Département de formation doctorale en informatique **École doctorale IAEM Lorraine**
UFR STMIA

Étude et réalisation d'un système d'extraction de connaissances à partir de textes

THÈSE

présentée et soutenue publiquement le 15 novembre 2004

pour l'obtention du

Doctorat de l'université Henri Poincaré – Nancy 1

(spécialité informatique)

par

HACÈNE CHERFI

Composition du jury

Président :	Marie-Christine Haton, Professeur à l'Université Henri Poincaré – Nancy 1
Rapporteurs :	Henri Briand, Professeur à l'École polytechnique de l'université de Nantes Pascale Sébillot, Maître de Conférences à l'Université de Rennes 1 – HDR
Examinateur :	Daniel Kayser, Professeur à l'Université de Paris Nord – XIII
Directeurs :	Amedeo Napoli, Directeur de recherche CNRS Yannick Toussaint, Chargé de recherche INRIA

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Mis en page avec la classe thloria.

Remerciements

Je tiens à adresser mes remerciements les plus chaleureux aux membres du jury qui ont bien voulu s'intéresser à mon travail.

À Yannick Toussaint pour le temps et l'énergie qu'il a consacrés à ce travail de thèse, mais également ses encouragements, son accompagnement, son soutien et ses conseils scientifiques et personnels durant ce travail.

À Amedeo Napoli, celui qui m'a fait confiance, qui m'a encadré et m'a aidé à avoir le recul et la vision claire nécessaires pour mener, au mieux, ce travail. Merci également d'avoir accepté que je fasse partie, le temps d'une thèse, de l'équipe ORPAILLEUR.

Aux rapporteurs et aux examinateurs : Henri Briand, Pascale Sébillot, Marie-Christine Haton et Daniel Kayser pour leurs nombreux commentaires très pertinents qui ont permis d'améliorer ce mémoire.

À tous les membres de l'équipe ORPAILLEUR un très grand merci ainsi qu'à Christelle, la très efficace assistante de l'équipe.

À l'équipe URI de l'INIST pour les textes de mes expérimentations mais également pour l'interprétation des résultats ; parmi eux : Xavier Polanco, Claire François, Jean Royauté et Alain Zasadzinski.

À ceux qui sont, naturellement et très vite, passés de collègues à ami(e)s. En premiers : Jérôme (babtou), Armelle et Hélène (deux muses) ; puis en vrac (que dis-je sans un ordre particulier) Jean (son chapeau, son fromage), Mathieu (à indexer en A.), Clara (farfarella), Laïka (petite luciole), Sylvain, BenJ, Joseph, Huy, Makram, Hend, Karima, Rim, Sandy, Sandra(s), Sandrine (son café), Irina, Suzanne, Fréd, Benoît, Yann, Évelyne, Alain, Laurent, Bernard(s), Nico(s) et une foultitude d'autres personnes du LORIA et d'ailleurs : Henri, Miriam, Kiki, Hania, Nadia, Yasmine, Lydia, Chida, Vincent(s), William, Jo.

À ma famille, mon énorme soutien indéfectible et sans qui ce travail n'aurait pu être mené. Merci pour leur patience.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

Cette thèse présente la fouille de textes (FdT) comme un processus d'extraction de connaissances dans des bases de données (ECBD) qui opère sur des *données textuelles.* La problématique générale en FdT est de tirer profit d'éléments d'information extraits afin d'exprimer des *connaissances* utilisables pour le domaine traité par les textes. Les nouvelles connaissances extraites servent à enrichir les connaissances actuelles d'un domaine contenues, par exemple, dans une base de connaissances. Ensuite, l'extraction de nouvelles connaissances permet de raisonner sur les connaissances actuelles pour modifier (réviser, spécifier, etc.) ou bien justifier les connaissances actuelles. La pertinence des nouvelles connaissances extraites par le processus de FdT est jugée par un *analyste* — un expert du domaine de fouille.

La FdT doit répondre à quatre besoins : (1) de taille et de structure des données textuelles à fouiller (un texte, plusieurs milliers de textes), (2) d'indépendance par rapport à la nature des données textuelles et des connaissances à extraire, (3) d'indépendance par rapport à l'ordre de traitement des textes et (4) d'indépendance par rapport au domaine de fouille, c'est-à-dire à un besoin de reproductibilité.

(1) Premièrement, une méthodologie de FdT doit permettre une caractérisation globale du contenu d'un ensemble de textes. Ce besoin se retrouve dans des applications de *constitution et filtrage d'une terminologie* ou une application de *recherche d'information.* Deuxièmement, la FdT permet de trouver des liens entre les textes, comme les régularités des contenus, difficiles à repérer par une lecture séquentielle de l'ensemble des textes. Le besoin de fouille sur plusieurs textes se retrouve dans une application d'*extraction d'information* qui porte sur plusieurs textes.

(2) Une méthodologie de FdT doit fonctionner indifféremment sur tout type de données non structurées, c'est-à-dire sur des textes, des listes de formules chimiques, de séquences ADN de protéines, etc. Le processus doit être robuste pour fonctionner sur ces différentes données textuelles (pas de gestion d'erreurs s'il y en a et pas de vérification de types de données par exemple), et ce, quelque soit le *type* de connaissance décrit par les textes (compte rendu, démonstration, etc.).

(3) Une méthodologie de FdT doit donner les mêmes résultats en partant d'un même ensemble de textes, indépendamment de l'ordre de la prise en compte des textes, c'est-à-dire que les textes sont donc analysés dans leur globalité.

(4) Une méthodologie de FdT ne doit pas être *ad hoc* à un domaine particulier, elle doit être reproductible pour un autre domaine. Seules les données textuelles et les connaissances du domaine changent. La méthodologie doit donc rester stable et générique.

Une méthodologie de FdT soulève un certain nombre de problèmes, étant données les contraintes liées aux besoins cités ci-dessus. Les difficultés liées à une méthodologie de FdT concernent (a) le choix d'une représentation des textes en vue de leur traitement pour en extraire des connaissances, (b) le choix de la technique de fouille de données à appliquer, (c) le choix de la méthode d'évaluation de la qualité des connaissances extraites.

Un premier problème est la pertinence et l'intérêt ou non de fouiller dans des textes hétérogènes. La finalité d'un processus de FdT est d'extraire des connaissances utiles et interprétables étant donnés un domaine et un objectif de fouille. Est-ce pertinent de fouiller un ensemble de textes collectés de façon quelconque ? Y a-t-il des régularités de contenu à trouver entre un texte de biologie moléculaire et une documentation aéronautique ? C'est peu probable sauf si on s'intéresse à identifier des éléments d'information de langage général indépendants du domaine. Il est donc important de disposer d'un ensemble homogène de textes afin de rendre possible et efficace la tâche d'extraction de connaissances potentiellement pertinentes.

Les nouvelles connaissances extraites doivent être réutilisables si besoin pour une autre tâche du domaine. Un modèle du domaine (par exemple une ontologie) ou, à défaut, des connaissances d'un analyste doivent être disponibles et pouvoir être enrichies de ces nouvelles connaissances. Ce problème se retrouve, par exemple, dans des applications de *structuration d'une ontologie* d'un domaine.

