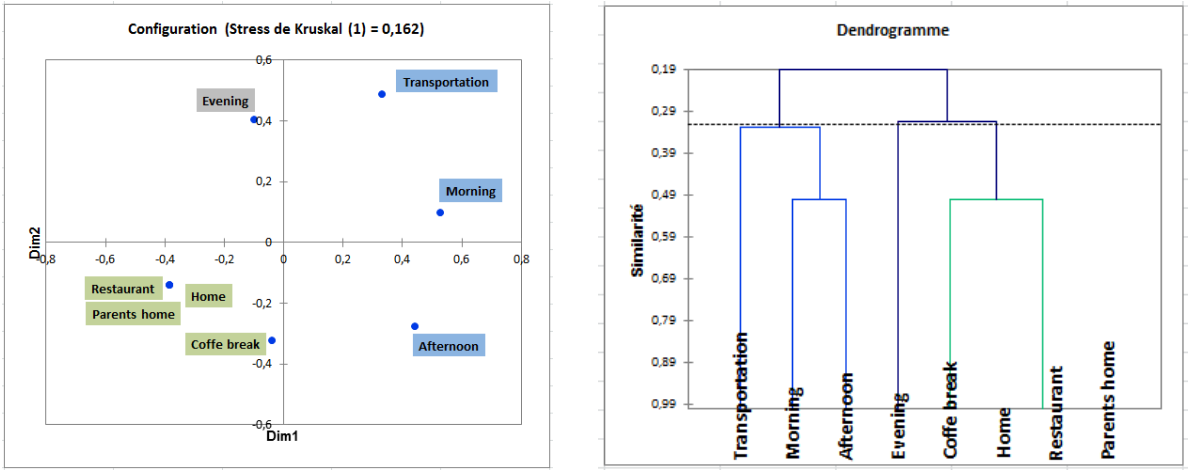

**Figure V-24 : 2e représentation de la similarité conceptuelle des éléments de contexte selon la stratégie à faible dépendance**

On observe que toutes les applications (faiblement) associées à l'élément de contexte *Restaurant* sont également (faiblement) associés à l'élément de contexte *Home* (Similarité conceptuelle = 100%) et à l'élément de contexte *parents' home* (similarité conceptuelle = 100%). Aucune conclusion n'avait pu être tirée sur ces trois éléments de contexte selon la stratégie à haute dépendance, car *Restaurant* et *Parents' home* avaient disparu du contexte formel.

Ainsi, la stratégie à faible dépendance permet de mettre en évidence la (forte) similarité conceptuelle entre des éléments de contexte faiblement associés à des applications.

La représentation sous forme de carte, fournie par le Tableau V-16, montre que les éléments de contexte *Morning* et *Home* sont éloignés dans la stratégie à faible dépendance, alors qu'ils appartenaient au même cluster selon la stratégie à haute dépendance ; on peut dire que *Morning* et *Home* sont fortement similaires en termes d'applications fréquemment utilisées, mais pas en termes d'applications peu utilisées.

**Tableau V-16 : Carte MDS et clusters déduits de la similarité conceptuelle des éléments de contexte selon la stratégie à faible dépendance**




La Figure V-25 représente l'impact mutuel calculé à partir du treillis de la Figure V-4, selon la stratégie à faible dépendance. Ceci représente l'impact mutuel entre les applications et les éléments de contexte auxquels elles sont faiblement associées (selon la stratégie à faible dépendance).

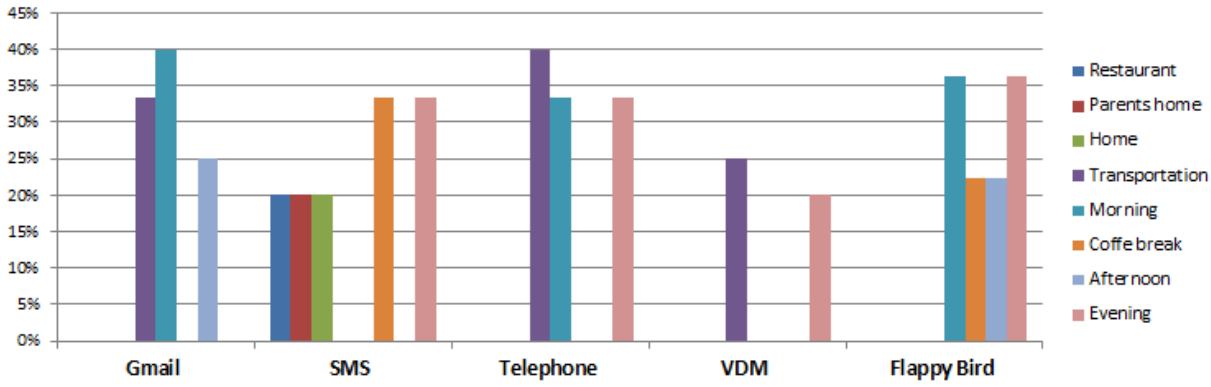



**Figure V-25 : Impact mutuel selon la stratégie à faible dépendance**

On observe que l'application *Gmail* est majoritairement associée à *Morning* avec un impact mutuel de 40%, à l'inverse de la stratégie à haute dépendance qui montre que l'impact mutuel entre *Gmail* et *Morning* est nul.

De la même manière, un fort impact mutuel est observé entre *Telephone* et *Transportation* selon la stratégie à faible dépendance, alors qu'aucune conclusion ne peut être

tirée selon la stratégie à haute dépendance car *Transportation* n'apparaît plus dans le contexte formel.

La Figure V-26 représente l'impact mutuel entre les applications et les éléments de contexte (où ces applications sont peu utilisées).

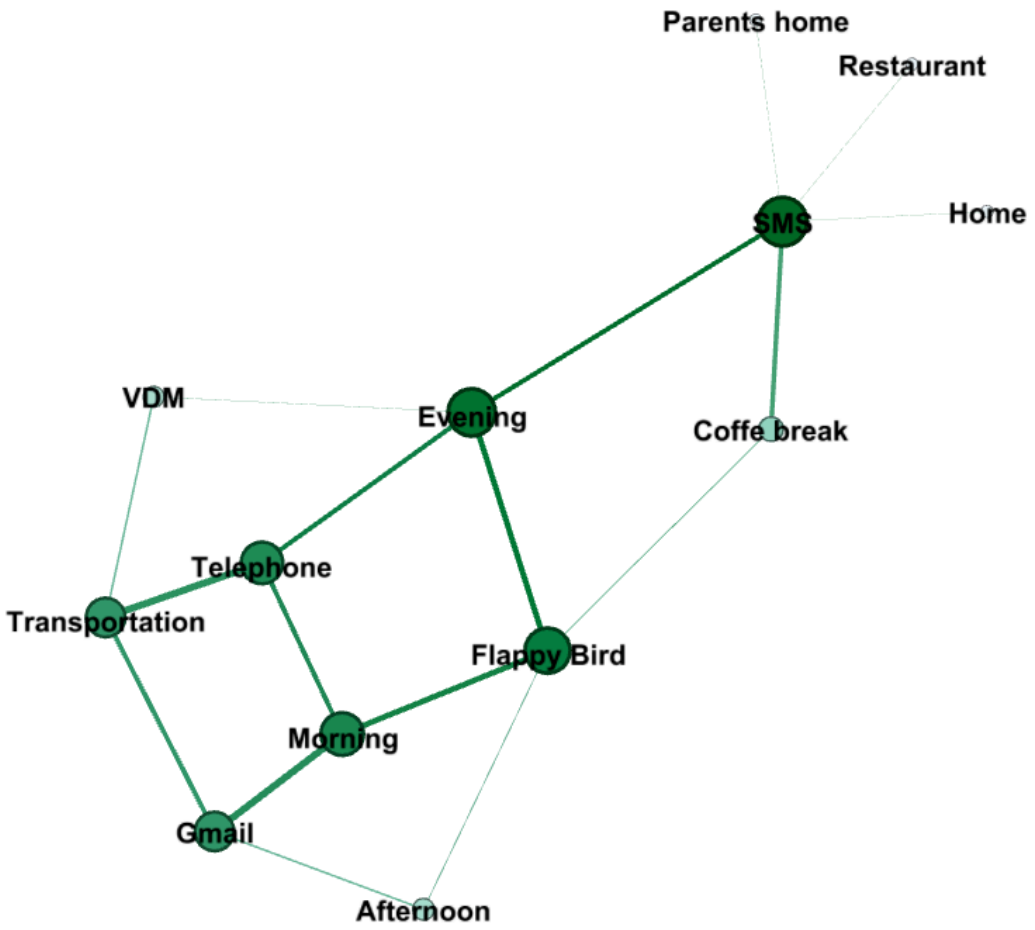



**Figure V-26 : Graphe d'impact mutuel selon la stratégie à faible dépendance**

Si l'on compare ce graphe avec celui de l'impact mutuel selon la stratégie à haute dépendance (Figure V-20), on voit que l'élément de contexte *Evening* réapparaît ici avec un impact élevé, alors qu'il n'était pas présent dans la stratégie à haute dépendance. La même observation peut être faite pour l'élément de contexte *Transportation*.

En conclusion, la stratégie à faible dépendance permet de tirer des conclusions complémentaires à celles que la stratégie à haute dépendance a permis d'obtenir. Cette stratégie peut être intéressante lorsque l'on veut chercher à comprendre finement l'usage non intensif des applications.

### V.5.3. **Stratégie inverse**

Nous procédons dans cette section à la même analyse que celle que nous avons effectuée pour la stratégie à faible dépendance. Il s'agit ici d'étudier les cas de non utilisation d'applications dans des contextes donnés. Ceci peut être intéressant par exemple pour détecter ce qu'un concepteur d'applications pourrait considérer comme des anomalies (i.e., des applications jamais utilisées dans des contextes pourtant propices a priori).

La Figure V-27, la Figure V-28 et le Tableau V-17 fournissent trois représentations de la similarité conceptuelle entre les applications selon la stratégie inverse.

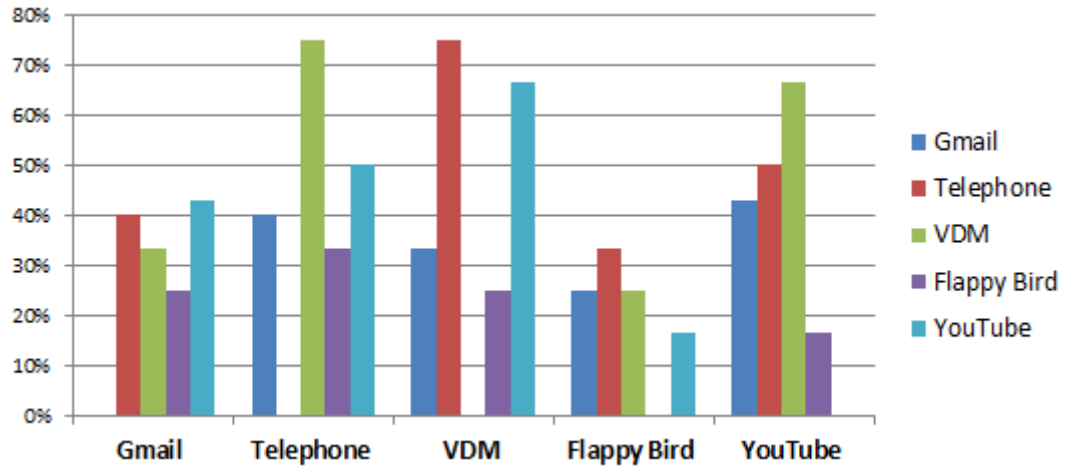



**Figure V-27 : 1ere représentation de la similarité conceptuelle des applications selon la stratégie inverse**

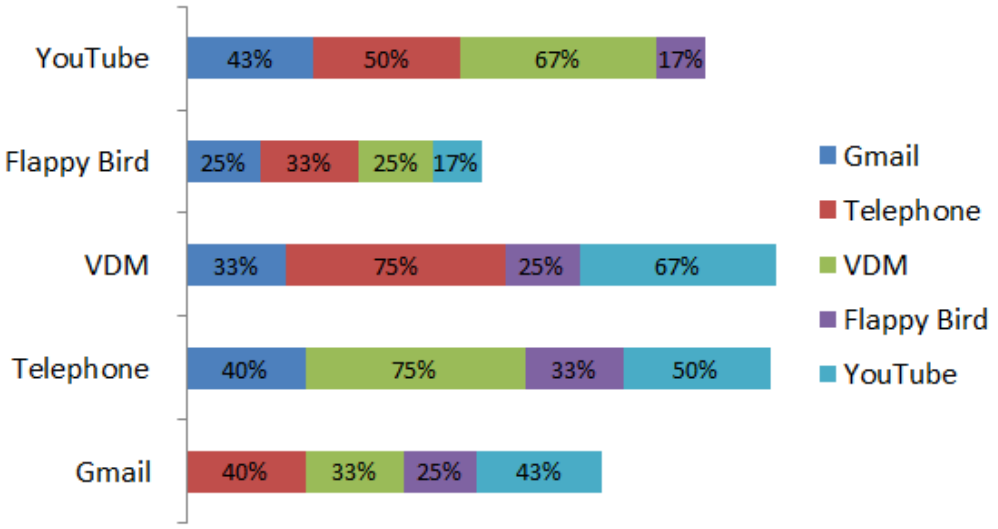
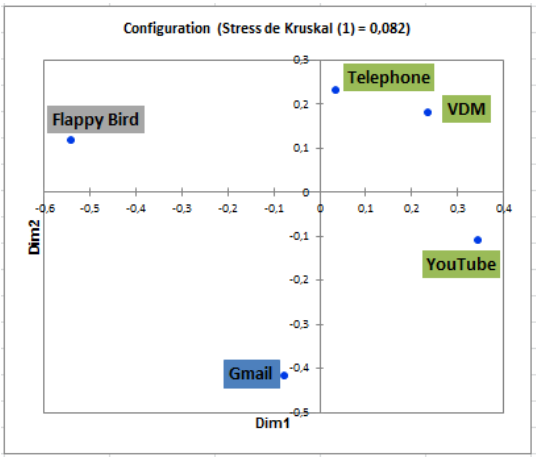
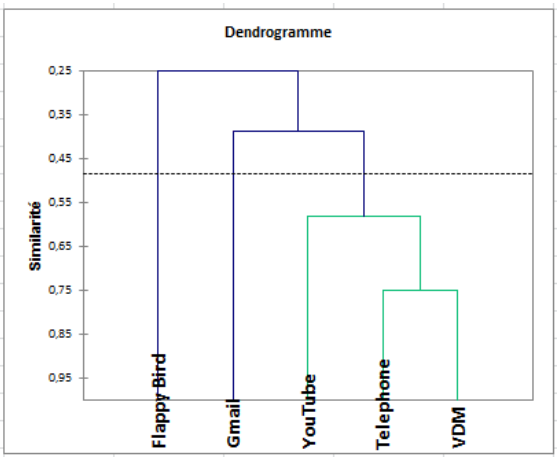



**Figure V-28 : 2e représentation de la similarité conceptuelle des applications selon la stratégie inverse**

**Tableau V-17 : Carte MDS et clusters déduits de la similarité conceptuelle des applications selon la stratégie inverse**




Ces résultats montrent par exemple que certaines applications sont similaires, à la fois dans leur contexte d'utilisation (stratégie à haute dépendance) que dans leur contexte de non utilisation. C'est le cas par exemple de *Telephone* et *VDM*.

A l'inverse, l'application *Youtube* est similaire à *VDM* en termes de contextes de non utilisation, alors que rien ne peut être conclu dans la stratégie à haute dépendance car elle n'est pas présente dans le contexte formel.

La Figure V-29, la Figure V-30 et le Tableau V-18 représentent la similarité conceptuelle des éléments de contexte selon la stratégie inverse.

