ATELIER

# QUALITÉ DES DONNÉES ET DES CONNAISSANCES

# 2009

JA, SG

# *QDC 2009*

Actes du 5e Atelier

# Qualité des Données et des Connaissances

En conjonction avec EGC 2009

27 Janvier 2009
Strasbourg, France

**Organisé par**
**Jérôme Azé et Sylvie Guillaume**

# Cinquième Atelier
# Qualité des Données et des Connaissances

27 janvier 2009, Strasbourg, France

## Préface

Après le succès des quatre premières éditions de l'atelier Qualité des Données et des Connaissances en conjonction avec la conférence EGC - 2005 à Paris, 2006 à Lille, 2007 à Namur et 2008 à Nice - nous proposons d'organiser la cinquième édition de l'atelier à l'occasion des journées EGC 2009 à Strasbourg.

Cet atelier se concentre sur les méthodes et techniques d'analyse et d'évaluation de qualité au sens large, tant en fouille de données qu'en gestion des connaissances :

- qualité des données (*nettoyage, méthodologies de prétraitement, métriques d'évaluation et approches algorithmiques, prise en compte de l'hétérogénéité des données*),
- qualité des modèles et de leurs résultats en fouille de données (*évaluation des méthodes et algorithmes, études des mesures d'intérêt, agrégation, post-traitement des résultats*),
- qualité des connaissances (*ontologies, alignements, typologie, visualisation, usages*).

La découverte de connaissances et la prise de décision à partir de données de qualité médiocre (*c'est-à-dire contenant des erreurs, doublons, incohérences, valeurs manquantes, ...*) ont des conséquences directes et significatives pour tous les utilisateurs, quelque soit le domaine d'application, gouvernemental, commercial, industriel ou scientifique. Pour cela, le thème de la qualité des données et des connaissances est devenu un des sujets d'intérêt tout à la fois émergent dans le domaine de la recherche et critique dans les entreprises.

Toutes les applications dédiées à l'analyse des données (*telles que la fouille de données textuelles par exemple*) requièrent différentes formes de préparation des données avec de nombreuses techniques de traitement, afin que les données passées en entrée aux algorithmes de fouille se conforment à des distributions relativement "sympathiques", ne contenant pas d'incohérences, de doublons, de valeurs manquantes ou incorrectes. Seulement, entre la réalité des données disponibles et toute la machinerie permettant leur analyse, un assez vaste fossé demeure.

*In fine*, l'évaluation des résultats issus du processus de traitement des données, est généralement effectuée par un spécialiste (*expert, analyste, ...*). Cette tâche de post-traitement est souvent très lourde et un moyen de la faciliter consiste à aider le spécialiste en lui fournissant des critères de décision sous la forme de mesures de qualité ou d'intérêt des résultats. Ces mesures doivent être conçues afin de combiner deux dimensions : l'une objective liée à la qualité des données, l'autre subjective liée aux intérêts du spécialiste.

Bien que les techniques utilisées en fouille de données et en gestion des connaissances soient très différentes, elles partagent l'objectif de produire des modèles de connaissances pertinents pour les décideurs, avec une préoccupation commune d'évaluation de la qualité des modèles produits. Cet atelier concerne donc tous les domaines qui participent à la chaîne de production des connaissances : données, méthodes de fouille et gestion des connaissances.

Nous encourageons la soumission d'articles de recherche et/ou d'études de cas industriels liés à tous les aspects de la qualité des données, des méthodes de fouille et de gestion des connaissances au sens large. La durée de l'atelier est d'une demi-journée dédiée à des présentations d'articles dans les thèmes d'intérêt indiqués ci-après :

- Métriques de qualité des données
- Techniques de nettoyage et préparation intelligente des données ; détection de données contradictoires, de données isolées, de doublons, d'incohérences, bruit

- Fouille et découverte de *patterns* de non-qualité ou de qualité médiocre
- Transformations, réconciliation, consolidation des données
- Correction d'erreurs
- Métriques de qualité pour les résultats de fouille ou d'analyse
- Métriques de qualité centrées utilisateurs, mesures subjectives et objectives, mesure d'intérêt des règles
- Validation de modèles de fouille de données
- Post-traitement des résultats
- Qualité de modèles de représentation de connaissances et d'ontologies
- Identification d'objets
- Appariement et alignement d'ontologies
- Application à tout type de données (*XML, données transactionnelles, numériques, catégorielles, multimédia, ontologies OWL*) dans différents contextes d'application (*Bioinformatique, Marketing, e-Commerce, etc*)

Jérôme Azé et Sylvie Guillaume
Organisateurs de QDC 2009

# Comités

## Comité d'organisation

- Jérôme Azé, LRI, Université Paris-Sud 11
- Sylvie Guillaume, LIMOS, Université Blaise Pascal

## Comité de programme

- Alexandre Aussem, LIESP, Université Lyon 1
- Jérôme Azé, LRI, Université Paris-Sud 11
- Laure Berti-Equille, IRISA - Université de Rennes 1
- Julien Blanchard, Polytech'Nantes, Laboratoire d'Informatique de Nantes Atlantique
- Marc Boullé, Orange Labs, TEACH/EASY
- Henri Briand, Polytech'Nantes, Laboratoire d'Informatique de Nantes Atlantique
- Martine Cadot, LORIA, Nancy
- Jean Diatta, IREMIA, Université de la Réunion
- Thanh-Nghi Do, Telecom Bretagne
- Béatrice Duval, LERIA - UFR Sciences
- Sylvie Guillaume, LIMOS, Université Blaise Pascal
- Fabrice Guillet, Polytech'Nantes, Laboratoire d'Informatique de Nantes Atlantique
- Xuan Hiep Huynh, Université de Cantho, Vietnam
- Ali Khenchaf, Laboratoire E3I2-EA3876, Brest
- Yves Kodratoff, CNRS, LRI, Université Paris-Sud 11
- Pascale Kuntz, Polytech'Nantes, Laboratoire d'Informatique de Nantes Atlantique
- Stéphane Lallich, ERIC, Université Lumière - Lyon 2
- Ludovic Lebart, CNRS, Département SES, ENST
- Alain Léger, Orange - France Telecom R&D
- Vincent Lemaire, Orange Labs, TEACH/EASY
- Philippe Lenca, Telecom Bretagne
- Israel-César Lerman, IRISA - Université de Rennes 1
- Engelbert Mephu Nguifo, LIMOS, Université Blaise Pascal
- Patrick Meyer, Telecom Bretagne
- Sorin Moga, Telecom Bretagne
- Annie Morin, IRISA - Université de Rennes 1
- Alain Morineau, INRIA Rocquencourt
- Amédéo Napoli, LORIA, Nancy
- Nathalie Pernelle, LRI, Université Paris-Sud 11
- Pascal Poncelet, LGI2P - Ecole des Mines d'Alès
- Ricco Rakotomalala, ERIC, Université Lumière - Lyon 2
- Chantal Reynaud, LRI, Université Paris-Sud 11
- Ansaf Salleb-Aouissi, Columbia University, Center for Computational Learning Systems
- André Totohasina, Université Nord Madagascar Antsiranana
- Benoit Vaillant, Cohéris - SPAD
- Christel Vrain, LIFO - Orléans

# Table des matières

**Analyse de concepts formels pour la construction d'ontologies à partir de textes : la question du corpus**
Thibault Mondary, Sylvie Després **1**

**Critères d'évaluation contextuelles pour le traitement automatique**
Adrien Revault d'Allonnes, Jérôme Besombes **9**

**Généralisation d'une propriété de monotonie de la *all-confidence* pour l'extraction de motifs intéressants non fréquents**
Yannick Le Bras, Philippe Lenca, Stéphane Lallich, Sorin Moga **17**

**Introduction de l'élagage pour l'extraction de règles d'association de classe sans génération de candidats**
Emna Bahri, Stéphane Lallich **25**

**Utilisation des règles d'association pour l'amélioration de la qualité des bases de données**
Soumaya Ben Hassine, Delphine Clément **33**

# Analyse de concepts formels pour la construction d'ontologies à partir de textes : la question du corpus

Thibault Mondary*, Sylvie Després*

*LIPN - UMR 7030
CNRS - Université Paris 13
F-93430 Villetaneuse, France
prenom.nom@lipn.univ-paris13.fr,

**Résumé.** Nous nous intéressons à l'évaluation d'une méthode permettant l'acquisition de hiérarchies à partir de textes. L'analyse de concepts formels semble donner de bons résultats dans les travaux de Cimiano et al. (2005) et plus récemment Bendaoud et al. (2007). Nous montrons que sur certains corpus les résultats semblent moins utilisables. Est-ce dû au style du texte ? Au domaine à modéliser ? Après une présentation de la méthode utilisée pour construire nos contextes formels, nous analysons les résultats obtenus sur différents corpus. Plusieurs variantes de construction des contextes formels sont proposées, axées sur la sélection des objets, avec pour objectif d'améliorer l'utilisabilité du treillis dans un contexte de conceptualisation semi-automatique.

## 1 Introduction

Le recours aux textes pour la construction d'ontologies est légitimé par l'hypothèse qu'ils représentent des connaissances partagées stabilisées d'un domaine. Ils ne se substituent pas à un expert du domaine, mais peuvent lui faciliter la tâche de modélisation s'ils sont accompagnés d'outils qui rendent saillantes les informations désirées. Dans cet article nous nous intéressons à l'évaluation des résultats obtenus en utilisant l'analyse de concepts formels (ACF), dans la lignée de (Cimiano et al., 2005) et (Bendaoud et al., 2007). L'ACF est ici destinée à aider le travail de conceptualisation par l'expert lors de la construction d'une ontologie. Les concepts formels et leurs relations de subsomption sont considérées comme des suggestions de concepts et de relations de hiérarchie pour l'ontologie. Le but est d'obtenir l'ACF la plus "suggestive" possible, c'est-à-dire celle qui constitue le meilleur guide pour le travail de conceptualisation. Après avoir rappelé les principes de l'ACF appliquée au texte, nous présentons les résultats des expérimentations menées sur des corpus de genres et de styles différents. Dans une troisième partie nous tentons de déterminer des conditions d'applicabilité de la méthode, et proposons des pistes pour remédier aux problèmes rencontrés. Ceux-ci sont notamment la taille des treillis, le manque de relations de subsomptions utilisables, la présence de bruit dans les objets et les attributs du contexte formel qui nuisent à l'utilisabilité du treillis comme guide pour le travail de conceptualisation.

## 2 Analyse de concepts formels appliquée aux textes

L'analyse de concepts formels (Ganter et Wille, 1999) est une méthode de classification symbolique qui vise à découvrir tous les regroupements possibles d'éléments ayant des traits en commun. La hiérarchie résultante qui regroupe des objets partageant les mêmes propriétés est appelée treillis de concepts. La notion centrale de l'ACF est le **contexte formel**. C'est un triplet $\mathbb{K} = (G, M, I)$ où $G$ est un ensemble d'objets, $M$ un ensemble d'attributs et $I$ une relation binaire entre $G$ et $M$ appelée relation d'incidence de $\mathbb{K}$ et vérifiant $I \subseteq G \times M$. Un couple $(g, m) \in I$ signifie que l'objet $g \in G$ possède l'attribut $m \in M$. Soit $\mathbb{K}$ un contexte formel. Pour tout $A \subseteq G$ et $B \subseteq M$, on définit $A' = \{m \in M | \forall g \in A, (g, m) \in I\}$ et $B' = \{g \in G | \forall m \in B, (g, m) \in I\}$. $A'$ est l'ensemble des attributs communs à tous les objets de $A$ et $B'$ est l'ensemble des objets possédant tous les attributs de $B$. Nous pouvons maintenant définir un **concept formel** comme étant un couple $(A, B)$ tel que $A \subseteq G$ et $B \subseteq M$, vérifiant $A' = B$ et $B' = A$. $A$ est l'**extension** du concept formel et $B$ son **intension**. L'ensemble des concepts formels associés au contexte $\mathbb{K} = (G, M, I)$ est noté $\mathfrak{B}(G, M, I)$. Les concepts de $\mathfrak{B}(G, M, I)$ sont ordonnés par une relation de subsomption $\sqsubseteq$ telle que $(A1, B1) \sqsubseteq (A2, B2) ssi A1 \subseteq A2$, ou de façon duale $B2 \subseteq B1$.

Cimiano et al. (2005) et ultérieurement Bendaoud et al. (2007) proposent de construire le contexte formel en s'appuyant sur les idées de Hindle (1990). En partant d'une analyse syntaxique des phrases obtenue automatiquement à l'aide

| | paralyser | paralyser-ABLE | refuser | refuser-ABLE |
|---|---|---|---|---|
| grève | × | | | × |
| entreprise | | × | | |
| travailleur | | | × | |
| dirigeant | | | × | |
| travail | | | | × |

TAB. 1 – *Contexte formel exemple*

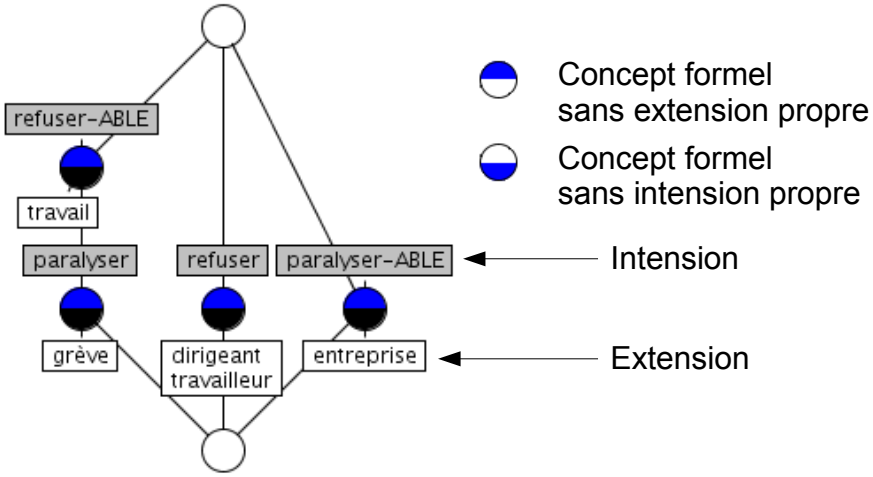


FIG. 1 – *Treillis exemple*

d'un analyseur syntaxique, ils construisent le contexte formel en extrayant de cette analyse les noms et les verbes. Les objets du contexte formel sont les noms apparaissant en tant que sujet, complément d'objet ou locution prépositionnelle. Les attributs du contexte formel sont les verbes de la phrase considérée, complétés du suffixe -ABLE lorsqu'ils proviennent d'un complément ou d'une locution.

Prenons par exemple le texte suivant : *Une grève paralyse l'entreprise. Les travailleurs refusent le travail dominical. Les dirigeants refusent la grève.* En ne gardant que la tête lemmatisée des sujets et des compléments, nous obtenons le contexte de la table 1. Le treillis résultant est présenté figure 1, il est destiné à l'expert. Ce dernier étiquette les concepts formels qu'il juge pertinents pour le domaine.

# 3 Expérimentations

Pour analyser l'applicabilité de l'ACF, nous considérons trois corpus. Les deux premiers corpus sont composés de documents en français rédigés par l'Organisation Internationale du Travail[1] : les conventions C87, C98 et leurs jurisprudences. Les conventions internationales du travail ont le statut juridique de traités internationaux. C87 s'intitule "Convention sur la liberté syndicale et la protection du droit syndical, 1948" et C98 est la "Convention sur le droit d'organisation et de négociation collective, 1949". Nous distinguons le corpus C87_C98 composé uniquement des deux conventions, 3605 mots, et le corpus C87_C98_Jurisp incluant également les jurisprudences, 99966 mots. Le troisième corpus est constitué de recettes de cuisines récoltées sur internet, en français également, pour 3644 mots.

La figure 2 montre le processus de construction du contexte formel. Du corpus nous extrayons en parallèle les candidats termes et les relations de dépendance syntaxique. L'analyseur syntaxique utilisé pour le français dans ce travail est Syntex (Bourigault et al., 2005). Il a obtenu de bons résultats lors de la campagne EASY[2] et fournit une analyse syntaxique en dépendances bien adaptée à notre objectif de repérage des sujets/compléments, sans utiliser de connaissances externes. YaTeA (Sophie Aubin (2006)) est l'extracteur de termes choisi. Il exploite à la fois des patrons d'extraction et une désambiguïsation endogène par la redondance pour extraire des groupes nominaux candidats-termes.

[1] http://www.ilo.org/ilolex/french/
[2] http://w3.erss.univ-tlse2.fr:8080/index.jsp?perso=bourigault&subURL=syntex.html

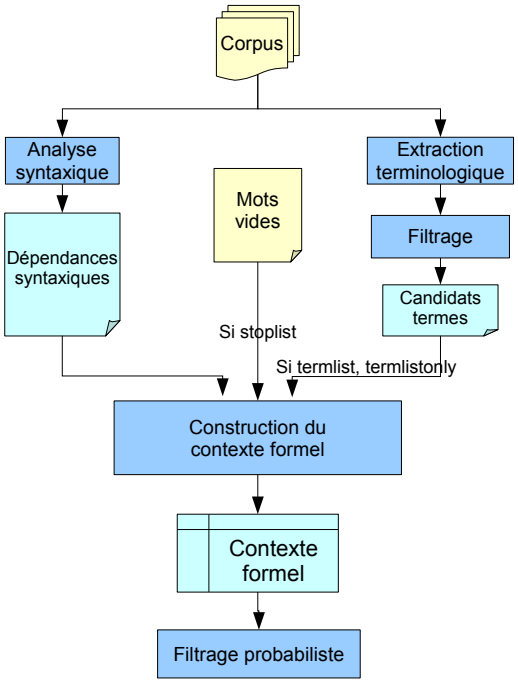


FIG. 2 – *Construction du contexte formel*

Nous construisons le contexte avec les dépendances sujet-verbe(+préposition si elle existe) et complément-verbe(+préposition si elle existe). Nous proposons six variantes de construction du contexte formel, qui diffèrent principalement sur le choix des objets à inclure dans le contexte. L'intuition sous-jacente est que les termes sont de meilleurs descripteurs sémantiques que les mots, mais qu'il y a plus de redondance (et donc plus de regroupements possibles) avec les mots.

**usuelle** : Inclure dans le contexte formel en tant qu'objets les têtes lemmatisées des sujets et des compléments d'objet et en tant qu'attributs leurs verbes associés (munis des prépositions le cas échéant), à condition que ces têtes soient des noms.

**stoplist** : Idem à la méthode usuelle, à condition que les candidats objets et attributs n'appartiennent pas à une liste de mots vides fournie par l'utilisateur et ne contiennent pas de chiffre.

**termlist** : Idem à la méthode usuelle, en tentant d'apparier les candidats objets avec une liste de termes lemmatisés fournie en entrée par l'utilisateur. Si l'appariement échoue, inclure seulement la tête. Par exemple dans l'exemple de la première partie, l'objet "travail" serait remplacé par "travail dominical".

**termlist+stoplist** : Comme ci-dessus, mais en rejetant au préalable (avant de tenter un appariement) les candidats objets et attributs qui appartiennent à une liste de mots vides ou qui contiennent des chiffres.

**termlistonly** : N'inclure dans le contexte que les sujets ou compléments d'objet qui s'apparient avec un terme de la liste fournie en entrée. Les verbes associés aux sujets ou compléments sont inclus dans le contexte comme attributs.

**termlistonly+stoplist** : Comme ci-dessus, mais filtrer les verbes selon une liste de mots vides. Les candidats attributs qui appartiennent à cette liste ou qui contiennent des chiffres sont rejetés.

Dans les méthodes utilisant la liste de termes en entrée une étape de filtrage de la liste de candidats termes est nécessaire. Cette étape est réalisée à l'aide de patrons (dans les expériences, nous utilisons notamment les patrons "supprimer les candidats termes qui contiennent un nombre", ceux qui contiennent "paragraphe ou article"...). La fabrication des patrons nécessite une étape de lecture manuelle des candidats-termes qui est moins coûteuse en temps qu'un filtrage exhaustif. Pour les recettes, nous éliminons 37 candidats-termes, ceux contenant des nombres (par exemple "ébullition 1 l"). Pour les corpus C87_C98, 13 candidats-termes sont exclus. Les patrons appliqués au corpus C87_C98_Cases nous suppriment 767 candidats-termes.

Un filtrage statistique probabiliste *a posteriori* est prévu dans la méthode, mais n'est pas appliqué dans les expériences présentées ici. Dans l'état actuel de notre implémentation nous reprenons celui utilisé par Cimiano et al. (2005), à savoir la probabilité conditionnelle $P(obj|attr) = f(obj, attr)/f(attr)$ pour chaque couple $(obj, attr)$ du contexte formel. $f(obj, attr)$ représente le nombre d'occurrences du couple $(obj, attr)$ dans le contexte[3] et $f(attr)$ le nombre d'occur-

[3]Ce n'est pas forcément une valeur booléenne, mais pour le calcul des concepts formels cela en devient une

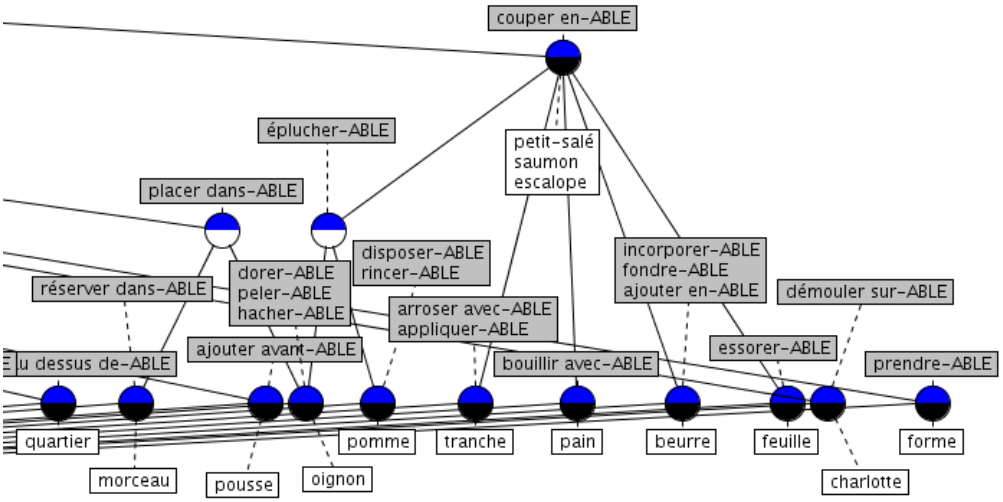


FIG. 3 – *Treillis pour les recettes, méthode stoplist*

rences de l'attribut $attr$. Ce calcul donne pour chaque couple une probabilité, dont il reste à déterminer un seuil pour décider si ce couple doit être conservé dans le contexte formel filtré.

La table 2 donne une analyse quantitative de l'application de l'ACF sur les trois corpus en utilisant la méthode usuelle. Nous pouvons remarquer que le corpus recettes donne deux fois plus de concepts que le corpus C87_C98, pour un nombre de mots similaire. Rapporté au nombre de mots, nous remarquons que le gros corpus C87_C98_Cases n'est pas beaucoup plus productif en concepts que C87_C98, dix fois plus petit. Un aperçu du treillis obtenu pour le corpus recettes est proposé figure 3. La figure 4 montre une partie du treillis pour le corpus C87_C98.