Un autre problème que nous soulevons est celui d'*identifier* les éléments d'information pertinents contenus dans un texte. L'identification de connaissances dans les textes pose des problèmes analogues à ceux d'applications en *recherche d'information,* en *extraction d'information* et en *structuration de terminologie*. D'autre part, comment peut-on caractériser puis classer un texte par rapport à d'autres textes ? Ces deux questions sont traitées par le choix d'une représentation convenable des textes — par exemple, une représentation des textes par un ensemble de *concepts* du domaine.

Ce chapitre introductif résume le contenu de notre mémoire en apportant nos réponses aux besoins exprimés pour la FdT et aux difficultés liées à ces besoins décrits ci-dessus. Le mémoire détaille nos choix de représentation des textes et d'analyse des résultats que nous justifierons. La faisabilité de la méthodologie de FdT que nous avons définie est validée par un outil opérationnel sur des bases de textes réalistes de plus d'un millier de textes.

Le processus de fouille de textes

La fouille de textes débute par la modélisation des textes en vue de leur préparation pour l'étape de *fouille de données* et s'achève par l'interprétation des résultats de la fouille pour l'enrichissement des connaissances d'un domaine. L'ensemble de ces trois tâches constitue une chaîne que nous appelons « processus de fouille de textes ». Le processus de FdT s'aligne sur le processus d'ECBD présenté par *Fayyad et al.* [Fayyad *et al.*, 1996a] mais possède des spécificités liées aux données textuelles manipulées par ce processus. Il suffit de parcourir un guide touristique, un manuel d'instructions, un brevet d'une molécule chimique ou un article scientifique pour se rendre compte qu'ils ne sont pas comparables en termes de structure et de connaissances véhiculées.

Contrairement aux données classiquement manipulées en ECBD (bases de données, données structurées, etc.), nous montrons que l'étape de modélisation des données textuelles a une grande influence sur la qualité des connaissances extraites à partir des textes.

Nous utilisons des techniques de fouille de données afin d'extraire des éléments d'information susceptibles de constituer des connaissances *pertinentes.* Les techniques de fouille de données ont montré leur capacité à traiter de grandes masses de données. Depuis une quinzaine d'années, la disponibilité de grandes *masses* de textes, principalement en provenance du Web et des bases de données bibliographiques, des domaines industriels (documentations techniques en aéronautique, automobile, etc.) ou médicaux (dossiers cliniques, études épidémiologiques, etc.) justifie l'utilisation de techniques de fouille de données pour la FdT.

Nous observons les éléments d'information grâce à l'extraction des *règles d'association.* Notre processus de fouille de textes cherche donc à extraire, d'un ensemble de textes, des règles d'association portant sur les termes contenus dans les textes. Plus particulièrement, la méthode de fouille de données que nous appliquons s'appuie sur la recherche de *motifs fréquents* qui permettent d'extraire un ensemble de *règles d'association.* Nous estimons que la facilité d'interprétation d'une règle d'association par un analyste est un point positif de la méthode de fouille de données par extraction de règles d'association. Un *motif* est un ensemble de termes utilisé pour décrire un texte. Un motif est *fréquent* s'il apparaît au moins un certain nombre de fois dans les textes. Nous utilisons l'algorithme *Close* pour l'extraction efficace des motifs *fermés* fréquents, c'est-à-dire des motifs qui ont la caractéristique d'être des motifs fermés.

Une règle d'association R est extraite à partir de deux motifs B et H telle que $R : B \overset{P}{\Longrightarrow} H$. La règle R signifie que tout texte qui possède le motif B possède aussi le motif H avec une probabilité P. Un sous-ensemble réduit de règles d'association dites *informatives* de l'ensemble des règles d'association possibles est extrait par notre processus.

Nous considérons que la représentation d'un texte par un ensemble de termes est bien adaptée au calcul des règles d'association car nous pouvons considérer cet ensemble de termes comme un motif qui permettra d'extraire une règle d'association. En revanche, le très grand nombre de règles extraites constitue un problème au sens où l'ensemble des règles devient trop difficile à appréhender par un analyste.

Les apports de la thèse

L'originalité de notre travail réside dans la mise au point d'une méthodologie opérationnelle de fouille de textes s'appuyant sur une approche symbolique. Nous proposons à l'analyste des éléments d'information extraits d'un ensemble de textes dans un ordre de *pertinence* établi et justifié. En ce sens, notre approche se démarque des approches classiques en ECBD qui placent l'intervention de l'analyste au centre du processus pour effectuer toutes les opérations de prétraitement des données et d'interprétation des résultats. Nous plaçons l'intervention de l'analyste dans le rôle de prise de décision finale pour interpréter et valider les connaissances préalablement extraites, filtrées et jugées pertinentes par rapport au domaine des textes fouillés.

Une étude d'un ensemble de *mesures de qualité* probabilistes qu'il est possible d'attacher aux règles d'association est menée afin de les *classer* et d'élaguer le nombre exponentiel de règles d'association extraites selon un critère de présence forte/rare des termes. Les mesures de qualité sont fondées sur le principe de la cooccurrence des *termes* dans un même texte. Un terme habituellement présent avec un autre terme ne peut pas être dû qu'au hasard. La cooccurrence de termes reflète souvent des liens sémantiques entre termes d'un texte. Nous suggérons une classification des règles selon différents « points de vue ». Pour ce faire, nous proposons un algorithme qui combine un ensemble de mesures de qualité. Le critère de sélection des règles pertinentes est

quelquefois désigné par la recherche de *pépites* de connaissances.

Dans ce mémoire, il est clairement montré quel rôle ces mesures de qualité apportent à l'interprétation des règles extraites, comment elles peuvent influer sur la qualité globale du processus de fouille de textes et comment elles peuvent être utilisées pour alimenter des ontologies du domaine des textes. L'interaction avec l'analyste est facilitée car il est guidé durant l'étape d'interprétation grâce au classement et la sélection préalables des règles d'association extraites.

L'utilisation des connaissances du domaine, en parallèle avec les mesures de qualité, renforce les possibilités de sélection des règles par ce qu'elles apportent comme connaissances nouvelles à l'analyste. Nous exploitons un *modèle de connaissances* qui exprime des relations de généralisation entre termes. Nous évaluons la qualité d'une règle d'association par rapport à un modèle de connaissances du domaine en définissant une *mesure de vraisemblance.* La méthodologie que nous proposons permet d'avoir une démarche incrémentale en fouille de textes car le modèle est progressivement enrichi et la valeur de la mesure de vraisemblance d'une règle est modifiée par cet enrichissement.

Deux applications

L'usage et l'interprétation appropriés des mesures probabilistes de qualité, l'exploitation des connaissances du domaine des textes ont été implantés dans un système informatique appelé TAMIS composé de deux modules.

Le premier module, dit *syntaxique,* est fondé sur l'extraction des règles d'association *informatives*. Le module TAMIS *syntaxique* classe les règles extraites selon différentes mesures de qualité. L'analyse des résultats produits par TAMIS *syntaxique* permet de trouver les pépites de connaissances par combinaison des ces mesures de qualité selon un algorithme que nous proposons.