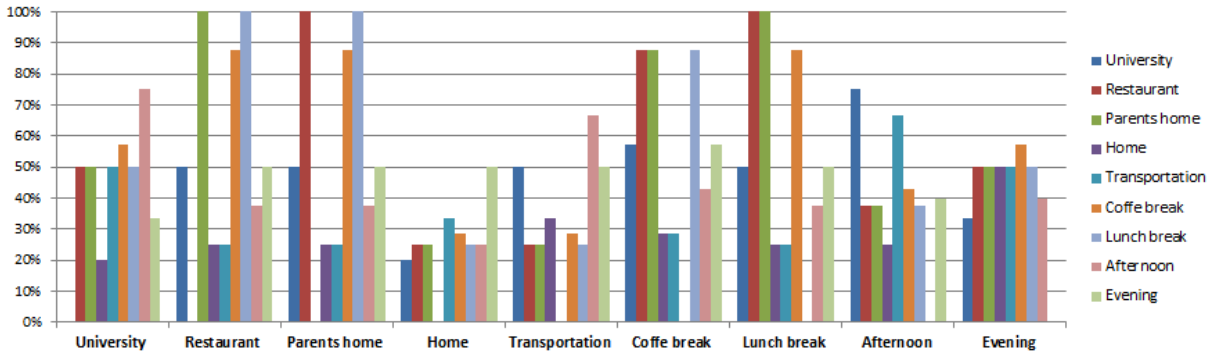



**Figure V-29 : 1ere représentation de la similarité conceptuelle des éléments de contexte selon la stratégie inverse**

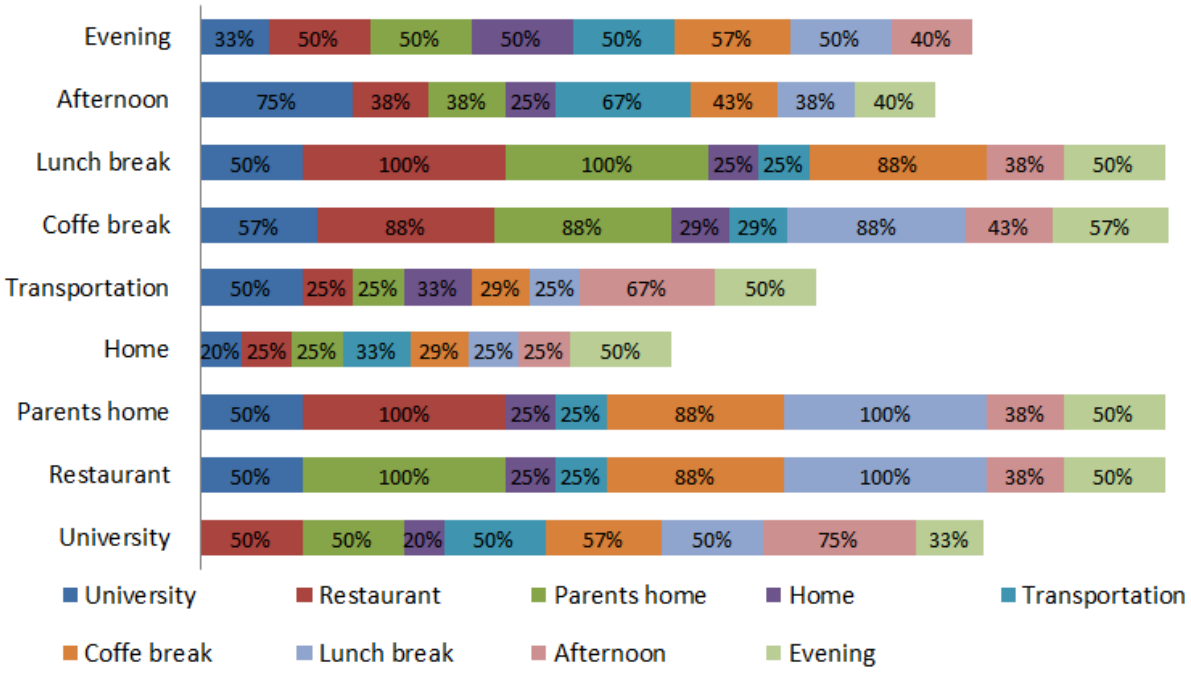
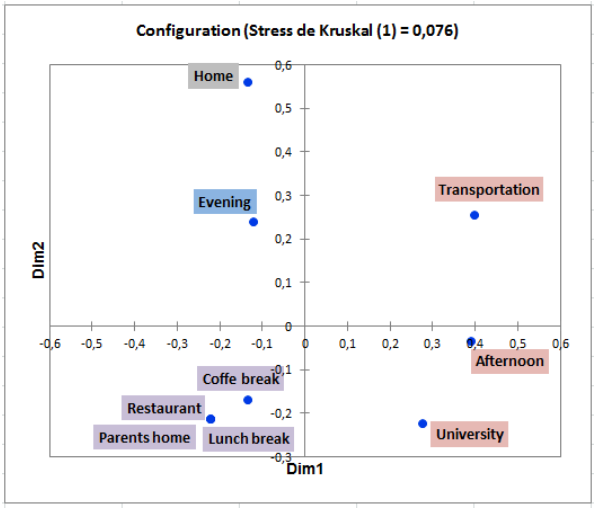
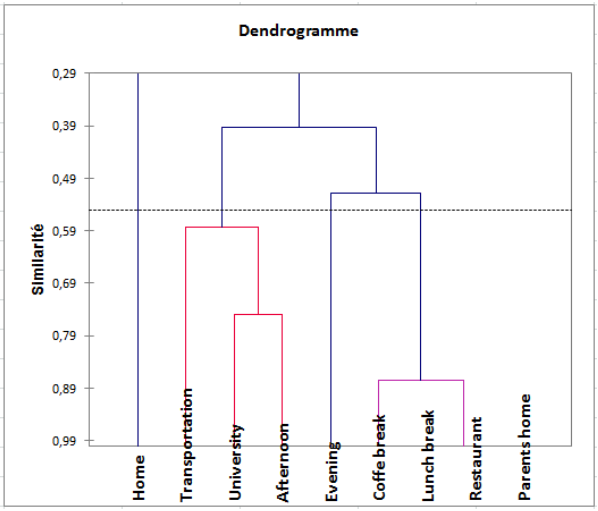



**Figure V-30 : 2e représentation de la similarité conceptuelle des éléments de contexte selon la stratégie inverse**

**Tableau V-18 : Carte MDS et clusters déduits de la similarité conceptuelle des éléments de contexte selon la stratégie inverse**




Ces résultats nous montrent que ce sont les mêmes applications qui ne sont jamais utilisées au restaurant, au domicile des parents et durant la pause déjeuner. (On avait obtenu un résultat équivalent pour les applications peu utilisées entre *Restaurant* et *Parents' home*).

On constate également que les éléments de contexte *Afternoon* et *University* restent similaires aussi bien en termes d'applications jamais utilisées qu'en termes d'applications fréquemment utilisées. Par contre, certains éléments de contexte apparaissent comme très similaires en termes d'applications non utilisées, alors qu'ils n'apparaissent pas dans le contexte formel de la stratégie à haute dépendance.

La Figure V-32 et la Figure V-32 représentent l'impact mutuel entre les applications et les éléments de contexte dans lesquels elles ne sont jamais utilisées.

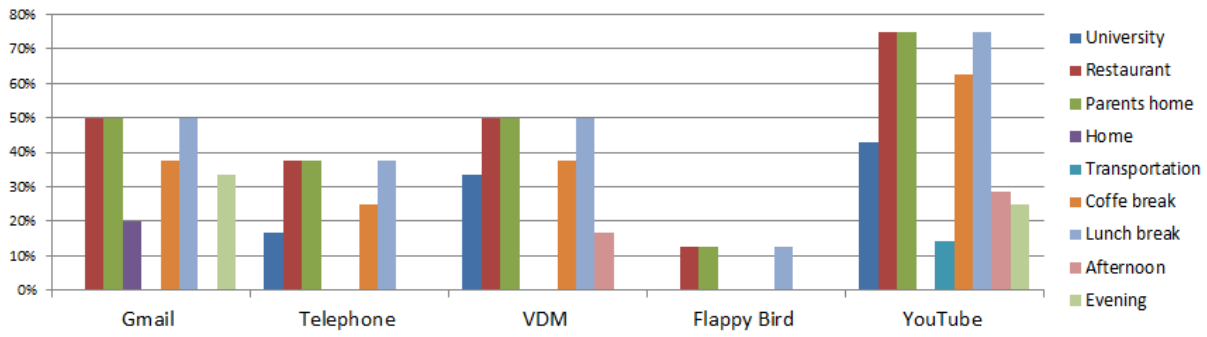



**Figure V-31 : Impact mutuel selon la stratégie inverse**

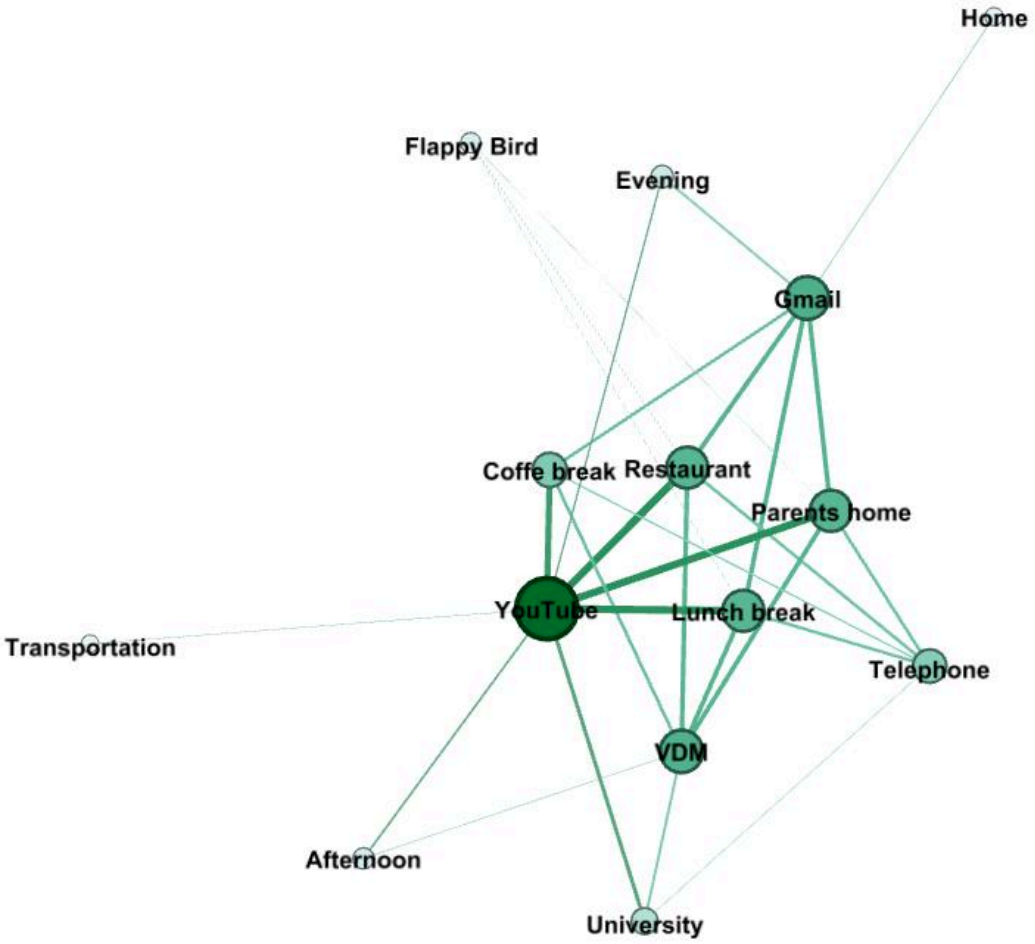



**Figure V-32 : Graphe d'impact mutuel selon la stratégie inverse**

On observe sur la Figure V-32 que les nœuds centraux sur le graphe correspondent aux nœuds périphériques sur le graphe de la Figure V-20 correspondant à la stratégie à haute dépendance (et réciproquement). De même, des nœuds de plus grande taille deviennent les plus petits et inversement.

On constate notamment ici un très fort impact mutuel entre *Youtube* et plusieurs éléments de contexte (*Restaurant*, *Parents' home* et *Lunch break*), non présents dans la stratégie à haute dépendance car ni cette application, ni ces trois éléments de contexte ne sont présents dans le contexte formel associé.

De la même manière, les impacts mutuels les plus élevés dans la stratégie à haute dépendance correspondent à des applications et éléments de contexte qui n'apparaissent pas dans le contexte formel de la stratégie inverse (par exemple *SMS*, *Network 3G*).

Dans cette section, nous avons détaillé l'interprétation des treillis de Galois obtenus selon différentes stratégies : à haute dépendance (comparée avec la stratégie simpliste du Chapitre III), à faible dépendance, et stratégie inverse. Pour les stratégies à haute et à faible dépendance, reposant sur une mesure de fréquence, nous avons ici utilisé la fréquence sémantique présentée dans la section V.3.1.

Afin de compléter notre étude, nous souhaitons évaluer la valeur ajoutée de l'utilisation d'une fréquence sémantique par rapport à une fréquence non sémantique ; c'est l'objet de la section suivante. Néanmoins, afin de ne pas alourdir ce mémoire nous nous focalisons pour cela sur une application particulière : *Gmail.*

### V.5.4. **Focus sur l'application *Gmail* pour l'étude de l'impact du choix de la fréquence**

Dans cette section, nous analysons plus spécifiquement les résultats concernant l'application de messagerie *Gmail*. Tout d'abord, nous calculons sa fréquence d'utilisation (Tableau V-19) en appliquant les deux mesures de fréquence définies dans la section V.3.1 : la fréquence sémantique et la fréquence non sémantique.

**Tableau V-19 : Fréquence d'utilisation de l'application *Gmail***

| | Localisation | | Réseaux | Temps | |
|---|---|---|---|---|---|
| | Univ | Transport | Network 3G | Morning | Afternoon |
| $f_s$ | 66 % | 33 % | 100 % | 50 % | 50 % |
| $f_{ns}$ | 28 % | 14 % | 28 % | 14 % | 14 % |

Dans le Tableau V-20, nous appliquons le calcul des poids conceptuels de l'application *Gmail* selon les différentes stratégies. Pour les stratégies dépendantes de la fréquence nous calculons les valeurs des poids avec et sans considérer les relations sémantiques (resp. $f_s$ et $f_{ns}$) entre les éléments de contexte. Encore une fois, nous calculons les seuils $S_h$ et $S_b$ en utilisant $\beta = 0{,}25$.

**Tableau V-20 : Poids conceptuel de l'application *Gmail* selon les différentes stratégies**

| Stratégie simpliste | Stratégie inverse | |
|---|---|---|
| 44 % | 50 % | |
| En considérant les relations sémantiques entre les éléments de contexte | | |
| Stratégie haute | Stratégie moyenne | Stratégie basse |
| 33 % | Disparu ! | 42 % |
| Sans considérer les relations sémantiques entre les éléments de contexte | | |
| Stratégie haute | Stratégie moyenne | Stratégie basse |
| 33 % | Disparu ! | 44 % |

Nous remarquons que l'application *Gmail* a disparu dans la stratégie à dépendance moyenne, c'est-à-dire que nous ne trouvons plus cette application dans le treillis de Galois résultant. Dans le cas de la stratégie à haute dépendance, nous remarquons que le poids conceptuel de l'application *Gmail* a une valeur de 33% avec les deux types de fréquence (sémantique ou non). On observe par contre une légère différence du poids de *Gmail* dans le cas de la stratégie à faible dépendance selon le type de fréquence utilisé (42% versus 44%). La prise en compte de la sémantique dans le calcul de la fréquence a donc un impact faible sur le poids conceptuel dans l'exemple étudié ici.

Nous avons également calculé la similarité conceptuelle entre l'application *Gmail* et les autres applications. Le Tableau V-21 montre les résultats du calcul de similarité conceptuelle en utilisant les différentes stratégies proposées pour la construction du contexte formel. L'objectif est surtout ici de comparer les résultats selon que la sémantique est prise en compte ou non dans le calcul de la fréquence.

**Tableau V-21 : Similarité conceptuelle de *Gmail* avec les autres applications selon les différentes stratégies**

| Stratégie simpliste | | | | | Stratégie inverse | | | |
|---|---|---|---|---|---|---|---|---|
| SMS | Telephone | VDM | Flappy Bird | Youtube | Telephone | VDM | Flappy Bird | Youtube |
| 44 % | 42 % | 33 % | 50 % | 20 % | 40 % | 33 % | 25 % | 42 % |
| **En considérant les relations sémantiques entre les éléments de contexte** | | | | | | | | |
| Stratégie haute | | | | | Stratégie basse | | | |
| SMS | Telephone | VDM | Flappy Bird | | SMS | Telephone | VDM | Flappy Bird |
| 67 % | 33 % | 25 % | 50 % | | 11 % | 40 % | 25 % | 30 % |
| **Sans considérer les relations sémantiques entre les éléments de contexte** | | | | | | | | |
| Stratégie haute | | | | | Stratégie basse | | | |
| SMS | Telephone | VDM | | | SMS | Telephone | VDM | |
| 66 % | 33 % | 25 % | | | 16 % | 33 % | 33 % | |

On constate que, pour la stratégie à haute dépendance, la similarité conceptuelle entre *Gmail* et *SMS*, *Telephone* et *VDM* est exactement la même avec les deux fréquences. Par contre, la forte similarité conceptuelle entre *Gmail* et *Flappy Bird* n'apparaît pas avec la fréquence non sémantique car l'application *Flappy Bird* a disparu du contexte formel.