Les figures 5 et 6 présentent une évaluation quantitative de l'impact des variantes de construction sur le treillis. La première figure représente, pour chaque variante, le nombre d'objets, d'attributs et de concepts formels obtenu. Nous pouvons remarquer que les variantes `xxx+stoplist` diminuent dans chaque cas le nombre d'éléments ; pour C87_C98, la diminution engendrée par les mots vides est plus importante que pour les recettes. Ce phénomène est notamment dû à une erreur de l'analyseur syntaxique qui ajoute les numéros des articles des conventions comme objets. Les variantes `termlist(+stoplist)` augmentent le nombre d'objets et de concepts. Ceci est dû au fait que les variantes `usuelles (+stoplist)` travaillent sur des mots et regroupent des termes composés qui ne le seraient pas avec les variantes `termlist(+stoplist)`. Par exemple si dans une phrase nous avons le terme "organisation de travailleurs" en sujet et dans une autre phrase le mot "organisation" en sujet également, la variante `termlist` produira deux objets distincts tandis que les variantes `usuelle(+stoplist)` regroupera les attributs de "organisation de travailleurs" et de "organisation" sous le même objet "organisation". Le nombre d'attributs ne varie pas en passant des variantes `usuelles(+stoplist)` aux variantes `termlist(+stoplist)`, ce qui est normal. Les variantes `termlistonly(+stoplist)`, en incluant uniquement les termes proposés par l'utilisateur, font diminuer notablement tant le nombre d'objets et d'attributs que le nombre de concepts. L'ajout des termes pour le corpus recettes est plus productif en objets que pour le corpus C87_C98, cela s'explique par le fait que les termes sont plus nombreux pour les recettes et s'apparient mieux, c'est-à-dire qu'ils sont davantage présents en tant que sujets ou compléments.

La figure 6 propose une mesure du nombre de concepts selon leur niveau dans chaque treillis. Les concepts de niveau 1 représentent ceux qui sont directement au-dessus de `bottom`. Sur cette figure on observe l'influence de l'injection des termes sur les regroupements : les variantes `termlistonly(+stoplist)` réduisent légèrement la hauteur du treillis. Il est intéressant de noter que pour la méthode usuelle le corpus recettes produit quasiment trois fois plus de concepts de niveau deux que pour le corpus C87_C98. De plus le rapport $Niveau1/Niveau2$ donne $4$ pour C87_C98 contre $2.48$ pour les recettes, ce qui indique que le treillis du corpus recettes propose quasiment deux fois plus de relations hiérarchiques que l'autre par objet. L'influence des mots vides sur le corpus C87_C98 laisse voir que ses regroupements sont essentiellement vides de sens, ce qui est vrai dans une moindre proportion pour le corpus recettes.

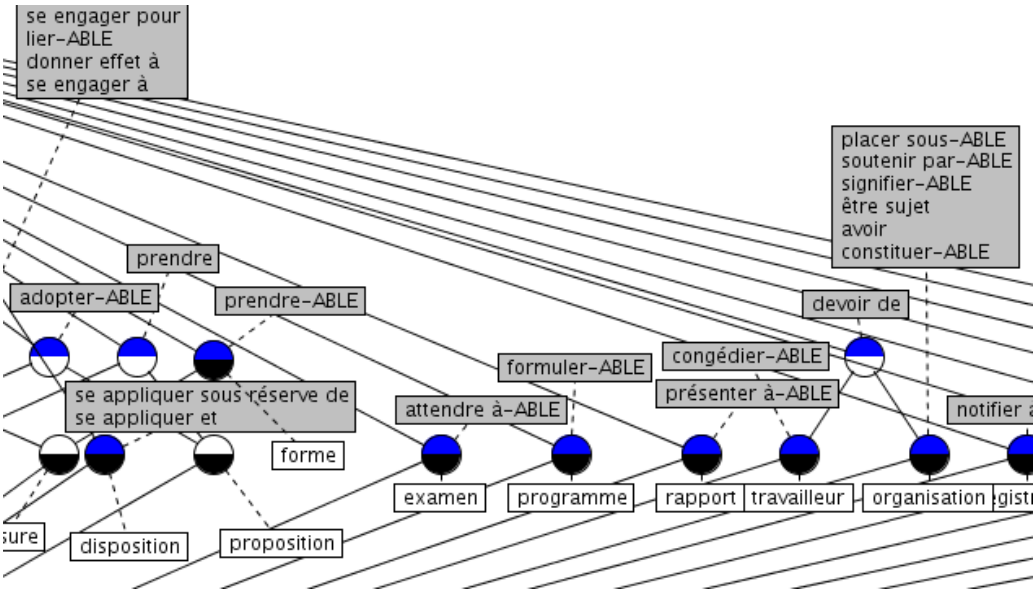


FIG. 4 – *Treillis pour le corpus C87_C98, méthode stoplist*

| | C87_C98 | C87_C98_Jurisp | Recettes |
|---|---|---|---|
| Nombre d'objets du contexte formel | 70 | 688 | 122 |
| Nombre d'attributs du contexte formel | 74 | 1379 | 131 |
| Nombre de concepts formels | 54 | 1564 | 107 |
| *Par rapport au nombre de mots du corpus, base 100* | 1.5 | 1.56 | 2.94 |
| Profondeur du treillis | 3 | 8 | 3 |

TAB. 2 – *Comparaison quantitative, méthode usuelle*

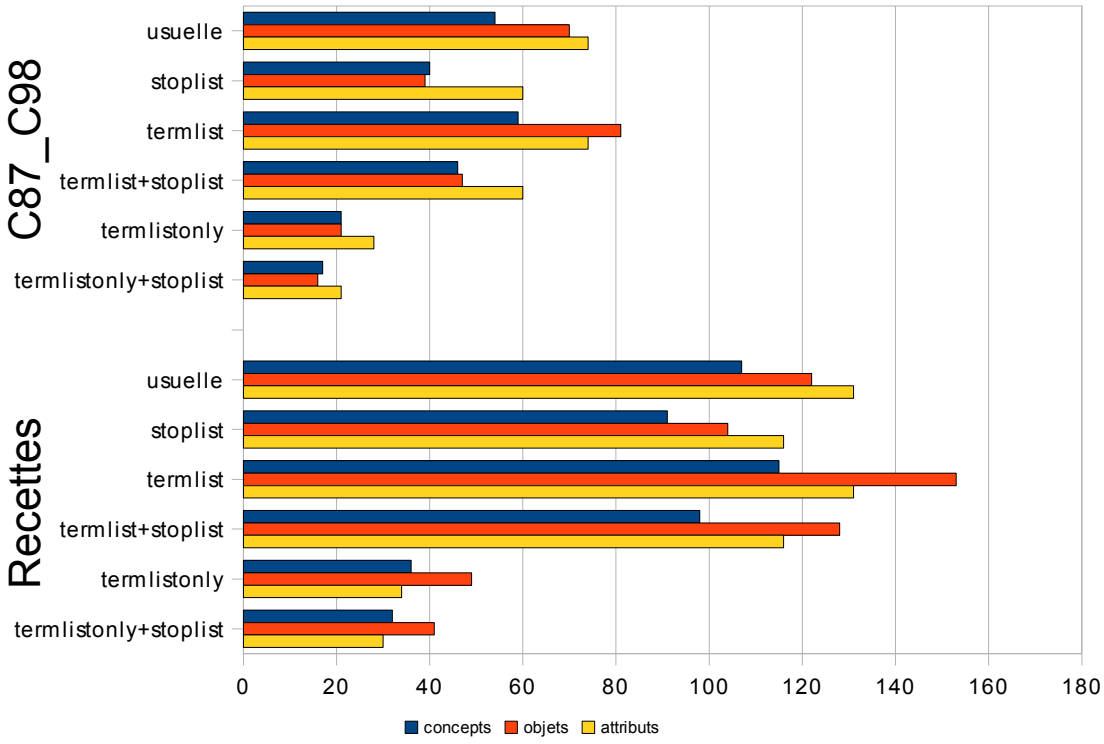


FIG. 5 – *Concepts, objets et attributs, en occurrences pour C87_C98 et Recettes*

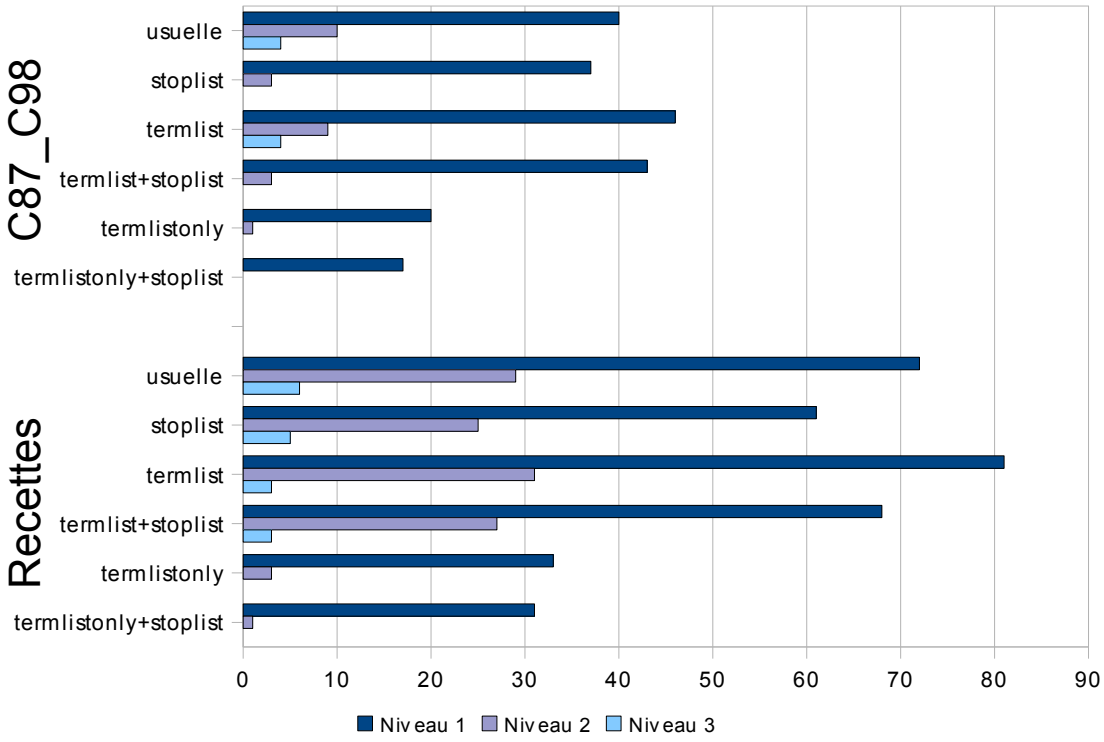


FIG. 6 – *Concepts de niveau 1, 2 et 3 (le maximum), en occurrences pour C87_C98 et Recettes*

| | C87_C98 | C87_C98_Jurisp | Recettes |
|---|---|---|---|
| Domaine | liberté syndicale | liberté syndicale | cuisine |
| Nombre de mots | 3605 | 99966 | 3644 |
| Nombre de phrases | 158 | 4533 | 299 |
| Phrase la plus longue (en mots) | 237 | 292 | 58 |
| Nombre moyen de mots par phrase | 22,82 | 22,05 | 12,18 |
| Nombre de verbes | 430 | 14137 | 666 |
| Dont modaux | 36 | 1195 | 6 |

TAB. 3 – *Caractéristiques des corpus au niveau des occurrences*

# 4 Interprétations, propositions et conditions d'applicabilité

Le treillis des recettes semble utilisable par l'expert du domaine. On peut y lire que le concept dont l'extension est `{pomme}` a pour intension `{disposer-ABLE, rincer-ABLE, éplucher-ABLE, couper-en-ABLE}`. Il est subsummé par le concept sans extension directe (il n'existe pas dans le texte d'objet ayant uniquement les attributs `éplucher-ABLE` et `couper-en-ABLE`) d'intension `{éplucher-ABLE, couper-en-ABLE}`. C'est à l'expert de décider comment étiqueter les concepts formels et quelles relations de subsomptions conserver dans l'ontologie. Comment interpréter le fait qu'un aliment qui est pelable et hachable soit aussi épluchable et coupable ?

*A contrario*, le treillis des conventions est très peu utilisable par l'expert : les intensions sont peu informatives (`prendre, placer-sous-ABLE, soutenir-par-ABLE...`) et la seule hiérarchie est constituée de verbes de modalité (en particulier `devoir`) vides de sens. Le treillis du gros corpus comporte sans doute plus de relations intéressantes mais est tellement fourni qu'il en devient illisible et inexploitable.

En considérant les tables 3 et 4, nous remarquons que les phrases du corpus recettes sont en moyenne presque deux fois plus courtes que celles du corpus C87_C98. Empiriquement, l'absence de modaux et la brièveté des phrases semble être un indicateur de la productivité en concepts du treillis. Les phrases du corpus C87_C98 sont beaucoup plus complexes, avec des ellipses et des modalités (par exemple : *Des mesures appropriées aux conditions nationales doivent, si nécessaire, être prises pour encourager et promouvoir le développement et l'utilisation les plus larges de procédures de négociation volontaire de conventions collectives entre les employeurs et les organisations d'employeurs d'une part, et*

| | C87_C98 | C87_C98_Jurisp | Recettes |
|---|---|---|---|
| Taille du vocabulaire lemmatisé | 868 | 5858 | 1198 |
| Fréquence moyenne du voc. lemmatisé | 4,15 | 17,06 | 3,04 |
| Taille du vocabulaire fléchi | 1102 | 9686 | 1422 |
| Fréquence moyenne du voc. fléchi | 3,27 | 10,32 | 2,56 |
| Nombre de candidats termes | 122 | 4064 | 199 |
| Nombre de termes retenus | 109 | 3297 | 162 |

TAB. 4 – *Caractéristiques des corpus au niveau du vocabulaire*

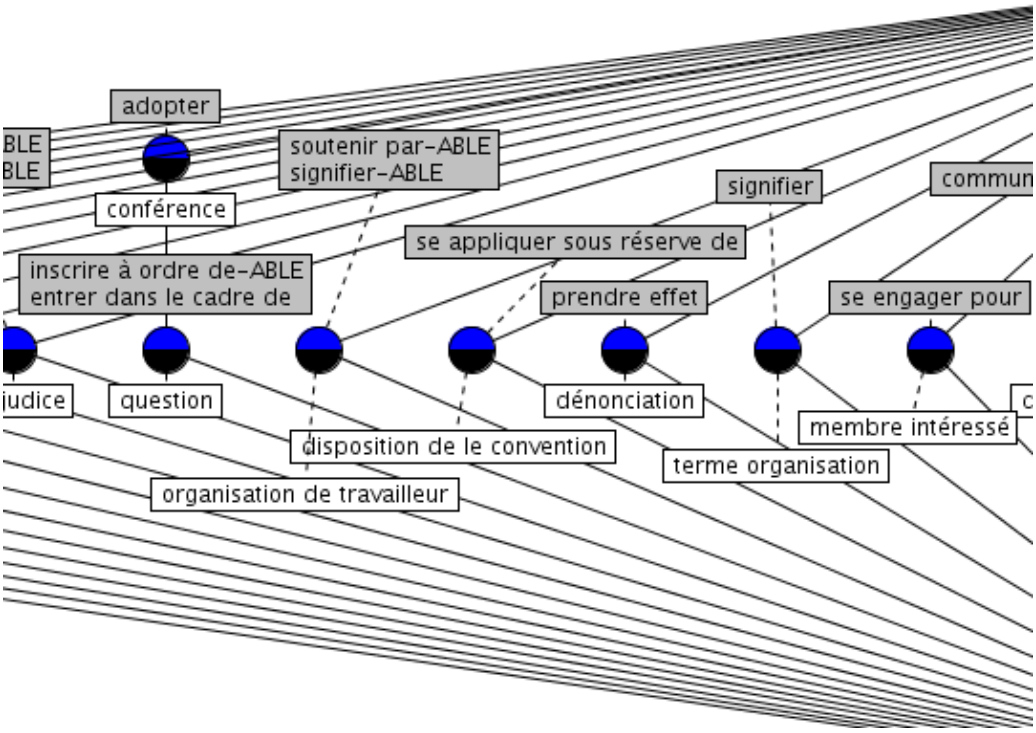


FIG. 7 – *Treillis pour le corpus C87_C98, méthode termlist+stoplist*

*les organisations de travailleurs d'autre part, en vue de régler par ce moyen les conditions d'emploi.*), ce qui soulève la question de la performance de l'analyseur syntaxique et celle concernant le traitement des modalités : une analyse plus fine que les simples dépendances sujet/verbe/complément semble nécessaire dans ce cas.

Nous avons remarqué après une lecture détaillée du corpus C87_C98 que seulement un tiers du texte est en rapport avec notre tâche de modélisation des violations, le reste servant de cadre à la ratification des conventions. Une ACF sur cette partie donne moins de bruit dans le treillis : il est donc nécessaire de bien sélectionner le corpus que nous voulons utiliser pour modéliser. Cette sélection pourrait être effectuée en analysant le vocabulaire utilisé et en le comparant avec une ressource existante, mais dans le domaine juridique cette procédure est délicate, du fait que les termes du domaine relèvent également du sens commun (par exemple "travail", "organisation") et devrait plutôt être envisagée manuellement, voire à l'aide d'un algorithme d'apprentissage supervisé.

Augmenter la taille du corpus afin d'augmenter la redondance pour extraire plus de concepts est une solution qui nécessite de mettre en place des stratégies de filtrage et d'exploration du treillis. Le filtrage peut s'envisager *a priori* dans la sélection des attributs et des objets intéressants (méthodes `termlistonly(+stoplist)`) ou *a posteriori* en utilisant des méthodes statistiques (comme dans Cimiano et al. (2005)). L'exploration du treillis nécessite des outils permettant une approche incrémentale de la construction, dans laquelle le candidat-treillis est visualisé au moment de la construction du contexte.

Les méthodes de construction `termlistonly` et `termlistonly+stoplist` diminuent le bruit du treillis au détriment des regroupements. Les méthodes `termlist` et `termlist+stoplist` sont celles qui ajoutent de la connaissance dans l'interprétabilité des résultats et maximisent les concepts dans le treillis, au détriment toutefois des regroupements (mais moins que pour `termlistonly(+stoplist)`. La figure 7 montre l'augmentation de l'interprétabilité

du treillis : l'extension "organisation de travailleur[4]" est plus parlante pour l'expert du domaine que simplement "organisation", idem pour "disposition de le convention".

# 5 Conclusion et perspectives

Nous avons tenté au travers de cet article de soulever la question de l'applicabilité de l'analyse de concepts formels dans la tâche de construction semi-automatique de hiérarchies ontologiques à partir de textes. Les expériences menées sur des corpus de tailles comparables mais de genres et de styles différents nous ont permis de pointer les limites de l'approche sujet/verbe/complément. La taille des phrases et la modalité semblent être des indicateurs de l'interprétabilité du treillis par l'expert. Nous avons proposé une méthode permettant de contenir la taille du treillis et d'augmenter son interprétabilité, mais se pose le problème du coût en temps du filtrage *a priori*. Notre priorité est maintenant d'explorer des méthodes permettant d'optimiser cette étape, en regroupant les synonymes par exemple. Nous nous posons également la question de la pertinence d'inclure dans un même contexte formel les relations sujet/verbe et complément/verbe. Ne sont-elles pas complémentaires pour la construction des concepts qui partagent des attributs de nature différentes ?

# Références

Bendaoud, R., M. Rouane Hacene, Y. Toussaint, B. Delecroix, et A. Napoli (2007). Construction d'une ontologie à partir d'un corpus de textes avec l'acf. In F. Trichet (Ed.), *Actes des 18eme journées francophones d'ingénierie des connaissances (IC2007)*. Cépaduès.

Bourigault, D., C. Fabre, C. Frérot, M.-P. Jacques, et S. Ozdowska (2005). Syntex, analyseur syntaxique de corpus. In *Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles*, Dourdan, France.

Cimiano, P., A. Hotho, et S. Staab (2005). Learning concept hierarchies from text corpora using formal concept anaylsis. *Journal of Artificial Intelligence Research (JAIR) 24*, 305–339.

Ganter, B. et R. Wille (1999). *Formal Concept Analysis*. Berlin : Springer.

Hindle, D. (1990). Noun classification from predicate-argument structures. In *Proceedings of the 28th annual meeting on Association for Computational Linguistics*, Morristown, NJ, USA, pp. 268–275. Association for Computational Linguistics.

Sophie Aubin, T. H. (2006). Improving term extraction with terminological resources. In T. Salakoski, F. Ginter, S. Pyysalo, et T. Pahikkala (Eds.), *Advances in Natural Language Processing 5th International Conference on NLP, FinTAL 2006*, Turku, Finland, pp. 380–387. Springer.

# Summary

This paper deals about semi-automatic induction of conceptual hierarchies from texts. Formal concept analysis seems to give good results in Cimiano et al. (2005) and more recently in Bedaoud et al. (2007). We show that on certain corpora (in particular from International Labour Office) results are less usable. Why ? We will test fca on different corpora and try to find why the results aren't as good as expected, focusing on a semi-automatic conceptualization task.

[4]Les extensions sont lemmatisées

# Critères d'évaluation contextuelle pour le traitement automatique

Adrien Revault d'Allonnes*, Jérôme Besombes**

*Laboratoire d'Informatique de Paris VI – LIP6
104 avenue du président Kennedy, 75016 Paris,
Adrien.Revault-d'Allonnes@lip6.fr,
**Office National d'Études et de Recherche Aérospatiales – ONERA
ONERA - BP72, 29 avenue de la Division Leclerc, 92322 Châtillon Cx
Jerome.Besombes@onera.fr

**Résumé.** Les traitements automatiques de gros volumes de données deviennent de plus en plus incontournables dans tous les domaines d'activité. Certains domaines sont, cependant, plus sensibles à la qualité des données produites. Si les mesures statistiques classiques offrent une bonne indication des performances d'algorithmes, elles qualifient rarement les données produites proprement dites. Nos travaux proposent des critères d'évaluation de ces données et les intègrent à un procédé de qualification de la qualité d'une information. Nous montrons comment intégrer ces différents paramètres à un système d'extraction automatique et comment celui-ci va évaluer sa production à la volée. Ce score de qualité servira à assister l'expert chargé de la vérification des données soit en tant qu'index de classement des informations en fonction de leur crédibilité, soit en mettant en avant les informations douteuses à vérifier en priorité. Afin de fournir le score le plus exhaustif et le plus interprétable possible, nous considérerons différents paramètres : la confiance faite à la source de l'information, la capacité de cette source à produire l'information considérée et l'information elle-même lorsque confrontée tant à des connaissances expertes qu'à d'autres informations produites.

## 1 Introduction

Les évolutions récentes des attributions des forces militaires, comme l'évolution de leurs contextes d'utilisation (e.g. opérations de maintien de la paix, guerres asymétriques, lutte contre le terrorisme, ...), offrent de nombreux défis pour la conception et la mise-en-œuvre de systèmes d'informations dédiés. Les volumes d'informations disponibles, en augmentation continue, rendent le besoin de systèmes automatisés incontournable. Néanmoins, de la sensibilité des informations manipulées, et de leurs conséquences, découle un besoin de manipuler des informations sûres et donc une méfiance envers les systèmes totalement automatisés. Cet inconvénient peut être minimisé en associant automatiquement un score de qualité à chaque information extraite ou produite. On pourra, par la suite, utiliser ce score pour ordonner les informations, voire fixer des seuils d'alertes sous lesquels il sera nécessaire d'effectuer des vérifications, avant toute prise de décision.