Le second module, appelé Sem-TAMIS, élague les règles d'association en mesurant l'apport de connaissances des règles extraites par rapport à un modèle donné *a priori* du domaine. L'analyse des résultats donnés par l'outil Sem-TAMIS permet d'enrichir effectivement le modèle de connaissances d'un domaine selon une stratégie de placement que nous proposons.

Plan de lecture du mémoire

Nous décrivons la structure des différents chapitres de notre mémoire en commençant, dans les deux premiers chapitres, par situer notre problématique de fouille et lier les différents points abordés aux les travaux similaires existant pour la FdT. Les deux chapitres suivants (4 et 5) constituent notre apport au domaine de FdT.

Le chapitre 2 définit le processus de fouille de textes tel que nous le concevons. Le processus de FdT est constitué de plusieurs étapes. La première étape est la modélisation des données textuelles. Nous présentons l'étape de modélisation des textes et nous justifions notre choix de la représentation des textes incluse dans l'étape de préparation des données pour la FdT. Le bilan que nous proposons à la fin de ce chapitre nous sert d'appui pour le développement de notre système de FdT.

Le chapitre 3 présente les approches de fouille de données (seconde étape du processus de FdT). Nous mettons en parallèle les approches existantes ou les travaux en cours et la technique que nous choisissons (l'extraction de règles d'association). Nous donnons le cadre théorique et

formel de l'étape de fouille de données. Nous présentons, ensuite, l'intérêt de cette étape pour des différentes applications opérant sur les données textuelles.

Dans le chapitre 4, nous décrivons l'outil ORPAILLEUR de FdT, que nous avons développé, appelé TAMIS : *Text Analysis by Mining Interesting_ruleS.* L'outil TAMIS automatise le processus d'extraction de connaissances à partir de textes. Dans un premier temps, nous présentons l'analyse des résultats de notre processus en les confrontant à l'avis de l'analyste. Puis nous décrivons des expérimentations portant sur un corpus de biologie moléculaire montre l'adéquation du classement calculé pour l'aide à l'interprétation des règles extraites.

Les critères de sélection des règles par les mesures de qualité portent sur les textes eux-mêmes. Cependant, nous n'utilisons pas encore les connaissances du domaine décrites *a priori* dans un modèle du domaine, par exemple une *ontologie* du domaine. Pour ce faire, nous décrivons au chapitre 5 une sélection des règles qui utilise un modèle de connaissances.

Le chapitre 5 constitue le second volet de notre outil pour une FdT dite *sémantique.* Nous définissons une *mesure de vraisemblance* qui permet d'évaluer l'adéquation des règles extraites au modèle de connaissances du domaine. En effet, Nous pouvons classer les règles en deux catégories. D'une part, les règles qui sont strictement conformes au modèle sont dites *triviales* et sont élaguées. D'autre part, les règles qui ne dérivent pas du modèle sont potentiellement porteuses de nouvelles connaissances. Nous montrons les propriétés et l'intérêt de cette mesure pour enrichir le modèle de connaissances.

Le chapitre 6 constitue la conclusion de cette thèse. L'apport de notre thèse au domaine de la FdT y est présenté. Le besoin d'applications en fouille de textes est de plus en plus important, nous décrivons brièvement ce besoin et nous donnons les spécificités souhaitables d'un outil de FdT que nous avons dégagées par retour d'expériences de ce travail de thèse.

Afin de simplifier la lecture de ce mémoire, certaines parties techniques sont mises en annexes. L'annexe A présente certaines notations issues des outils de traitement automatique de la langue que nous avons utilisés dans notre processus de FdT. Cette annexe présente également le codage que nous avons choisi pour représenter les données textuelles en entrée de notre processus de FdT. L'annexe B contient certaines justifications et démonstrations de la formalisation mathématique des règles d'association qui sont présentées en chapitre 3. L'annexe C décrit l'outil TAMIS dans sa version actuelle. Enfin, l'annexe D donne le détail de l'interprétation de certaines règles durant la confrontation de nos résultats à l'avis l'analyste. Le mémoire termine par une bibliographie, un glossaire qui donne certaines définitions, sigles et acronymes, suivis d'un index des termes importants du domaine utilisés dans les chapitres de notre mémoire.

Chapitre 2

Définition de la fouille de textes

« An example of datamining might be to extract objects from a database that have the attribute *female* and the relationship *child* to another object. This would establish the class *daughter*. Database marketing might establish a class "yuppie conservative with a guilty conscience". Such a class could be used by a charitable organisation to solicit donations. If the datamining is done well, the organisation can expect higher returns. » **Anonyme**

Sommaire

Introduction

Le terme « fouille de textes » distingue, dans la littérature, des méthodologies et des outils très différents. Selon la culture scientifique des chercheurs qui s'intéressent à la fouille de textes, ce terme recouvre des travaux en recherche d'information, en extraction d'information, en extraction de terminologies, en structuration d'ontologies, pour les systèmes de questions/réponses etc. Dans ce mémoire, nous définissons (en § 2.2) la fouille de textes (FdT) par le contenu comme étant un processus d'extraction de connaissances dans des bases de données (ECBD) appliqué à des *données textuelles.* Notre finalité est de définir une méthodologie de fouille pour ces données

textuelles. Pour ce faire, nous discutons des propriétés des données textuelles en soulevant des problèmes concernant l'analyse et la représentation des contenus des textes. Cette problématique relève des travaux existants en traitement automatique de la langue qui peuvent, selon le cas, répondre ou non à nos besoins pour la modélisation des données textuelles. La modélisation des textes constitue l'étape de préparation des données textuelles en vue de la mise en œuvre des étapes suivantes du processus de FdT— l'utilisation des techniques de fouille de données, l'interprétation des connaissances extraites. Ce chapitre s'attache à caractériser les données textuelles de façon non exhaustive, notamment en ce qui concerne l'application d'outils de traitement automatique des langues. L'étude exclusivement linguistique d'un texte ne constitue pas la préoccupation majeure de notre travail. En revanche, les caractéristiques des données textuelles que nous développons dans ce chapitre (§ 2.3) sont issues d'un besoin et, surtout, d'un retour d'expériences que nous avons menées pour extraire des connaissances à partir d'un corpus de textes.

Dans la seconde partie de ce chapitre (§ 2.4), nous décrivons le processus de modélisation du contenu des textes que nous avons mis en œuvre. Nous nous focaliserons dans les chapitres suivants de ce mémoire sur l'étape de fouille de données et l'interprétation par l'analyste qui mènent à une mise à jour d'une base de connaissances (*i.e.* un modèle terminologique dans un domaine).

2.1 Extraction de connaissances dans des bases de données (ECBD)

L'*Extraction de Connaissances dans des Bases de Données* (ECBD) est une activité qui consiste à analyser un ensemble de données brutes pour en extraire des connaissances exploitables. Les connaissances sont des éléments qui possèdent une syntaxe et une sémantique, formalisées dans un langage de représentation de connaissances. Les connaissances sont manipulées dans un Système à Base de Connaissances (SBC) pour résoudre des problèmes et effectuer des raisonnements. Un raisonnement permet d'inférer de nouvelles connaissances à partir de connaissances existantes.