De la même manière, la similarité conceptuelle entre *Gmail* et *SMS, Telephone* et *VDM* est assez proche quelle que soit la fréquence, selon la stratégie à faible dépendance, mais là aussi, l'application *Flappy Bird* disparaît avec la fréquence non sémantique.

On peut donc conclure que l'utilisation de la fréquence non sémantique permet de trouver l'essentiel des résultats obtenus avec la fréquence sémantique, mais qu'il manque

certaines informations. Par contre, dans tous les cas la fréquence non sémantique fournit des résultats plus pertinents que la stratégie simpliste, qui n'utilise aucune mesure de fréquence.

Nous étudions ci-dessous si cette conclusion est également vraie en ce qui concerne le calcul de l'impact mutuel.

Le Tableau V-22 indique les valeurs d'impact mutuel entre l'application *Gmail* et les différents éléments de contexte selon les différentes stratégies et notamment avec les deux types de fréquences.

**Tableau V-22 : Impact mutuel entre *Gmail* et les éléments de contexte selon les différentes stratégies**

| Stratégie simpliste | | | | | Stratégie inverse | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Univ | Transport | Network 3G | Morning | Afternoon | Restaurant | Parents Home | Home | Coffe break | Lunch break | Evening |
| 17 % | 37 % | 44 % | 44 % | 28 % | 50 % | 50 % | 20 % | 37 % | 50 % | 33 % |
| **Avec les fréquences sémantiques** | | | | | | | | | | |
| Stratégie haute | | | | | Stratégie basse | | | | | |
| Univ | | Network 3G | | | Transport | | Morning | | Afternoon | |
| 25 % | | 33 % | | | 33 % | | 40 % | | 25 % | |
| **Avec les fréquences non sémantiques** | | | | | | | | | | |
| Stratégie haute | | | | | Stratégie basse | | | | | |
| Univ | | Network 3G | | | Transport | | Morning | | Afternoon | |
| 25 % | | 33 % | | | 33 % | | 33 % | | 20 % | |

Nous observons là aussi que les résultats obtenus avec la fréquence non sémantique sont très proches des résultats obtenus avec la fréquence sémantique. Cela confirme qu'il est toujours plus pertinent de calculer une fréquence, par rapport au choix de la stratégie simpliste sans tenir compte de la fréquence.

## V.6. **Conclusion**

Nous avons proposé dans ce chapitre des stratégies pour la construction du contexte formel pour mieux refléter l'intensité des relations entre les objets et les attributs des données initiales. Tout d'abord nous avons présenté la stratégie simpliste qui consiste à construire un contexte formel en remplaçant par 1 toute valeur non nulle dans le tableau contenant le nombre d'occurrence des attributs associés à chaque objet (par exemple ici, le nombre de fois qu'une application a été utilisée dans chaque contexte). Cette stratégie ne nécessite pas de calcul, par contre elle peut produire une représentation non fiable des données initiales et biaiser l'ensemble des connaissances qui peuvent être extraites du treillis de Galois. Nous avons proposé par la suite des stratégies alternatives de construction de contexte formel plus représentatives des données initiales et permettant de se focaliser sur des relations plus ou moins fortes entre les données, selon les besoins d'information de l'utilisateur. Nous avons utilisé la fréquence des relations entre les objets et les attributs dans les données initiales et nous avons défini deux types de fréquence : l'une ne tenant compte que de la proportion relative des occurrences, et l'autre tenant compte en plus de la sémantique des attributs. A partir de cette fréquence, nous avons proposé trois stratégies :

- Une stratégie à haute dépendance entre les objets et les attributs, où les fréquences élevées sont transformées en 1 et toutes les autres en 0 pour obtenir une matrice binaire.

- Une stratégie à faible dépendance, où les fréquences faibles sont transformées en 1 et toutes les autres en 0.

- Une stratégie à dépendance moyenne, où les fréquences moyennes sont transformées en 1 et toutes les autres en 0.

Ces trois stratégies sont basées sur deux seuils : seuil-bas (Sb) et seuil-haut (Sh). Une perspective de ces travaux consistera à travailler sur le paramétrage de ces seuils au travers du paramètre $\beta$, afin de les adapter aux besoins de l'utilisateur.

Nous avons également proposé une stratégie dite inverse pour identifier les objets et les attributs qui n'ont aucune relation de dépendance entre eux. Nous avons illustré les stratégies proposées et utilisé les mesures définies dans le Chapitre III pour interpréter et comparer les résultats obtenus.

Nous pouvons conclure que, dans le cas d'une relation non binaire, la stratégie à haute dépendance est toujours préférable à la stratégie simpliste, que la fréquence soit sémantique ou non.

L'exploitation de la sémantique des données fournit bien sûr des résultats plus pertinents, mais nous avons vu que la fréquence non sémantique fournissait malgré tous des résultats bien plus proches de ceux de la fréquence sémantique que de ceux de la stratégie simpliste. Nous avons également illustré l'utilisation des fréquences faible et la stratégie inverse, qui peuvent être intéressantes aussi pour étudier les relations peu intenses (ou inexistantes) entre des objets et des attributs.

Dans le chapitre suivant, nous proposons une autre application de nos travaux, qui permet d'illustrer le choix de la stratégie de construction du contexte formel la plus adaptée.

# CHAPITRE VI :
# UNE APPLICATION À L'ÉTUDE D'UNE COMMUNAUTÉ DE CHERCHEURS

# Sommaire

## VI.1. **Introduction**

Nous présentons dans ce chapitre une autre application de nos travaux, visant à utiliser l'ACF pour étudier la communauté des chercheurs dans le domaine des lignes de produits logiciels. Nous avons travaillé sur des données réelles collectées par Danillo Sprovieri, un collègue de notre laboratoire. Ces données concernent des articles et les auteurs associés, au sein de la communauté Software Product Line (SPL), et précisément issus de la conférence SPLC2016[19] et de l'atelier DSPL2016[20]. Pour construire une grande base de connaissances, Danillo Sprovieri a collecté, pour chaque auteur identifié comme appartenant à la communauté SPL, l'ensemble de ses articles récents. Un total de 1275 auteurs et 2011 articles ont été sélectionnés pour ce travail.

Un ensemble de 17 thématiques a été défini, pour caractériser ces articles, comme *l'ingénierie des exigences*, la *gestion de contextes* et *l'auto-adaptation*. La méthodologie que nous détaillerons dans la section VII.2, consiste d'abord en une collecte de données, suivie par le choix d'une stratégie pour la construction du contexte formel, puis en déduire un treillis, auquel nos mesures sont appliquées pour en extraire des connaissances.

## VI.2. **Collecte des données et identification des besoins**

L'ensemble des données est une collection d'auteurs, d'articles et des thématiques associées. Les données ont été collectées sous forme d'une matrice. Chaque ligne de la matrice représente un auteur et le nombre d'articles recensés pour chacune des 17 thématiques. Par la suite, nous avons identifié la liste des besoins suivante :

[19] The 20th International Systems and Software Product Line Conference

[20] 9th International Workshop On Dynamic Software Product Lines

1. Constitution d'une vue globale d'une thématique spécifique.
2. Sélection des auteurs pertinents et importants sur une thématique donnée.
3. Identification de l'auteur le plus marquant pour une thématique spécifique.
4. Comparaison de plusieurs thématiques.
5. Identification des thématiques qui ne sont pas ou peu abordées dans la littérature.
6. Classification des thématiques les plus traitées dans les articles publiés.
7. Exploration des collaborations entre les auteurs travaillant sur des thématiques similaires.
8. Suggestion d'extension des collaborations existantes et de nouvelles collaborations possibles entre les auteurs.

Pour pouvoir répondre aux besoins listés ci-dessous, deux problématiques majeures doivent être traitées :

1. Le volume des données et la difficulté de leur analyse manuelle. Les données collectées comprennent en effet 1275 auteurs, 2011 articles et 17 thématiques.
2. La difficulté à identifier l'impact mutuel entre les auteurs et les thématiques, ainsi que la similarité entre les thématiques d'une part, et les auteurs travaillant sur une même thématique d'autre part.

Pour effectuer une revue de la littérature, plusieurs méthodes peuvent être appliquées, comme : *Structured Literature Review, Survey, Systematic literature review, Systematic mapping study, State of the art...* Par contre, ces méthodes ne répondent pas aux besoins mentionnés précédemment. De plus, ces méthodes fournissent seulement une vue d'ensemble des concepts, et ne considèrent pas les relations entre les auteurs et les sujets. Notre proposition étend les solutions existantes en ajoutant une compréhension des relations entre les thématiques et les auteurs de cette communauté, ainsi que des relations entre les thématiques.

Pour appliquer notre méthodologie, nous avons considéré les 8 besoins et les deux problématiques mentionnés. Pour sélectionner la stratégie la plus adaptée nous avons utilisé la matrice de départ qui représente le nombre d'articles de chaque auteur sur chaque thématique.

## VI.3. **Choix d'une stratégie de construction du contexte formel**

Les besoins identifiés sont très importants pour appliquer notre méthodologie.

Le besoin (1) était d'avoir un vue d'ensemble du domaine. Toutes les stratégies que nous avons proposées aboutissent à la construction de treillis qui représente une vue d'ensemble du domaine ; par conséquent pour bien choisir la stratégie, nous devons prendre en compte les autres besoins.

Les besoins (2) et (3) ont plus d'influence sur le choix de la stratégie. Pour identifier les auteurs significatifs pour un domaine spécifique, nous allons mettre en œuvre la stratégie à haute dépendance pour nous concentrer sur les relations fortes entres les auteurs et les thématiques. Ensuite, nous utiliserons les mesures de poids conceptuel pour les auteurs et les mesures d'impact entre les auteurs et les thématiques. Dans la section suivante, nous montrons les résultats de cette étude pour obtenir la liste des auteurs pertinents pour une thématique donnée.

Le besoin (4) consiste à trouver des connaissances similaires entre les thématiques, ce qui peut être évalué en utilisant la mesure de similarité entre les thématiques.

Concernant le besoin (5) qui sert à identifier les thématiques les moins étudiées, il sera lié aux thématiques qui ne sont pas très présentes dans le treillis ; celles-ci seront identifiées grâce à la mesure de poids conceptuel pour les thématiques.

Le besoin (6) est lié à une étude spécifique sur l'ensemble des thématiques. Pour cela, la stratégie à haute dépendance sera plus adaptée pour réaliser une étude des thématiques fréquentes, et étudier la similarité et la relation entre les thématiques.

Le besoin (7) consiste à étudier les relations entre les auteurs et les thématiques pour construire des groupes d'auteurs qui pourraient collaborer. Les résultats pour ce besoin seront présentés après avoir appliqué la stratégie à haute dépendance afin d'identifier les auteurs les plus marquant dans une thématique.

Le besoin (8) consiste à proposer des collaborations entre les auteurs. Pour cela nous allons faire une étude spécifique sur les auteurs selon la stratégie à haute dépendance pour recommander de nouvelles collaborations.

L'analyse des besoins nous amène ainsi à sélectionner la stratégie à haute dépendance, dont nous présentons les résultats dans la section suivante.

Comme nous l'avons expliqué dans le Chapitre V, nous devons fixer la valeur du paramètre $\beta$ pour le seuil-haut. Dans ce chapitre, nous allons fixer la valeur de $\beta$ à 0 pour conserver toutes les relations fortes entre les auteurs et les thématiques. Nous rappelons qu'une valeur de 0 pour la variable $\beta$ signifie que le seuil de la fréquence haute est égal à la moyenne des fréquences.

## VI.4. **Expérimentation avec la stratégie à haute dépendance**

Nous avons utilisé les données collectées et appliqué notre stratégie à haute dépendance, afin de construire le treillis de Galois à partir d'une matrice binaire, pour appliquer nos mesures pour l'interprétation du treillis résultant.

Le Tableau VI-1 ci-dessous présente un exemple pour l'auteur Bosch Jan : nous avons considéré 63 articles rédigés par cet auteur et indiquons le nombre d'articles liés à chaque thématique, sachant qu'un article peut être lié à plusieurs thématiques.

**Tableau VI-1 : Exemple de l'auteur Bosch Jan dans la matrice initiale**

| | Id | NoPaper | Ecosystem | Evolution | Architecture | Framework | RequirementsEngineering | ContextManagement | ProcessMethod | Social | Ontology | SelfAdaptive | PILsSPLs | KnowledgManagement | ServiceScience | BPM | IoT | Green | Cloud |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Bosch Jan** | **151** | **63** | 8 | 5 | 14 | 8 | 5 | 0 | 2 | 0 | 0 | 0 | 46 | 0 | 0 | 0 | 0 | 0 | 0 |

La stratégie à haute dépendance avec $\beta$ =0 permet de conserver, pour chaque auteur, uniquement les thématiques représentées plus que la moyenne dans ses articles. Nous nous concentrons donc ici sur les relations fortes entre les auteurs et les thématiques de recherche. Les auteurs et les thématiques éliminés par cette stratégie sont liés par des fréquences faibles

et ne sont pas significatifs pour notre étude de cas selon les besoins que nous avons identifiés. De plus leur présence pourrait biaiser les résultats et l'étude des relations fortes.

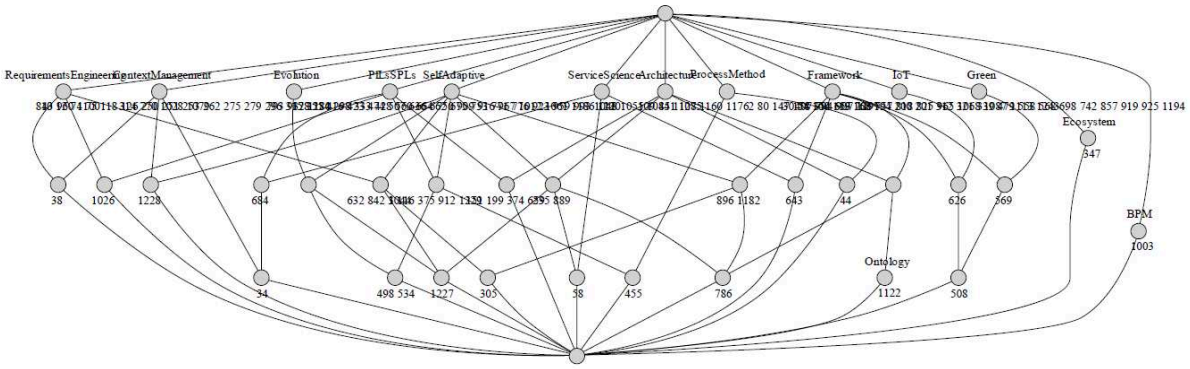



**Figure VI-1 : Treillis de Galois**

La Figure VI-1 montre le treillis de Galois obtenu, qui comprend 39 concepts, où chaque concept contient un ensemble d'auteurs et de thématiques liés entre eux. Nous appliquons les mesures définies dans le Chapitre III afin de répondre aux besoins définis précédemment. Tous ces résultats seront interprétés et évalués dans la section VI.5.

### VI.4.1. **Analyse orientée auteurs**

Les besoins (2) et (3) se focalisent sur l'objet « auteurs ». Dans cette section, notre objectif est d'obtenir, à partir de l'ensemble des auteurs, un sous-ensemble contenant les auteurs les plus influents. Pour cela, nous appliquons la mesure de poids conceptuel à chaque auteur dans le treillis, afin de calculer le pourcentage de concepts auxquels il appartient. Dans la section VI.4.3, nous reprendrons les besoins (2) et (3) en appliquant une autre mesure pour étudier la relation entre les auteurs et les thématiques.

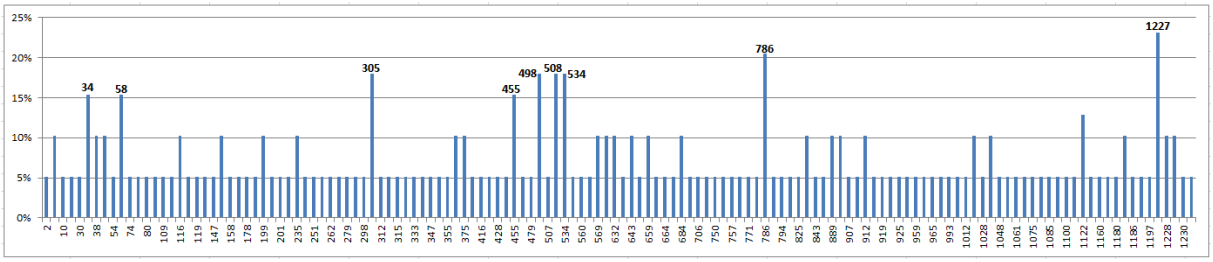



**Figure VI-2 : Poids conceptuel des objets « auteurs » dans le treillis**

La Figure VI-2 présente le poids conceptuel des auteurs (*objets*) dans le treillis, qui varie entre 5% et 23%. La plupart des auteurs ont un poids de 5%, tandis que seulement 9 auteurs ont un poids supérieur à 15%. Dans le tableau suivant, nous présentons ces 9 auteurs par poids décroissant.