Pour fournir le score le plus exhaustif et le plus interprétable possible, celui-ci devra tenir compte de différents paramètres : la confiance faite à la source, la capacité de cette source à produire l'information considérée et l'information elle même lorsque confrontée tant à des connaissances expertes qu'à d'autres informations produites.

## 2 Évolution de la proposition

L'annexe du STANAG2022 OTAN (1997) a défini pour la première fois la grille de cotation du renseignement pour les forces de l'OTAN. Elle proposait une échelle à deux dimensions et six degrés de vérité (cf. tableau 1), dans laquelle la qualité d'une information était représentée par une combinaison de sa plausibilité et de la fiabilité de sa source. La fiabilité caractérise donc la source, indépendamment de l'information considérée. Toute information fournie par une même source bénéficierait donc de la même fiabilité.

Le second axe de cotation du STANAG2022 évalue la certitude de l'information. Le degré supérieur de cet axe est affecté aux informations confirmées par d'autres sources. Ceci tend à suggérer que la certitude de l'information se mesure comme un degré de confirmation : plus une information est confirmée, plus elle est certaine et, à l'inverse, plus elle est infirmée, moins elle est crédible.

De cette échelle de lecture nous retenons au moins les deux critiques suivantes :

1. L'évaluation sur deux axes, au lieu d'augmenter la lisibilité du score, le rend plus obscur. En effet, quelle information est la plus probable si l'une est cotée $B3$ et l'autre $C2$ ? Si l'on ajoute à ceci que la plausibilité découle de la confirmation par des sources *distinctes*, la simple lecture de la fiabilité devient problématique.

| Fiabilité | Définition |
|---|---|
| F | Fiabilité ne peut être estimée |
| E | Pas fiable |
| D | Habituellement pas fiable |
| C | Plutôt fiable |
| B | Habituellement fiable |
| A | Complètement fiable |

| Crédibilité | Définition |
|---|---|
| 6 | Crédibilité ne peut être estimée |
| 5 | Peu probable |
| 4 | Douteuse |
| 3 | Possiblement vraie |
| 2 | Probablement vraie |
| 1 | Confirmée par d'autre sources |

TAB. 1 – *STANAG2022 : Fiabilité de la source et crédibilité de l'information*

2. Comme nous l'avons dit, les critères d'évaluation de la plausibilité d'une information se réduisent à son degré de confirmation / infirmation. D'autres critères seraient sans doute indiqués pour mieux estimer la qualité d'une information. Matheus et al. (2005) présentent un point de vue intéressant sur les 'pédigrées' d'informations. Le contexte de production de deux informations par la même source permettrait, par exemple, de contrebalancer la fiabilité de la source et introduirait donc de la flexibilité dans son impact sur sa production.

Ces remarques peuvent paraitre, de prime abord, contradictoires. D'un côté, nous voulons offrir à l'utilisateur un score unique exprimant la qualité d'une information et ainsi la confiance qu'il doit avoir en elle. En particulier, ceci simplifierait les comparaisons entre informations, ce qui serait une nette amélioration de l'existant. D'autre part, nous souhaitons conserver les critères d'évaluation existants et en ajouter des nouveaux ou, du moins, clarifier les cas où les domaines de définition étaient mal définis ou confus.

Pour parer à ces problèmes nous allons :
– définir deux critères additionnels : la compétence de la source et la plausibilité de l'information
– redéfinir les critères existants
– intégrer chaque critère à une chaîne de calcul qui définira un protocole général pour le calcul d'un score unique

# 3 Définition de critères

Nous allons maintenant donner notre point de vue sur les critères existants et introduire les critères additionnels que nous proposons d'intégrer à notre score de qualité.

Pour nous conformer à l'esprit d'origine, nous avons choisi d'exprimer les valeurs de vérité de chaque critère sur une échelle à six niveaux, mais ce choix n'est en rien limitatif.

## 3.1 Fiabilité

La fiabilité de la source est toujours définie selon la grille initiale (cf. tableau 1). Néanmoins, nous limitons maintenant l'étendue de ce que ce critère juge, cf. Rogova et Nimier (2004). Dans la documentation clairsemée traitant du sujet (l'AJP-2.1 OTAN (2002)) la fiabilité de la source ne se contente pas de représenter la confiance faite à la source, mais également des notions de la compétence de la source à fournir l'information ainsi que la confiance exprimée par cette source pour son information.

Nous définissons la fiabilité comme totalement indépendante de l'information. De cette manière cet 'avis' aura le même impact sur toutes les informations provenant d'une même source. Bien entendu, la fiabilité de la source n'est pas un critère définitif. Le temps et l'expérience pourront la faire varier, comme illustré au paragraphe 6. Une source fiable donnant systématiquement des informations fausses perdra de la confiance et, de la même manière, une source inconnue se verra réévaluée à l'aune de sa production. Néanmoins, à un instant donné, toutes les informations produites par une même source bénéficieront du même taux de fiabilité.

Supposons, par exemple, que notre source soit un capteur électronique. Nous prétendons que la fiabilité de la source correspond à l'état du capteur. De cette manière quelle que soit l'information produite, la fiabilité de la source aura la même influence sur le calcul de sa qualité.

## 3.2 Compétence

La capacité qu'une source a à fournir une information peut dépendre de l'information elle-même. Il est possible qu'une source donne une information hors de son domaine d'expertise. Une telle information ne doit pas, pour autant, être rejetée

pour cette unique raison, mais le destinataire doit en tenir compte lors de la classification de l'information ou lors de l'évaluation de sa qualité.

Si nous revenons à notre exemple, un capteur électronique est étalonné pour fonctionner sur une certaine plage. Ceci n'empêche pas ce capteur de fournir des informations en dehors de cette plage. Dans un tel cas de figure, la qualité devrait intégrer cette estimation de la compétence pour pondérer la confiance.

Une autre interprétation est le contexte d'utilisation. Prenons, par exemple, les conditions météorologiques. Ces conditions vont impacter toutes les observations faites pendant qu'elles s'appliqueront.

Nous définissons donc la compétence de la source comme une fonction de la source en regard de l'information considérée, et mesurée sur l'échelle donnée par le tableau 2.

| Compétence | Définition |
|---|---|
| 6 | La compétence ne peut être appréciée |
| 5 | Incompétent |
| 4 | Insuffisamment compétent |
| 3 | Partiellement compétent |
| 2 | Compétent |
| 1 | Expert |

TAB. 2 – *Compétence de la source, fonction de l'information considérée*

| Plausibilité | Définition |
|---|---|
| 6 | La plausibilité ne peut être appréciée |
| 5 | Impossible |
| 4 | Peu probable |
| 3 | Possible |
| 2 | Réaliste |
| 1 | Certaine |

TAB. 3 – *Plausibilité de l'information : conformité à la connaissance extérieure*

## 3.3 Plausibilité

Supposons maintenant que nous disposons de connaissances extérieures, par exemple stockées dans une ontologie. Nous désignons par connaissances extérieures notre connaissance sur l'état du monde, indépendamment de nos informations extraites. Si une information est produite en désaccord avec ces connaissances extérieures, nous prétendons que cette contradiction est différente d'une contradiction avec des informations construites. Il est vraisemblable qu'un doute en l'information émerge, dans la mesure où notre connaissance du monde est supposée sûre.

Si notre capteur est, par exemple, un radar de vitesse braqué sur une route. En connaissant la route et quelques rudiments des lois de la physique, il est possible de contredire certaines mesures. Ceci est différent de la comparaison de mesures, parce que notre connaissance du monde étant certaine, elle ne sera pas remise en cause par les informations que nous évaluons. En revanche, les informations acquises peuvent – et doivent – être réévaluées.

La plausibilité mesure donc l'adéquation entre l'information et notre point de vue sur l'état du monde. Un exemple de degrés de vérité pour ce critère est présenté dans le tableau 3.

## 3.4 Crédibilité

En raison du degré maximal de crédibilité de la grille de cotation existante, nous choisissons de l'évaluer comme un indice de confirmation par les informations acquises. Une fois encore, nous insistons sur la distinction faite entre les comparaisons entre informations acquises et leur compatibilité avec le modèle de raisonnement. Les informations que nous comparons en évaluant la crédibilité sont toutes construites et ont donc toutes un indice de qualité mouvant, par opposition aux connaissances du monde dont la confiance est fixe et maximale. Le tableau 4 présente les degrés de vérité de la crédibilité.

## 3.5 Qualité

En sortie, notre système proposera un degré de qualité associé à l'information considérée. Celui-ci exprime, en un score unique et comparable, la combinaison des critères ci-dessus. Un exemple de degrés de vérité associés à la qualité est présenté dans le tableau 5.

| Crédibilité | Definition |
|---|---|
| 6 | La crédibilité ne peut étre appréciée |
| 5 | Contredite par des informations fiables |
| 4 | Partiellement contredite |
| 3 | Insuffisamment confirmée |
| 2 | Partiellement confirmée |
| 1 | Confirmée par des informations fiables |

TAB. 4 – *Degrés de crédibilité : indice de confirmation entre informations extraites*

| Confiance | Définition |
|---|---|
| 6 | La qualité ne peut être appréciée |
| 5 | Peu probable |
| 4 | Douteuse |
| 3 | Possible |
| 2 | Probable |
| 1 | Extrêmement probable |

TAB. 5 – *Sortie du système : degrés de qualité associé à une information*

# 4 La chaîne de cotation

Maintenant que nous avons défini les critères que nous souhaitons inclure dans notre méthode d'évaluation, nous allons préciser les tendances d'impact de chaque facteur sur l'évolution de la qualité de l'information et présenter notre point de vue sur leur utilisation.

## 4.1 Présentation et origine

Nous représentons, dans la figure 1, le processus de cotation de manière séquentielle. L'ordre d'intégration des différents critères provient d'une interprétation naturelle et intuitive du procédé partant de l'évaluation de la source, suivie d'un critère traitant source *et* information et se concluant par les confrontations de l'information aux autres connaissances. La figure 1 indique également les pentes générales de la qualité, pour chaque critère. Chaque étape consiste en la réévaluation du score précédent – celui de l'étape précédente – lorsque confrontée au nouveau critère. Nous présenterons également l'idée de stratégies d'évaluation permettant de modéliser différentes postures de crédulité de manière semblable à ce que nous présentions dans Revault d'Allonnes et al. (2007). Les notions de stratégies seront illustrées d'exemples, dans la section 5.

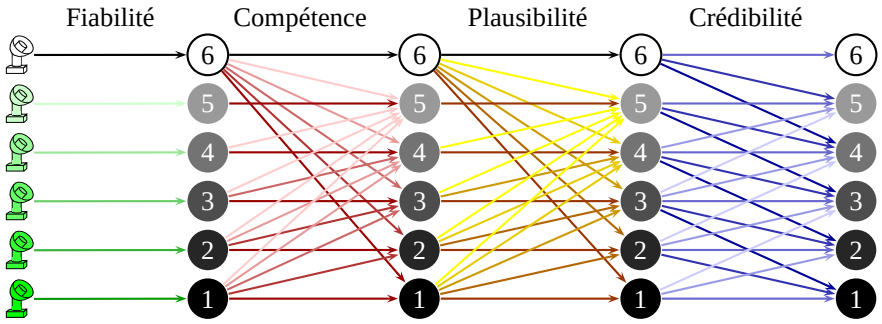


FIG. 1 – *La chaîne de cotation*

## 4.2 Degrés de vérité et incertitude

Avant d'entrer dans le détail de notre méthode, il faut noter que chaque évaluation de critère s'opère sur une même échelle, i.e. de 1≡'haute' à 5≡'basse', avec 6≡'inqualifiable'. Cependant, chaque évaluation possède sa propre interprétation, comme les tableaux de la section précédente le précisent. Nous allons maintenant construire un degré de qualité d'une information en fonction de ces critères.

Dans chaque grille de lecture des degrés de vérité, le score 'inqualifiable' est une valeur neutre. C'est à dire que la combinaison de n'importe quel score de confiance $s$ à une évaluation de critère de niveau 6 renverra $s$. De même, n'importe quelle évaluation de critère $e$ combinée avec un degré de confiance de 6 renverra $e$ ou, dans le cas de la fiabilité de la source, l'interprétation de $e$ fonction de la stratégie. Ceci explique que la figure 1 montre que n'importe quel score est atteignable depuis une confiance de 6. D'un autre côté, en raison de l'interprétation intuitive de la valeur 'l'indice

ne peut être exprimé', aucun retour vers cette valeur n'est possible, une fois que la confiance l'a quittée. Ceci explique pourquoi aucune flèche ne pointe vers une confiance de 6.

Ce score, ou degré de confiance, ressemble en partie à des théories existantes et usitées. En effet, ce que nous proposons est l'évaluation de données incertaines et potentiellement imprécises. Nous proposions par exemple, dans Revault d'Allonnes et al. (2007), une interprétation multi-valuée du problème. Nous pensons donc que ces théories offrent de nombreux outils pour combiner les évaluations de nos critères, voire pour leur évaluation. Comme nous l'avons dit, notre échelle à six niveaux est un héritage direct de l'existant. Chaque critère pourrait, cependant, être défini et calculé de manière continue, puis exprimé à l'utilisateur à l'aide de variables linguistiques (cf. Zadeh (1975)). Le calcul de la confirmation, infirmation ou conflit pourrait être modélisé par la théorie de Dempster-Shafer Shafer (1976, 1990). Notre degré de confiance, lui-même, pourrait provenir d'une sorte de Transferable-Belief Model Smets et al. (1991), et associé à son interprétation pignistique.

# 5 Principes de la chaîne de cotation et stratégies

## 5.1 Fiabilité de la source

Supposons que nous ayons une nouvelle information entrante. Elle n'est donc pas encore évaluée. Puisque cette information est inconnue, nous allons la croire à la hauteur de la confiance faite à sa source. La première étape de notre processus va, ainsi, attribuer à une information inconnue le niveau de confiance (tableau 5, niveaux de gris dans la figure 1) équivalent à la fiabilité de sa source. La figure 1 présente la 'stratégie par défaut' qui associe A⇔1, …, F⇔6. À part la nécessité d'avoir une équivalence entre les degrés 'inqualifiables' – i.e. F⇔6 – toute autre stratégie acceptable – i.e. où A⩾B⩾C⩾D⩾E, voir tableau 6 pour des exemples – pourrait être utilisée.

| | $S_1$ | $S_2$ | $S_3$ |
|---|---|---|---|
| A | 1 | 2 | 3 |
| B | 2 | 2 | 4 |
| C | 3 | 3 | 4 |
| D | 4 | 5 | 5 |
| E | 5 | 5 | 5 |
| F | 6 | 6 | 6 |

TAB. 6 – *3 stratégies différentes pour le calcul de la fiabilité de la source. La stratégie $S_1$ est la stratégie par défaut de la figure 1.*

| | $S_1$ | $S_2$ | $S_3$ |
|---|---|---|---|
| 1 | $c$ | $c$ | $c$ |
| 2 | $\min(c+1,5)$ | $\min(c+2,5)$ | $\min(c+1,5)$ |
| 3 | $\min(c+2,5)$ | $\min(c+3,5)$ | $\min(c+3,5)$ |
| 4 | $\min(c+3,5)$ | $\min(c+4,5)$ | $\min(c+3,5)$ |
| 5 | $\min(c+4,5)$ | $\min(c+4,5)$ | $\min(c+5,5)$ |
| 6 | $c$ | $c$ | $c$ |

TAB. 7 – *3 stratégies différentes pour l'évaluation de la compétence de la source et la plausibilité de l'information, où $c$ est l'indice de qualité courant.*

## 5.2 Compétence et plausibilité

L'étape d'intégration de la 'compétence' de notre processus a un impact négatif sur le score de qualité. Ceci est dû au fait que l'initialisation de la confiance s'opérant au niveau de la confiance faite à la source, on aura sans doute supposé que ladite source 'sait de quoi elle parle' ou, du moins, que nous lui faisons confiance à hauteur de ceci. Il est donc logique que le seul impact de la compétence soit négatif. Une source de toute confiance fournissant une information hors de son domaine d'expertise nous incitera à prendre l'information avec circonspection.

De manière semblable, l'impact de la plausibilité de l'information aura tendance à faire baisser le score global. Si nous croyons une information à hauteur de sa source ; si sa source est compétente en la matière, le fait que l'information soit *possible* ne doit pas augmenter notre confiance, puisque nous avions commencé par la supposer au moins possible. En revanche, si elle paraît improbable, nous commencerons à en douter.

Pour déterminer de combien chacune de ces étapes va réduire la confiance, nous nous tournons maintenant vers les stratégies. La figure 1 montre les tendances générales attendues pour chaque niveau de confiance. Une fois encore, elle présente la stratégie par défaut pour laquelle l'impact croît proportionnellement à l'évaluation du critère. Cette stratégie et deux autres sont présentées dans le tableau 7.

Puisque chaque étape de la chaîne de cotation peut être vue comme une fusion du score courant avec l'évaluation d'un nouveau critère, toute méthode probabiliste Nimier (2004); Rogova et Nimier (2004) ou tout autre opérateur de fusion Bloch (1996) peut être envisagé. Par exemple, des opérateurs conjonctifs[1] pourraient modéliser une stratégie méfiante et les opérateurs disjonctifs représenteraient une stratégie confiante, de manière semblable à ce que nous proposions

[1] Un opérateur de fusion est dit conjonctif (resp. disjonctif) si et seulement si le résultat de la fusion est inférieur (resp. supérieur) ou égal au minimum (resp. maximum) des valeurs fusionnées.

dans Revault d'Allonnes et al. (2007). Le choix des opérateurs, et donc des stratégies, sera laissé à l'utilisateur en fonction de l'application visée.

### 5.3 Crédibilité

La dernière étape de la chaîne de cotation est le calcul de l'indice de confirmation, hérité du STANAG2022. Puisque nous avons séparé les notions d'évaluation d'information et de comparaison d'informations acquises, cette dernière étape compare des informations possibles bien qu'incertaines et cotées. Ceci est une différence majeure avec l'étape précédente de plausibilité. En effet, ce que nous considérons à l'étape de la plausibilité est notre connaissance du monde. Nous l'avons appelée 'connaissance extérieure' pour la distinguer de la connaissance construite, et cotée. Nous insistons sur le fait que la connaissance extérieure ne sera pas remise en cause lorsque confrontée à nos informations construites. La confiance en sera toujours maximale.

À cette étape de crédibilité, cependant, nous comparons des informations incertaines. Chacune de ces informations a un degré de confiance associé, degré qui pourra être mis à jour lorsque nous comparerons les informations. L'évolution de la confiance en fonction des confirmations ou infirmations est un procédé naturel. Ceci étant, puisque nous considérons à la fois confirmation et infirmation, il coule de source que ce critère peut faire augmenter la confiance aussi bien qu'il peut la faire baisser. Puisque toutes les informations comparées sont cotées, leur score pourra être utilisé pour pondérer leur degré de confirmation ou d'infirmation (cf. Besombes et Cholvy (2007)).

## 6 Mise à jour et alertes

Après l'étape de crédibilité de notre chaîne de cotation, l'information considérée a été cotée en fonction de la fiabilité de sa source ainsi que de sa compétence sur le domaine de définition de l'information, la plausibilité de cette information en regard de notre perception du monde et, enfin, un indice de confirmation entre informations incertaines. À ce moment, nous pouvons envisager de mettre à jour la fiabilité de la source de l'information, cf. figure 2. Cette rétroaction de la confiance en l'information sur la fiabilité de sa source ne fait pas partie de la chaîne, mais a un effet central. Si une source de confiance fournit systématiquement des informations cotées peu probables, il pourrait être judicieux de revoir notre confiance. Il n'y a pas de règles de mise à jour, notamment parce qu'elle dépend beaucoup de la source, ainsi que de l'application.

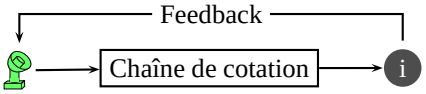


FIG. 2 – *Mise à jour de la fiabilité de la source*

Une alternative à la mise à jour de la fiabilité de la source serait l'alerte à l'utilisateur lors de comportements imprévus de la chaîne de cotation. En effet, d'après notre description et selon nos attentes, une source sûre aura plutôt tendance à fournir des informations de qualité. Dans le cas contraire, il pourrait être enrichissant de faire remonter l'abération à l'utilisateur pour vérification. Suite à son jugement on pourra revoir la fiabilité de la source mais, dans certains cas, on aura détecté une information inattendue.

## 7 Application symbolique : l'extraction d'information

Afin de pouvoir traiter le nombre croissant d'informations textuelles (sources ouvertes, interceptions, ...), des systèmes automatiques à partir de reconnaissance d'entités nommées peuvent assister un opérateur dans sa tâche d'extraction d'information. De tels systèmes peuvent extraire de l'information de textes de manière automatique et peuvent encore classifier et combiner ces informations pour créer de la connaissance plus élaborée. L'inconvénient de ce genre de technique est la perte de contrôle que l'utilisateur a sur le résultat. En effet, le système va extraire de l'information et la fournir à l'utilisateur, par exemple sous forme d'annotations. À ce moment l'utilisateur peut :

- faire totalement confiance au système en supposant qu'une information extraite est nécessairement vraie
- ou, se souvenant qu'aucun système n'est parfait, vérifier toutes les informations lui-même

De toute évidence, la seconde option est éliminée d'office puisqu'elle annule tout avantage de l'utilisation du système automatique. La confiance aveugle peut également être dangereuse, en particulier pour tout utilisateur manipulant des données sensibles.

Nous allons maintenant présenter comment notre chaîne de cotation, définie en section 4, s'intègre naturellement à une évaluation des informations produites par un système automatique.

Donnons-nous, tout d'abord, un rapport textuel, émanant d'une source fiable ainsi qu'une date automatiquement extraite de ce rapport (cf. Makkonen et Ahonen-Myka (2003)). La fiabilité de la source entraine une confiance *a priori* que nous évaluerons à 1. Ce premier score est l'initialisation de notre chaîne de cotation.

Étant donné le texte et divers types d'information (une date, un sujet, un nom d'organisation, ..., voir National Institute of Standards and Technology (2007)), un algorithme d'extraction spécifique sera utilisé. Chaque algorithme pourra être assorti d'une évaluation (rappel, précision, f-mesure, ..., voir LDC (2001); NIST (2007) pour une liste exhaustive) constituant la compétence de l'algorithme à produire une information.

Supposons que l'algorithme ait une précision de 90% convertie en une compétence de niveau 2. Le choix de préférer la précision au rappel ou à la f-mesure pour l'évaluation de la compétence se justifie par le fait que la précision mesure la probabilité a priori qu'une information extraite soit pertinente, c'est à dire la compétence de l'algorithme en regard d'une information donnée. Selon la stratégie choisie, cf. section 5, la confiance sera maintenant d'un niveau 2.