Un expert du domaine relatif aux données, l'*analyste*, est chargé de diriger l'extraction. Ces nouvelles connaissances viennent compléter le savoir de l'analyste sur le domaine. En fonction de ses objectifs, l'analyste va sélectionner les données et utiliser les outils de *Fouille de Données* (FdD) pour construire des modèles du domaine expliquant les données. L'analyste peut ensuite sélectionner et exploiter les modèles qui représentent un *point de vue* « satisfaisant ».

L'ECBD fédère trois grands domaines de recherche :

- l'apprentissage — trouver des relations entre les propriétés possédées par des instances. Une instance est une entité ou un individu possédant un certain nombre de propriétés. Par exemple, Jean est une instance de personne qui peut être caractérisée par son nom, son date de naissance, sa taille, sa profession, etc., qui, lorsqu'elles sont définies (*i.e.* lorsqu'elles prennent des valeurs) constituent ses propriétés ;
- la rétro-conception dans les bases de données relationnelles — extraire des dépendances fonctionnelles entre les propriétés dans un schéma d'une base de données ;
- la représentation de connaissances — donner une représentation logique aux éléments d'information manipulés pour effectuer un raisonnement, par exemple une inférence par déduction ou par induction (*i.e.* ou généralisation).

Un système d'ECBD [Simon, 2000] s'articule autour de quatre composantes :

(a) Une ou plusieurs bases de données et leurs systèmes de gestion. Un système d'ECBD doit être capable de traiter des masses de données volumineuses. Le passage à l'échelle d'une petite à une grande application doit se faire de façon transparente pour l'analyste ;

(b) Un système à base de connaissances qui permet à la fois la gestion des connaissances et la résolution de problèmes liés au domaine des données. Le SBC utilise une base de connaissances (par exemple une ontologie du domaine) qui est enrichie grâce aux nouvelles connaissances inférées par le SBC ;

(c) Un système de fouille de données (FdD) pouvant s'appuyer sur des techniques symboliques comme l'extraction des règles d'association [Agrawal et Srikant, 1994], la classification par treillis de Galois [Barbut et Monjardet, 1970; Davey et Priestley, 1994] ou l'induction par des arbres de décision [Breiman *et al.*, 1984; Quinlan, 1986]. La FdD peut également s'appuyer sur des techniques numériques telle que l'analyse des données ou les statistiques ;

(d) Une interface se chargeant des interactions avec l'analyste et de la visualisation des résultats. L'analyste est chargé de guider les recherches et de valider les connaissances extraites. Il est donc au centre de ces quatre composantes.

Il y a parfois confusion, pour certains auteurs, entre FdD et ECBD. Ces auteurs considèrent que l'utilisation d'outils de FdD suffit à extraire des connaissances. Or, la FdD est une étape contenue dans le processus d'ECBD.

Dans la section § 2.2 qui suit, nous nous appuyons sur le processus d'ECBD pour définir la fouille de textes comme étant un processus d'ECBD prenant en entrée des données textuelles.

2.2 Fouille de textes : un paradigme de l'ECBD

La fouille de textes, ou *text mining*, est introduite au milieu des années quatre-vingt-dix sous le terme *Knowledge Discovery in Textual Databases* (KDT) [Feldman et Dagan, 1995] ou *Text Data Mining* (TDM) [Hearst, 1999], puis traduit en français dans [Kodratoff, 2000b] par *Extraction des Connaissances à partir de Textes* (ECT). Nous gardons le terme « fouille de textes » car c'est le plus usité dans la littérature, bien que le terme ECT nous paraît plus approprié. Dans le texte introductif de l'atelier « Text Mining » de la conférence KDD2000, les organisateurs [Grobelnik *et al.*, 2000] écrivent que : « l'objectif de la fouille de textes est d'exploiter l'information contenue dans les documents textuels de différentes manières, incluant les analyses classiquement faites en fouille de données : découvrir des patrons et des tendances dans les données, trouver des associations entre les notions, construire des règles de prédiction, etc. ». Dans [Hearst, 1999] : « la fouille de données textuelles est un processus d'analyse exploratoire de données qui permet de révéler de nouvelles connaissances[1] ou de permettre de répondre, de façon pertinente, à des questions. ». Pour [Kodratoff, 2000a] : « le but d'un processus de fouille de textes est de trouver des relations intéressantes (...) impossibles ou difficiles à détecter par une analyse séquentielle de l'information. ». Toutes ces définitions sont en accord avec notre vision de la FdT.

Nous considérons la *fouille de textes* (FdT) comme un paradigme de l'ECBD au sens où le processus de FdT prend modèle sur celui de l'ECBD, c'est-à-dire que c'est une instance de

[1] Le mot *information* est employé dans la version originale de cette citation.

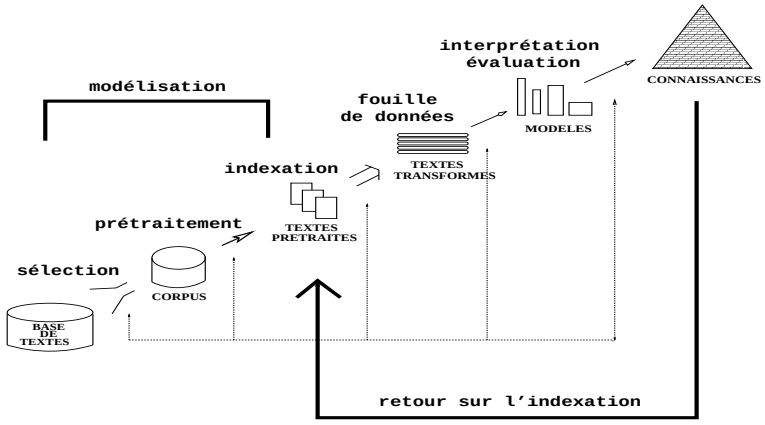


FIG. 2.1 – La chaîne de traitement pour le processus de fouille de textes.

l'ECBD appliquée aux textes. Les textes sont des données peu structurées comparées aux données qui sont modélisées et décrites dans des bases de données. Nous montrons comment extraire, à partir des textes, des éléments d'information qui deviennent par la suite des connaissances. Nous faisons l'hypothèse qu'un texte contient des connaissances explicites (descriptions, faits, etc.) et implicites (domaine du discours, renvois vers faits annexes non décrits dans le texte). Il faut savoir représenter les connaissances explicites et les exploiter pour inférer les connaissances implicites contenues dans les textes.

Nous présentons les différentes étapes d'un système de FdT en § 2.2.1, puis nous nous focalisons, en § 2.2.2, sur la place que tient l'étape de validation et d'interprétation des connaissances potentielles extraites durant ce processus. L'étape de validation et d'interprétation des connaissances extraites d'un processus de FdT n'a pas fait l'objet d'une grande attention dans les travaux antérieurs en FdT. En effet, la diversité des définitions et des travaux se réclamant de la FdT nous amène à nous préoccuper surtout de l'utilisation des connaissances extraites afin de résoudre des problèmes ou pour apporter des connaissances nouvelles à un analyste dans un domaine de spécialité.