**Tableau VI-2 : Liste des auteurs ayant le poids conceptuel le plus élevé**

| Id Auteur | Nom Auteur | Poids[21] |
|---|---|---|
| 1227 | Michel Wermelinger | 23,08% |
| 786 | Daniel A. Menasce | 20,51% |
| 305 | Scott A. DeLoach | 17,95% |
| 498 | Svein O. Hallsteinsen | 17,95% |
| 508 | Thomas Hartmann | 17,95% |
| 534 | Mike G. Hinchey | 17,95% |
| 34 | Germán H. Alférez | 15,38% |
| 58 | Jesper Andersson | 15,38% |
| 455 | Hassan Gomaa | 15,38% |

Selon le Tableau VI-2, Michel Wermelinger est l'auteur qui apparait dans le plus grand nombre de concepts avec d'autres auteurs sur des thématiques communes. Michel Wermelinger est associé à de nombreuses thématiques et il appartient à des concepts de la partie supérieure du treillis.

### VI.4.2. **Analyse orientée thématiques**

Concernant les besoins (5) et (6), nous cherchons à obtenir des informations sur les thématiques en fonction de leur présence dans le treillis. En nous appuyant sur la mesure de poids conceptuel, nous allons identifier les thématiques les plus répandues, c'est-à-dire celles qui ont une forte présence dans les concepts du treillis. Nous nous intéressons également aux thématiques qui ont une faible présence dans le treillis, ce qui signifie qu'elles sont peu abordées par les auteurs de cette communauté de recherche.

[21] Le calcul du poids se fait en divisant le nombre de concepts contenant un auteur par le nombre total de concepts dans le treillis.

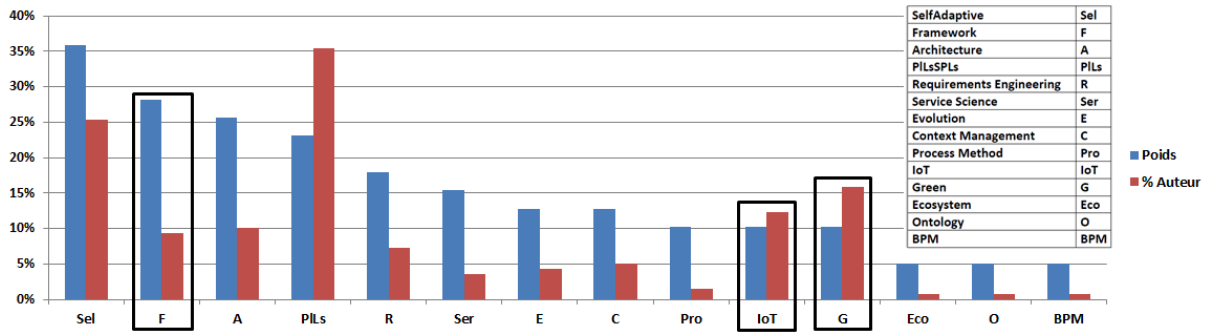



**Figure VI-3 : Poids conceptuel des thématiques dans le treillis et dans les données initiales**

Dans la Figure VI-3, le poids conceptuel d'une thématique *t* en bleu représente la proportion de concepts contenant *t* dans le treillis, alors que la valeur en rouge représente le pourcentage d'auteurs ayant contribué à cette thématique selon les données initiales. Les thématiques représentées dans le diagramme sont triées par poids décroissant. Par exemple, la thématique *SelfAdaptive* est la plus répandue ; elle appartient à 14 concepts et, sachant que le treillis contient 39 concepts, le poids conceptuel de cette thématique est donc de 36%. D'autre part, 25% des auteurs ont fait des publications dans cette thématique. Les thématiques qui sont le moins présentes (conceptuellement parlant) sont *Ecosystem*, *Ontology* et *BPM*, car chacune d'elles appartient à deux concepts du treillis seulement, ce qui équivaut à un poids de 5%.

Nous avons comparé, pour chaque thématique, son poids conceptuel et le pourcentage des auteurs y ayant contribué par des publications. Nous avons remarqué que le pourcentage d'auteurs ayant des publications dans les deux thématiques *IoT* et *Green* est élevé, bien que les poids de ces thématiques soient faibles. En effet, la thématique *IoT* appartient à 10% des concepts, alors que 12% des auteurs ont des publications dans cette thématique. Par contre, seulement 9% des auteurs ont des publications dans la thématique *Framework*, bien que cette thématique ait un poids élevé, vu qu'elle est présente dans environ 27% des concepts de treillis.

Par conséquent, nous pouvons constater qu'une thématique très répandue n'est pas forcément associée à un pourcentage élevé d'auteurs, car ces auteurs peuvent avoir publié simultanément dans plusieurs thématiques. D'autre part, nous pouvons remarquer que les thématiques les moins répandues ont un pourcentage d'auteurs élevé, ce qui peut signifier que ces thématiques ne sont pas abordées simultanément avec une autre.

### VI.4.3. **Étude de l'impact mutuel entre auteur et thématique**

Nous reprenons maintenant les besoins (2) et (3), déjà traités dans la section VI.4.1, en y rajoutant le besoin (7), pour étudier l'impact entre les auteurs et les thématiques sur lesquelles ils publient en utilisant l'impact mutuel. Nous présentons par la suite un ensemble de combinaisons possibles entre les auteurs et les thématiques, ainsi que leur impact mutuel. La valeur de celui-ci varie ici entre 0% et 38%. Pour donner un exemple représentatif, nous avons choisi dans le Tableau VI-3 un sous-ensemble contenant les pourcentages les plus significatifs (entre 17% et 38%).

**Tableau VI-3 : Impacts mutuels les plus élevés**

| Thématique | Auteur | Impact mutuel |
|---|---|---|
| Green | Thomas Hartmann | 38% |
| IoT[22] | Thomas Hartmann | 38% |
| BPM[23] | Colette Rolland | 33% |
| Ecosystem | Holger Eichelberger | 33% |
| Evolution | Mike G. Hinchey | 33% |
| Evolution | Svein O. Hallsteinsen | 33% |
| ProcessMethod | Eduardo Almeida | 33% |
| ServiceScience | Germán H. Alférez | 33% |
| Architecture | Daniel a. Menascé | 29% |
| ContextManagement | Danny Weyns | 29% |
| ContextManagement | Raian Ali | 29% |
| Framework | Thomas Hartmann | 29% |
| SelfAdaptive | Michel Wermelinger | 28% |
| RequirementsEngineering | Scott A. DeLoach | 27% |
| PlLsSPLs | Germán H. Alférez | 25% |
| PlLsSPLs | Hassan Gomaa | 25% |
| Ontology | Vijayan Sugumaran | 17% |

[22] Internet of Things

[23] Business Process Management

Par définition, l'impact mutuel est la probabilité d'occurrence d'une thématique et un auteur simultanément dans le treillis. Dans le tableau ci-dessus, nous remarquons que Thomas Hartmann est l'auteur qui a l'impact mutuel le plus élevé avec les trois thématiques *Green*, *IoT* et *Framework*. En regardant l'auteur Colette Rolland, nous remarquons qu'elle a l'impact mutuel le plus élevé (33%) avec la thématique *BPM*.

Par conséquent, l'impact mutuel d'un auteur *A* et une thématique *T* donne une représentation quantitative de l'activité de recherche exercée par *A* précisément dans la thématique *T*. Nous notons que l'auteur *A* a une faible présence dans les autres thématiques vu qu'il est plus spécialisé dans la thématique *T*. En observant l'impact mutuel d'un autre point de vue, nous remarquons qu'une thématique est souvent abordée exclusivement par un auteur déterminé.

Nous avons utilisé l'outil Gephi pour présenter l'impact mutuel entre les auteurs et les thématiques, comme le montre la Figure VI-4. Ce graphe permet de mettre en évidence les thématiques les plus significatives en termes d'impact mutuel.

La Figure VI-4 permet de voir les thématiques sur lesquelles de nombreux auteurs ont publié, comme *PlLsSPLs* et *SelfAdaptive*. De plus, à partir d'une thématique donnée, nous pouvons identifier l'auteur ayant l'impact mutuel maximum ; c'est le cas par exemple de l'auteur numéro 508 (Thomas Hartmann), qui a l'impact le plus élevé avec les trois thématiques *Green, IoT* et *Framework*. D'autres auteurs ont publié dans plusieurs thématiques comme l'auteur numéro 786 (Daniel A. Menascé), impliqué dans les trois thématiques *SelfAdaptive, Architecture* et *Framework* avec un impact mutuel élevé. A l'inverse, nous voyons facilement les auteurs qui publient dans une seule thématique, comme l'auteur 80 (Atzori Luigi) spécialisé dans la thématique *IoT*.

Ce graphe permet également de repérer aisément les thématiques auxquelles très peu d'auteurs de cette communauté ont publié, telles que *BPM* et *Ecosystem*. Nous pouvons recommander à des auteurs de publier dans ces deux thématiques très peu traitées.

Nous pouvons aussi observer les combinaisons de thématiques qui sont partagées par peu d'auteurs, comme par exemple *IoT*, qui ne partage que deux auteurs avec les deux thématiques *Framework* et *Green.* A l'inverse on identifie les thématiques qui concernent de nombreux auteurs, comme par exemple *Architecture.*

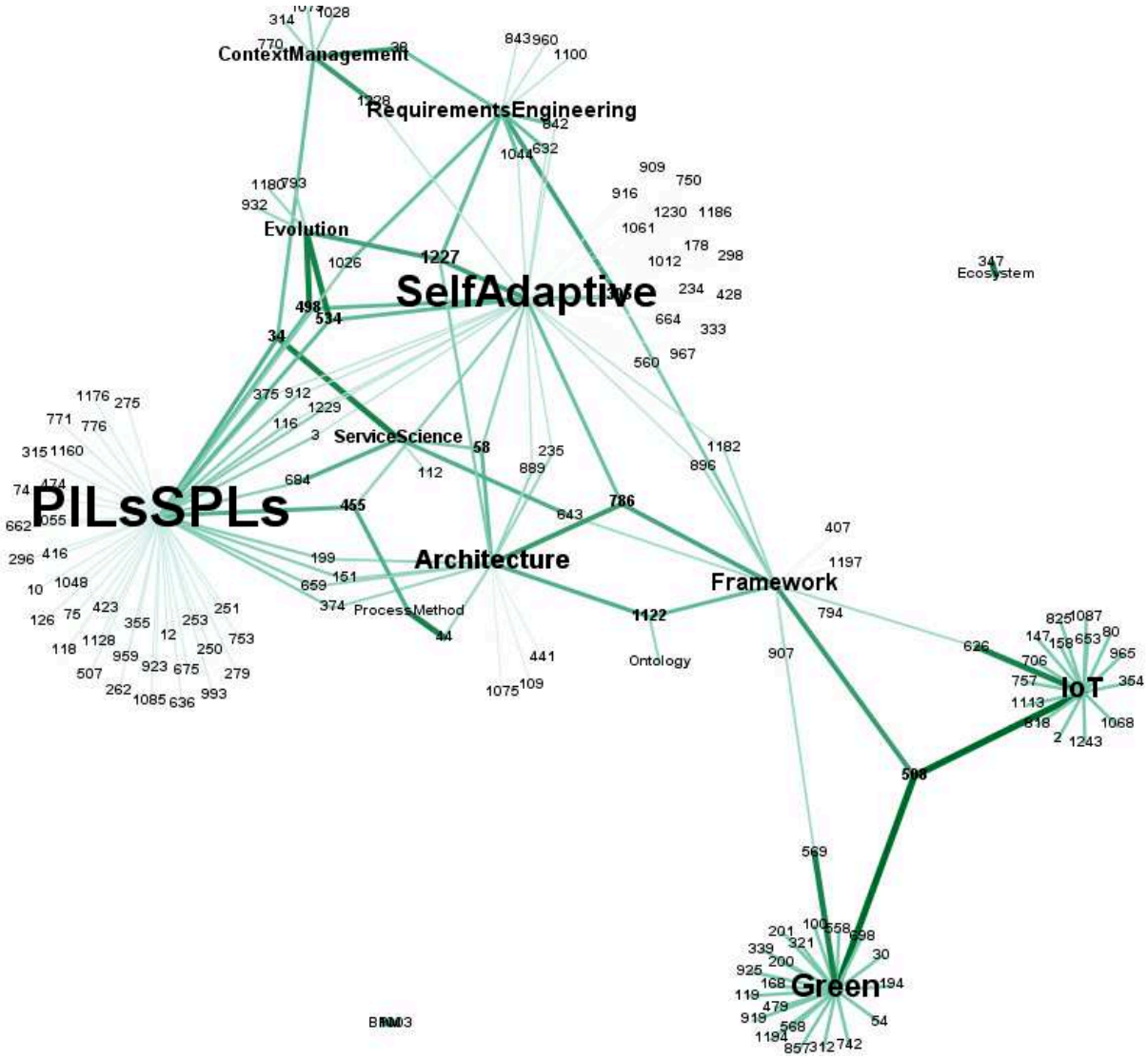



**Figure VI-4 : Graphe d'impact mutuel entre les auteurs et les thématiques**

## VI.4.4. **Étude de la similarité conceptuelle entre les auteurs**

Les besoins (4) et (8) consistent à identifier des collaborations possibles entre les auteurs ayant des activités de recherche dans la même thématique. Pour cela, nous calculons la similarité conceptuelle entre les auteurs.

**Tableau VI-4 : Extrait de la matrice de similarité conceptuelle entre les auteurs**

| | 2 | 3 | 10 | 12 | 30 | 34 | 38 | 44 | 54 | 58 | 74 | 75 | 80 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | | 20% | 33% | 33% | 33% | 14% | 20% | 20% | 33% | 14% | 33% | 33% | 100% | 33% |
| 3 | 20% | | 50% | 50% | 20% | 25% | 14% | 14% | 20% | 25% | 50% | 50% | 20% | 20% |
| 10 | 33% | 50% | | 100% | 33% | 33% | 20% | 20% | 33% | 14% | 100% | 100% | 33% | 33% |
| 12 | 33% | 50% | 100% | | 33% | 33% | 20% | 20% | 33% | 14% | 100% | 100% | 33% | 33% |
| 30 | 33% | 20% | 33% | 33% | | 14% | 20% | 20% | 100% | 14% | 33% | 33% | 33% | 100% |
| 34 | 14% | 25% | 33% | 33% | 14% | | 25% | 11% | 14% | 20% | 33% | 33% | 14% | 14% |
| 38 | 20% | 14% | 20% | 20% | 20% | 25% | | 14% | 20% | 11% | 20% | 20% | 20% | 20% |
| 44 | 20% | 14% | 20% | 20% | 20% | 11% | 14% | | 20% | 25% | 20% | 20% | 20% | 20% |
| 54 | 33% | 20% | 33% | 33% | 100% | 14% | 20% | 20% | | 14% | 33% | 33% | 33% | 100% |
| 58 | 14% | 25% | 14% | 14% | 14% | 20% | 11% | 25% | 14% | | 14% | 14% | 14% | 14% |
| 74 | 33% | 50% | 100% | 100% | 33% | 33% | 20% | 20% | 33% | 14% | | 100% | 33% | 33% |
| 75 | 33% | 50% | 100% | 100% | 33% | 33% | 20% | 20% | 33% | 14% | 100% | | 33% | 33% |
| 80 | 100% | 20% | 33% | 33% | 33% | 14% | 20% | 20% | 33% | 14% | 33% | 33% | | 33% |
| 100 | 33% | 20% | 33% | 33% | 100% | 14% | 20% | 20% | 100% | 14% | 33% | 33% | 33% | |

Le Tableau VI-4 ci-dessus représente une partie de la matrice de similarité conceptuelle concernant certains auteurs (identifiés par leur numéro). Les valeurs de similarité conceptuelle varient entre 10% et 100%. Par exemple, le profil de recherche de l'auteur Mathieu Acher (numéro 12) est très similaire à celui de Patrizia Asirelli (numéro 75), vu que leur similarité conceptuelle est 100%.

Nous avons utilisé la similarité conceptuelle pour identifier des collaborations possibles entre des auteurs ayant des profils de recherche similaires. Nous rappelons que les auteurs ayant une valeur de similarité conceptuelle élevée les uns avec les autres sont ceux qui apparaissent simultanément dans les mêmes concepts du treillis. Nous avons cherché les groupes d'auteurs ayant entre eux une similarité conceptuelle de 100%. Cela nous a permis de proposer un ensemble de collaborations possibles, correspondant à chacun de ces groupes, comme le montre le Tableau VI-5.