La date extraite pourra ensuite être comparée à d'autres informations. Supposons, par exemple, que nous cherchions à dater une manifestation dans un pays et que la date initialement extraite soit le '04/07/2008'. L'extraction va produire l'information 'Vendredi quatre juillet 2008'. Si notre connaissance du pays place une fête nationale, ou l'anniversaire d'un événement important, le 4 juillet dans le pays étudié, cette connaissance extérieure va confirmer qu'une manifestation est possible à cette date. La plausibilité de l'information sera donc estimée à 1 mais la confiance restera à son niveau actuel.

Enfin, si le mot 'vendredi' est repéré plus loin dans le même texte, cette information tendra à confirmer la date du quatre juillet, et augmentera la crédibilité de l'information et, par là même, la confiance en notre information. D'autres informations extraites pourraient, également, contredire ou corroborer la date.

Cette implémentation de la chaîne de cotation pour l'extraction d'information est illustrée par la figure 3.

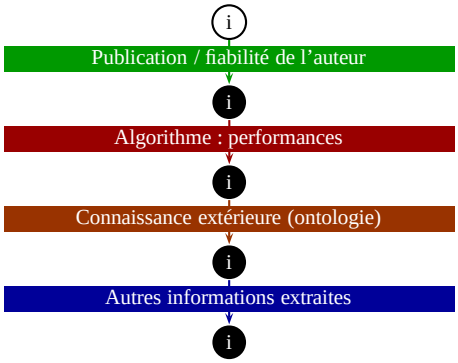


FIG. 3 – *La chaîne de cotation pour l'extraction d'information*

# 8 Conclusion

Parce que le contexte du renseignement militaire évolue vers le traitement de vastes quantités de données, structurées comme des mesures techniques de capteurs, ou déstructurées comme les sources ouvertes (presse, internet, rapports diplomatiques, ...) nous pensons que le besoin de systèmes automatisés augmente. Pour proposer des systèmes lisibles et utiles, il faut prévoir le filtrage de données en fonction de leur vertu informative et de leur qualité.

À partir des critères existants, nous proposons des critères additionnels pour évaluer automatiquement, un taux de confiance. Nous classons les facteurs utilisés selon leurs domaines de définitions et d'influence, intégrant deux nouveaux critères et en en redéfinissant un autre.

Nous expliquons ensuite comment nous avons construit une chaîne de cotation utilisant de manière intuitive nos nouvelles définitions.

# Références

Besombes, J. et L. Cholvy (2007). Information fusion : using an ontology to information evaluation. In *Proceedings of the International Colloquium on Information Fusion 2007*, Xi'an, China, pp. 416–422.

Bloch, I. (1996). Information combination operators for data fusion : A comparative review with classification. *Systems, Man and Cybernetics – Part A 26*(1), 52–67.

LDC (2001). *Message Understanding Conference 7 (MUC7)*, Philadelphia. LDC.

Makkonen, J. et H. Ahonen-Myka (2003). Utilizing temporal information in topic detection and tracking. In T. Koch et I. T. Solvberg (Eds.), *Proceedings of 7th European Conference on Digital Libraries (ECDL 2003)*, pp. 393–404. Springer-Verlag.

Matheus, C. J., D. Tribble, M. M. Kokar, M. G. Ceruti, et S. C. McGirr (2005). Towards a formal pedigree ontology for level-one sensor fusion. In *Proceedings of the 10th International Command & Control Research and Technology Symposium*.

National Institute of Standards and Technology (2007). *Sixteenth Text REtrieval Conference Proceedings (TREC 2007)*, Gaithersburg, Maryland. National Institute of Standards and Technology.

Nimier, V. (2004). Information evaluation : a formalisation of operational recommendations. In *Proceedings of the Seventh International Conference on Information Fusion*, Stockholm, Sweden, pp. 1166–1171.

NIST (2007). Automatic Content Extraction Evaluation (ACE07).

OTAN (1997). Annex to stanag 2022 (edition 8).

OTAN (2002). Allied Joint Publication – 2.1, Intelligence Procedures.

Revault d'Allonnes, A., H. Akdag, et O. Poirel (2007). Trust-moderated information-likelihood. A multi-valued logics approach. In *Computation and Logic in the Real World, CiE 2007*, Sienna, Italy.

Rogova, G. L. et V. Nimier (2004). Reliability in information fusion : literature survey. In *Proceedings of the Seventh International Conference on Information Fusion*, Stockholm, Sweden.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press.

Shafer, G. (1990). Perspectives on the theory and practice of belief functions. *International Journal on Approximate Reasoning 4*(5–6), 323–362.

Smets, P., Y.-T. Hsia, A. Saffiotti, R. Kennes, et E. Umkehrer (1991). The transferable belief model. In R. Kruse et P. Siegel (Eds.), *ECSQARU*, Volume 548 of *Lecture Notes in Computer Science*, Marseille, France, pp. 91–96. Springer.

Zadeh, L. A. (1975). The concept of a linguistic variable and its application to approximate reasoning – i. *Information Sciences 8*(3), 199–249.

## Summary

Automatic processing of large quantities of data has become increasingly unavoidable in most areas. Some areas are, however, more sensitive to the quality of the produced data. Classical statistical measures are a cornerstone of algorithm évaluation, yet they rarely qualify the individual results. Our work offers criteria to evaluate these produced data and integrates them in a confidence-scoring chain. We show how to use our criteria in an automatic extraction system and how this system will score its production on the fly. This confidence score can then be used by the expert in charge of data verification either as an ranking index or as a highlighting on doubtful information. To produce the most comprehensive reliability score we can, we consider different parameters: source quality; its ability to produce considered information; the information itself in the light of expert knowledge and other constructed information.

# Généralisation de la propriété de monotonie de la *all-confidence* pour l'extraction de motifs intéressants non fréquents

Yannick Le Bras*,***, Philippe Lenca*,***, Stéphane Lallich** et Sorin Moga*,***

*Institut TELECOM ; TELECOM Bretagne ;
UMR CNRS 3192 Lab-STICC;
Technopôle Brest Iroise CS 83818
29238 Brest Cedex 3
{yannick.lebras || philippe.lenca || sorin.moga}@telecom-bretagne.eu
**Université de Lyon,
Laboratoire ERIC, Lyon 2, France
stephane.lallich@univ-lyon2.fr
***Université Européenne de Bretagne, France

**Résumé.** Différentes études ont montré les limites du couple support/confiance dans les algorithmes de type Apriori, tant du point de vue quantitatif (quantité des règles), que qualitatif (redondance, intérêt, pépites de connaissance). Une solution consiste à concentrer au plus tôt la recherche sur les règles intéressantes en utilisant des mesures d'intérêt possédant des propriétés algorithmiques, mais aussi des capacités à mettre en évidence des règles d'un intérêt certain. Celles-ci permettent alors de trouver les règles d'intérêt élevé, sans utiliser un élagage préalable par la condition de support. Elles rendent également possible la recherche efficace de pépites de connaissance. C'est le cas de la all-confidence (ou h-confidence), transformation antimonotone de la confiance. Nous nous intéressons ici à la possibilité d'appliquer une transformation semblable à d'autres mesures, au travers d'une condition nécessaire s'appuyant sur un cadre formel que nous définissons. Nous montrons cependant que parmi les 27 mesures étudiées ici, seules 5 d'entre elles peuvent être transformées en une mesure antimonotone. Ainsi bien que très prometteuse, cette propriété d'antimonotonie n'est à l'heure actuelle applicable qu'à peu de mesures.

## 1 Introduction

La recherche de règles a attiré l'attention de nombreux chercheurs en apprentissage automatique et fouille de données. La principale justification de cet intérêt se trouve dans leur capacité à représenter les connaissances de manière compréhensible. Dans le cadre de l'apprentissage supervisé, les arbres de décision, par exemple CART (Breiman et al. (1984)) et C4.5 (Quinlan (1993)), ou bien les listes de décision, par exemple CN2 (Clark et Niblett (1989)), sont deux approches efficaces pour les tâches de classification. Dans le cadre non supervisé, la recherche de règles d'association est probablement la méthode la plus populaire pour établir des relations entre les différentes variables d'une base de données. Depuis sa première formulation et la proposition de l'algorithme AIS par Agrawal et al. (1993), le problème de la découverte de règles d'association, et par conséquent le problème sous-jacent de la recherche de motifs fréquents, ont fait l'objet de nombreux travaux.

Rappelons dans un premier temps les fondements de la recherche de règles d'association, ainsi que ses principaux défis. Une règle d'association $\mathtt{A} \rightarrow \mathtt{B}$ est la donnée de deux ensembles d'attributs $\mathtt{A}$ et $\mathtt{B}$ (aussi appelés motifs) non vides tels que $\mathtt{A} \cap \mathtt{B} = \emptyset$. Étant donnée un ensemble d'individus (chaque individu étant décrit par un ensemble d'attributs), la règle $\mathtt{A} \rightarrow \mathtt{B}$ signifie que lorsqu'un individu contient le motif $\mathtt{A}$, alors, probablement, il contient aussi le motif $\mathtt{B}$. Le problème de la découverte de règles d'association, conformément à Agrawal et al. (1993), consiste en la génération de toutes les règles ayant un support et une confiance supérieurs à des seuils fixés au préalable par l'utilisateur. Le support définit la proportion de transactions de la base de données incluant les motifs $\mathtt{A}$ et $\mathtt{B}$ (noté $supp(\mathtt{A} \rightarrow \mathtt{B})$), tandis que la confiance représente cette proportion, mais au sein des transactions contenant $\mathtt{A}$. Elle est notée $conf(\mathtt{A} \rightarrow \mathtt{B})$ . La recherche de règles d'association se déroule en deux étapes. Dans un premier temps, la contrainte de support est appliquée pour extraire de la base les motifs fréquents, c'est-à-dire des motifs dont le support est supérieur au seuil fixé. Dans un second temps, la contrainte de confiance minimale est appliquée aux motifs fréquents pour former des règles intéressantes. Cette stratégie implique deux problèmes majeurs : l'un concerne la complexité de la méthode, et l'autre la qualité des règles produites.

Trouver tous les motifs fréquents dans une base de données comportant $k$ attributs est une tâche coûteuse en temps de calcul, car le nombre de motifs présents dans la base peut atteindre $2^k$. Cependant, la propriété de clôture du support,

appelée aussi antimonotonie, a permis l'élaboration d'algorithmes très efficaces tels que APRIORI (Agrawal et Srikant (1994)), PARTITION (Savasere et al. (1995)), ECLAT (Zaki (2000)) ou encore FP-GROWTH (Han et al. (2000)). D'intéressantes synthèses ont été réalisées par Hipp et al. (2000), Goethals et Zaki (2004) et Goethals (2005). Toutes ces approches soulignent le rôle prépondérant du support et du seuil associé. S'affranchir du support et/ou du seuil a été identifié comme un problème majeur, et pour le résoudre, deux grandes lignes se démarquent. L'une d'entre elles consiste à se libérer de l'utilisation du support (Cohen et al. (2001)), quand l'autre tente d'éliminer la tâche de fixation du seuil (Koh (2008)).
Trouver exactement toutes les règles intéressantes est également un problème, car les ensembles de règles générés sont très grands, en particulier avec le couple support/confiance. Une stratégie consiste à augmenter le seuil de support, ce qui entraîne la disparition de règles intéressantes, telles que les règles de faible support et de confiance élevée, appelées pépites de connaissance. Une autre stratégie, qui consiste à augmenter le seuil de confiance, favorise les règles ayant un conséquent fréquent, ce qui donne un grand nombre de règles inintéressantes, au sens où la connaissance de l'antécédent ne caractérise en rien la présence du conséquent.
On peut encore améliorer la recherche en ajoutant des contraintes sur l'apparition des motifs (Pei et Han (2000)), ou bien en utilisant les mesures d'intérêt au moment de la génération de candidats pour réduire à la fois le nombre de règles générées et le temps de parcours des bases de données. Cette technique est exploitée par exemple par Bayardo et al. (1999) dans l'algorithme DENSEMINER. En effet, comme le soulignent Yao et al. (2003), les mesures d'intérêt peuvent jouer des rôles différents : elles peuvent servir à diminuer l'espace de recherche (par exemple le support), sélectionner les règles intéressantes (par exemple la confiance), ou encore quantifier l'utilité et l'efficacité des règles (par exemple le coût). Nous nous limitons aux mesures objectives de l'intérêt des règles, qui ne dépendent que de la contingence des données, et parmi celles-ci, nous nous intéressons plus particulièrement à celles qui permettent à la fois de réduire l'espace de recherche et d'évaluer efficacement la qualité des motifs découverts.
De notre point de vue, cette approche est certainement la plus prometteuse. Pour dépasser le primat de la condition de support, différentes solutions ont été proposées au niveau algorithmique, plus spécialement pour la confiance.
Le reste de l'article est organisé de la façon suivante. Dans la section 2, nous passons en revue de récents travaux se concentrant sur les propriétés algorithmiques. Puis dans la section 3, nous introduisons un nouveau cadre formel permettant de mettre en relation les propriétés algorithmiques et analytiques des mesures. Dans la section 4, nous appliquons ce cadre à la recherche d'une généralisation de la *all-confidence* de Omiecinski (2003), qui permet de rendre antimonotone la confiance, et démontrons une condition suffisante de non-existence de cette généralisation. Forts de ce résultat, nous montrons qu'un grand nombre de mesures sont incompatibles avec cette propriété. Finalement, nous offrons une conclusion de notre travail dans la section 5.

# 2 Etat de l'art

Pour s'affranchir du support, quelques auteurs ont proposé des travaux se concentrant sur les propriétés algorithmiques de la confiance, le but annoncé étant de parvenir à obtenir toutes les règles de confiance élevée, et plus particulièrement les pépites de connaissance. Nous passons en revue ici certains des travaux proposant une alternative au support. Une partie d'entre eux agissent sur le type de motifs considérés, en minimisant leur nombre. D'autres se focalisent sur les mesures et exploitent leurs propriétés intrinsèques.
Wang et al. (2001) introduisent la notion de *Universal Existential Upward Closure* qui exploite une propriété de monotonie de la confiance. Cette propriété s'applique aux règles de classification, et permet une exploration par le haut du treillis des motifs. Cette stratégie est efficace en termes de génération de candidats, et permet de résoudre le problème des pépites de connaissance. En se concentrant sur la mesure de Jaccard, Cohen et al. (2001) proposent un algorithme d'approximation fondé sur les tables de hachage. L'apparition de faux positifs ou négatifs peut être contrôlée par les paramètres de l'algorithme. Sa complexité dépend de la force de ce contrôle, qui ne peut cependant jamais être complet. Zimmermann et de Raedt (2004) adaptent quant à eux une technique développée par Morishita et Sese (2000), et proposent l'algorithme CORCLASS pour la classification. Cette technique exploite pleinement une propriété d'antimonotonie liée à la convexité du $\chi^2$. Li (2006) introduit la notion d'ensemble de règles optimales pour les règles de classification. Un tel ensemble contient toutes les règles, à l'exception de celles qui n'ont pas un intérêt plus grand que l'une de leurs règles plus générales. Ces ensembles de règles ont des propriétés qui sont à mettre en rapport avec les motifs fermés et les règles non redondantes étudiées par Zaki (2004). Leur principal avantage algorithmique se trouve dans la stratégie d'élagage que ces propriétés permettent d'adopter, stratégie qui s'applique à un grand nombre de mesures. L'auteur donne une preuve cas par cas pour 13 mesures. Dans Le Bras et al. (2009) nous généralisons ce résultat en proposant une condition nécessaire et suffisante d'existence de cette stratégie pour les mesures objectives. Cela nous permet alors de vérifier qu'elle peut s'appliquer à 26 des 32 mesures que nous y étudions.
Bonchi et Lucchese (2005) introduisent pour leur part une nouvelle notion d'antimonotonie appelée la *Loose-Antimonotonie*, qui s'applique efficacement à la plupart des contraintes statistiques. L'avantage d'une telle propriété est la

possibilité d'exploiter les travaux menés sur les algorithmes de type Apriori. Omiecinski (2003) introduit les mesures de *any-confidence*, *all-confidence* et *bond*. Les deux dernières ont l'avantage d'être antimonotones, comme le support, et facilement calculables. Il est alors possible de s'appuyer sur les nombreux efforts faits dans le domaine des algorithmes de type Apriori pour utiliser ces deux mesures. La *all-confidence*, définie pour un itemset $I$ sur une base $\mathcal{D}$ par $\frac{|\{d\in\mathcal{D}|I\subseteq d\}|}{\max\limits_{l\subsetneq I, l\neq\emptyset}\{|\{d\in\mathcal{D}|l\subseteq d\}|\}}$, sera en particulier renommée *h-confidence* par Xiong et al. (2003) et étudiée plus en détail. Un motif est intéressant au sens de la *h-confidence* si toutes les règles qu'il peut engendrer ont une confiance supérieure à un seuil donné. La *h-confidence* permet aussi de filtrer les motifs, jugés inintéressants, ayant la propriété de *cross-support*, c'est-à-dire les motifs dont les attributs ont des valeurs de support très différentes. L'algorithme HYPERCLIQUE-MINER (Xiong et al. (2003)) exploite ces deux propriétés.
Notons que les mesures objectives peuvent être considérées comme des fonctions de $\mathbb{R}^3 \rightarrow \mathbb{R}$. Lenca et al. (2008) ont proposé une étude poussée des domaines d'arrivée pour certaines mesures. D'autres proposent de contrôler l'espace d'arrivée en normalisant les mesures (Diatta et al. (2007); Slowiński et al. (2008)).
Nous introduisons ici un nouveau cadre formel proche de celui de Hébert et Crémilleux (2007) et l'utilisons pour faire le lien entre les propriétés analytiques des mesures et la propriété de monotonie, inspirée de la définition de la *all-confidence*, que nous proposons dans cet article.

# 3 L'espace des règles d'association

Nous proposons ici d'établir un cadre formel qui nous permet d'établir une étude analytique des mesures d'intérêt. Les mesures d'intérêt objectives, sur lesquelles porte notre étude, s'expriment en fonction de la table de contingence en fréquences relatives, et donc de trois variables. L'étude de leurs variations en fonction de ces variables nous permettra par la suite de faire le lien entre les mesures et leurs propriétés algorithmiques, mais elle implique la description d'un domaine de définition nous permettant de n'étudier que des cas réels.

## 3.1 Règles d'association

Une base de données booléenne $\mathcal{DB}$ est décrite par un triplet $(\mathcal{A}, R, \mathcal{T})$, où $\mathcal{A}$ est un ensemble d'attributs, $\mathcal{T}$ un ensemble de transactions et $R$ une relation binaire sur $\mathcal{A} \times \mathcal{T}$. Une règle d'association est définie par une base de données $\mathcal{DB}$, un ensemble $\mathtt{A} \subset \mathcal{A}$ non vide appelé motif antécédent (ou tout simplement antécédent), et un ensemble $\mathtt{B} \subset \mathcal{A}$ non vide appelé motif conséquent (ou tout simplement conséquent), tel que $\mathtt{A} \cap \mathtt{B} = \emptyset$. On note une règle d'association $\mathtt{A} \xrightarrow{\mathcal{DB}} \mathtt{B}$ ou simplement $\mathtt{A} \rightarrow \mathtt{B}$ s'il n'y a aucune ambiguïté possible.
Le support d'un motif traduit la fréquence d'apparition du motif dans la base de données. Nous le noterons $supp_{\mathcal{DB}}(\mathtt{A})$, ou, lorsqu'il n'y a pas d'ambiguïté, simplement $supp(\mathtt{A})$. Le support d'une règle $\mathtt{A} \rightarrow \mathtt{B}$ est le support du motif $\mathtt{AB}$.

## 3.2 Tables de contingence

Soit $\mathtt{A}$ et $\mathtt{B}$ deux motifs dans une base de données $\mathcal{DB}$. La table de contingence en fréquences relatives conjointes de $\mathtt{A}$ et $\mathtt{B}$ renseigne sur la coprésence de ces deux motifs (figure 1).

| | $\mathtt{B}$ | $\bar{\mathtt{B}}$ | |
|---|---|---|---|
| $\mathtt{A}$ | $supp(\mathtt{AB})$ | $supp(\mathtt{A}\bar{\mathtt{B}})$ | $supp(\mathtt{A})$ |
| $\bar{\mathtt{A}}$ | $supp(\bar{\mathtt{A}}\mathtt{B})$ | $supp(\bar{\mathtt{A}}\bar{\mathtt{B}})$ | $supp(\bar{\mathtt{A}})$ |
| | $supp(\mathtt{B})$ | $supp(\bar{\mathtt{B}})$ | 1 |

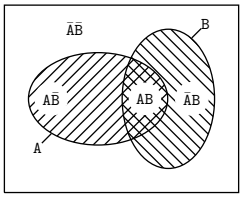


FIG. 1 – *Table de contingence de* $\mathtt{A}$ *et* $\mathtt{B}$*, sous forme de tableau et sous forme graphique.*

La table de contingence a 3 degrés de liberté : il faut au moins trois de ses valeurs pour la décrire, et trois valeurs suffisent pour retrouver les autres valeurs. Par exemple, la table de contingence est entièrement décrite par les deux fréquences relatives marginales $supp(\mathtt{A})$ et $supp(\mathtt{B})$, et la fréquence relative conjointe $supp(\mathtt{AB})$.
Une règle d'association sur une base de données étant définie par deux motifs, on peut faire le raccourci de parler de table de contingence d'une règle d'association, ce qui nous amène à proposer la notion de système descripteur.

**Définition 1.** Nous appelons *système descripteur* de la table de contingence un triplet de fonctions $(f, g, h)$ sur les règles d'association qui permet de décrire entièrement la table de contingence des règles d'association.

**Exemple 1.** Définissons les fonctions suivantes :

$$ant(\mathtt{A} \rightarrow \mathtt{B}) = supp(\mathtt{A});\ cons(\mathtt{A} \rightarrow \mathtt{B}) = supp(\mathtt{B}); conf(\mathtt{A} \rightarrow \mathtt{B}) = \frac{supp(\mathtt{AB})}{supp(\mathtt{A})}$$

Le triplet $(conf, ant, cons)$ est un système descripteur de la table de contingence. De même, si l'on définit les fonctions $ex(\mathtt{A} \rightarrow \mathtt{B}) = supp(\mathtt{AB})$ et $c\text{-}ex(\mathtt{A} \rightarrow \mathtt{B}) = supp(\mathtt{A}\bar{\mathtt{B}})$, les triplets $(ex, ant, cons)$ et $(c\text{-}ex, ant, cons)$ sont des systèmes descripteurs.

Une mesure d'intérêt objective est une fonction de l'espace des règles d'association dans l'espace des réels étendus ($\mathbb{R} \cup \{-\infty, +\infty\}$). Elle sert à quantifier l'intérêt des règles. Il existe un grand nombre de mesures objectives (Tan et al. (2004); Yao et al. (2006); Geng et Hamilton (2006); Lenca et al. (2008)). Celles-ci s'expriment en fonction de la table de contingence, et peuvent donc être considérées comme des fonctions de trois variables (un système descripteur de la table de contingence). Dans ce travail, nous nous intéressons uniquement à ce type de mesures en approfondissant notamment le travail de Omiecinski (2003) et Xiong et al. (2003) pour la *all-confidence*.