2.2.1 Chaîne de traitement pour le processus de FdT

Nous décrivons le processus de FdT par la figure 2.1 qui est calquée sur le schéma de l'ECBD présenté dans [Fayyad *et al.*, 1996a] et montre les différentes étapes de traitement dans un processus de FdT. Les données traitées sont constituées d'un ensemble de textes. Chaque texte est représenté par un ensemble de mots-clés. Cette représentation est stockée dans une base de données.

Nous considérons un texte comme une entité porteuse d'une information qu'il faut préparer, représenter et organiser pour que nous puissions utiliser des outils de fouille de données et valider les résultats de la fouille. La transformation des données textuelles en connaissances se compose donc de trois principales étapes :

(1) La modélisation du contenu des textes ;

(2) Les outils de fouille de données proprement dits ;

(3) Le module d'analyse des résultats et leur validation.

La modélisation du contenu des textes permet d'extraire les données à partir des textes. Nous nous appuyons sur une représentation de type : un texte = {un ensemble de mots-clés} qui est une représentation également communément utilisée en *recherche d'information* car cette représentation nous permet d'utiliser, par la suite, des outils de FdD.

De la même façon que pour un processus d'ECBD, les outils de FdD constituent le module calculatoire d'un système de FdT. Les algorithmes de fouille de données que nous réutilisons et adaptons ont montré leur intérêt par la capacité à traiter de grandes masses de données, ce qui nous permet d'envisager de traiter les données très volumineuses extraites des textes. Par conséquent, l'utilisation des techniques existantes de nous semble pertinente FdD dans le processus de FdT.

La contribution de l'analyste est indispensable pour les étapes d'analyse et la validation des connaissances potentielles extraites car ces deux étapes ne peuvent pas se faire de façon automatique. Le processus de FdT est semi-automatique. Ce n'est qu'une fois les résultats validés qu'ils prennent le statut de connaissances. Ces connaissances peuvent alimenter une base de connaissances ou être exploitées à nouveau par le processus de FdT afin d'affiner la modélisation des textes. Nous appelons, par la suite, cette base de connaissances l'*ontologie* du domaine. De notre point de vue, une ontologie du domaine est une hiérarchie de concepts d'un domaine de spécialité. Chaque concept est représenté par un terme[2]. L'ontologie est une description valable pour une tâche ciblée et dans un domaine restreint. Une ontologie du domaine (*domain ontology*) est différente de la définition classique d'une ontologie (*top-level ontology*) qui sert à représenter des structures conceptuelles et méta-structures applicables à des points de vues philosophiques et logiques de l'univers [Maedche et Staab, 2000].

2.2.2 Acquisition itérative et incrémentale de connaissances

Le processus de FdT n'est pas linéaire comme le suggère à première vue FIG. 2.1. Il est possible d'effectuer un retour entre deux ou plusieurs étapes successives de la chaîne de traitement afin d'améliorer le résultat de chaque étape et d'affiner, au final, le résultat du processus. Pour ce faire, nous avons ajouté à FIG. 2.1 une étape importante que nous avons observée, appelée « retour sur l'indexation », afin de montrer que le processus de FdT est *itératif.*

Nous adoptons l'hypothèse forte qu'un système de fouille de textes doit s'appuyer sur l'utilisation de connaissances du domaine lors de l'extraction de connaissances à partir de textes. Nous exploitons cette hypothèse que nous décrivons au chapitre 5. La FdT est ainsi vue comme le processus alimentant un système à base de connaissances : les connaissances extraites sont stockées dans la base pour être réutilisées dans d'autres applications et mises à jour le cas échéant. Chaque étape du processus de fouille de textes s'appuie sur l'étape qui la précède. La chaîne de traitement pour le processus de FdT (*cf.* § 2.2.1) est fortement *incrémentale.* Les éléments d'information identifiés dans une étape servent à enrichir l'étape suivante et ainsi de suite. L'analyste choisit un point de vue à analyser sur tout ou une partie des données de départ.

La chaîne de traitement pour le processus de FdT que nous décrivons en § 2.2.1 rend le travail de l'analyste efficace en lui donnant accès prioritairement à des connaissances *rares* et/ou potentiellement *nouvelles.* L'analyste doit avoir le rôle de prise de décision finale pour valider ou pas les connaissances extraites, filtrées et jugées pertinentes par rapport au domaine des textes

[2]Nous définissons, ici, un terme comme étant une suite d'un ou plusieurs mots. Nous définirons plus précisément un terme en § 2.4.2.

fouillés. Le jugement de pertinence se fait selon des critères de qualité que nous nous fixons et que nous développons aux chapitres 4 et 5. Les critères de qualité sont : (1) d'ordre *syntaxique* pour la recherche de « pépites » de connaissances parmi l'ensemble des informations présentes dans les textes, (2) d'ordre *sémantique* pour la recherche de nouvelles connaissances inconnues de l'analyste et non décrites dans une base de connaissances du domaine, par exemple une ontologie. En cela, le processus de FdT décrit en § 2.2.1 se démarque du processus classique d'ECBD qui place l'analyste au centre du processus pour effectuer toutes les opérations de prétraitement des données, d'interprétation et de validation de connaissances extraites à partir de données.

Pour commencer la description détaillée des étapes d'un processus de FdT, nous caractérisons les données textuelles en entrée de ce processus. Par la suite, nous donnons des notions de l'utilisation d'une technique de FdD qui extrait des motifs fréquents et des règles d'association et sur laquelle nous revenons en détail au § 3, Nous faisons ensuite un bilan de la FdT.

2.3 Modélisation du contenu des textes : des liens avec le TAL

Nous proposons une étude, plus précisément une réflexion, sur la caractérisation de notre point de vue des données textuelles afin d'utiliser ces données dans un processus de FdT par le contenu. Nous soulignons durant cette réflexion certains travaux existants en Traitement Automatique de la Langue (TAL). Les travaux de TAL que nous citons correspondent à nos besoins de représentation et d'utilisation des données textuelles pour la FdT. La réflexion que nous proposons dans cette section n'est pas exhaustive car elle résulte d'un besoin d'exprimer des éléments d'information contenus dans les textes afin d'appliquer des outils de FdD.

Les travaux en TAL ont pour but, dès l'apparition de l'Intelligence Artificielle (IA) dans les années cinquante, de traiter automatiquement des textes en s'appuyant sur des théories linguistiques et un modèle formel de la langue afin de produire une syntaxe, plus précisément une grammaire générative. Cette grammaire, dite opérationnelle, sert de modèle à la compréhension du langage [Chomsky, 1957]. Il s'avère que l'utilisation d'une grammaire ne suffit pas car toute grammaire ne peut être exhaustive et reconnaître tous les cas possibles de construction de phrases correctes dans une langue. Il apparaît donc que le traitement automatique de textes ne peut s'appuyer uniquement sur des analyses linguistiques mais également sur des connaissances psychologiques et sociologiques. Les théories psycholinguistiques sont nécessaires pour analyser les connaissances véhiculées par un texte.