Ci-dessous, nous présentons les 11 premières listes de collaborations possibles qui contiennent, dans notre cas d'étude, au plus cinq auteurs, sachant que nous pouvons avoir un nombre supérieur dans d'autre cas d'étude.

**Tableau VI-5 : Listes de collaborations possibles**

| | | | | | |
|---|---|---|---|---|---|
| 1 | Walamitien H. Oyenan, | Sebastian Vansyckel | | | |
| 2 | Svein O. Hallsteinsen, | Mike G. Hinchey | | | |
| 3 | Shang-Wen Cheng Cheng, | P. Oreizy | | | |
| 4 | Juan C. Muñoz-Fernández, | N. Qureshi, | Amina Souag | | |
| 5 | Ronald Beaubrun, | John C. Georgas, | Michael E. Shin | | |
| 6 | Kim Mens, | G. Perrouin, | Rob van Ommering | | |
| 7 | Minseong Kim, | Javier Muñoz, | Pete Sawyer | | |
| 8 | François Fouquet, | Andreas Menychtas, | Sooyong Park, | Thomas Vogel | |
| 9 | Anind K. Dey, | Rene Mayrhofer, | G. Salvaneschi, | Stephan Sigg | |
| 10 | Jan Bosch, | Rafael Capilla, | Julie Street Fant, | A. Krishna | |
| 11 | Nadeem Abbas, | David Benavides, | Alessandro Fantechi, | Gustavo G. Pascual, | Jules White |

Nous avons aussi représenté sur la Figure VI-5 la similarité conceptuelle entre les auteurs, sous forme de carte 2D.

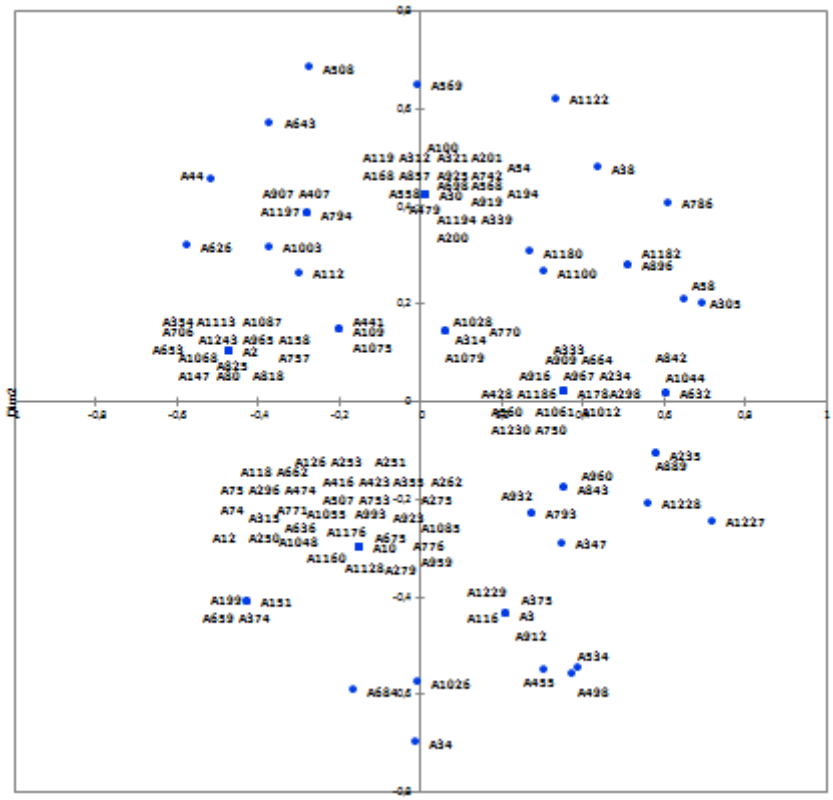


**Figure VI-5 : Carte MDS pour la similarité conceptuelle des auteurs**

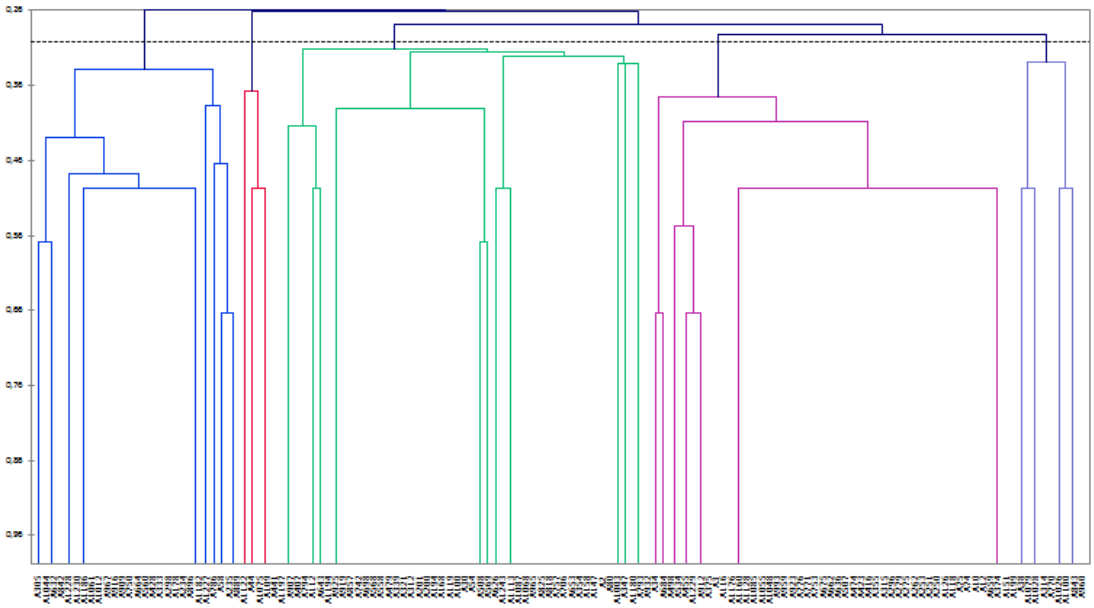


**Figure VI-6 : CAH pour la similarité conceptuelle des auteurs**

D'après la Figure VI-6, la classification ascendante hiérarchique a identifié cinq clusters d'auteurs, dont le contenu est présenté dans le Tableau VI-6.

**Tableau VI-6 : Contenu des cinq clusters d'auteurs identifiés par la CAH**

| | | | | |
|---|---|---|---|---|
| A2 | A3 | A58 | A38 | A44 |
| A30 | A10 | A178 | A314 | A109 |
| A54 | A12 | A234 | A770 | A441 |
| A80 | A34 | A235 | A843 | A1075 |
| A100 | A74 | A298 | A960 | A1122 |
| A112 | A75 | A305 | A1026 | |
| A119 | A116 | A333 | A1028 | |
| A147 | A118 | A428 | A1079 | |
| A158 | A126 | A560 | A1100 | |
| A168 | A151 | A632 | | |
| A194 | A199 | A664 | | |
| A200 | A250 | A750 | | |
| A201 | A251 | A786 | | |
| A312 | A253 | A842 | | |
| A321 | A262 | A889 | | |
| A339 | A275 | A896 | | |
| A347 | A279 | A909 | | |
| A354 | A296 | A916 | | |
| A407 | A315 | A967 | | |
| A479 | A355 | A1012 | | |
| A508 | A374 | A1044 | | |
| A558 | A375 | A1061 | | |
| A568 | A416 | A1182 | | |
| A569 | A423 | A1186 | | |
| A626 | A455 | A1227 | | |
| A643 | A474 | A1228 | | |

| | | |
|---|---|---|
| A653 | A498 | A1230 |
| A698 | A507 | |
| A706 | A534 | |
| A742 | A636 | |
| A757 | A659 | |
| A793 | A662 | |
| A794 | A675 | |
| A818 | A684 | |
| A825 | A753 | |
| A857 | A771 | |
| A907 | A776 | |
| A919 | A912 | |
| A925 | A923 | |
| A932 | A959 | |
| A965 | A993 | |
| A1003 | A1048 | |
| A1068 | A1055 | |
| A1087 | A1085 | |
| A1113 | A1128 | |
| A1180 | A1160 | |
| A1194 | A1176 | |
| A1197 | A1229 | |
| A1243 | | |

### VI.4.5. **Étude de la combinaison de thématiques**

Finalement, nous avons étudié les thématiques les plus similaires pour pouvoir proposer des combinaisons possibles comme il est décrit dans le besoin (4). Pour cela nous avons calculé la similarité conceptuelle entre les thématiques, puis nous avons identifié les valeurs les plus élevées dans le Tableau VI-7. Nous avons classé les combinaisons des thématiques par poids des thématiques décroissant (section VI.4.2), afin de donner plus d'importance aux thématiques les plus présentes dans le treillis, autrement dit celles qui ont le plus de relations avec les auteurs.

**Tableau VI-7 : Similarité conceptuelle entre les thématiques**

| Poids | | | Similarité |
|---|---|---|---|
| 36% | SelfAdaptive | Evolution | 27% |
| | SelfAdaptive | Architecture | 26% |
| | SelfAdaptive | RequirementsEngineering | 24% |
| | SelfAdaptive | PlLsSPLs | 21% |
| 28% | Framework | Green | 25% |
| | Framework | IoT | 25% |
| | Framework | Architecture | 24% |
| 26% | Architecture | SelfAdaptive | 26% |

| | | | |
|---|---|---|---|
| | Architecture | Framework | 24% |
| | Architecture | Ontology | 20% |
| 23% | PlLsSPLs | ServiceScience | 25% |
| | PlLsSPLs | SelfAdaptive | 21% |
| 18% | RequirementsEngineering | SelfAdaptive | 24% |
| | RequirementsEngineering | ContextManagement | 20% |
| | RequirementsEngineering | Evolution | 20% |
| 15% | ServiceScience | PlLsSPLs | 25% |
| | ServiceScience | ContextManagement | 22% |
| 13% | Evolution | SelfAdaptive | 27% |
| | Evolution | RequirementsEngineering | 20% |
| 13% | ContextManagement | ServiceScience | 22% |
| | ContextManagement | RequirementsEngineering | 20% |
| 10% | ProcessMethod | BPM | 20% |
| | ProcessMethod | Ecosystem | 20% |
| | ProcessMethod | Ontology | 20% |
| 10% | IoT | Green | 33% |
| | IoT | Framework | 25% |
| | IoT | BPM | 20% |
| | IoT | Ecosystem | 20% |
| | IoT | Ontology | 20% |
| 10% | Green | IoT | 33% |
| | Green | Framework | 25% |
| | Green | BPM | 20% |
| | Green | Ecosystem | 20% |
| | Green | Ontology | 20% |
| 5% | Ecosystem | BPM | 33% |
| | Ecosystem | Ontology | 33% |
| 5% | Ontology | Ecosystem | 33% |
| | Ontology | BPM | 33% |
| 5% | BPM | Ecosystem | 33% |
| | BPM | Ontology | 33% |

Nous avons identifié les clusters par CAH sur la Figure VI-8 et indiqué ces clusters via des couleurs sur la carte en 2D de la Figure VI-7.

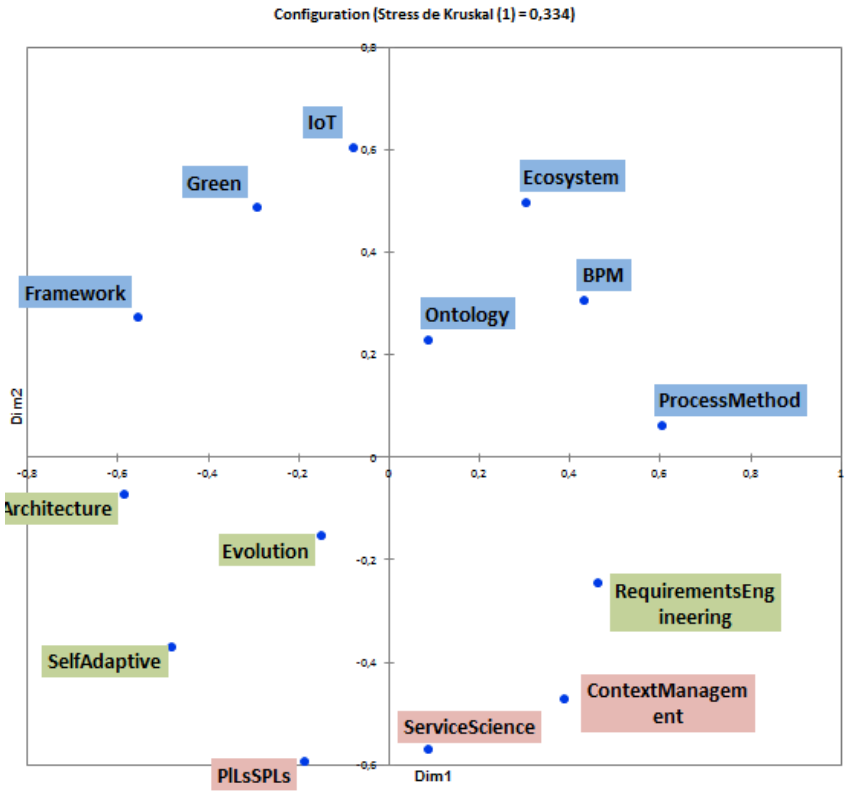



**Figure VI-7 : Carte 2D pour la similarité conceptuelle des thématiques**

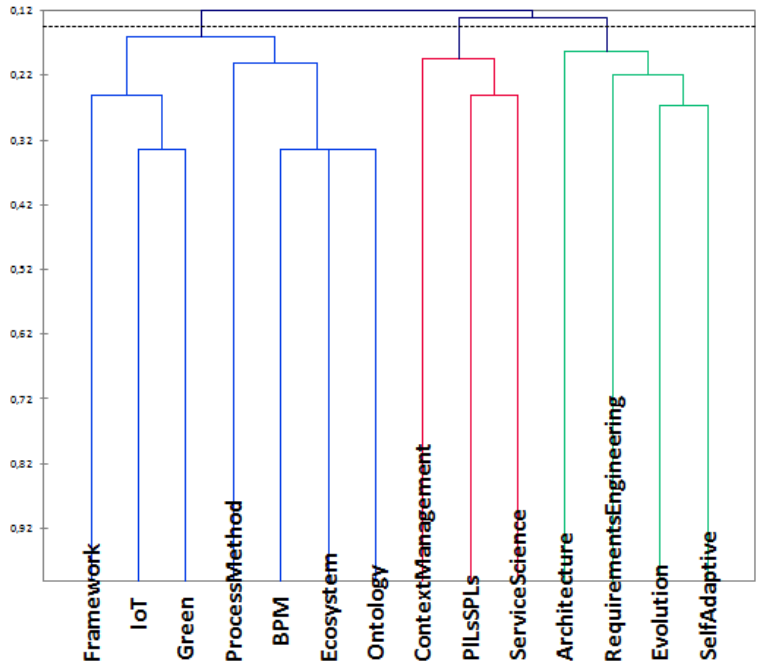



**Figure VI-8 : CAH pour la similarité conceptuelle des thématiques**

La CAH a permis d'identifier trois clusters de thématiques, dans le domaine des lignes de produits logiciels.

Les clusters de thématiques, identifiés grâce à la mesure de la similarité conceptuelle, permettent de proposer aux auteurs d'une thématique donnée de traiter des problématiques de recherche en relation avec plusieurs thématiques combinées. Par exemple, les auteurs publiant dans la thématique *SelfAdaptive* peuvent aussi traiter des problématiques liées aux thématiques *Evolution*, *Architecture* ou *RequirementsEngineering*, car ces quatre thématiques constituent un cluster selon la mesure de similarité conceptuelle.

## VI.5. **Conclusion**

Dans ce chapitre, nous avons appliqué, sur un ensemble de données collectées au sein de notre laboratoire, notre méthodologie d'analyse reposant sur l'ACF, nos mesures et une stratégie de construction de contexte formel. Nous avons tout d'abord identifié des besoins pour pouvoir choisir la stratégie la plus adaptée et sélectionner les mesures adéquates pour répondre à ces besoins.

Nous synthétisons ici les connaissances extraites de notre analyse.

Tout d'abord, nous avons obtenu une vue orientée auteur d'un domaine de recherche spécifique, qui est plus avancée qu'une simple étude bibliographique. D'une part, l'étude de l'impact mutuel entre les auteurs et les thématiques constitue une source de connaissance utile pour pouvoir proposer de nouvelles contributions de recherche pertinentes. D'autre part, l'étude de la similarité conceptuelle entre les thématiques d'un même domaine permet de suggérer aux auteurs de rédiger des publications qui traitent des problématiques de recherche en relation avec plusieurs thématiques combinées. Notre approche peut fournir de base pour des études prospectives et pour promouvoir de nouvelles thématiques qui sont peu (ou pas) traitées.