## 3.3 Domaine minimal conjoint

Comme il existe des variables aléatoires sur un univers probabiliste, il est possible de définir des variables sur l'espace des règles d'association. Un système descripteur $d$ de la table de contingence est alors un triplet de variables sur cet espace, et une mesure d'intérêt $m$ peut être exprimée à l'aide d'une fonction $\phi_m$ du triplet. Si l'on veut étudier cette fonction, il suffit de se restreindre à l'étude sur le domaine de variation conjoint de ce triplet. Cette étude n'aura d'ailleurs pas de sens en dehors de ce domaine, où les points ne correspondent à aucune situation réelle.

**Définition 2.** Nous appelons le couple $(\phi_m, \mathcal{D}_d)$, formé par cette fonction et le domaine de variation conjoint (associé à un système descripteur), la *fonction de mesure $d$-adaptée* à la mesure $m$.

Il est important de voir que l'écriture de la partie fonctionnelle de cette fonction de mesure dépend du système descripteur choisi. Cependant, une fois ce système fixé, la fonction de mesure adaptée est définie de manière unique. Nous nous permettrons donc, lorsque toute ambiguïté sera levée, d'omettre de préciser le système descripteur choisi.
Si $d$ est un système de descripteurs de la table de contingence, le domaine de variation conjoint associé à ce système est défini par les contraintes imposées par les valeurs entre elles.

**Exemple 2.** Soit la fonction $ex : (\mathtt{A} \xrightarrow{\mathcal{DB}} \mathtt{B}) \mapsto supp_{\mathcal{DB}}(\mathtt{AB})$. On définit, sur la base des fonctions introduites précédemment, le système descripteur $d_{ex} = (ex, ant, cons)$. Nous avons sur ce système l'ensemble de contraintes suivant :

$$\begin{array}{rcccl} 0 & < & ant & < & 1 \\ 0 & < & cons & < & 1 \\ \max\{0, ant + cons - 1\} & \leq & ex & \leq & \min\{ant, cons\} \end{array}$$

Détaillons ici la moins triviale de ces contraintes, $ant + cons - 1 \leq ex$ :

$$supp(\mathtt{AB}) = supp(\mathtt{A}) - supp(\mathtt{A}\bar{\mathtt{B}}) \geq supp(\mathtt{A}) - supp(\bar{\mathtt{B}}) \geq supp(\mathtt{A}) - 1 + supp(\mathtt{B})$$

Définissons alors le domaine :

$$D = \left\{ (x, y, z) \in \mathbb{Q}^3 \middle| \left\{ \begin{array}{rcccl} 0 & < & y & < & 1 \\ 0 & < & z & < & 1 \\ \max\{0, y + z - 1\} & \leq & x & \leq & \min\{y, z\} \end{array} \right. \right\}.$$

Nous avons pour le moment l'affirmation : $\mathcal{D}_{d_{ex}} \subseteq D$ ($D$ est complet). Pour montrer l'inclusion réciproque (i.e. $D$ est minimal), il faut prouver que pour tout élément de $D$, il existe une règle d'association (et donc une base de données) qui lui correspond.
Soit à cet effet $(x, y, z) \in D$, posons $n \in \mathbb{N}$ tel que $(x \cdot n, y \cdot n, z \cdot n) \in \mathbb{N}^3$ (un tel $n$ existe car $x$, $y$ et $z$ sont rationnels). Construisons la base de données $\mathcal{B}$ du tableau 1. La définition du domaine donne un sens à cette base de données en assurant les inégalités : $0 < x < y < y + z - x < 1$. Alors dans $\mathcal{B}$, on a les égalités $supp(\mathtt{AB}) = x$; $supp(\mathtt{A}) = y$; $supp(\mathtt{B}) = z$. Ainsi $(x, y, z)$ est bien une valeur possible du triplet $(ex, ant, cons)$. Finalement, nous avons montré que le domaine de variation conjoint du triplet $d_{ex}$ est exactement $D$. L'étude des variations des fonctions $\phi_m$ peut donc se limiter à ce domaine.

| | 1 | | $x \cdot n$ | | | $y \cdot n$ | | | $(y+z-x) \cdot n$ | | | $n$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1 | $\cdots$ | 1 | 1 | $\cdots$ | 1 | 0 | $\cdots$ | 0 | 0 | $\cdots$ | 0 |
| B | 1 | $\cdots$ | 1 | 0 | $\cdots$ | 0 | 1 | $\cdots$ | 1 | 0 | $\cdots$ | 0 |
| | | AB | | | A¬B | | | | B¬A | | | |

TAB. 1 – *Base de données représentant le triplet $(x, y, z)$. Pour des raisons de place, les individus sont donnés en colonne, les attributs en ligne. On a $supp(AB) = x$; $supp(A) = y$; $supp(B) = z$.*

# 4 Généralisation de la all-confidence

Dans la suite, nous nous concentrons sur le système de descripteurs $d_{ex}$. Il ne sera donc plus précisé. Lorsque nous étudions les variations d'une fonction $f$ définie sur le domaine $D$ "par rapport à sa 1ère variable", nous étudions en fait les variations de la fonction définie pour tout couple $(y, z)$ sur le domaine $\{x|(x, y, z) \in D\}$ par $x \mapsto f(x, y, z)$ .

**Définition 3.** (all-mesure) Soit $m$ une mesure d'intérêt des règles d'association. Nous définissons la mesure $all\text{-}m$ sur un motif $I$ à partir de $m$ de la façon suivante :

$$all\text{-}m(I) \quad = \quad \min_{\mathtt{AB}=I, \mathtt{A}\neq\emptyset, \mathtt{B}\neq\emptyset} \{m(\mathtt{A} \rightarrow \mathtt{B})\}.$$

La all-mesure d'un motif $I$ est donc la plus petite valeur que puisse prendre la mesure considérée sur une règle extraite du motif $I$.

**Définition 4.** (all-monotonie) $m$ est dite all-monotone si la mesure $all\text{-}m$ est antimonotone.

Cette propriété très intéressante du point de vue algorithmique a été démontrée par Omiecinski (2003) pour la confiance. Nous indiquons que ce résultat est vérifié pour les transformées monotones de la confiance :

**Propriété 1.** *La confiance, et toutes les mesures fonctions croissantes de la confiance seule sont all-monotones.*

*Démonstration.* En effet, soit $m$ une telle mesure, et $I \subset I'$ deux motifs. Puisque $m$ est croissante en fonction de la confiance, $\min_{\mathtt{AB}=I} m(conf(\mathtt{A} \rightarrow \mathtt{B})) = m\left(\min_{\mathtt{AB}=I} conf(\mathtt{A} \rightarrow \mathtt{B})\right)$. Et comme la confiance est elle même all-monotone, $m\left(\min_{\mathtt{AB}=I} conf(\mathtt{A} \rightarrow \mathtt{B})\right) \geq m\left(\min_{\mathtt{AB}=I'} conf(\mathtt{A} \rightarrow \mathtt{B})\right)$. Finalement, $m$ est all-monotone :

$$\min_{\mathtt{AB}=I} m(conf(\mathtt{A} \rightarrow \mathtt{B})) = m\left(\min_{\mathtt{AB}=I} conf(\mathtt{A} \rightarrow \mathtt{B})\right) \geq m\left(\min_{\mathtt{AB}=I'} conf(\mathtt{A} \rightarrow \mathtt{B})\right) = \min_{\mathtt{AB}=I'} m(conf(\mathtt{A} \rightarrow \mathtt{B}))$$

□

En revanche, nos résultats (théorèmes 1 et 2) montrent qu'un grand nombre de mesures classiquement étudiées dans la littérature ne vérifient pas cette propriété.

**Théorème 1.** *Soit $m$ une mesure d'intérêt des règles d'association, et $(\phi_m, \mathcal{D}_{d_{ex}})$ la fonction de mesure adaptée à $m$. Si $\phi_m$ est strictement décroissante en la deuxième (antécédents) et troisième (conséquents) variable, alors $m$ n'est pas all-monotone.*

*Démonstration.* **Supposons que cette fonction de mesure soit décroissante strictement en la deuxième et la troisième variable.** Nous allons montrer que pour de telles mesures, il est possible de construire une base de données simple qui sert de contre-exemple à la all-monotonie. Choisissons de travailler sur le nombre minimal d'attributs dont nous avons besoin pour étudier une situation de all-monotonie, c'est à dire 3. Nommons les A, B et C. Nous voulons donc comparer dans une certaine base de données les règles issues du motif AB et du motif ABC. Nous allons donner la définition de valeurs rationnelles $\{x, y, z, t, y', z'\}$ en fonction de diverses contraintes, avec pour but d'obtenir dans notre base de données les relations :

$$supp(\mathtt{AB}) = x,\ supp(\mathtt{A}) = y,\ supp(\mathtt{B}) = z$$
$$supp(\mathtt{ABC}) = x,\ supp(\mathtt{C}) = t,\ supp(\mathtt{AC}) = y',\ supp(\mathtt{BC}) = z'$$

Les propriétés des fréquences impliquent :
- $0 < supp(\mathtt{A}) < 1$, c'est-à-dire $0 < y < 1$ ;
- $0 < supp(\mathtt{B}) < 1$, c'est-à-dire $0 < z < 1$ ;
- $0 < supp(\mathtt{ABC}) \leq \min(supp(\mathtt{AC}), supp(\mathtt{BC}))$, c'est-à-dire $0 < x \leq \min(y', z')$ ;
- $y' = supp(\mathtt{AC}) < supp(\mathtt{A}) = y$ ;
- $z' = supp(\mathtt{BC}) < supp(\mathtt{B}) = z$ ;
- $t = supp(\mathtt{C}) \geq supp(\mathtt{AC}) + supp(\mathtt{BC}) - supp(\mathtt{ABC}) = y' + z' - x$.

Nous posons des inégalités strictes pour pouvoir exploiter pleinement la stricte décroissance de la mesure. Supposons de plus par convention que l'attribut C apparaît moins fréquemment que B, c'est à dire $t \leq z$.
Résumons donc les contraintes qui permettent de définir nos rationnels :

$$y \in ]0,1[ \quad (1) \qquad z \in ]0,1[ \quad (2)$$

$$0 < y' < y \quad (3) \qquad 0 < z' < z \quad (4) \qquad 0 < x \leq \min\{z', y'\} \quad (5)$$

$$y' + z' - x < t \quad (6) \qquad t \leq z \quad (7) \qquad t + (y - y') + (z - z') \leq 1 \quad (8)$$

**Remarque 1.** Notons immédiatement que les valeurs $y = \frac{4}{16}$, $z = \frac{8}{16}$, $y' = \frac{2}{16}$, $z' = \frac{3}{16}$, $x = \frac{1}{16}$, $t = \frac{7}{16}$ vérifient ces contraintes et permettent d'engendrer la base de données décrite table 2.

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| B | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| C | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 |

TAB. 2 – *Base de données à 16 transactions*

**Remarque 2.** D'une manière plus générale, on s'aperçoit que pour un entier $n$ donné, toute base de données à trois attributs $\{\mathtt{A}, \mathtt{B}, \mathtt{C}\}$ et $n+3$ transactions $\mathtt{T}_1, \cdots, \mathtt{T}_{n+3}$ réparties comme suit : $\mathtt{T}_1 = \{\mathtt{A}\}, \mathtt{T}_2 = \{\mathtt{B}\}, \mathtt{T}_3 = \{\mathtt{C}\}, \mathtt{T}_4 = \mathtt{T}_5 = \cdots = \mathtt{T}_{n+3} = \{\mathtt{A}, \mathtt{B}, \mathtt{C}\}$, vérifie les contraintes.

Sur chacune de ces bases de données, les contraintes nous permettent d'écrire les inégalités suivantes (où on trouve au dessus du signe d'inégalité la contrainte qui permet le passage), en prenant en compte les hypothèses de variations :

$$\begin{array}{rcl}
m(\mathtt{A} \rightarrow \mathtt{B}) = \phi_m(x,y,z) & \overset{3}{<} & \phi_m(x,y',z) = m(\mathtt{AC} \rightarrow \mathtt{B}) \\
& \overset{4}{<} & \phi_m(x,y,z') = m(\mathtt{A} \rightarrow \mathtt{BC}) \\
& \overset{5+3}{<} & \phi_m(x,x,z) \overset{7}{\leq} \phi_m(x,x,t) = m(\mathtt{AB} \rightarrow \mathtt{C}) \\
m(\mathtt{B} \rightarrow \mathtt{A}) = \phi_m(x,z,y) & \overset{4}{<} & \phi_m(x,z',y) = m(\mathtt{BC} \rightarrow \mathtt{A}) \\
& \overset{3}{<} & \phi_m(x,z,y') = m(\mathtt{B} \rightarrow \mathtt{AC}) \\
& \overset{5+4}{<} & \phi_m(x,z,x) \overset{7}{\leq} \phi_m(x,t,x) = m(\mathtt{C} \rightarrow \mathtt{AB})
\end{array}$$

Ces inégalités nous donnent la conclusion : $all\text{-}m(\{\mathtt{A}, \mathtt{B}\}) < all\text{-}m(\{\mathtt{A}, \mathtt{B}, \mathtt{C}\})$, **ce qui contredit l'antimonotonie**. □

**Exemple 3.** La fonction de mesure adaptée à la mesure du facteur de Bayes à la forme suivante : $\phi_{BF}(x,y,z) = \frac{x}{y-x} \times \frac{1-z}{z}$. Elle possède les propriétés de décroissance énoncées précédemment. Elle doit donc ne pas être all-monotone. Calculons $BF$ sur la base de la table 2 :

$$\begin{array}{llll}
BF(\mathtt{A} \rightarrow \mathtt{B}) = \frac{1}{3} & BF(\mathtt{AC} \rightarrow \mathtt{B}) = 1 & BF(\mathtt{B} \rightarrow \mathtt{AC}) = 1 & BF(\mathtt{BC} \rightarrow \mathtt{A}) = \frac{3}{2} \\
BF(\mathtt{B} \rightarrow \mathtt{A}) = \frac{3}{7} & BF(\mathtt{A} \rightarrow \mathtt{BC}) = \frac{13}{9} & BF(\mathtt{AB} \rightarrow \mathtt{C}) = \infty & BF(\mathtt{C} \rightarrow \mathtt{AB}) = \infty
\end{array}$$

Ces calculs mettent en évidence l'inexistence de la all-monotonie, puisque $all\text{-}BF(\{\mathtt{A}, \mathtt{B}\}) < all\text{-}BF(\{\mathtt{A}, \mathtt{B}, \mathtt{C}\})$.

**Remarque 3.** L'hypothèse de stricte monotonie est trop forte. Voyons la mesure de Loevinger :$L(\mathtt{A} \rightarrow \mathtt{B}) = 1 - \frac{supp(\mathtt{A}\bar{\mathtt{B}})}{supp(\mathtt{A})supp(\bar{\mathtt{B}})}$. Sa fonction de mesure adaptée sur $\mathcal{D}$ est $\phi_L(x,y,z) = 1 - \frac{y-x}{y \cdot (1-z)}$. Et sur le plan défini par $x - y = 0$, sa dérivée par rapport à la troisième variable est nulle. Il n'y a donc pas stricte monotonie par rapport à la troisième variable. Pourtant, sur la base de données définie précédemment :

$$\begin{array}{llll}
L(\mathtt{A} \rightarrow \mathtt{B}) = -\frac{1}{2} & L(\mathtt{AC} \rightarrow \mathtt{B}) = 0 & L(\mathtt{B} \rightarrow \mathtt{AC}) = 0 & L(\mathtt{BC} \rightarrow \mathtt{A}) = \frac{1}{9} \\
L(\mathtt{B} \rightarrow \mathtt{A}) = -\frac{1}{6} & L(\mathtt{A} \rightarrow \mathtt{BC}) = \frac{1}{13} & L(\mathtt{AB} \rightarrow \mathtt{C}) = 1 & L(\mathtt{C} \rightarrow \mathtt{AB}) = \frac{3}{35}
\end{array}$$

La mesure de Loevinger n'est donc pas all-monotone. D'autres mesures présentent la même particularité en la deuxième variable sur le plan $x - z = 0$.

**Théorème 2.** *Soit $m$ une mesure d'intérêt des règles d'association, et $(\phi_m, \mathcal{D}_{d_{ex}})$ la fonction de mesure adaptée à $m$. Si $\phi_m(x,y,z)$ est strictement décroissante en la deuxième et troisième variable, ailleurs que sur les plans $x - z = 0$ et $x - y = 0$, où elle peut éventuellement être constante respectivement en la deuxième et la troisième variable, alors $m$ n'est pas all-monotone.*

*Démonstration.* La démonstration est la même que pour le théorème 1, les inégalités $\overset{7}{\leq}$ devenant des égalités. □

**Remarque 4.** La réciproque du théorème 2 n'est pas vérifiée et nous n'avons donc pas de condition nécessaire et suffisante. Le cas de la mesure *contramin* (Azé et Kodratoff (2004)) justifie cette affirmation. On l'applique à une base de données de taille $n$ décrite remarque 2. Puisque $contr(\mathtt{A} \rightarrow \mathtt{B}) = \frac{supp(\mathtt{AB}) - supp(\mathtt{A\overline{B}})}{supp(\mathtt{B})}$, on mesure alors : $all\text{-}contr(\{\mathtt{A}, \mathtt{B}\}) = \frac{n-1}{n+1}$ et $all\text{-}contr(\{\mathtt{A}, \mathtt{B}, \mathtt{C}\}) = \frac{n-1}{n}$. C'est-à-dire $all\text{-}contr(\{\mathtt{A}, \mathtt{B}\}) < all\text{-}contr(\{\mathtt{A}, \mathtt{B}, \mathtt{C}\})$ dès que $n > 1$. Ainsi $contr$ n'est pas all-monotone. Sa fonction de mesure adaptée est $\frac{2x-y}{z}$, dont le sens de variation par rapport à $z$ dépend du signe de $2x - y$ : $contr$ n'est pas décroissante en fonction de sa troisième variable, pas même au sens large.

Ces travaux nous permettent d'apporter trois types de conclusion sur les mesures issues de la littérature (Tan et al. (2004); Geng et Hamilton (2006); Lenca et al. (2008)), la première s'appuyant sur le théorème 2, la deuxième sur la propriété 1 concernant les mesures fonctions de la confiance, et la dernière mettant en avant le lien entre *recall* et *confiance* :

- En combinant le résultat du théorème 2 et, pour certains cas, l'utilisation des différents contre-exemples des remarques 1 et 2, nous observons qu'un certain nombre de mesures ne vérifient pas la propriété de all-monotonie : *spécificité*, *précision*, *lift*, *levier*, *confiance centrée*, *Jaccard*, *confiance positive*, *odds ratio*, *Klosgen*, *contramin*, *valeur ajoutée*, *conviction*, *one way support*, $J_1$*-mesure*, *Piatetsky-Shapiro*, *cosine*, *Loevinger*, *gain informationnel*, *facteur bayésien*, *Zhang*, *index d'implication* et *kappa* ;
- A l'opposé, les mesures de *confiance*, *Sebag-Shoenauer*, *Ganascia* et le *taux d'exemples-contre exemples* vérifient cette propriété, car elles sont des fonctions croissantes de la confiance (elles classent les règles de la même façon mais ces mesures se différencient selon des objectifs utilisateurs, Lenca et al. (2008)), il est donc possible de les utiliser dans un algorithme de type Apriori pour réaliser un élagage du treillis des motifs ;
- Enfin, concernant la mesure de *recall*, on remarque que pour deux itemsets $\mathtt{A}$ et $\mathtt{B}$, $rec(\mathtt{A} \rightarrow \mathtt{B}) = \frac{supp(\mathtt{AB})}{supp(\mathtt{B})} = conf(\mathtt{B} \rightarrow \mathtt{A})$. Ainsi, pour un itemset $I$, $\{rec(\mathtt{A} \rightarrow \mathtt{B}) | \mathtt{AB} = I\} = \{conf(\mathtt{A} \rightarrow \mathtt{B}) | \mathtt{AB} = I\}$. Donc $all\text{-}rec(I) = all\text{-}conf(I)$, et la mesure de *recall* est all-montone.

# 5 Conclusion et perspectives

Les mesures objectives d'intérêt des règles possèdent quelques propriétés algorithmiques qui n'ont pour l'instant été que peu exploitées, mis à part pour la confiance. Nous proposons dans cet article un cadre formel pour l'étude analytique des mesures. À l'aide de ce cadre, nous pouvons mettre en évidence les propriétés intrinsèques des mesures, et faire le lien avec des propriétés algorithmiques existantes. La all-confidence est l'une de ces propriétés, puisqu'elle permet de rendre antimonotone la confiance. Nous appliquons ce cadre pour répondre à la question suivante : existe-t-il d'autres mesures pouvant être rendues antimonotones de la même manière ? Nos résultats montrent qu'un grand nombre de mesures ne vérifient pas la propriété de all-monotonie que nous avons définie par extension de la all-confidence. Cependant, quelques mesures peuvent être rendues antimonotones par ce procédé.
Nous montrons également que certaines propriétés algorithmiques dépendent en fait de propriétés analytiques des mesures. Notre cadre formel fournit un moyen systématique d'étude des mesures, le but principal étant, pour une mesure donnée, de pouvoir déterminer les propriétés algorithmiques qu'elle vérifie, et donc les algorithmes qui seront utilisables avec cette mesure.

# Références


Agrawal, R., T. Imieliski, et A. Swami (1993). Mining association rules between sets of items in large databases. In *ICMD*, pp. 207–216.

Agrawal, R. et R. Srikant (1994). Fast algorithms for mining association rules in large databases. In *VLDB*, pp. 478–499.

Azé, J. et Y. Kodratoff (2004). Extraction de "pépites" de connaissance dans les données : une nouvelle approche et une étude de la sensibilité au bruit. *RNTI 1*, 247–270.

Bayardo, Jr, R. J., R. Agrawal, et D. Gunopulos (1999). Constraint-based rule mining in large, dense databases. In *ICDE*, pp. 188–197.

Bonchi, F. et C. Lucchese (2005). Pushing tougher constraints in frequent pattern mining. In *PAKDD*, pp. 114–124.

Breiman, L., J. Friedman, R. Olshen, et C. Stone (1984). *Classification and Regression Trees*. Wadsworth and Brooks.

Clark, P. et T. Niblett (1989). The CN2 induction algorithm. *ML 3*, 261–283.

Cohen, E., M. Datar, S. Fujiwara, A. Gionis, P. Indyk, R. Motwani, J. D. Ullman, et C. Yang (2001). Finding interesting associations without support pruning. *TKDE 13*(1), 64–78.

Diatta, J., H. Ralambondrainy, et A. Totohasina (2007). Towards a unifying probabilistic implicative normalized quality measure for association rules. In *QMDM*, pp. 237–250. Springer.

Geng, L. et H. J. Hamilton (2006). Interestingness measures for data mining: A survey. *ACM 38*(3, Article 9).

Goethals, B. (2005). Frequent set mining. In *The Data Mining and Knowledge Discovery Handbook*, pp. 377–397. Springer.

Goethals, B. et M. J. Zaki (2004). Advances in frequent itemset mining implementations: report on FIMI'03. *SIGKDD 6*(1), 109–117.

Han, J., J. Pei, et Y. Yin (2000). Mining frequent patterns without candidate generation. In *ICMD*, pp. 1–12.

Hébert, C. et B. Crémilleux (2007). A unified view of objective interestingness measures. In *MLDM*, pp. 533–547.