Le problème de la représentation et de l'utilisation des connaissances se pose donc à un niveau méta-linguistique, c'est-à-dire au niveau conceptuel. Le niveau conceptuel permet d'unifier les différentes graphies de mots utilisés en langage naturel dans les textes ainsi que leurs synonymes autour d'un terme unique dit *attesté* par un analyste du domaine. Ce terme servira de *concept* pour désigner ces différents mots. D. Kayser [Kayser, 1988] parle d'ailleurs d'une famille ouverte d'entités, chacune d'elle dénote une interprétation possible du concept. Par exemple, les termes en biologie moléculaire « DNA topoisomerase IV», « topoisomerase II » et « DNA gyrase » sont indifféremment utilisés par les auteurs pour désigner l'enzyme « gyrase ». Selon l'expérience faite en laboratoire, un des termes qui dénotent le concept de « gyrase » sera utilisé. L'analyste, expert du domaine, sait suivant le terme utilisé le type d'expérience réalisée (dosage, *in vivo, in vitro*, etc.). L'analyste joue donc un rôle plus classique de *terminologue* puisqu'il relie plusieurs termes

du domaine à un seul concept.

Nous insistons sur le fait que notre problématique porte sur la fouille de textes par le contenu. Il s'agit de prendre en compte une représentation conceptuelle de contenu d'un texte pour permettre un raisonnement sur les connaissances qui sont véhiculées par ce texte. Nous reprenons, à ce titre, la définition de [Lévy, 1994] : « Le but d'un traitement *sémantique*, c'est-à-dire le traitement du contenu du texte, est de trouver pour un corpus et un objectif, un système d'interprétation convenable, c'est-à-dire les faits et règles qui produisent à partir du texte les éléments utiles à l'usage projeté ». Contrairement à la vision chomskienne qui considère le traitement sémantique comme la production de primitives constituant une grammaire (*cf.* les travaux de *sémantique grammaticale* de Montague [Montague, 1974] fondés sur une logique d'ordre supérieur), la démarche en FdT, s'appuyant sur la *sémantique lexicale*, est moins ambitieuse en ce sens que la formalisation de connaissances se fait à partir d'un corpus donné et pour une tâche donnée. Néanmoins, une telle démarche est courante dans beaucoup de travaux en TAL, notamment en *recherche d'information* et en *extraction d'information*, que nous définissons plus tard, parce qu'elle est opérationnelle. Les données textuelles sont une forme particulière de données au sens où les données textuelles ne sont pas délimités, structurées et étiquetés sémantiquement de façon explicite contrairement aux informations dans une base de données.

Nous examinons les caractéristiques d'une donnée textuelle en § 2.3.1 à travers les connaissances que véhicule ce texte. Nous discutons des différents niveaux d'analyse d'un texte en vue de sa compréhension en § 2.3.2, classiquement en TAL, puis ensuite, de notre point de vue. Nous expliquons quels types de textes nous voulons traiter en § 2.3.3. Différentes représentations possibles pour analyser le contenu d'un texte sont données en § 2.3.4 et enfin, la motivation du choix de notre représentation des textes y est expliquée.

2.3.1 Caractéristiques d'une donnée textuelle

Un *texte* peut être vu comme une suite de mots (ou lexèmes) séparés par des espaces et par un ensemble de caractères de ponctuation. Les mots sont groupés dans des phrases, elles-mêmes groupées en un paragraphe. Une séquence de paragraphes constitue une section, une suite de sections constitue un texte. La caractérisation linguistique en TAL d'un texte passe par l'étude de ses constituants au niveau structurel (mots, phrases, paragraphes, etc.). La dimension linguistique exprimée dans un texte est considérée comme un moyen d'accès à l'information que représente un texte. L'étude exclusivement linguistique d'un texte n'est d'ailleurs pas l'objet de notre travail. En effet, nous définissons un texte par l'ensemble des connaissances qu'il véhicule car un texte est plus qu'une suite de sections juxtaposées. Un texte véhicule des connaissances qui peuvent être énoncées sous différentes formes :

- Propositionnelle (fait, description, etc.) qui prend une valeur de vérité `vrai` ou `faux`. Par exemple, « le chat de la voisine boit du lait. » ;
- Prédicative. Par exemple, « le prénom du chanteur du groupe The Cure est Robert. » dont le prédicat et sa valeur sont : $\texttt{Prénom}(\texttt{chanteur}, \text{“}\texttt{The Cure}\text{”}) = \text{“}\texttt{Robert}\text{”}$;
- Modale (croyance, évolution, intention, etc.). Par exemple, « je crois de plus en plus ... » ;
- Temporelle, par exemple, « une augmentation brusque de la température est significative si elle ne survient pas après la mise en route d'un dispositif de chauffe. » [Ligozat, 1996] ;
- Spatiale, par exemple, « le campus est au sud de la ville. » ;
- Liée à la quantification. Par exemple, « tous les hommes sont mortels. » ;

– Ou des combinaisons entre ces différentes formes. Par exemple, « je crois de plus en plus que le prénom du chanteur du groupe The Cure est Robert. ».

Des marqueurs linguistiques comme des adverbes (avant, après, etc.), des verbes (croire, espérer, etc.) ou des groupes nominaux (à proximité de, le prénom du, etc.), ainsi que les temps des verbes (passé, présent, futur, etc.) peuvent renseigner sur les formes de connaissances et les relations entre connaissances. Ces formes de connaissances constituent ce qu'il est possible de trouver dans un texte par une analyse TAL du contenu d'un texte.

De plus, les connaissances véhiculées par un texte obéissent à un ordre et à un enchaînement logiques que propose l'auteur du texte. Par exemple, soit le résumé suivant extrait du corpus de biologie moléculaire que nous utilisons dans nos expérimentations. Les textes du corpus sont choisis dans un domaine de spécialité très restreint qui concerne la mutation de gènes dans les bactéries. La mutation rend ces bactéries résistantes aux antibiotiques.

Titre : A Corynebacterium glutamicum gene conferring multidrug resistance in the heterologous host Escherichia coli.
Résumé : A chromosomal DNA fragment from the erythromycin-sensitive bacterium Corynebacterium glutamicum ATCC 13032 was shown to mediate resistance against erythromycin, tetracycline, puromycin, and bleomycin in Escherichia coli.

Multicopy cloning of the fragment did not cause a resistance phenotype in C. glutamicum. The corresponding gene encodes a hydrophobic protein with 12 potential transmembrane-spanning ex-helical segments showing similarity to drug-H+ antiporters.

Le titre nous informe qu'un gène de la bactérie *Corynebacterium glutamicum* la rend résistante à plusieurs antibiotiques dans un environnement bactérien donné (Escherichia coli). C'est un schéma classique de mutation mais il est incomplet. La première phrase de ce résumé précise le schéma de la résistance, à savoir :

– Le nom de la protéine, qui est une séquence de gènes, de la bactérie Corynebacterium glutamicum qui est l'objet de l'étude ATCC 13032 ;
– Les antibiotiques auxquels la bactérie résiste : erythromycin, tetracycline, puromycin et bleomycin ;
– Le caractère restreint et local de la mutation du gène dans un fragment de l'ADN : DNA fragment.