Notre approche offre une nouvelle façon d'explorer la littérature, en introduisant les notions suivantes :

**Auteur impliqué dans le plus de disciplines :** connaître un auteur ayant des publications dans plusieurs thématiques de recherche peut aider à sélectionner un nouveau chef de projet dans un domaine spécifique.

**Thématiques les moins répandues** : en identifiant les thématiques les moins traitées dans un domaine spécifique, nous pouvons faire des recommandations pour exploiter de nouvelles pistes et combler le manque de recherche dans ce domaine.

**Auteurs experts** : en étudiant l'impact mutuel entre les auteurs et les thématiques, nous pouvons identifier des auteurs comme des référents pour une thématique spécifique vu que la majorité de leurs travaux de recherche sont spécialisés dans cette thématique.

**Collaboration entre les auteurs** : ceci consiste à proposer des collaborations entre les auteurs travaillant sur les mêmes thématiques, afin de développer de nouveaux travaux de recherche combinant les thématiques identifiées.

**Thématiques similaires** : la mesure de similarité conceptuelle peut aider à faire des nouvelles innovations en combinant deux ou plusieurs thématiques. Par exemple, la combinaison de *PlLsSPLs* et *SelfAdaptive* peut être considérée comme une nouvelle tendance de recherche qui se concentre sur l'ingénierie des systèmes d'information.

# CHAPITRE VII :
# CONCLUSION GÉNÉRALE ET PERSPECTIVES

# Sommaire

## VII.1. **Conclusion générale**

L'Analyse de Concepts Formels (ACF) est utilisée dans de nombreux travaux de recherche pour la représentation des connaissances, l'analyse, la visualisation et l'interprétation des données. L'objectif de cette thèse a été de proposer une **méthodologie** afin de permettre à des **utilisateurs non experts** de construire les données d'entrée nécessaires à l'ACF et d'interpréter eux-mêmes les treillis de Galois qu'ils génèrent.

Dans ce manuscrit, nous avons tout d'abord dressé un état de l'art des travaux de recherche qui utilisent l'ACF, en étudiant les articles publiés dans les principales conférences du domaine. : CLA, ICFCA, FCA4AI, ICCS.

Nous avons ensuite défini **trois nouvelles mesures pour l'interprétation des treillis de Galois**, afin d'analyser les objets, les attributs et les concepts d'un treillis : le poids conceptuel d'un objet ou d'un attribut, la similarité conceptuelle entre deux objets ou deux attributs, ainsi que l'impact mutuel entre un objet et un attribut.

Par la suite, nous avons proposé des **stratégies pour la construction du contexte formel** adaptées aux relations non naturellement binaires. Nous avons en particulier proposé de tenir compte de la force des relations entre objets et attributs, afin de construire un contexte formel fidèle aux données et représentatif des relations auxquelles l'utilisateur souhaite s'intéresser en priorité (relations fortes ou au contraire relations faibles, voire inexistantes).

Nous avons implémenté toutes ces mesures et stratégies, et les avons **expérimentées** sur plusieurs exemples : l'analyse d'usage d'applications sur smartphone et la réalisation d'une revue de la littérature.

## VII.2. **Méthodologie d'utilisation et d'interprétation de l'ACF**

Nous résumons ici nos contributions sous la forme d'une méthodologie d'utilisation et d'interprétation de l'ACF, à destination d'utilisateurs non experts de l'analyse de données. Il s'agit notamment d'aider les utilisateurs à construire des données d'entrées pertinentes et de les assister pour l'interprétation des treillis de Galois obtenus en sortie. Notre approche

consiste à rendre les résultats de ces traitements aisément compréhensibles par des non-spécialistes. Cette automatisation partielle du processus permettant l'interprétation des treillis de Galois vise à rendre cette méthode accessible à un plus grand nombre d'utilisateurs et à favoriser l'exploration des connaissances ainsi que le passage à l'échelle.

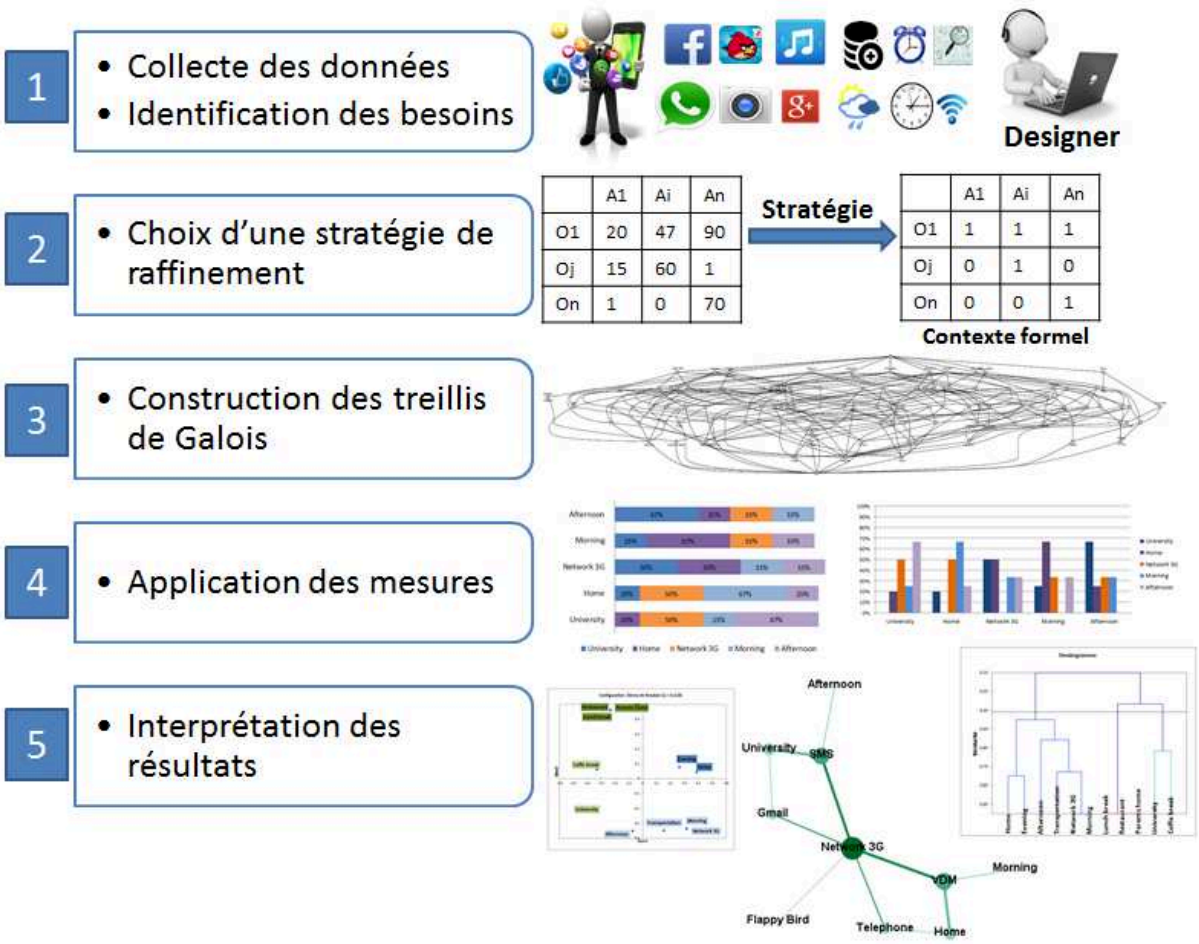



**Figure VII-1 : Méthodologie d'utilisation et d'interprétation de l'ACF**

Notre méthodologie, illustrée sur la Figure VII-1 consiste à collecter des données et identifier les besoins de l'utilisateur, puis à sélectionner une stratégie de construction du contexte formel. Viennent ensuite la construction des treillis de Galois et l'application des mesures proposées pour enfin interpréter les résultats obtenus. Nous détaillons ci-dessous chacune de ces étapes.

### 1) Collecte des données et identification des besoins

La première étape consiste à disposer des données collectées et de connaitre les besoins de l'utilisateur. Nous n'avons pas détaillé dans ce manuscrit le processus de collecte des données, propre à chaque cas d'étude. L'objectif est de construire des données d'entrée permettant de caractériser des *objets* par le nombre d'occurrences de chacun de leurs *attributs*. Les données peuvent être de différents types, par exemple, des articles décrits par leurs auteurs, des maladies décrites par leurs symptômes, des diplômes par des compétences associées. Ces données peuvent provenir d'applications, de formulaires, de questionnaires, de logiciels ou de capteurs. Le format attendu est une matrice à deux dimensions avec les objets sur les lignes, les attributs sur les colonnes, et des valeurs numériques dans la matrice.

### 2) Choix d'une stratégie de construction du contexte formel

Selon les besoins des utilisateurs, une ou plusieurs stratégies sont sélectionnées en fonction des attentes de l'utilisateur, selon par exemple qu'il s'intéresse aux relations fortes entre les données indépendamment des relations faibles, ou inversement.

### 3) Construction des treillis

La construction des treillis est basée sur les contextes formels construits durant la phase précédente. Tout algorithme de construction de treillis de Galois peut être utilisé pour cette étape. Les contributions de cette thèse concernent les phases 2, 4 et 5 de cette méthodologie.

Nous avons choisi dans cette thèse d'utiliser l'outil *Lattice Miner* pour générer une représentation visuelle des treillis, et le module *Concepts* en Python pour générer des treillis au format xml.

### 4) Application des mesures

Nous avons présenté dans le Chapitre III les mesures que nous avons proposées pour interpréter les treillis. Selon l'objectif de l'utilisateur, l'une ou plusieurs des mesures est adaptée comme nous l'avons illustré dans le Chapitre IV, le Chapitre V et le Chapitre VI. Les mesures sont implémentées indépendamment de la taille de chaque treillis. D'autres mesures de la littérature peuvent être appliquées durant cette phase.

### 5) Interprétation des résultats

Il s'agit ici d'interpréter les résultats des mesures calculées lors de l'étape précédente. Nous avons proposé des visualisations intuitives des résultats des calculs, notamment une représentation sous forme de carte en 2D pour la similarité conceptuelle et sous forme de graphe pour l'impact mutuel. Nous avons pu, sur les exemples étudiés, extraire des informations et comparer les résultats obtenus avec différentes stratégies. Nous avons aussi, dans le Chapitre III et en annexe, pu comparer plusieurs treillis, ce qui n'est habituellement pas aisé.

La méthodologie présentée dans cette section a été appliquée à plusieurs cas d'étude, et ouvre de nombreuses perspectives, que nous décrivons dans la suite.

## VII.3. **Perspectives**

Nous décrivons ici les perspectives de nos travaux, en complément de celles que nous avons déjà mentionnées au fil du manuscrit.

La première perspective consistera à rédiger un manuel d'utilisateur pour notre méthodologie, dans lequel nous décrirons en détail toutes les étapes à suivre, que nous illustrerons sur un exemple simple. Ce manuel expliquera tous les programmes développés et tous les outils utilisés.

Nous indiquerons en particulier à l'utilisateur comment sélectionner une stratégie de construction de contexte formel et choisir ensuite les mesures appropriées selon ses données et ses objectifs.

Nous procéderons ensuite à une évaluation auprès d'utilisateurs non experts de l'analyse de données et nous établirons pour cela des critères de satisfaction.

La deuxième perspective consistera à concevoir et à implémenter des mesures supplémentaires afin d'aller encore plus loin dans l'interprétation des treillis de Galois. Nous nous sommes concentrés jusqu'ici sur l'analyse des concepts, des objets et des attributs ; nous proposerons des mesures permettant d'exploiter les liens présents dans les treillis, i.e., les relations de généralisation et de spécialisation entre les concepts formels. En effet, la

sémantique de ces relations peut s'avérer très riche et mériterait d'être prise en compte pour l'interprétation du treillis.

Une idée pour exploiter les liens du treillis est d'utiliser les mesures définies par l'analyse de réseaux sociaux. Par exemple, il existe différentes mesures pour évaluer l'importance d'un nœud dans un réseau en fonction du nombre de ses voisins (centralité de degré) ou en fonction de sa proximité aux autres nœuds (centralité de proximité). Il est également possible de regrouper les nœuds d'un graphe dans des clusters appelés communautés (dans le vocabulaire de l'analyse des réseaux sociaux). Les nœuds appartenant à une même communauté sont fortement liés les uns aux autres, alors qu'ils sont faiblement liés aux nœuds des autres communautés.

L'utilisation de ces mesures pourrait constituer un complément intéressant à celles que nous avons proposées dans cette thèse, en travaillant non seulement au niveau des objets et des attributs, mais également au niveau des concepts du treillis.

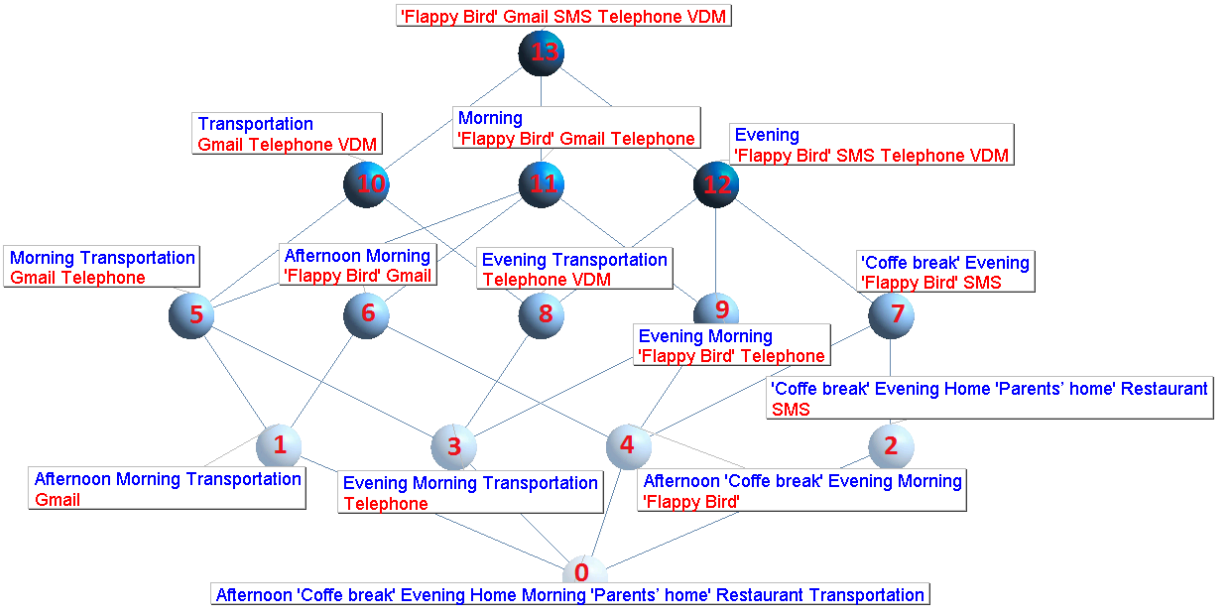



**Figure VII-2 : Exemple de treillis de Galois**

Nous fournissons sur la Figure VII-3 un exemple de représentation du treillis de Galois de la Figure VII-2 avec un outil dédié à l'analyse des graphes (Gephi), déjà utilisé dans les chapitres précédents.

Nous avons considéré ce treillis comme un graphe où les nœuds sont les concepts, et les liens entre les concepts sont les arcs du graphe.