Hipp, J., U. Güntzer, et G. Nakhaeizadeh (2000). Algorithms for association rule mining — a general survey and comparison. *SIGKDD 2*(1), 58–64.

Koh, Y. S. (2008). Mining non-coincidental rules without a user defined support threshold. In *PAKDD*, pp. 910–915.

Le Bras, Y., P. Lenca, et S. Lallich (2009). Optimal rules discovery: a framework and a necessary and sufficient condition of antimonotonicity. In *PAKDD (accepted)*.

Lenca, P., P. Meyer, B. Vaillant, et S. Lallich (2008). On selecting interestingness measures for association rules: user oriented description and multiple criteria decision aid. *EJOR 184*(2), 610–626.

Li, J. (2006). On optimal rule discovery. *TKDE 18*(4), 460–471.

Morishita, S. et J. Sese (2000). Transversing itemset lattices with statistical metric pruning. In *PODS*, pp. 226–236.

Omiecinski, E. (2003). Alternative interest measures for mining associations in databases. *TKDE 15*(1), 57–69.

Pei, J. et J. Han (2000). Can we push more constraints into frequent pattern mining ? In *KDD*, pp. 350–354.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

Savasere, A., E. Omiecinski, et S. B. Navathe (1995). An efficient algorithm for mining association rules in large databases. In *VLDB*, pp. 432–444.

Slowiński, R., S. Greco, et I. Szczęch (2008). Analysis of monotonicity properties of new normalized rule interestingness measures. In *HCP*, Volume 1, pp. 231–242.

Tan, P.-N., V. Kumar, et J. Srivastava (2004). Selecting the right objective measure for association analysis. *IS 4*(29), 293–313.

Wang, K., Y. He, et D. W. Cheung (2001). Mining confident rules without support requirement. In *IKM*, pp. 89–96.

Xiong, H., P.-N. Tan, et V. Kumar (2003). Mining strong affinity association patterns in data sets with skewed support distribution. In *ICDM*, pp. 387–394.

Yao, Y., Y. Chen, et X. Yang (2003). A measurement-theoretic foundation for rule interestingness evaluation. In *ICDM WS FDM*, pp. 221–227.

Yao, Y., Y. Chen, et X. D. Yang (2006). A measurement-theoretic foundation of rule interestingness evaluation. In *FNADM*, pp. 41–59. Springer.

Zaki, M. J. (2000). Scalable algorithms for association mining. *TKDE 12*(3), 372–390.

Zaki, M. J. (2004). Mining non-redundant association rules. *DMKD 9*(3), 223–248.

Zimmermann, A. et L. de Raedt (2004). CorClass: Correlated association rule mining for classification. In *DS*, pp. 60–72.

## Summary

Many studies have shown the limits of the support/confidence framework used in Apriori-like algorithms, both from the quantitative (quantity of rules) and from the qualitative (redundance, interest, nuggets of knowledge) viewpoints. One solution is to focus as soon as possible on interesting rules, by using interestingness measures possessing algorithmic properties, but also a high power of highlighting very interesting rules. They allow rules of high interest to be found without a preliminary support pruning. They thus make the mining of nuggets of knowledge possible. This is the case for all-confidence, or h-confidence, an antimonotonic transformation of confidence. We here study the possibility of applying such a transformation to a large panel of measures thanks to a necessary condition based on a formal framework that we proposed. However, we show that among the 27 measures studied, only five of them are antimonotone. Then even though very interesting, this antimonotone property is not applicable to a large set of measures.

# Introduction de l'élagage pour l'extraction de règles d'association de classe sans génération de candidats

Emna Bahri*, Stéphane Lallich*

*Laboratoire ERIC, Université de Lyon, 5, avenue Pierre Mendès-France, 69500, Bron
emna.bahri | stephane.lallich@univ-lyon2.fr, http://eric.univ-lyon2.fr

**Résumé.** Ce travail porte sur l'extraction des règles d'association prédictive en apprentissage supervisé. On sait que les algorithmes d'extraction de règles d'association développés en apprentissage non-supervisé extraient toutes les règles dont le support et la confiance dépassent des seuils préfixés. L'extraction des règles d'association prédictive à l'aide de ces algorithmes, avec un filtrage a posteriori des règles, pose plusieurs problèmes. Dans la première étape, on extrait des itemsets fréquents inutiles, ceux qui ne contiennent pas la classe, alors que la seconde étape peut être simplifiée, puisqu'un itemset contenant la classe ne donne lieu qu'à une seule règle de classe. Pour travailler avec les seuils de support les plus bas possibles, nous proposons FCP-Growth-P une adaptation de FP-Growth qui a différents avantages. Tout d'abord, FCP-Growth-P élimine les itemsets fréquents qui ne contiennent pas d'item de classe. D'autre part, pour ne pas désavantager la classe la moins nombreuse, nous adaptons le seuil de support afin de travailler avec le même seuil de support à l'intérieur de chaque classe. Enfin, pour un meilleur élagage, FCP-Growth-P intègre la condition d'élagage de Li sur la spécialisation des règles.

## 1 Introduction

Cet article porte sur l'apprentissage supervisé par règles à partir des tableaux attributs-valeurs issus des grandes bases de données. Ce type d'apprentissage est particulièrement intéressant car il fournit des connaissances directement interprétables, ce qui est très utile dans certains domaines, par exemple la médecine ou le scoring bancaire. Parmi les méthodes d'apprentissage supervisé à partir de règles, on citera notamment les arbres de décision, pour une présentation desquels nous renvoyons à Rakotomalala (2005), et la classification associative initiée par Liu et al. (1998), qui organise la prédiction de la classe à partir des règles d'association dont le conséquent est l'une des modalités de la variable de classe, appelées règles d'association de classe. C'est à cette dernière méthode qu'est consacré notre travail.

Comme nous le verrons en section 2, pour construire les règles de classe, de nombreux algorithmes de classification associative utilisent les algorithmes d'extraction de règles d'association conçus dans le cadre de l'apprentissage non-supervisé. Ces algorithmes d'extraction de règles d'association, par exemple Apriori (Agrawal et Srikant (1994)) ou FP-Growth (Han et al. (2000)), dont le principe sera rappelé en section 2, extraient toutes les règles dont le support et la confiance dépassent des seuils préfixés. Dans une première étape, ces algorithmes extraient tous les itemsets fréquents, ceux dont le support dépasse le seuil de support, noté $minsupp$. C'est l'étape la plus coûteuse en termes de temps d'exécution, car le nombre de ces items fréquents dépend exponentiellement du nombre d'items manipulés (pour $p$ items, on a $2^p$ itemsets possibles). Pour que cette étape soit exécutable, on doit parfois augmenter le seuil de support, ce qui donne un rôle central à la condition de support et rend plus difficile l'extraction de pépites, à savoir les règles ayant une très forte confiance et un très faible support. Lors de la seconde étape, ces algorithmes subdivisent les itemsets fréquents en deux parties disjointes, pour construire toutes les règles possibles, en ne gardant que les règles dont la confiance dépasse le seuil de confiance choisi, noté $minconf$.

Cette utilisation des algorithmes non-supervisés pour déterminer les règles de classe, assortie d'un filtrage ultérieur des règles, présente différents inconvénients. D'une part, dans la première étape, on extrait des itemsets fréquents inutiles, à savoir ceux qui ne contiennent pas une modalité de classe, ce qui empêche de baisser le seuil de support autant qu'on le voudrait. D'autre part, la seconde étape peut être simplifiée, puisqu'un itemset contenant l'une des modalités de la classe (ou itemset de classe) ne peut donner lieu qu'à une seule règle de classe. Par ailleurs, l'utilisation d'un seuil de support s'appliquant sur la totalité de la base désavantage la classe minoritaire qui est généralement la plus intéressante et la plus difficile à prédire correctement. Enfin, ces différents algorithmes se fondent seulement sur le support et la confiance pour extraire les règles intéressantes, alors que différents travaux (Guillet et Hamilton (2007)) ont montré qu'on peut améliorer la qualité des règles extraites en utilisant d'autres mesures d'intérêt, plus pertinentes que la confiance et le support.

Face à ces problèmes, nous présentons FCP-Growth une adaptation de FP-Growth qui dans l'étape de construction des itemsets fréquents permet de ne générer que les itemsets fréquents de classe. En outre, pour tenir compte d'un éventuel

déséquilibre des classes, le même seuil de support est utilisé à l'intérieur de chaque classe. Enfin, nous introduisons une condition d'élagage dans FCP-Growth pour éviter d'extraire les spécialisations de règles qui n'ont pas d'intérêt, définissant ainsi l'algorithme FCP-Growth-P. Cet article présente d'abord un état de l'art des méthodes de construction de règles d'association prédictive, en s'intéressant plus particulièrement aux différents algorithmes de recherche d'itemsets fréquents et aux différentes méthodes utilisées pour l'élagage (Section 2). Puis, nous expliquons les améliorations apportées à FP-Growth (Section 3) et nous comparons les résultats de FCP-Growth-P et de FCP-Growth à ceux de FP-Growth, la méthode standard, après les avoir appliqués à sept bases de données différentes (Section 4). Nous terminons en dressant les perspectives qu'ouvre ce travail (Section 5).

# 2 Etat de l'art

Après avoir présenté les principaux algorithmes de classification associative, nous rappelons les différentes méthodes d'extraction de règles d'association en non-supervisé et les stratégies d'élagage.

## 2.1 Principaux algorithmes de classification associative

Depuis l'article fondateur de Liu et al. (1998), différents travaux ont mis en évidence les bons résultats de la classification associative en termes de diminution du taux d'erreur. La classification associative est fondée sur la prédiction de la classe à partir de règles d'association particulières, dites règles d'association de classe (*class association rules*) ou règles d'association prédictives. Une règle d'association de classe est une règle dont le conséquent doit être la variable indicatrice de l'une des modalités de la classe. Une telle règle s'écrit donc $A \rightarrow c_i$, où $A$ est une conjonction de descripteurs booléens et $c_i$ est la variable indicatrice de la $i_e$ modalité de classe. L'intérêt des règles de classe est de permettre la focalisation sur des groupes d'individus, éventuellement très petits, homogènes du point de vue des descripteurs et présentant la même classe. Par rapport aux arbres de décision, la classification associative a l'intérêt de ne pas pratiquer de stratégie gloutonne, ce qui permet d'explorer l'ensemble des règles construites.

Les méthodes de classification associative procèdent en deux phases, une phase de construction des règles d'association de classe suivie d'une phase de prédiction à partir des règles de classe obtenues. La première phase se décompose elle-même en deux étapes, l'une consacrée à l'extraction des itemsets fréquents, compte tenu du seuil de support choisi, l'autre portant sur la construction des règles d'association de classe satisfaisant le seuil de confiance préfixé.

Parmi ces méthodes, on citera les algorithmes CBA (*Classification Based on Association*, Liu et al. (1998)), qui utilise Apriori pour générer toutes les règles d'association satisfaisant aux seuils de support et de confiance choisis au départ, CPAR (*Classification based on Predictive Association Rules*, Yin et Han), qui s'inspire du principe de la méthode FOIL (*First Order Inductive Learner*, Quinlan et Cameron-Jones (1995)) pour générer les règles, évitant ainsi la répétition des calculs pour les règles redondantes et CMAR (*Classification based on Multiple Association Rules*, Li et al. (2001a)) qui utilise FP-Growth pour extraire les règles. Ces méthodes de classification associative ont donné de bons résultats. Cependant, elles se fondent sur une première phase d'extraction des règles d'association qui n'est pas en relation avec la seconde phase de classification.

## 2.2 Extraction des règles d'association

Durant la première phase de leur mise en oeuvre, celle d'extraction des règles, les algorithmes usuels de classification associative ont recours aux algorithmes d'extraction de règles d'association utilisés en apprentissage non-supervisé. Parmi ceux-ci, les plus utilisés sont Apriori et FP-Growth.

La stratégie utilisée par l'algorithme Apriori (Agrawal et Srikant (1994)) pour générer les itemsets fréquents est simple. Il détermine ceux-ci dans un ordre croissant de longueur, en se fondant sur la propriété d'antimonotonie de la condition de support. Selon cette propriété, tous les sur-itemsets d'un itemset non fréquent sont non fréquents et tous les sous-itemsets d'un itemset fréquent sont fréquents. Au niveau $k$ du treillis des itemsets, en fusionnant deux par deux tous les itemsets de longueur $k-1$, Apriori génère tous les itemsets candidats de longueur $k$ et ne retient parmi ceux-ci que ceux dont la fréquence est supérieure ou égale au seuil de support $minsupp$. Ensuite, à partir de chaque itemset fréquent $X$, Apriori examine toutes les règles du type $X - Y \rightarrow Y$, où $Y \subset X$, $Y \neq \oslash$ et ne retient que celles dont la confiance est au moins égale à $minconf$, le seuil de confiance choisi. Ces règles sont alors ajoutées à ER, l'ensemble des règles d'association qui satisfont aux seuils de support et de confiance préfixés.

A l'opposé d'Apriori qui génère des itemsets candidats et qui les teste pour ne conserver que les itemsets fréquents, FP-Growth (Han et al. (2000)) construit les itemsets fréquents sans génération de candidats. En fait, il utilise la stratégie "diviser et dominer" (*divide-and-conquer*). Tout d'abord, il compresse les itemsets fréquents représentés dans la base de données à l'aide d'un FP-Tree (*frequent-pattern tree*) dont les branches contiennent les associations possibles des items.

Chaque association peut être divisée en fragments (*pattern fragment*) qui constituent les itemsets fréquents. La méthode FP-Growth transforme le problème de la recherche de l'itemset fréquent le plus long par la recherche du plus petit et sa concaténation avec le suffixe correspondant (le dernier itemset fréquent de la branche aboutissant à l'item considéré). Ceci permet de réduire le coût de la recherche.

**Exemple explicatif de FP-Growth** : Pour expliquer le principe de FP-Growth, on choisit la base de données simplifiée décrite dans le tableau 1 qui contient des items descriptifs ($I$) et des items de classe ($C$).

| Id Transaction | Liste d'item-TID |
|---|---|
| T1 | $I1, I2, I5, C1$ |
| T2 | $I2, I4, C2$ |
| T3 | $I2, I1, I3, C1$ |
| T4 | $I1, I2, I4, C2$ |

TAB. 1: Exemple

| ITEMS | Support |
|---|---|
| $I2$ | 4 |
| $I1$ | 3 |
| $I4$ | 2 |
| $C1$ | 2 |
| $C2$ | 2 |
| $I3$ | 1 |
| $I5$ | 1 |

TAB. 2: Items associés à leur support

– Premièrement, FP-Growth parcourt la base de donnée et lit tout les items existant en calculant leur support et les compare avec le seuil de support préfixé. Ainsi, il n'accepte que les items ayant un support supérieur ou égal au seuil appelés items fréquents. On choisit ici un seuil de support absolu égal à 1. FP-Growth enregistre les items fréquents dans une liste $L$ où ils figurent par ordre décroissant de support. Dans notre exemple, la liste correspondante est : $L = I2 : 4, I1 : 3, I4 : 2, C1 : 2, C2 : 2, I3 : 1, I5 : 1$

– Deuxièmement, un FP-Tree est construit par la création d'une racine vide et un second parcours de la base de données où chaque transaction est décrite dans l'ordre des items donné par la liste L. Par exemple, la première transaction $T1 : I1, I2, I5, C1$ sera sauvegardée en $T1 : I2, I1, C1, I5$. Cette première transaction, une fois ordonnée, va construire la première branche du FP-Tree avec 4 noeuds $(I2 : 1), (I1 : 1), (C1 : 1)$ et $(I5 : 1)$. La seconde transaction T2, contenant T2 dans l'ordre de $L : I2, I4, C2$, construit une branche où le noeud I2 est lié à la racine, le noeud I4 lié à I2 et le noeud C2 lié à I4. Cette branche partage un préfixe commun I2, avec T1. Ainsi, lors de construction du noeud I4 :1, doit-on incrémenter de 1 le compteur du noeud I2 (I2 :2).

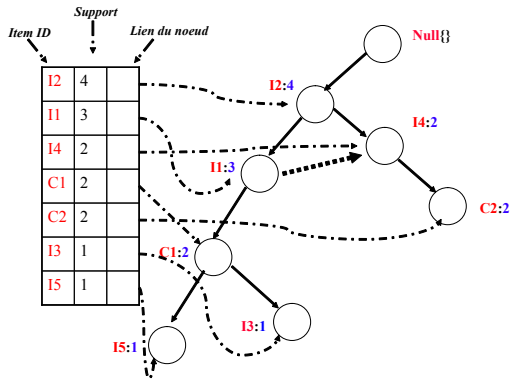


FIG. 1: Construction du FP-Tree

– En dernier lieu, le FP-Tree est fouillé par la création des (sub-)fragment conditionnels de base. En fait, pour trouver ces fragments, on extrait pour chaque fragment de longueur 1 (*suffix pattern*) l'ensemble des préfixes existant dans le chemin du FP-Tree (*conditional pattern base*). L'itemset fréquent est obtenu par la concaténation du suffixe avec les fragments fréquents extraits des FP-Tree conditionnels (voir tableau 3)

| Item | Motifs conditionnels | FP-Tree conditionnels | Motifs fréquents |
|---|---|---|---|
| I5 | $I1, I2, C1, I2, I1, I3$ | $(I2 : 2, I1 : 1)$ | $I2, I5 : 2, I1, I5 : 2, I2, I1, I5 : 2$ |
| I4 | $I2 : 1, I2, I1 : 1$ | $(I2 : 2)$ | $I2, I4 : 2$ |

TAB. 3: Fouille du FP-Tree

Par ailleurs, il faut noter que les règles d'association de classe obtenues par filtrage des règles extraites à partir des algorithmes non supervisés se trouvent confrontées au problèmes des classes déséquilibrées. En effet, dans l'étape de génération des itemsets fréquents, ces algorithmes sélectionnent les itemsets avec un seuil de support qui ne prend pas en compte le poids des différentes modalités de la classe à prédire, ce qui désavantage mécaniquement les règles de classe associées aux modalités les moins fréquentes.

Dans notre étude, nous avons choisi de retenir FP-Growth, en raison de sa structuration (FP-Tree), qui est unanimement considéré comme plus efficace qu'Apriori. Nous proposons ainsi FCP-Growth, une variante de FP-Growth qui ne construit que les itemsets de classe, dont le principe est exposé dans (Bahri and Lallich (2009)).

### 2.3 Elagage

Lors de l'extraction des itemsets fréquents, les différents algorithmes cités se basent sur un seul principe d'élagage, celui de la condition de support. Dans la littérature, il existe plusieurs façon d'élaguer qui donnent des bons résultats. On trouve notamment :

1. La règle la plus intéressante (principe du CSA) : étant données deux règles R1 et R2, R1 est plus intéressente que R2 si $conf(R1) > conf(R2)$, ou $conf(R1) = conf(R2)$ et $sup(R1) > sup(R2)$, ou enfin $conf(R1) = conf(R2)$, $sup(R1) = sup(R2)$ et R1 possède moins d'items. Le principe d'élagage correspondant consiste à supprimer la règle la moins générale. Ce principe est notamment utilisé par CMAR.
2. La règle la plus performante : On peut citer le critère FOIL (Quinlan et Cameron-Jones (1995)), utilisé par CPAR (Yin et Han), qui se base sur la mesure "$FOIL - Prune(R) = (pos - neg) \div (pos + neg)$" où $pos$ et $neg$ sont les nombres de t-uples positifs et négatifs couverts par la règle R. Un autre critère possible est la corrélation de l'antécédent d'une règle avec la classe mesurée par le $\chi^2$. Ce critère est utilisé par CMAR (Li et al. (2001a)).
3. Comportement de la règle vis-à-vis des contre-exemples : les travaux de Li (Li (2006), Li et Zhang (2003) et Li et al. (2001b)) montrent que si la spécialisation d'une règle ne diminue pas le nombre de contre-exemples, alors ni cette spécialisation, ni aucune sur-spécialisation n'est plus intéressante au sens de la confiance ou de différentes mesures d'intérêt. C'est ce dernier critère d'élagage, utilisé par Li avec Apriori, que nous avons choisi d'introduire dans l'algorithme FCP-Growth pour obtenir FCP-Growth-P.

## 3 La méthode proposée : FCP-Growth

Afin de diminuer le coût des règles d'association prédictives en termes de temps d'exécution et de stockage, on doit se limiter à la génération des itemsets fréquents de classe, c'est-à-dire les itemsets fréquents qui contiennent l'une des modalités de classe. A cette fin, nous avons proposé FCP-Growth (*Frequent Class Pattern*) une variante de FP-Growth qui ne stocke un itemset fréquent dans le FCP-Tree (*Frequent Class pattern Tree*) que si celui-ci est un itemset de classe, du type $c_i A$. Un tel itemset ne génère qu'une seule règle de classe, la règle $A \rightarrow c_i$. Ceci nous permet un double gain, en temps d'exécution et en espace de stockage.

Simultanément, pour résoudre le problème des classes déséquilibrées, nous proposons de fixer un seuil de support ajustable, qui tient compte de l'effectif de chaque modalité de la classe à prédire. Si $n_i$ désigne l'effectif de $c_i$, la $i^e$ modalité de classe, une fois choisi le seuil $\sigma$ sur la base entière, le seuil de support adapté à la classe $c_i$ s'écrit $\sigma_i$, où $n \times \sigma_i = n_i \times \sigma$. Le seuil de support adapté est donc proportionnel à l'effectif de classe, de telle sorte que si l'on exprime ce seuil en proportion de l'effectif de classe et non plus de l'effectif de la base, celui-ci soit le même pour chaque classe. De la sorte, on évite que les classes les moins nombreuses soient systématiquement désavantagées lors de la construction des itemsets de classe et des règles d'association prédictives associées.

**Algorithme**. Notre adaptation a pour principe d'ajouter des modules à la version de base FP-Gowth.

1. Prétraitement de la base de données : Une fois la classe à prédire choisie par l'utilisateur parmi les attributs, les données sont réorganisées sous forme d'un tableau dont la première colonne contient les différentes modalités assignées à la classe. Cette méthode nous assure de construire le FCP-Tree des seuls itemsets de classe fréquents. Pour expliquer notre méthode et la comparer avec la version d'origine FP-Growth, nous gardons le même exemple. Au terme de la phase de prétraitement, la base de données sera organisée comme indiqué dans le tableau 4.

| TID | List of item-TID |
|---|---|
| T1 | $C1, I1, I2, I5$ |
| T2 | $C2, I2, I4$ |
| T3 | $C1, I2, I1, I3$ |
| T4 | $C2, I1, I2, I4$ |

TAB. 4: Base de données après le prétraitement

| ITEMS | Supp pour C1 | Supp pour C2 |
|---|---|---|
| $I2$ | 4 | 4 |
| $I1$ | 3 | 3 |
| $I4$ | 2 | 2 |
| $I3$ | - | 1 |
| $I5$ | - | 1 |

TAB. 5: Items fréquents accordés au support de classe

2. Ajustement du seuil de support : Cette étape nous permet de définir, non pas un seuil de support fixe comme dans la version de base de FP-Growth, mais un seuil de support adapté à chaque modalité $c_i$, qui dépend de $n_i$, le nombre de transactions validant la modalité. Par exemple, pour la modalité $c_i$, on aura $\sigma_i = \sigma \times n_i/n$. Durant cette phase,

au lieu d'une seule liste L qui contient les items fréquents par ordre décroissant de support comme FP-Growth, FCP-Growth construit autant de listes $L$ que de modalités de classe et attribue aux items de classe un poids $P$ plus grand que le support maximal des autres items, considéré comme support. Ainsi, chaque liste commencera par l'item de classe correspondant à son support mininum de classe. Pour notre exemple, on supposera que le support de $C1$ est 200 et C2 de 100. Si on définit $\sigma$ à $1\%$ alors on aura un support minimun de 2 pour $C1$ et de 1 pour $C2$.