Pour avoir une analyse complète des connaissances contenues dans ce résumé d'article, nous ne pouvons pas nous satisfaire uniquement de ces informations. Il nous faut repérer le nom du gène. Si nous poursuivons la lecture du résumé, l'information sur le gène est-elle renseignée dans la troisième phrase du résumé ? « The corresponding gene encodes a hydrophobic protein with 12 potential transmembrane-spanning ex-helical segments showing similarity to drug-H+ antiporters ». Malheureusement, nous trouvons seulement une description des caractéristiques du gène, c'est-à-dire qu'il participe à la composition de protéines hydrophobes ayant des segments de forme particulière. Ces protéines sont proches d'une famille, les « drug-H+ antiporters », connues dans le domaine. Il nous faut recourir à la lecture de la suite (*i.e.* du corps) du texte pour trouver le nom du gène identifié par l'étude. Le nom de gène *cmr* est effectivement cité plusieurs fois dans le corps du texte.

Les problèmes que nous soulevons dans le résumé ci-dessus sont nombreux. Nous nous posons la question de la pertinence des informations contenues dans un texte. Ces informations sont-elles complètes et suffisantes ? Nous discutons ce point au § 2.3.2 en décrivant les différents niveaux

d'analyse d'un texte. Quel statut faut-il donner à ces informations lorsqu'elles sont décrites par parties en plusieurs endroits du texte (dans plusieurs phrases du résumé, une partie dans le résumé et la suite dans le corps du texte). Nous discutons ce point au § 2.3.3 en décrivant la typologie d'un texte. Par conséquent, nous nous interrogeons sur la façon d'identifier et de représenter ces informations qui permette, d'une part, d'analyser des textes en rendant compte de leur contenu et, d'autre part, d'activer des outils de FdD. Nous discutons ce point au § 2.3.4 en décrivant les différentes représentations du contenu des textes.

2.3.2 Niveaux d'analyse pour la compréhension d'un texte

Plusieurs niveaux d'analyse permettent d'analyser un texte au niveau TAL pour la compréhension de son contenu. Même s'il n'y a pas un consensus sur la dénomination des différents niveaux d'analyse d'un texte, les travaux en TAL décomposent un texte en au moins cinq niveaux :

(1) L'analyse lexicale établit des liens entre mots et permet de construire une représentation de la phrase ;

(2) La syntaxe de la phrase, *i.e.* l'analyse de la construction grammaticale, l'analyse morpho-syntaxique ;

(3) La sémantique de la phrase, *i.e.* le sens de la phrase en dehors du contexte et en dehors de la composition de sens avec les autres phrases du même texte ;

(4) La structure du discours, *i.e.* ce qui gère l'articulation entre les paragraphes et l'enchaînement des phrases dans les paragraphes ;

(5) Enfin, la structure logique du texte (introduction, hypothèse, fait, commentaire, conclusion, etc.).

Chaque niveau s'appuie sur l'analyse faite au niveau qui le précède. Nous retrouvons, dans ces cinq niveaux, la nécessité de formaliser le *sens* du texte (*i.e.* les connaissances véhiculées par le texte). Les problèmes qui se posent sur les quatre premiers niveaux selon G. Sabah [Sabah, 2000] sont :

Au niveau (1) Lexical : quel lien existe-t-il entre les mots et leurs sens ?

Au niveau (2) Syntaxique : quel sens est porté par les structures grammaticales dans lesquelles interviennent ces mots ?

Au niveau (3) Sémantique : comment sont représentées, obtenues et traitées ces significations ?

Au niveau (4) Pragmatique : quelles sont les influences des connaissances sur le monde et la situation pour déterminer le sens ?

Ces quatre niveaux font des textes des données plus complexes à traiter que des données d'une base de données dont la sémantique des relations est généralement plus simple et a fait l'objet d'une modélisation au préalable.

Comme il n'existe pas à l'heure actuelle de sémantique unifiée pour la représentation de l'intégralité du contenu d'un texte, la première étape de notre processus est donc de définir une modélisation de son contenu. Cette complexité d'analyse des textes fait l'intérêt de la FdT comparée à la fouille dans des bases de données. En effet, le contenu qui est extrait d'un texte peut être différent en fonction du niveau de représentation choisi.

Dans un premier niveau d'analyse, le processus de compréhension de textes consiste à relier des termes dans une même phrase. Par exemple, lorsque nous lisons la suite des deux termes : « Pierre Dupont », nous savons que cela désigne une personne, que *Pierre* en est le prénom et *Dupont* le nom. Comment le sait-on ? À partir de ses connaissances implicites, tout lecteur francophone sait que *Pierre* est soit un prénom généralement attribué à un humain de sexe masculin, soit un objet concret (issu de la roche terrestre). *Dupont* est un nom propre car il ne correspond à aucun objet (concret ou abstrait) particulier. L'apparition des deux termes ensemble fait référence à un schéma de relation (`nom1` =*prénom*+majuscule — `nom2` =*patronyme*) que le lecteur connaît : l'identification d'une personne[3]. Le même processus est répété pour la phrase, et ainsi de suite pour plusieurs phrases, paragraphes, sections.

Le niveau supérieur de compréhension du texte consiste à relier entre elles des parties de paragraphes d'un même texte, ou des parties venant de textes différents. L'action de lier des parties de textes détermine l'environnement du discours et relève d'un processus de détermination et de maintien de cohérence au sens logique. Un processus informatique ne dispose pas de ces connaissances implicites. Pour cela, la compréhension automatique de textes nécessite que les connaissances soient explicitées dans une base de connaissances.

Nous illustrons la suite de cette section par des exemples de textes afin de mettre en évidence certains problèmes liés à d'autres niveaux d'analyse du contenu des textes. Ces niveaux d'analyse sont apparus lors de la phase de préparation et de modélisation des données textuelles servant à activer le processus de FdT. Par conséquent, les niveaux que nous décrivons dans la suite de cette section sont typiques d'une analyse TAL tout en étant transversaux ou correspondant à la fois à plusieurs des cinq niveaux classiques de TAL (lexical, syntaxique, etc.). Nous soulignons la difficulté de la compréhension d'un texte à travers des spécificités liées au traitement des dimensions multilingue et culturelle en § 2.3.2.1, au repérage de concepts et d'entités nommées en § 2.3.2.2, au traitement de l'ambiguïté en § 2.3.2.3 et au traitement des présupposés d'interprétation en § 2.3.2.4.

2.3.2.1 Traitement des dimensions multilingue et culturelle

Les textes en langage naturel véhiculent une part de connaissances relatives à des aspects autres qu'intrinsèquement linguistiques. Des connaissances supplémentaires d'ordre multilingue, culturel, sociologique, etc. peuvent modifier la compréhension d'un texte.

Par exemple, dans les actes de la conférence francophone TALN relative au traitement automatique du langage naturel, nous avons extrait le résumé d'un article présentant un système de génération de question / réponse. Si l'article est rédigé en anglais, alors les auteurs doivent rédiger une traduction en français du résumé de leur article (cas de l'exemple qui suit) et *vice-versa*.