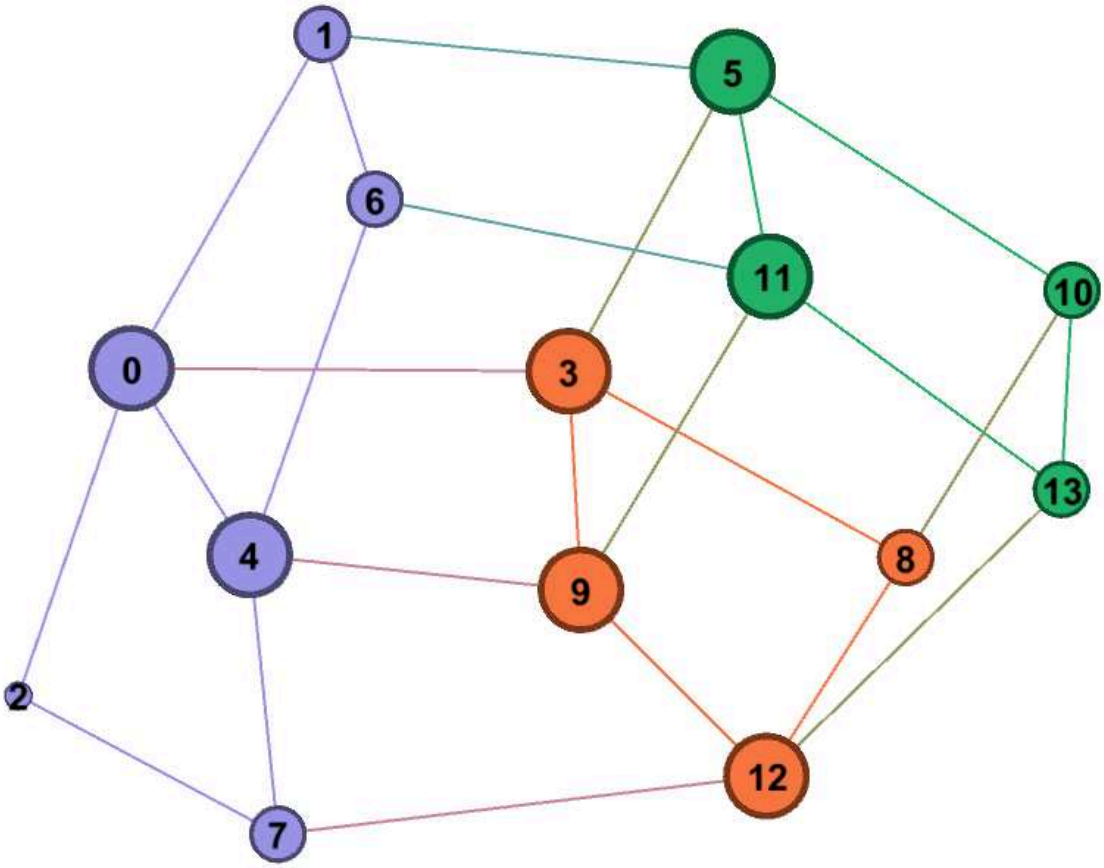



**Figure VII-3 : Représentation sous forme de graphe du treillis réalisée avec Gephi**

Sur la Figure VII-3, la taille des nœuds est liée au nombre de leurs voisins dans le graphe (appelé degré). Un algorithme de détection de communautés a été appliqué (reposant sur l'optimisation d'une fonction appelée modularité). Trois communautés ont été détectées associées à des couleurs différentes : bleu, rouge et vert.

Une manière d'exploiter ce graphe pourrait être, par exemple :

- de choisir un concept « important » pour représenter chaque communauté, par exemple celui de plus fort degré.

- s'intéresser aux nœuds qui sont proches de plusieurs communautés, ou au contraire liés uniquement à des nœuds de leur communauté. De nombreuses analyses peuvent donc être menées dans cette direction.

# Bibliographie

Alam, M., Buzmakov, A., Codocedo, V., & Napoli, A. (2015). Bridging DBpedia Categories and DL-Concept Definitions using Formal Concept Analysis. In *Proceedings of the 4th International Workshop "'What can FCA do for Artificial Intelligence?'", FCA4AI 2015, co-located with the International Joint Conference on Artificial Intelligence (IJCAI 2015), Buenos Aires, Argentina.* (Vol. 1430). Buenos Aires, Argentina. Retrieved from https://hal.inria.fr/hal-01186330

Alam, M., Le, T. N. N., & Napoli, A. (2016a). LatViz: A New Practical Tool for Performing Interactive Exploration over Concept Lattices. In *CLA 2016 - Thirteenth International Conference on Concept Lattices and Their Applications* . Moscow, Russia. Retrieved from https://hal.inria.fr/hal-01420751

Alam, M., Le, T. N. N., & Napoli, A. (2016b). Steps Towards Interactive Formal Concept Analysis with LatViz. In *Proceedings of the 5th International Workshop "'What can FCA do for Artificial Intelligence'"? co-located with the European Conference on Artificial Intelligence ECAI 2016*. The Hague, Netherlands. Retrieved from https://hal.inria.fr/hal-01420753

Alam, M., & Napoli, A. (2014). Defining Views with Formal Concept Analysis for Understanding SPARQL Query Results. In *Proceedings of the Eleventh International Conference on Concept Lattices and Their Applications CLA*. Košice, Slovakia. Retrieved from https://hal.inria.fr/hal-01089770

Alam, M., Napoli, A., & Osmuk, M. (2015). RV-Xplorer: A Way to Navigate Lattice-Based Views over RDF Graphs. In *Proceedings of the Twelfth International Conference on Concept Lattices and Their Applications* (Vol. 1466). Clermont-Ferrand, France. Retrieved from https://hal.inria.fr/hal-01186344

Andrews, S., Brewster, B., & Day, T. (2016). Organised Crime and Social Media: Detecting and Corroborating Weak Signals of Human Trafficking Online. In O. Haemmerlé, G. Stapleton, & C. Faron Zucker (Eds.), *Graph-Based Representation and Reasoning* (pp. 137–150). Cham: Springer International Publishing.

Antoni, L., Gunis, J., Krajci, S., Krıdlo, O., & Snajder, L. (2014). The Educational Tasks and Objectives System within a Formal Context. In *Concept Lattices and Their Applications CLA*.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. https://doi.org/10.13140/2.1.1341.1520

Buzmakov, A., Kuznetsov, S. O., & Napoli, A. (2014). Scalable Estimates of Concept Stability. In C. V. Glodeanu, M. Kaytoue, & C. Sacarea (Eds.), *Formal Concept Analysis* (pp. 157–172). Cham: Springer International Publishing.

Carbonnel, J., Bertet, K., Huchard, M., & Nebut, C. (2016). FCA for Software Product Lines Representation: Mixing Configuration and Feature Relationships in a Unique Canonical Representation. In M. Huchard & S. Kuznetsov (Eds.), *CLA: Concept Lattices and their Applications* (pp. 109–122). Moscow, Russia: CEUR Workshop Proceedings. Retrieved from https://hal-lirmm.ccsd.cnrs.fr/lirmm-01354971

Ciobanu, G., Horne, R., & Vuaideanu, C. (2014). Extracting Threshold Conceptual Structures from Web Documents. In N. Hernandez, R. Jäschke, & M. Croitoru (Eds.), *Graph-Based Representation and Reasoning* (pp. 130–144). Cham: Springer International Publishing.

Coste, F., Garet, G., Groisillier, A., Nicolas, J., & Tonon, T. (2014). Automated Enzyme classification by Formal Concept Analysis. In *ICFCA - 12th International Conference on Formal Concept Analysis*. Cluj-Napoca, Romania: Springer. Retrieved from https://hal.inria.fr/hal-01063727

Coulet, A., Domenach, F., Kaytoue, M., & Napoli, A. (2013). Using pattern structures for analyzing ontology-based annotations of biomedical data. In S. Konieczny, N. Maudet, D. Habet, & V. Risch (Eds.), *Septièmes Journées d'Intelligence Artificielle Fondamentale* (pp. 97–106). Aix-en-Provence, France: LSIS and LIF Marseille. Retrieved from https://hal.inria.fr/hal-00922392

Cox, T. F., & Cox, M. A. A. (2000). *Multidimensional scaling*. Chapman and hall/CRC.

Dolques, X., Mondal, K. C., Braud, A., Huchard, M., & Le Ber, F. (2014). RCA as a Data Transforming Method: A Comparison with Propositionalisation. In C. V. Glodeanu, M. Kaytoue, & C. Sacarea (Eds.), *Formal Concept Analysis* (pp. 112–127). Cham: Springer International Publishing.

Ganter, B., & Glodeanu, C. V. (2014). Factors and Skills. In C. V. Glodeanu, M. Kaytoue, & C. Sacarea (Eds.), *Formal Concept Analysis* (pp. 173–187). Cham: Springer International

Publishing.

Grissa, D., Comte, B., Pujos-Guillot, E., & Napoli, A. (2016a). A Hybrid Approach for Mining Metabolomic Data. In *FCA4AI - 5th Workshop "'What can FCA do for Artificial Intelligence?'"* (Vol. 1703). La Haye, Netherlands: CEUR-WS.org. Retrieved from https://hal.archives-ouvertes.fr/hal-01422050

Grissa, D., Comte, B., Pujos-Guillot, E., & Napoli, A. (2016b). A Hybrid Data Mining Approach for the Identification of Biomarkers in Metabolomic Data. In *Proceedings of the Thirteenth International Conference on Concept Lattices and Their Applications, Moscow, Russia, July 18-22, 2016.* (pp. 161–174). Retrieved from http://ceur-ws.org/Vol-1624/paper13.pdf

Guénoche, A., & Mechelen Van, I. (1993). Galois approach to the induction of concepts. In *Categories and concepts : Theoretical views and inductive data analysis, R. Michalski et al. (Eds.), Academic Press* (pp. 287–308).

Howe, J. (2006). The Rise of Crowdsourcing. *Wired Magazine*, *14*(6). Retrieved from http://www.wired.com/wired/archive/14.06/crowds.html

Ignatov, D. I., Kaminskaya, A. Y., Bezzubtseva, A. A., Konstantinov, A. V, & Poelmans, J. (2013). FCA-Based Models and a Prototype Data Analysis System for Crowdsourcing Platforms. In H. D. Pfeiffer, D. I. Ignatov, J. Poelmans, & N. Gadiraju (Eds.), *Conceptual Structures for STEM Research and Education* (pp. 173–192). Berlin, Heidelberg: Springer Berlin Heidelberg.

Ignatov, D. I., Kaminskaya, A. Y., Konstantinova, N., Malyukov, A., & Poelmans, J. (2014). FCA-Based Recommender Models and Data Analysis for Crowdsourcing Platform Witology. In N. Hernandez, R. Jäschke, & M. Croitoru (Eds.), *Graph-Based Representation and Reasoning* (pp. 287–292). Cham: Springer International Publishing.

Ikeda, M., Otaki, K., & Yamamoto, A. (2014). Formal Concept Analysis for Process Enhancement Based on a Pair of Perspectives. In *Proceedings of the Eleventh International Conference on Concept Lattices and Their Applications, Košice, Slovakia, October 7-10, 2014.* (pp. 59–70). Retrieved from http://ceur-ws.org/Vol-1252/cla2014_submission_8.pdf

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, *32*(3), 241–254.

https://doi.org/10.1007/BF02289588

Kis, L. L., Sacarea, C., & Troanca, D. (2016). FCA Tools Bundle - A Tool that Enables Dyadic and Triadic Conceptual Navigation. In *FCA4AI@ECAI*.

Kuznetsov, S. O. (2013). Fitting Pattern Structures to Knowledge Discovery in Big Data. In P. Cellier, F. Distel, & B. Ganter (Eds.), *Formal Concept Analysis* (pp. 254–266). Berlin, Heidelberg: Springer Berlin Heidelberg.

Kuznetsov, S. O., & Makhalova, T. P. (2015). Concept Interestingness Measures: a Comparative Study. In S. Ben Yahia & J. Konecny (Eds.), *Proceedings of the Twelfth International Conference on Concept Lattices and Their Applications, Clermont-Ferrand, France, October 13-16* (pp. 59–72). CEUR-WS.org.

Makhalova, T., Ilvovsky, D., & Galitsky, B. (2015). Pattern Structures for News Clustering. In *Proceedings of the 4th International Conference on What Can FCA Do for Artificial Intelligence? - Volume 1430* (pp. 35–42). Aachen, Germany, Germany: CEUR-WS.org. Retrieved from http://dl.acm.org/citation.cfm?id=2907112.2907117

Mimouni, N., Nazarenko, A., & Salotti, S. (2015). A Conceptual Approach for Relational IR: Application to Legal Collections. In J. Baixeries, C. Sacarea, & M. Ojeda-Aciego (Eds.), *Formal Concept Analysis* (pp. 303–318). Cham: Springer International Publishing.

Neznanov, A. A., Ilvovsky, D. A., & Kuznetsov, S. O. (2013). Fcart: A new fca-based system for data analysis and knowledge discovery. *Contributions to the 11th International Conference on Formal Concept Analysis*, 31–44.

Neznanov, A. A., & Parinov, A. (2014). About universality and flexibility of FCA-based software tools. *CEUR Workshop Proceedings*, *1257*, 59–65.

Neznanov, A. A., & Parinov, A. A. (2016). Unified External Data Access Implementation in Formal Concept Analysis Research Toolbox. In *Proceedings of the Thirteenth International Conference on Concept Lattices and Their Applications, Moscow, Russia, July 18-22, 2016.* (pp. 285–297). Retrieved from http://ceur-ws.org/Vol-1624/paper22.pdf

Pattison, T., & Ceglar, A. (2014). Interaction Challenges for the Dynamic Construction of Partially-Ordered Sets. In *Concept Lattices and Their Applications CLA*.

Priss, U. (2006). Formal Concept Analysis in Information Science. *Annual Review of*

*Information Science and Technology*, *40*(1), 521–543. https://doi.org/10.1002/aris.v40:1

Priss, U. (2013). Using FCA to Analyse How Students Learn to Program. In P. Cellier, F. Distel, & B. Ganter (Eds.), *Formal Concept Analysis* (pp. 216–227). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/https://doi.org/10.1007/978-3-642-38317-5_14

Stanley, R., & Astudillo, H. (2015). A conceptual-KDD Tool for Ontology Construction from a Database Schema. In *Proceedings of the 4th International Conference on What Can FCA Do for Artificial Intelligence? - Volume 1430* (pp. 17–26). Aachen, Germany, Germany: CEUR-WS.org. Retrieved from http://dl.acm.org/citation.cfm?id=2907112.2907115

Tang, M. T., Buzmakov, A., Toussaint, Y., & Napoli, A. (2016). Building a Domain Knowledge Model Based on a Concept Lattice Integrating Expert Constraints. In *Proceedings of the Thirteenth International Conference on Concept Lattices and Their Applications CLA* (Vol. 1624, pp. 349–362). Moscow, Russia. Retrieved from https://hal.inria.fr/hal-01400452

Trabelsi, M., Meddouri, N., & Maddouri, M. (2016). New Taxonomy of Classification Methods Based on Formal Concepts Analysis. In *What can FCA do for Artificial Intelligence FCA4AI@ECAI* (pp. 113–120).

Wille, R. (2005). Formal Concept Analysis as Mathematical Theory of Concepts and Concept Hierarchies. In B. Ganter, G. Stumme, & R. Wille (Eds.), *Formal Concept Analysis: Foundations and Applications* (pp. 1–33). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/11528784_1

Wray, T., & Eklund, P. (2014). Using formal concept analysis to create pathways through museum collections. *CEUR Workshop Proceedings*, *1257*, 9–16.

Zenou, E., & Samuelides, M. (2005). Approche décentralisée des treillis de Galois pour la localisation topologique. In *Revue I3 (Information, Interaction, Intelligence)* (Vol. 5, pp. 85–108).

# Annexe : Comparaison entre plusieurs utilisateurs

## Comparaison des usages applicatifs de plusieurs utilisateurs

Nous effectuons ici quelques expérimentations similaires à celles du Chapitre III (sections III.3 et III.4), cette fois selon la stratégie haute.

Nous nous concentrons sur les connaissances que nous pouvons extraire sur le comportement d'un utilisateur en fonction de sa manière d'utiliser des applications. Nous avons évalué notre proposition sur des données collectées suite à un sondage que nous avons effectué dans notre université sur une population de 28 étudiants. L'objectif de ce sondage était de savoir dans quel contexte les étudiants utilisent des applications (réseaux sociaux, jeux, emails, etc.).

Les mesures définies dans le Chapitre III permettent d'effectuer des comparaisons plus facilement (autrement il faudrait comparer 28 treillis).

Nous avons étudié les résultats des mesures sur l'ensemble des 28 treillis correspondant aux 28 étudiants. Nous avons choisi pour cette section d'utiliser la stratégie à haute dépendance avec $\beta$ = 0,25 pour pouvoir comparer ces treillis de Galois selon chaque mesure d'interprétation.

Nous présentons d'abord, quelques comparaisons des valeurs de similarité conceptuelle et d'impact mutuel impliquant l'application *Gmail*. Nous montrons ensuite une sélection de quelques résultats supplémentaires. Nous indiquons en vert l'étudiant E6 dont nous avions étudié plus spécifiquement le treillis dans le Chapitre III.

La Figure 1 représente la comparaison des valeurs de la similarité d'usage entre les applications *Gmail* et *SMS* pour les 28 étudiants selon la stratégie à haute dépendance. On remarque que tous les étudiants ne sont pas représentés, car certains n'ont pas utilisée *Gmail* ou *SMS*. La valeur moyenne de la similarité conceptuelle est de 35% et on note une similarité conceptuelle élevée de 67% pour l'étudiant E6. La Figure 2 et la Figure 3 présentent respectivement la comparaison de la similarité d'usage entre les applications *Gmail* et *telephone*, et entre les applications *Flappy bird* et *telephone* pour les 28 étudiants selon la stratégie à haute dépendance.