Pour l'exemple, on aura deux listes, $L1$ pour la classe $C1$ et $L2$ pour la classe C2 :
– $L1 = (C1 : P1), (I2 : 4), (I3 : 3), (I4 : 2)$
– $L2 = (C2 : P2), (I2 : 4), (I3 : 3), (I4 : 2), (I3 : 1), (I5 : 1)$
Donc d'après $L1$, la première transaction sera enregistrée comme suit : $C1, I2, I1$

3. Construction du FCP-Tree et des motifs fréquents : Cette étape se déroule de la même façon que pour FP-Growth. Cependant, seules les transactions qui commencent par une modalité de la classe à prédire seront traitées. La phase de prétraitement nous a facilité cette tâche. Ainsi aura-t-on autant de fils de la racine du FCP-Tree que de modalités de la classe. On commence alors par la construction de la racine vide, puis on parcourt la base de données et à chaque fois qu'il y a une modalité de classe différente, on construit un nouveau noeud lié à la racine du FCP-Tree. Pour notre exemple, on construit deux noeud fils, $C1$ et $C2$. La première branche du FCP-Tree est construite par la première transaction $T1$. Par exemple, la première branche correspondant à $T1$ sera lié au noeud $C1$ par deux noeuds $I2 : 1$ et $I1 : 1$. Ainsi, on est sûr que pour chaque item fréquent, on a un motif conditionnel qui contient un item de classe.

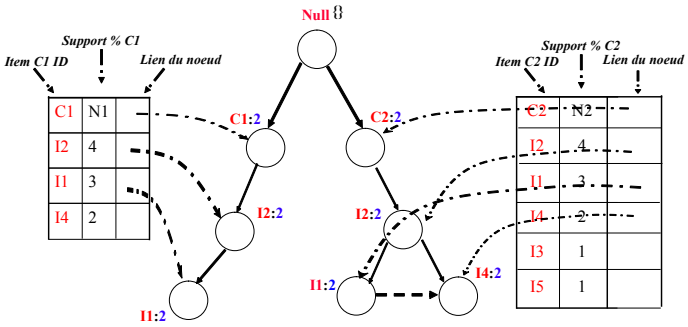


FIG. 2: Construction de FCP-tree

4. Fouille du FCP-Tree et génération des items fréquents : On procède de la même façon que FP-Growth. On extrait les motifs suffixes et on construit pour chacun ses motifs conditionnels. On extrait alors tous les itemsets de classe plus fréquents que le seuil. Les itemsets fréquents sont obtenus par l'association de chaque item avec ses motifs fréquents correspondants. Pour notre exemple, on est sûr d'éliminer l'itemset $I4I1$ trouvé dans la table 3 grâce à FCP-Growth et à la modification de construction du FCP-Tree où chaque item doit être attaché à une branche qui contient un item de classe

5. Elagage : Nous avons introduit dans FCP-Growth la méthode d'élagage précitée de Li (Li, 2006) qui repose sur le fait que toute spécialisation d'une règle qui ne diminue pas le nombre de contre-exemples est moins intéressante que la règle de départ. Cette méthode a été proposée dans le cadre d'Apriori, un algorithme d'extraction d'itemsets fréquents avec génération de candidats. Pour l'introduire dans FCP-Growth qui est un algorithme d'extraction d'itemsets fréquents sans génération de candidats, nous avons dû procéder à quelques adaptations. A chaque transaction et la construction d'un noeud (item $I$), on met à jour et on calcule le nombre de transactions qui valident l'association des items depuis le début de la branche jusqu'au noeud de l'item $supp(Ic)$ et le nombre de transactions qui contredit cette association $Supp(I\bar{c})$.

   Soit l'itemset de classe $Ac$, la règle de classe associée est $A \rightarrow c$. Les contre-exemples de cette règle correspondent à l'itemset $A\bar{c}$. Considérons l'item $x$ auquel on associe la spécialisation $Ax \rightarrow c$. La condition d'élagage est la suivante : pour une mesure $m$ optimale au sens de Li, si $supp(Ax\bar{c}) = Supp(A\bar{c})$, alors $m(Ax \rightarrow c) \leq m(A \rightarrow c)$. Pour intégrer cette contrainte dans un algorithme, il faut être capable de tenir à jour le nombre de contre-exemples de la règle de classe associée à l'itemset de classe courant.

   Dans FCP-Growth, chaque fils de la racine vide correspond à une modalité de classe. Par exemple, si on a les itemsets de classe $ca_1a_2$, puis $ca_1a_2a_3$, la condition d'élagage conduit à comparer les supports des contre-exemples de chacune des deux règles de classe $a_1a_2 \rightarrow c$ et $a_1a_2a_3 \rightarrow c$, c'est-à-dire $p_{\bar{c}a_1a_2}$, et $p_{\bar{c}a_1a_2a_3}$.

FCP-Growth tient à jour les supports $p_{ca_1a_2}$, et $p_{ca_1a_2a_3}$. Pour calculer $p_{\overline{c}a_1a_2} = p_{a_1a_2} - p_{ca_1a_2}$ et $p_{\overline{c}a_1a_2a_3} = p_{a_1a_2a_3} - p_{ca_1a_2a_3}$, il faut connaître $p_{a_1a_2}$ et $p_{a_1a_2a_3}$. De façon générale, à chaque noeud $cA$ de l'arbre il suffit de calculer non seulement $p_{cA}$ mais aussi $p_A$, ce qui permet d'en déduire $p_{\overline{c}A} = p_A - p_{cA}$. Dans l'exemple proposé, si $p_{\overline{c}a_1a_2a_3} = p_{\overline{c}a_1a_2}$, on élague ce noeud, car ni la règle spécialisée $a_1a_2a_3 \rightarrow c$ ni aucune de ses sur-spécialisations ne sont plus intéressantes que la règle $a_1a_2 \rightarrow c$ au sens de la confiance.

# 4 Expérimentation et résultats

FCP-Growth et FCP -Growth-P, les algorithmes que nous proposons pour construire directement les itemsets de classe fréquents et les règles qui en découlent, reposent sur plusieurs principes :

- FCP-Growth est une adaptation de FP-Growth, réputée plus rapide qu'Apriori.
- au cours de la construction du FCP-Tree, FCP-Growth élimine les itemsets non pertinents, ceux qui ne contiennent pas d'item de classe, alors que les méthodes usuelles construisent toutes les règles et éliminent ensuite les règles qui ne sont pas des règles de classe. Cette caractéristique devrait améliorer le temps d'exécution et permettre de diminuer le seuil de support, ce qui est particulièrement important en classification associative, où l'on est très intéressé par les pépites.
- FCP-Growth utilise un seuil de support adaptatif au sens où le seuil de support utilisé dans chaque classe n'est pas le même ; dans le but de ne pas pénaliser les petites classes, ce seuil est choisi proportionnel à la taille de la classe, ce qui revient à fixer comme seuil le même pourcentage de chaque classe.
- FCP-Growth-P résulte de l'adjontion à FCP-Growth d'une procédure d'élagage qui applique la condition de Li sur le support des contre-exemples, ce qui permet d'éliminer les spécialisations des règles qui n'ont pas d'intérêt.

Pour évaluer l'efficacité de FCP-Growth et FCP-Growth-P, nous les avons testés sur 7 bases de données réelles volumineuses (tableau 6) et nous avons calculé (tableaux 7 et 8), avec un seuil de support de 1%, le nombre total d'itemsets (# total Itemsets), le nombre d'itemsets de la classe $Ci$ (# Itemsets Ci) et le nombre d'itemsets non pertinents (# Itemsets non pertinents).

| Bases de données | #Exp | #Attr | #Classes | Poids des classes |
|---|---|---|---|---|
| D100T20I6 | 100000 | 20 | 2 | 58%-42% |
| D100T20I2 | 100000 | 20 | 2 | 55%-45% |
| Retail data | 63000 | 10 | 2 | 69%-31% |
| Adult | 48842 | 14 | 2 | 75,22%-24,78% |
| Connect-4 | 67557 | 42 | 3 | 65,83%-24,62%-9,55% |
| Abalone | 4177 | 8 | 3 | 53%-25%-22% |
| Census | 48842 | 14 | 2 | 75.22%-24.78% |

TAB. 6: Description des bases de données

La comparaison des résultats de FCP-Growth et de sa variante FCP-Growth-P, avec ceux de la version d'origine FP-Growth (tableaux 7 et 8) est menée sur la base de 4 critères : la diminution du nombre total d'itemsets construits, l'augmentation du nombre d'itemsets de classe, la bonne représentation de la classe minoritaire et l'amélioration du taux de couverture (c'est-à-dire la proportion d'exemples couverts par un itemset de classe intéressant).

Les résultats montrent, tout d'abord, l'importance du poids des itemsets non pertinents (les itemsets ne contenant pas la classe). Sur l'ensemble des sept bases traitées, il apparaît qu'entre le tiers et la moitié des itemsets générés par FP-Growth ne sont pas pertinents, ce qui alourdit inutilement la procédure. De plus, la mesure Variation1(obtenu en divisant # Itemsets FCP-Growth par # Itemsets FP-Growth) montre bien que, grâce à sa procédure d'élimination des itemsets qui ne sont pas de classe et à son seuil adaptatif, FCP-Growth ainsi que sa variante FCP-Growth-P permettent d'extraire plus de règles de classes que FP-Growth, tout en évitant que la classe minoritaire soit trop défavorisée. Sur les sept bases de données, ce gain varie entre 17% et 23%.

Pour évaluer les performances de FCP-Growth et FCP-Growth-P, nous avons retenu le taux de couverture qui indique la proportion d'exemples qui sont couverts par au moins un itemset de classe, c'est-à-dire qui vérifient la règle de classe correspondant à cet itemset. Les résultats du tableau 8 montrent très clairement la supériorité de FCP-Growth. En effet, pour chacune des bases, le taux de couverture est nettement augmenté pour arriver environ à plus de 96% de couverture. De plus, pour une meilleure qualité et grâce à sa procédure d'élagage, la variante FCP-Growth-P, héritant déjà des avantages de FCP-Growth en termes d'augmentation d'itemsets de classes, permet de réduire la quantité des règles de classe extraites en améliorant la qualité. En effet, la mesure variation2 du tableau 7 (obtenu en divisant # Itemsets FCP-Growth-P par # Itemsets FCP-Growth)montre que FCP-Growth-P diminue le nombre de règles de classe extraites de 6,67% à 11,37% suivant les bases, tout en assurant le même taux de couverture.

# 5 Conclusion et perspectives

Nous avons proposé FCP-Growth une adaptation de FP-Growth à la recherche des itemsets de classe et des règles d'association prédictives dans une perspective de classification associative, ainsi que FCP-Growth-P, sa variante avec élagage (au sens de Li). Tant FCP-Growth que FCP-Growth-P ne construisent que des itemsets fréquents de classe. En outre, le seuil de support utilisé dans chaque classe est proportionnel à la taille de chaque classe pour éviter de défavoriser les classes minoritaires. Les résultats trouvés montrent que FCP-Growth et plus encore FCP-Growth-P permettent un gain de temps grâce à la génération des seuls itemsets fréquents de classe. Ce gain peut aussi être exprimé en termes de facilité de construction et de stockage. De plus, l'examen des taux de couverture montre que nous obtenons une qualité de règles supérieure à celle des règles générées par FP-Growth. Au vu des résultats obtenus, nous allons utiliser FCP-Growth-P comme algorithme de génération de règles pour la première phase d'une procédure de classification associative performante.

# Références

Agrawal, R. et R. Srikant (1994). Fast algorithms for mining association rules. In J. B. Bocca, M. Jarke, et C. Zaniolo (Eds.), *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pp. 487–499. Morgan Kaufmann.

Bahri, E. et S. Lallich (2009). FCP-Growth, une adaptation de FP-Growth pour générer des règles d'association de classe. *Poster, 9èmes Journées Francophones d'Extraction et Gestion des Connaissances, EGC 2009, Strasbourg*.

Guillet, F. et H. Hamilton (2007). Quality measures in data mining. *Springer-Verlag*.

Han, J., J. Pei, et Y. Yin (2000). Mining frequent patterns without candidate generation. In W. Chen, J. Naughton, et P. A. Bernstein (Eds.), *2000 ACM SIGMOD Intl. Conference on Management of Data*, pp. 1–12. ACM Press.

Li, J. (2006). On optimal rule discovery. *IEEE Transformation on Knowldge and Data Engeneering. 18*(4), 460–471.

Li, J., H. Shen, et R. Topor (2001b). Mining the smallest association rule set for predictions. *IEEE International Conference on Data Mining*, 361–368.

Li, J. et Y. Zhang (2003). Direct interesting rule generation. *IEEE International Conference on Data Mining*, 155–162.

Li, W., J. Han, et J. Pei (2001a). CMAR : Accurate and efficient classification based on multiple class-association rules. In *ICDM*, pp. 369–376.

Liu, B., W. Hsu, et Y. Ma (1998). Integrating classification and association rule mining. In *Knowledge Discovery and Data Mining*, pp. 80–86.

Quinlan, J. R. et R. M. Cameron-Jones (1995). Induction of logic programs : FOIL and related systems. *New Generation Computing 13*, 287–312.

Rakotomalala, R. (2005). Arbres de décision. *Revue MODULAD 33*, 163–187.

Yin, X. et J. Han. CPAR : Classification based on predictive association rules. In *3rd SIAM International Conference on Data Mining (SDM'03)*, San Francisco, CA, pp. 331–335.

# Summary

In this search, we focused on supervised learning task using association rules algorithms (association based classification). These algorithms, developed in unsupervised learning, extract all the rules whose the support and confidence exceed a prefixed threshold support. The extraction of class association rules, using these algorithms, have several problems, because of the rules' a posteriori filtering. In the first stage, one extracts useless frequent itemsets, those which do not contain class, whereas the second stage can be simplified, since an itemset containing the class gives place only to only one class rule. In order to be able to work with a low threshold support, we propose FCP-Growth and an alternative FCP-Growth-P, an adaptation of FP-Growth which have three main advantages. First, it eliminates the frequent itemsets not containing a class. Secondly, to make the minority class be in advantage during the construction of the class itemsets, we adapt the threshold support, in order to use the same threshold support inside each class. Thirdly, or a better pruning, FCP-Growth-P joined the pruning's condition of Li on the rules' specialization.

| **Bases de données** | **Variation1** | **FP-Growth** | **FCP-Growth** | **FCP-Growth-P** | **Variation2** |
|---|---|---|---|---|---|
| *Seuil du support absolu* | *1,00%* | *1,00%* | *1,00%* | *1,00%* | *1,00%* |
| ***D100T20I6*** | | | | | |
| # Itemsets | -24,77% | 1090 | 820 | 730 | -10,97% |
| # Itemsets C1 | 12,22% | 452 | 508 | 438 | -13,72% |
| # Itemsets C2 | 28,46% | 243 | 312 | 292 | -6,48% |
| # Itemsets non pertinents | - | 395 | 0 | 0 | - |
| # Total itemsets de classe | 17,98% | 695 | 820 | 730 | ,-10,97% |
| ***D100T20I2*** | | | | | |
| # Itemsets | -24,62% | 1415 | 1075 | 967 | -10,04% |
| # Itemsets C1 | 23,78% | 565 | 699 | 610 | -12,73% |
| # Itemsets C2 | 23,06% | 306 | 376 | 357 | -6,48% |
| # Itemsets non pertinents | - | 545 | 0 | 0 | - |
| # Total itemsets de classe | 23,56% | 870 | 1075 | 967 | -10,04% |
| ***Retail data*** | | | | | |
| # Itemsets | -25,77% | 330 | 245 | 224 | -8,57% |
| # Itemsets C1 | 5,13% | 121 | 127 | 117 | -8,51% |
| # Itemsets C2 | 92,16% | 40 | 118 | 107 | -8,59% |
| # Itemsets non pertinents | - | 169 | 0 | 0 | - |
| # Total itemsets de classe | 52,17% | 161 | 245 | 224 | -8,57% |
| ***Abalone*** | | | | | |
| # Itemsets | -14,19% | 665 | 562 | 518 | -7,89% |
| # Itemsets C1 | 7,25% | 209 | 224 | 206 | -12,43% |
| # Itemsets C2 | 36,50% | 144 | 196 | 180 | -5,19% |
| # Itemsets C3 | 12,89% | 126 | 142 | 132 | -4,14% |
| # No relevant itemsets | - | 176 | 0 | 0 | - |
| # Tolal class itemsets | 17,57% | 478 | 562 | 518 | -7,89% |
| ***Adult*** | | | | | |
| # Itemsets | -14,60% | 890 | 760 | 675 | -11,18% |
| # Itemsets C1 | 2,8% | 392 | 403 | 398 | -14,53% |
| # Itemsets C2 | 43,33% | 249 | 357 | 277 | -7,82% |
| # No relevant itemsets | - | 249 | 0 | 0 | - |
| # Total class itemsets | 18,56% | 641 | 760 | 675 | -11,18% |
| ***Connect-4*** | | | | | |
| # Itemsets | -11,04% | 6543 | 5820 | 5184 | -11,37% |
| # Itemsets C1 | 11,12% | 2094 | 2328 | 2160 | -15,80% |
| # Itemsets C2 | 14,15% | 1439 | 2037 | 1657 | -8,84% |
| # Itemsets C3 | 17,03% | 1243 | 1455 | 1341 | -7,82% |
| # No relevant itemsets | - | 1767 | 0 | 0 | - |
| # Total class itemsets | 21,85% | 4776 | 5820 | 5184 | -11,37% |
| ***Cencus*** | | | | | |
| # Itemsets | -15,51% | 780 | 659 | 615 | -6,67% |
| # Itemsets C1 | 10,63% | 328 | 362 | 335 | -11,67% |
| # Itemsets C2 | 52,07% | 195 | 297 | 280 | -4,55% |
| # No relevant itemsets | - | 257 | 0 | 0 | - |
| # Total class itemsets | 26% | 523 | 659 | 615 | -6,67% |

TAB. 7: Tableau récapitulatif des résultats des 7 bases de données

| **Bases de données** | **FP-Growth** | **FCP-Growth et FCP-Growth-P** |
|---|---|---|
| D100T20I6 | 63% | 91% |
| D100T20I2 | 75% | 93% |
| Retail Data | 67% | 89% |
| Adult | 69% | 92% |
| Connect-4 | 73% | 90% |
| Abalone | 79% | 96% |
| cencus | 76% | 92% |

TAB. 8: Taux de couverture

# Utilisation des règles d'association pour l'amélioration de la qualité des bases de données

Soumaya Ben Hassine*, Delphine Clément**

*Laboratoire ERIC, Université Lyon2, 5 avenue Pierre Mendès France, 69676 Bron Cedex France
Soumaya_ben_hassine@voila.fr
**Hewlett-Packard, 1 avenue du Canada, ZA de Courtaboeuf, 91947 Les Ulis
Delphine_clement@hp.com

**Résumé.** Face à la croissance accrue des volumes de données dans les bases informatiques, l'évaluation de leur qualité est devenue une tâche d'autant plus ardue que les experts ont de plus en plus recours aux techniques automatiques de fouille pour la qualité des données (Data Quality Mining). Parmi ces techniques nous citons les règles d'association que nous utilisons dans le cadre de ce travail. Notre objectif étant double : documenter des règles métier, et détecter et/ou corriger des erreurs dans les données (enregistrements non conformes à des règles produites). Nous proposons, dans cet article, de dresser un état de l'art sur l'utilisation des règles d'association comme outil d'amélioration de la qualité des données ; de décrire notre approche basée sur l'utilisation du lift comme critère principal de génération des règles ; et, de présenter la mise en œuvre sur des données de Hewlett Packard, les différents résultats ainsi que les perspectives à envisager quant à la mise en production de ce type de méthodes.

## 1 Introduction

Depuis la fin des années 90, le problème de la qualité des données n'a cessé de se poser avec de plus en plus d'acuité face à la croissance accrue des volumes à traiter et des capacités de stockage. En effet, la communauté des chercheurs, soit-elle académique ou professionnelle, propose de plus en plus de paradigmes, modèles et méthodologies afin d'évaluer voire même d'améliorer la qualité des données. Ce problème intéresse, en particulier, les spécialistes du décisionnel qui trouvent dans la détection et la prévention des problèmes de la qualité des données un nouveau champ d'application des techniques de fouille ; et depuis 2001, on parle désormais de techniques de fouille pour la qualité des données ou encore Data Quality Mining (Hipp et al. 2001).
Dans ce contexte, nous proposons de montrer l'intérêt d'utiliser les règles d'association comme outil automatique d'audit des données réelles moyennant l'analyse de certains fichiers générés par le système de facturation et de gestion de commandes (Order Management Customer Master System) utilisé par l'entreprise Hewlett-Packard.
Notre objectif est double :

- le premier est curatif ; relatif à la détection et correction des données erronées
- le second est préventif ; relatif à l'analyse de l'origine des problèmes qualité sous jacents.

Mais avant de détailler le principe de notre approche, nous préludons par un bref état de l'art détaillant l'utilisation des règles d'association dans le domaine de la qualité des données en rappelant brièvement les bases de cette technique. Nous illustrons, par la suite, les différents résultats et nous concluons par rapport à l'efficacité de notre approche ainsi que les différentes perspectives envisagées.

## 2 Etat de l'art

Depuis leur formalisation par Agrawal (Agrawal et al. 1993) (Agrawal et al. 1994) dans le cadre de l'analyse du panier de la ménagère dans les données des grandes distributions, les règles d'association n'ont cessé d'intéresser les spécialistes décisionnels vu la simplicité de leur algorithme générateur (l'algorithme **Apriori**) et la facilité de lecture et d'interprétation de leur modèle de décision (des règles du type SI achat=couches ALORS achat=bière).
Cette technique de fouille consiste à générer des règles de type A → B par exécution de l'algorithme Apriori tout en utilisant, comme critère de validation, le couple (support[1] ,confiance[2]) respectivement pour le filtrage des nombres d'itemsets fréquents et la réduction de la taille de la base de règles générée.

[1]Etant donné la règle A → B ; le Support= $n(AB) \div n$ mesure la fréquence de la règle
[2]Etant donné la règle A → B ; la confiance=$P(AB) \div P(A)$ mesure le taux d'efficacité (fiabilité) de la règle

Grâce à cette simplicité de manipulation, les règles d'association ont vu leur champ d'application s'élargir à la détection des anomalies dans les bases informatiques et la résolution des problèmes de qualité des données surtout dans les systèmes de gestion de la relation client (CRM), de gestion intégrée (ERP) ou encore de business intelligence (BI) où les mauvaises connaissances induites des données de faible qualité, conséquence justifiée par le fameux dicton anglais "garbage in, garbage out", influent directement sur les décisions stratégiques de l'entreprise.

Plusieurs travaux ont été effectués dans ce contexte de résolution des problèmes qualité. Nous en citons quelques exemples en nous concentrant sur la méthodologie d'utilisation des règles d'association.

Tout d'abord, en 1999, Sapia propose d'utiliser les techniques de fouille pour améliorer les performances des techniques de nettoyage des données (Data Cleaning) dans le contexte de la conception et de l'implémentation des entrepôts de données ; le processus étant complexe (traite un grand nombre de bases de données opérationnelles) et gourmand en terme de temps d'exécution. En effet, dans *"On supporting the Data Warehouse Design by Data Mining Techniques"* (Sapia et al. 1999), Sapia liste pour chacune des activités de conception d'un datamart les techniques de Data Mining candidates, en expliquant leurs apports quant à l'amélioration de la tâche en question. Le tableau suivant résume ses différents propos.

| **Activité de construction d'un entrepôt de données** | **Activité de Data Mining associée** |
|---|---|
| Analyse de la source des données | **les règles d'association**, la régression multivariée, les Arbres de décisions, les arbres d'induction, les Réseaux bayésiens |
| Intégration structurelle des sources de données | **les règles d'association**, la régression linéaire |
| Le nettoyage des données | **les règles d'association** (utiliser un seuil élevé de confiance) |
| La conception multidimensionnelle des données | Les outils OLAP combinés avec les techniques d'analyse de corrélations |
| La conception physique de l'entrepôt | **les règles d'association** (pour identifier les séquences type de requêtes) |

TAB. 1 – *Les techniques de fouille pour améliorer les performances de conception d'un entrepôt.*

Ensuite, en 2001, Jochen Hipp (Hipp et al. 2001) propose d'utiliser les règles d'association pour l'évaluation de la qualité globale des données. Les règles étant, tout d'abord, filtrées par un seuil de confiance de 75%, puis, utilisées pour calculer une fonction score sur les transactions dans le but d'établir un ordre entre les transactions. Cette fonction score est définie par la formule suivante :

$$Score = \sum_{(} r \in R) confiance(r) up\tau \times violates(T, r),$$

où Violates(T,r)=1 si la règle est violée par la transaction T, 0 sinon

Le facteur $\tau$ appartient à l'ensemble $\{0.5; 0.7\}$ et est choisi de sorte à calibrer la fonction score selon l'exactitude de certaines règles par rapport à d'autres. Ainsi, une règle de confiance égale à 90% l'emporterait sur 3 règles de 75% et vice versa.
Malgré sa simplicité, cette proposition présente, tout de même, quelques inconvénients. En effet, l'approche utilisée pour l'appréciation de la qualité des données d'une base donnée suppose l'application de la fonction score à toutes les transactions de la base, ceci s'avère fastidieux quand il s'agit d'une base de plus d'1 million de transactions. Par ailleurs, la fonction score a uniquement été appliquée dans un cadre théorique sur des données expérimentales ; l'efficacité de la méthode sur des données réelles n'est donc pas garantie.

Dans ce même contexte d'évaluation de la qualité des données, Laure Berti-Equille (Berti-Equille 2007) se base sur la sémantique des règles d'association pour la construction de modèles permettant la détection des anomalies et des contradictions dans les données.
Elle définit, pour ce faire, 4 types de patterns d'anomalies :

1. Pattern des règles à conclusions nulles où il s'agit de s'intéresser aux règles A→B où A est une conjonction d'items et B une conclusion à valeurs nulles.
2. Pattern de détermination des doublons où on s'intéresse aux règles de la forme Similaire(X)→[Duplicat=Vrai]. Similaire(X) est l'ensemble des valeurs de X qui sont plus ou moins similaires et Duplicat est une fonction booléenne identifiant les enregistrements redondants en se basant sur l'appartenance de X à un certain ensemble de valeurs.
3. Pattern de détection de contradictions où les règles visées sont du type Assortiment(X)→ Non_Similaire(Y). Assortiment(X) est un ensemble d'attributs qui font référence au même objet et qui ont, de ce fait, des valeurs similaires ; ainsi pour l'ensemble de ces attributs on devrait retrouver les mêmes valeurs de Y.
4. Pattern de détection des anomalies par identification des valeurs anormales ne figurant pas dans l'ensemble des règles extraites.

Cependant, cette proposition, bien que très efficace dans la détection des anomalies, nécessite un temps de post-traitement considérable pour la construction des différents patterns à partir des règles extraites.

En 2007, Weijie Wei propose une autre approche permettant d'apprécier la qualité des données moyennant les règles d'association (Wei et al. 2007). Elle se base sur l'algorithme DCAR (Data Cleaning method based on Association Rules) visant à enrichir les règles métier par les règles d'association générées. Ces règles métier serviront, par la suite, au nettoyage de la base où l'on procède transaction par transaction en testant sa cohérence avec l'ensemble des règles produites. Ainsi, si la transaction satisfait toutes les règles, la transaction est dite correcte, sinon elle est considérée erronée même si elle ne contredit qu'une seule règle métier. Wei montre, par ailleurs, l'efficacité de cette technique surtout quand il s'agit de traiter un grand volume de données.

Nous terminons cette étude bibliographique par une approche récente proposée par Craig McClanahan en 2008 (McClanahan 2008) de BASF corporation qui propose d'utiliser les règles d'association pour l'analyse des corrélations négatives, du type B est rarement acheté quand A est acheté, données par la règle A→7B ; dans le contexte de l'analyse de la qualité d'une base de formules pour le mélange des couleurs ; et montre, de part cette méthode, la puissance des règles d'association à détecter les données isolées et en particulier les données anciennes non encore archivées.

Pour conclure, toutes les méthodes présentées ci-dessus montrent l'efficacité des règles d'association pour l'amélioration de la qualité des données dans les bases bien que leur intérêt n'est pas toujours démontré quand il s'agit de traiter de grands volumes de données. Nous proposons de détailler, dans ce qui suit, le principe notre méthodologie en présentant les outils utilisés pour la mise en place de la solution et en commentant les différents résultats obtenus.

# 3 Principe de l'approche

Comme mentionné dans le premier paragraphe introductif, notre travail porte sur l'analyse d'un fichier de $1\,077\,428$ enregistrements et 42 attributs hétérogènes (nominaux et continus), décrivant les données du système de facturation et de gestion de commandes utilisé par Hewlett-Packard, à savoir :

- le nom de l'entreprise cliente : Company_name
- son adresse : Adr
- le numéro de téléphone du contact : Tel_nbr
- la région : Region
- le nombre de jours restants avant le paiement de la commande par le client : Terms_payment_key
- le réseau de vente auquel appartient le contact : Sales_Org
- la date de création de l'enregistrement : Date_creat
- ...

La méthodologie utilisée est résumée par la figure ci-après Les différentes tâches sont réalisées par le logiciel de fouille de données WEKA. Le choix de cet outil a été conclu suite à une étude comparative réalisée sur certains outils libres, tels que Tanagra et R-Gui Project, et se justifie principalement par sa capacité à traiter les fichiers volumineux ainsi que la grande panoplie des méthodes proposées autant pour le pré-traitement des données (outils d'analyse des données, de visualisation et de filtrage des données et des connaissances) que pour le filtrage des règles produites et la visualisation

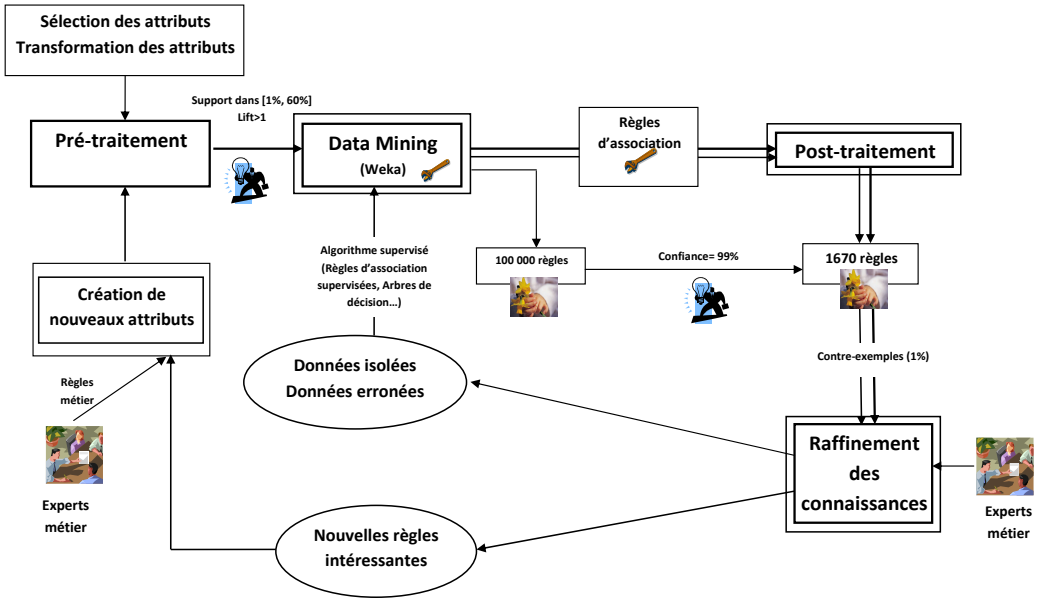


FIG. 1 – *Approche proposée d'utilisation des règles d'association pour améliorer la qualité des données.*

des connaissances produites outre leur insensibilité au bruit. Par ailleurs, les fonctionnalités offertes sont compréhensibles et faciles d'utilisation grâce à l'interface graphique que propose WEKA, sans oublier son grand avantage de portabilité vu qu'il est entièrement implémenté en Java.

## 3.1 Préparation des données

Une première exploration du fichier nous permet de repérer un certain nombre d'enregistrements aberrants (où l'on trouve par exemple une date dans le champ *Company_name*). Ces quelques lignes ne sont pas concernées par notre analyse. Par ailleurs, notre fichier comprend des données manquantes, comme c'est le cas dans la plupart des bases réelles. Cette problématique de traitement des données manquantes est pour le moment écartée de notre étude. Ainsi, une valeur non renseignée est remplacée par un " ? " et va être gérée par WEKA comme une modalité du champ concerné.
Cette première étape s'articule, alors, autour de 3 axes principaux :

– **la discrétisation des données numériques en données catégorielles :** Plusieurs méthodes de discrétisation sont définies dans la littérature utilisant différentes fonctions supervisées (basées sur les notions de gain d'entropies) et non supervisées (telles que les méthodes de préservation des corrélations (Mehta et al. 2005) et les méthodes classiques de discrétisation à fréquences ou amplitudes égales). WEKA ne proposant que le dernier type de discrétisation non supervisées, nous privilégions la discrétisation à fréquences égales à la discrétisation à amplitudes égales pour avoir une distribution plus ou moins homogène des exemples sur chacun des intervalles créés.
Cependant, le résultat de discrétisation semble parfois assez superflu pour certaines variables telles que le champ *date_creat*. Nous procédons, pour ces attributs, à une discrétisation basée sur les connaissances métier. Ainsi, au lieu de travailler sur les toutes les dates de création possibles (*date_creat*) s'étendant sur la dernière décennie, on

définira des tranches de segmentation en nous basant sur des connaissances métiers. Par exemple, si, en 2007, Hewlett-Packard a fait appel à un nouveau système de facturation et de gestion de commandes, une segmentation ciblant ce genre d'informations serait de la sorte : *Record_Avant_2007*, *Janv_2007*, *Fev_2007*

- **la transformation des données normales en données plus significatives prêtes à l'analyse :** Concrètement, il s'agit de transformer certains attributs non significatifs de façon à améliorer leurs expressivités tel que le fait d'utiliser le champ *long_tel_nbr* au lieu de *tel_nbr*. En effet, un numéro de téléphone tel que *long_tel_nbr*<8 est un numéro de téléphone erroné. Ainsi, les attributs décrivant le nom de la société, adresse et numéro de téléphones des entreprises clientes sont remplacés par leurs longueurs.
- **la sélection et la construction des attributs intéressants ainsi que l'élimination des données bruitées :** Dans cette étape, il s'agit de sélectionner les attributs sur lesquels nous allons nous baser pour effectuer nos analyses.

De cette manière, nous sommes parvenus à réduire nos données en sélectionnant 25 attributs intéressants au lieu des 42 initiaux.
Par ailleurs, une agrégation des données par type de commande/type d'adresse, nous a permis de réduire la taille de notre fichier de départ de 24% pour ne traiter au final que quelques 800 000 transactions.

## 3.2 Extraction des règles d'association

Une fois les données préparées et mises sous le format arff, nous appliquons l'algorithme Apriori en utilisant la paire de mesures (support,lift) comme critère d'extraction et de validation des règles, au lieu de l'utilisation classique du couple (support,confiance) ; le support traduisant la fréquence de la règle, la confiance son exactitude et le lift son intérêt.
Nous fixons un support compris entre 1 et 60%. Le seuil de 1% est choisi de sorte à permettre la double découverte des relations improbables et données isolées pouvant être intéressantes. Ensuite, nous utilisons le lift comme premier critère de génération des règles.
Il s'agit ainsi d'une mesure faisant intervenir à la fois la prémisse et la conclusion de la règle, mais qui est donc, toutefois, symétrique (ne tient pas contre du sens de la règle). Le lift peut prendre des valeurs comprises entre zéro et l'infini. Son interprétation dépend de la valeur qu'il prend. Une valeur inférieure à 1 indique que la règle est plutôt rare, tandis qu'une valeur supérieure à 1 correspond à une règle plutôt fréquente. Si $A \rightarrow B$ est notre règle telle que A et B sont complètement indépendants alors le lift vaut 1. En d'autres termes, la connaissance de A n'apporte pas d'information sur celui de B. Le lift est donné par la formule suivante : $Lift = (P(AB)) \div (P(A) * P(B))$; étant donné la règle$A \rightarrow B$ Le choix du lift est principalement justifié par le caractère sélectif d'une telle mesure, notamment quand on utilise un seuil minimum de 2 ou plus.
A l'issue de cette étape, 100 000 règles ont été générées. Face à ce volume envahissant, un deuxième filtrage s'avère nécessaire.

## 3.3 Filtrage des connaissances

Différentes techniques de filtrage ont été proposées par la littérature variant des techniques d'élagage aux techniques de généralisation (regroupement) des règles sans oublier la visualisation et les mesures d'intérêt.
Nous nous concentrons sur les deux dernières techniques que nous tentons d'utiliser pour réduire la taille de notre base des règles :

- **Utilisation des mesures d'intérêt :** Les mesures d'intérêt ont pour principal avantage la réduction immédiate de la base des règles ; contrairement aux autres techniques qui nécessitent une étude approfondie de l'ensemble des règles.
  Plusieurs indices et critères de filtrage sont définis. Une première classification permet de distinguer les mesures objectives des mesures subjectives. Les mesures objectives s'intéressent à l'aspect des données (leur distribution, la structure des règles, ...) et peuvent être soit symétriques (indépendantes du sens de la règle), soit asymétriques. Parmi ces mesures, nous citons la confiance, la conviction[3] (asymétriques) et le lift (symétrique)...
  Quant aux mesures subjectives, elles sont plutôt orientées utilisateur, et dépendent de ses objectifs ainsi que du contexte de l'analyse.
  Nous envisageons, dans un premier temps, un filtrage moyennant une mesure asymétrique permettant d'apprécier le sens adéquat de la règle.
- **Utilisation des techniques de visualisation** où l'on définit un certain nombre d'indicateurs sur la base desquels on va effectuer nos visualisations et décider de la pertinence des règles.
  Nous utilisons l'outil libre ARViewer[4] en visualisant les différents critères de confiance, lift et conviction. Cepen-

[3] La conviction est une mesure similaire au lift, elle mesure, cependant, l'effet des contre exemples sur l'efficacité de la règle
[4] http ://www2.lifl.fr/ jourdan/download/arv.html

dant, les résultats ne sont pas intéressants vu qu'au-delà de 50 règles, la qualité de l'image devient médiocre au point qu'il est impossible d'identifier les différentes règles dans les représentations.

Ainsi, nous privilégions les mesures d'intérêt et utilisons, comme premier abord, le critère de confiance étant un critère asymétrique qui complète le lift. Nous considérons, ainsi, les règles de confiance=99% (les règles les plus correctes qui admettent, à la fois, quelques contre exemples) ; et nous nous proposons de les valider par les experts métier en vue de les exploiter, par la suite, pour le nettoyage de la base. Grâce à ce filtrage, nous parvenons à réduire la taille de notre base de règles de 98,33% pour n'analyser que 1 670 règles.

Une étape supplémentaire de filtrage, basée soit sur l'utilisation des mesures d'intérêt, soit sur les techniques d'élagage, est planifiée en perspective à ce travail en fonction des conclusions des experts métier, mais avant cela nous proposons d'illustrer les principaux résultats de validation.

# 4 Résultats

L'ensemble des règles obtenues est analysé et validé par les experts de Hewlett-Packard. Cependant, le nombre des règles et contre exemples à analyser étant toujours assez important, nous commençons par analyser les règles comportant un seul item en conclusion tout en privilégiant celles dont la conclusion est relative à un item intéressant pour l'expert du métier ; telle que l'échéance de paiement (*terms_payment_key*) ou encore le réseau de vente (*dit_channel*, *sales_org*). Nous réduisons, par ce moyen, notre échantillon à moitié pour n'analyser qu'un ensemble de 800 règles.

Les résultats liés à l'utilisation des règles d'association sur le système de gestion de commandes sont à classer en trois catégories :

1. **Les règles d'association confirmant des règles métier connues :** Cette phase est utile car les règles métier, même si elles sont connues, ne sont pas toujours documentées de manière exhaustive. En outre, les règles métier peuvent être codées pour un résultat plus efficace lors d'une phase successive de fouilles de données.
   *Exemple : sales_org=US00 → dist_channel=S1*
   **Interprétation métier :** Le canal de distribution (*dist_channel*) est un champ obligatoire et est utilisé par les applications en aval du système de référence.
2. **Les règles d'association permettant la détection de données obsolètes non-conformes, ce qui déclenche une procédure d'archivage :** Cette phase est intéressante vu que les utilisateurs des données du système de référence n'ont pas la visibilité sur l'existence de ces données obsolètes. Le maintien de ces données dans un statut " actif " a un coût pour Hewlett-Packard.
   *Exemple : filled_name2=O dist_channel=ZZ ==>sales_org=Z9US*
   **Interprétation métier :** Le code du secteur de vente pour la région concernée par notre fichier de données est Z900 ou US00 ; tout enregistrement affecté d'un autre code est non-conforme. Cette règle met donc en lumière des données erronées ; après analyse, il en résulte que les enregistrements sont tous anciens et proviennent d'une migration d'un système legacy. Ces enregistrements seront archivés.
3. **Les contre-exemples des règles d'association permettant de mettre à jour un problème dont l'impact métier est coûteux :** Les utilisateurs des données pensent que les dispositifs nécessaires sont en place pour garantir la qualité de l'information sur les conditions de paiement. Jusqu'ici, aucun système d'alerte pour données non-conformes n'est mis en place. Les données non conformes ont un coût élevé pour la compagnie.
   *Exemple : sales_org=US00 dist_channel=S1 person_creat_obj=RSIBATCH → terms_payment_key2=NT30*
   **Interprétation métier :** Les conditions de paiement à 30 jours (NT30) sont une valeur par défaut dans le système de référence. La règle d'association, et surtout ses contre-exemples, a permis de détecter que les utilisateurs pouvaient manuellement modifier la valeur par défaut et augmenter les conditions de paiement à 45 jours ou purement et simplement effacer la valeur. Cette découverte a engendré une recherche des causes et aboutira à une correction et à une prévention du problème.

# 5 Discussion et Perspectives

Ce travail est un cas pratique d'application des règles d'association pour la détection automatique des problèmes qualité dans les grandes bases de données opérationnelles.

Notre approche se caractérise par une utilisation non classique de l'algorithme Apriori qui a permis d'identifier différents types d'anomalies (données non encore archivées, données cruciales erronées). En effet, les règles automatiquement découvertes ont intéressé les experts d'autant plus qu'elles ont confirmé certaines règles métiers. Cependant, ces interprétations ne peuvent être prises en compte que si le fichier de départ possède un niveau de

qualité minimal ; dans ce sens, plusieurs travaux incitent à la conception de modèles probabilistes pour évaluer la qualité des règles produites (Laure Berti-Equille 07) avant de les exploiter de point de vue métier.
Pour finir, malgré la difficulté d'utilisation des règles d'association dans un contexte réel autant sur le plan technique que sur le plan opérationnel, Hewlett-Packard considère l'expérience de l'utilisation des règles d'association sur ses données de gestion de commandes comme positive. Dès lors, Hewlett-Packard souhaite réaliser d'autres tests d'utilisation de ces techniques pour détecter des problèmes qualité, notamment dans le cadre des migrations de données de type fusion/acquisition et également dans le cadre de l'analyse du système d'information sur les clients particuliers, celui-ci contenant plusieurs millions d'enregistrements.
De point de vue technique, nous visons à améliorer notre processus d'extraction des règles :

– en réinjectant les règles pertinentes extraites afin de construire d'autres attributs intéressants et de déduire, par la suite, des règles plus pertinentes,
– et en améliorant notre procédure de filtrage en combinant à la fois l'utilisation des mesures d'intérêt et des techniques d'élagages tout en automatisant ces dernières, en fonction des besoins des experts.

## Références

(Agrawal et al. 1993) Agrawal R., Imielinski T., Swami A. (1993) "Mining Association Rules Between Sets of Items in Large Databases" In Proceedings of the ACM SIGMOD International Conference on Management of Data, May 1993.
(Agrawal et al. 1994) Agrawal R., Srikant R. (1994) "Fast algorithms for mining association rules" In Proceedings of the 20th VLDB Conference Santiago, Chile, 1994.
(Berti-Equille 2007) Berti-Equille L. (2007) "Quality awareness for managing and mining data", 2007.
(Giudici 2003) Giudici P. (2003) "Applied Data Mining : Statistical Methods for Business and Industry",2003.
(Hipp et al. 2001) Hipp J., Güntzer U., Grimmer U. (2001) "Data quality mining : Making a Virtue of Necessity",2001.
(Melha et al. 2005) Mehta S., Parthasarathy S. , and Yang H. (2005) "Toward Unsupervised Correlation Preserving Discretization", 2005.
(McClanahan et al. 2008) C.J. (2008) "Cleaning a formulation database using rule discovery techniques" In Proceedings of the 13th ICIQ Conference, Boston, US, 2008.
(Sapia et al. 1999) Sapia C., Höfling G., Müller M., Hausdorf C., Stoyan H., Grimmer U. (1999) "On supporting the data warehouse design by data mining techniques" In proceedings of the GI-Workshop data mining and data warehousing, 1999.
(Wei et al. 2007) Wei W., Zhng M., Zhang B., Tang X. (2007) "A Data Cleaning Method Based on Association Rules", 2007.

## Summary

Since the last decades and due to the high capacity of storage, data is being increasingly available in information society, which has led to the need for valid tools for its modelling and analysis such as Knowledge Discovery in Databases (KDD) methods (known as data mining methods). However, many problems go along with the business application of data mining. In fact, the quality of the generated models is strongly correlated to the data quality. This paper deals with the data mining contribution in the data quality domain, commonly known as DQM; where we used the association rules as a method to improve poor data quality in business databases. We explain our approach and discuss about the results and the possible perspectives.