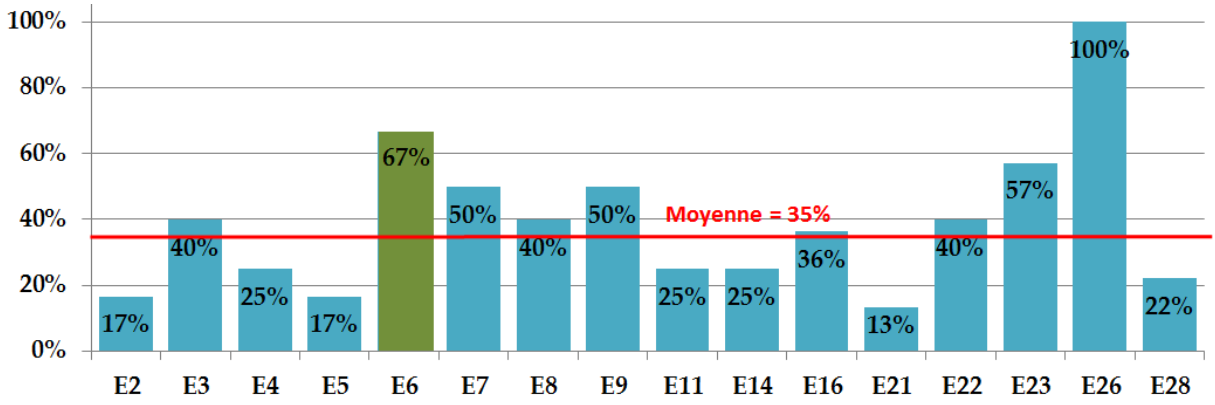



**Figure 1 : Comparaison des valeurs de la similarité conceptuelle (d'usage) entre les applications *Gmail* et *SMS* pour tous les étudiants selon la stratégie à haute dépendance**

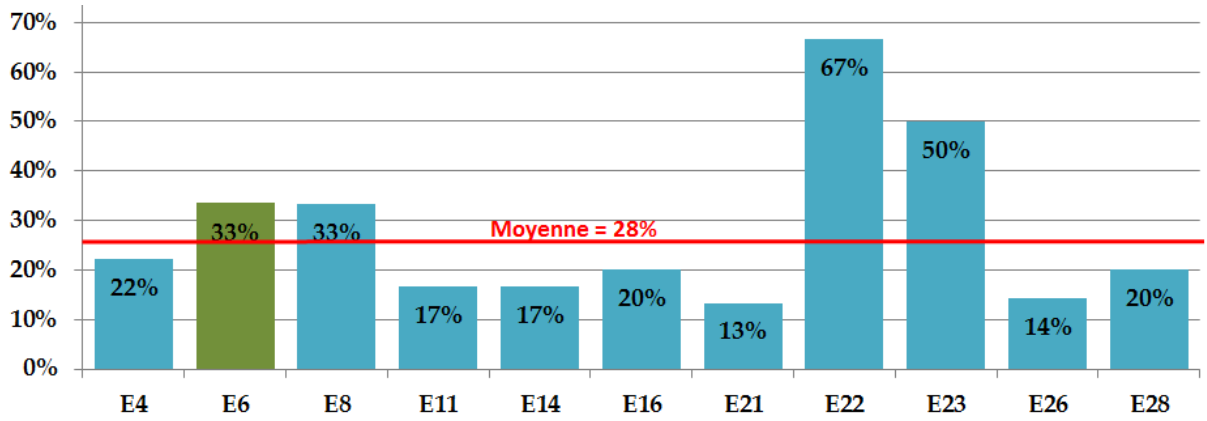



**Figure 2 : Comparaison des valeurs de la similarité conceptuelle (d'usage) entre les applications *Gmail* et *telephone* pour tous les étudiants selon la stratégie à haute dépendance**

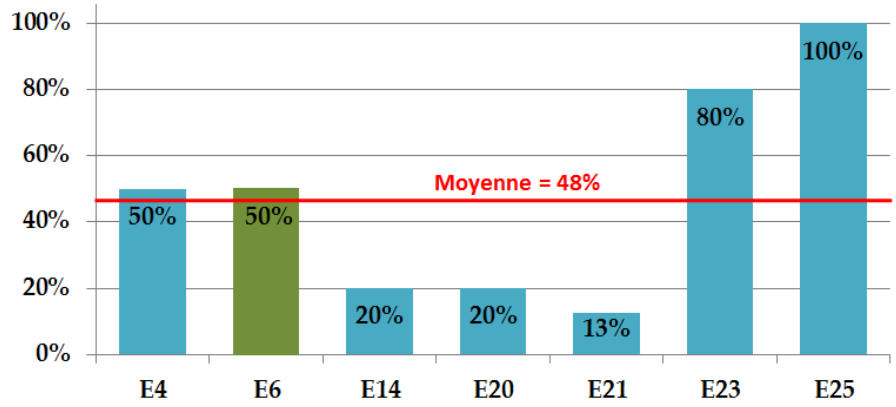



**Figure 3 : Comparaison des valeurs de la similarité conceptuelle (d'usage) entre les applications *Flappy bird* et *telephone* pour tous les étudiants selon la stratégie à haute dépendance**

La Figure 4 représente la comparaison de l'impact mutuel entre l'application *Gmail* et l'élément de contexte *university*, pour les 28 étudiants. L'impact mutuel moyen est de 15% et la valeur d'impact mutuel pour l'étudiant E6 est 25%. Avec la stratégie simpliste dans le Chapitre III, la moyenne des valeurs d'impact mutuel 35% et la valeur de l'impact mutuel pour l'étudiant E6 était de 17%. Cela confirme l'importance de tenir compte de la fréquence d'utilisation des applications dans les différents contextes, comme nous l'avons montré dans la section V.5.1.

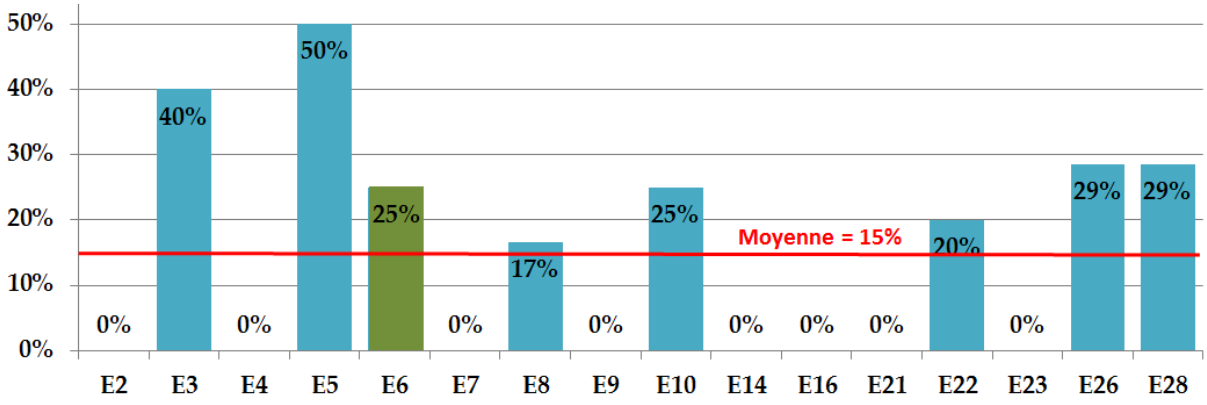



**Figure 4 : Impact mutuel entre *Gmail* et *university* pour tous les étudiants selon la stratégie à haute dépendance**

La Figure 5 représente la comparaison des valeurs de similarité conceptuelle entre les éléments de contexte *3G* et *transportation* ; l'étudiant E6 a disparu pour la comparaison de ces

deux éléments de contexte. On peut noter que la similarité conceptuelle moyenne entre ces deux éléments de contexte (24%) est diminuée par rapport à la moyenne de la similarité conceptuelle pour la stratégie simpliste (29%) du Chapitre III.

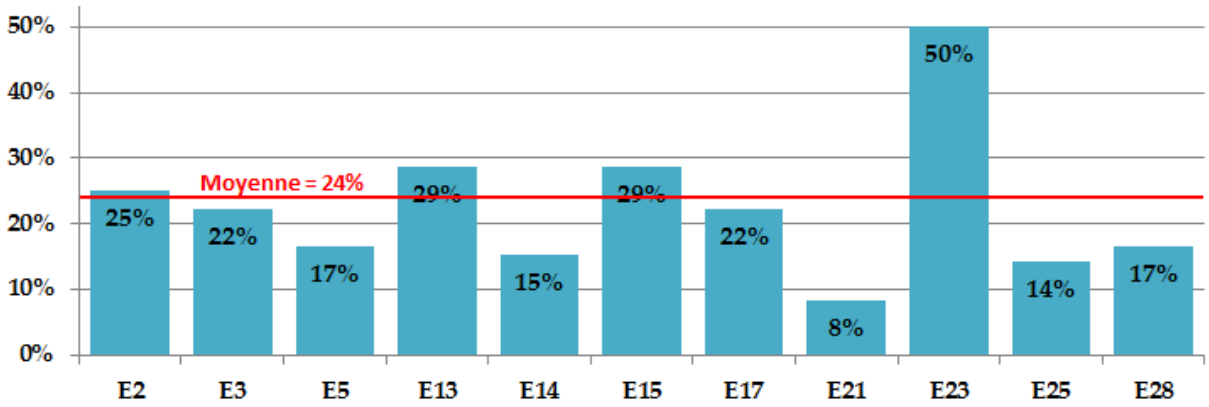



**Figure 5 : Comparaison des valeurs de similarité conceptuelle entre les éléments de contexte *3G* et *transportation* pour tous les étudiants selon la stratégie à haute dépendance**

La Figure 6 représente la comparaison entre la similarité conceptuelle des éléments de contexte *university* et *afternoon*. La similarité conceptuelle moyenne selon la stratégie à haute dépendance est diminuée par rapport la moyenne de similarité conceptuelle selon la stratégie simpliste, par contre la similarité conceptuelle de ces deux éléments de contexte pour l'étudiant E6 augmente de 60% à 67%.

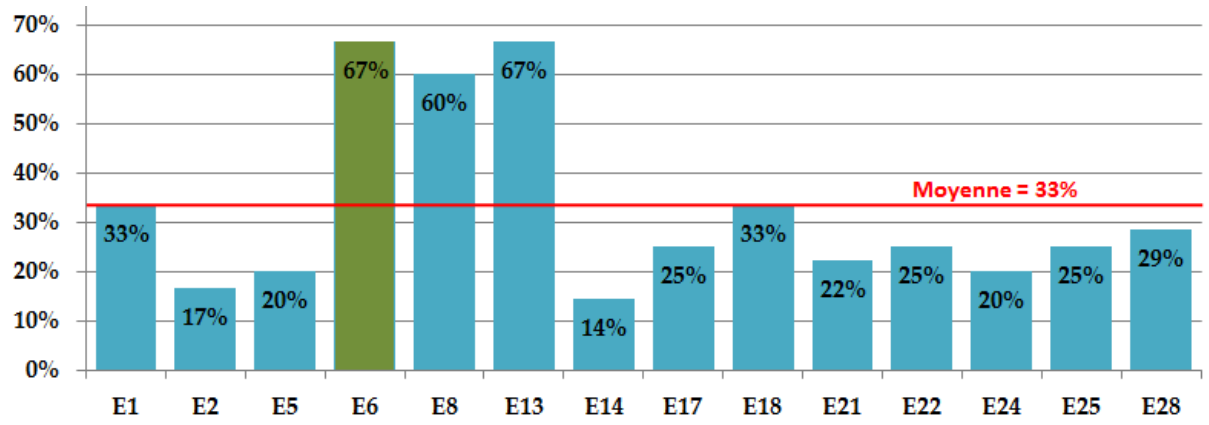



**Figure 6 : Comparaison des valeurs de similarité conceptuelle entre les éléments de contexte *university* et *afternoon* pour tous les étudiants selon la stratégie à haute dépendance**

La Figure 7 représente la comparaison de la similarité conceptuelle entre *university* et *morning*. La moyenne de 27% est bien inférieure à celle de la stratégie simpliste (45%).

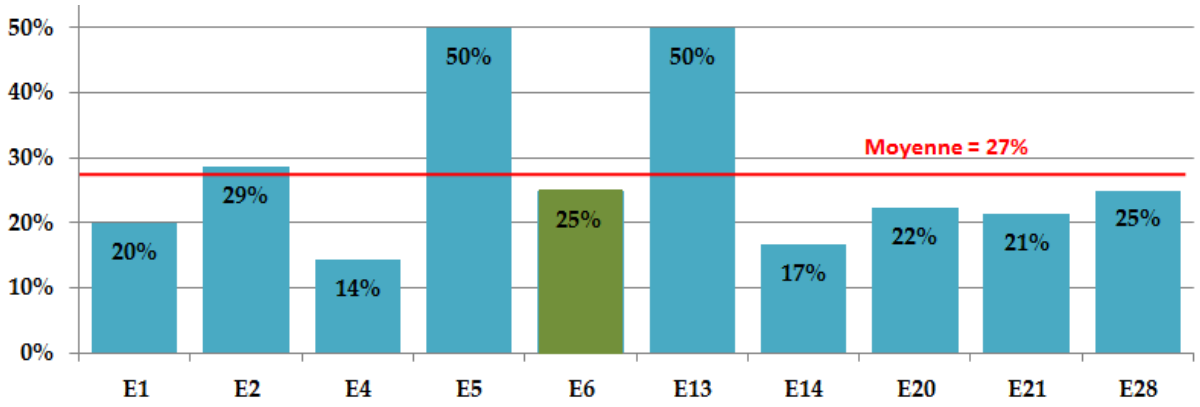



**Figure 7 : Comparaison des valeurs de similarité conceptuelle entre les éléments de contexte *university* et *morning* pour tous les étudiants selon la stratégie à haute dépendance**

Ces résultats confirment l'impact de la prise en compte de la fréquence d'usage des applications. La stratégie à haute dépendance fournit des résultats plus conformes à la réalité. La stratégie simpliste est à réserver pour des relations naturellement binaires.

## Aide à l'utilisation et à l'exploitation de l'Analyse de Concepts Formels pour des non spécialistes de l'analyse de données

**Résumé** – De nombreuses approches ont été élaborées pour extraire des connaissances à partir des données. On distingue traditionnellement l'analyse descriptive et l'analyse prédictive. Nous nous focalisons dans cette thèse sur l'analyse descriptive des données et plus particulièrement sur l'Analyse de Concepts Formels (ACF), qui permet de construire des clusters recouvrants (appelés *concepts formels*) dont la signification est explicite. Il existe une relation d'ordre partiel entre les concepts formels résultant de l'ACF, qui sont organisés en une structure mathématique appelée *treillis de Galois*. Malgré ses nombreux avantages, l'ACF est peu accessible à des utilisateurs non experts de l'analyse de données. En effet, malgré les représentations graphiques des treillis de Galois, ceux-ci restent difficiles à interpréter, notamment lorsque les données sont volumineuses. De plus, la construction des données d'entrée de l'ACF sous la forme d'un *contexte formel* peut être délicate. Pour cela, nous avons proposé une méthodologie d'interprétation des treillis de Galois reposant sur un ensemble de métriques simples, dont les résultats sont présentés sous une forme visuelle aussi intuitive que possible. Nous avons également développé des stratégies pour construire des contextes formels qui, non seulement, dénaturent le moins possible les données initiales, mais permettent aussi de tenir compte des besoins de l'utilisateur en termes de recherche d'information.

**Mots clés** : Analyse de données, Analyse de Concepts Formels, Treillis de Galois, Exploration de connaissances.

---

## Assistance in the use and exploitation of Formal Concepts Analysis for non-specialists in data analysis

**Abstract** – Many data analysis techniques have been developed to extract knowledge from the data. The two traditional approaches are descriptive analysis and predictive. We focus in this thesis on descriptive data analysis, and in particular on Formal Concepts Analysis (FCA). This approach builds overlapping clusters (called *formal concepts*) whose meaning is explicit. There is a partial order relationship between the formal concepts resulting from FCA, which are organized in a mathematical structure called a *Galois lattice*. Despite its advantages, FCA is poorly accessible to users who are not experts in data analysis. Although graphical representations of Galois lattices exist, their interpretation remains difficult for large data. Moreover, the construction of FCA input data, called *formal context*, can be tricky. For this, we have proposed a methodology for Galois lattice interpretation based on a set of simple metrics, the results of which are presented in a visual form as intuitive as possible. We have also developed strategies for constructing formal contexts that not only remain as close to the initial data as possible, but also take into consideration the user's needs in terms of information retrieval.

**Keywords**: Data Analysis, Formal Concepts Analysis, Galois Lattice, Knowledge Exploration.