Nous pouvons montrer la différence multilingue par la lecture parallèle des résumés en versions française et anglaise.

> Le but des systèmes de question-réponse est [...] Notre recherche vise plutôt à générer des réponses complètes, sous forme de phrases, étant donnée la réponse *exacte*.
>
> The goal of Question-Answering (QA) systems is [...] The subject of this research is to formulate complete and natural answer-sentences to questions, given the *short* answer.

[3] L'introduction de la majuscule et de la virgule dans ce schéma de relation permet de lever l'ambiguïté pour « Je ne te jette pas la pierre, Dupont ! ».

Nous remarquons la présence de l'adjectif "short" dans la version anglaise qui est remplacé dans la version française par l'adjectif « exacte ». Est-ce que ces informations sont équivalentes pour un spécialiste du domaine ? Plus loin dans les deux résumés, nous trouvons une phrase supplémentaire en anglais :

> The answer-sentences are meant to be self-sufficient ; that is, they should contain enough context to be understood without needing the original question.

Pourquoi cette précision n'est pas présente dans la version française ? Est-ce par omission ? Par manque de vocabulaire des auteurs car "self-sufficient" est difficilement traduisible en français ? Enfin, nous trouvons une phrase supplémentaire sur l'expérimentation dans la version française qui ne figure pas dans la version anglaise.

> Suite à une étude de corpus de phrases-réponses, nous avons développé un ensemble de patrons syntaxiques de réponses correspondant à chaque patron syntaxique de question.

Est-ce parce que l'expérimentation est largement discutée dans le corps de l'article en anglais ?

Nous devons considérer ces différences afin de se donner une représentation adéquate des textes que nous traitons. Le but est d'éviter que l'ordre des mots, l'agencement des idées n'influe sur la représentation des textes que nous prenons en entrée du processus de FdT.

2.3.2.2 Repérage de concepts et d'entités nommées

Le repérage de concepts et d'entités nommées est une tâche importante pour une analyse des textes par le contenu. Cette tâche relève de travaux en *extraction d'information.* En effet, il faut réduire le texte à un ensemble de traits caractéristiques qui représentent son contenu, et ainsi pouvoir manipuler ce texte dans un processus de FdT. Afin de comprendre des textes dans un domaine de spécialité, il faut pouvoir également repérer des éléments d'information particuliers contenus dans les textes Par exemple, dans les textes en biologie moléculaire traitant de la résistance des antibiotiques aux bactéries dues à des mutation de gènes, les éléments suivants doivent être trouvés dans les textes : le type de la mutation, le nom du gène, le nom de la bactérie, le nom de l'antibiotique, etc. Certains de ces noms sont des entités nommées : GyrA, ParC (pour les noms de gènes), Escherichia coli, Corynebacterium glutamicum (pour les noms de bactéries), erythromycin, tetracycline (pour les noms d'antibiotiques). Ser83–>Ile substitution (pour le type de mutation). En revanche, le repérage des concepts est plus délicat à faire. Il s'agit de retrouver un concept commun à partir de différentes graphies de termes présents dans les textes. Nous rappelons l'exemple donnée en § 2.3 des trois termes « DNA topoisomerase IV», « topoisomerase II » et « DNA gyrase » qui dénotent le même concept (*ie.*, la protéine « gyrase »).

En TAL, le repérage de concepts et d'entités nommées permet aussi de réaliser une application d'*extraction d'information* [Grishman, 1997] qui est également vue comme une application de filtrage de l'information contenue dans une grande masse de textes. Plus particulièrement, l'extraction d'information consiste à identifier des instances d'un prédicat appelée patron (ou *template*) et composé d'un ensemble d'arguments, ainsi que des relations qui existent entre les patrons. Soit l'exemple, adapté de [Nédellec *et al.*, 2001], de la phrase suivante :

> [. . .] Previously, it was shown that the *gerE protein inhibits* transcription in vitro of the *sigK gene* encoding *sigmaK*, and leading to a *Ser83–>Ile substitution* [. . .]

qui peut être représentée par la structure de TAB. 2.1.

TAB. 2.1 – Exemple de structure de type objet pour une phrase

Interaction :	**Type :**	inhibition	
	Agent :	gerE protein	
		Source :	GerE gene
		Cible :	SigK gene
		Expression :	substitution Ser83–>Ile
	Produit :	sigmaK protein	

2.3.2.3 Traitement de l'ambiguïté

Un texte rédigé en langage naturel est par nature ambigu. La désambiguïsation d'un texte est donc une tâche essentielle et doit être explicite lorsqu'il s'agit d'automatiser le traitement du langage naturel. La désambiguïsation lexicale (*word sense disambiguation*) [Wilks, 1997] d'un texte passe par l'affectation du sens en contexte à chaque terme du texte. La connaissance du monde permet à un lecteur de désambigüer de façon naturelle un texte, c'est-à-dire de suppléer aux données manquantes ou ambigües présentes dans le texte. Dans ce cas, la réalisation de la tâche de désambiguïsation est faite en confrontant les propres connaissances du lecteur à celles acquises grâce aux nouvelles données présentes dans un texte.

Le rattachement sujet–verbe, c'est-à-dire identifier un groupe nominal qui correspond à un verbe dans une phrase fait partie du traitement de la désambiguïsation. Par exemple : « D'après les assureurs, le gel du bonus qui a entraîné des changements du comportement des assurés est une aubaine ». Qui a entraîné le changement ? le bonus ? le gel ? le gel du bonus ? Qu'est-ce qui est une aubaine ? le changement ? le comportement ? le changement de comportement ? Pour qui est-ce une aubaine ? pour les assureurs ? pour les assurés ?

2.3.2.4 Traitement des présupposés d'interprétation

Les présupposés sont des connaissances, dites préalables, que possède le lecteur avant la lecture du texte. En plus de la compréhension du texte par désambiguïsation comme nous l'avons dit au paragraphe ci-dessus, ces connaissances orientent le raisonnement du lecteur pour inférer de nouvelles connaissances. Nous illustrons les présupposés par l'exemple suivant :

Les compagnies d'assurances automobiles collectent des compte rendus d'accidents de la circulation. Un texte au verso de la feuille du constat amiable d'un accident de la route est rédigé par un des automobilistes. Par convention, pour les assureurs, l'auteur du texte possède le véhicule A. Le texte décrit les faits pour établir les responsabilités dans l'accident. Par exemple dans [Gayral *et al.*, 1994] :

> Je roulais sur la partie droite de la chaussée quand un véhicule arrivant en face dans le virage a été complètement déporté. Serrant à droite au maximum, je n'ai pu éviter la voiture qui venait à grande vitesse.

Un lecteur déduit, avec peu d'efforts de raisonnement, un nombre de connaissances dont :

(a) Le récit comporte deux véhicules A et B ;

(b) Il y a un choc entre les véhicules A et B ;

(c) Les véhicules A et B roulent en sens inverse ;

(d) Le véhicule B roule à droite puis a été déporté vers la gauche ;

(e) Le choc a eu lieu sur la voie à droite du véhicule A ;

(f) Le véhicule B a été déporté parce qu'il a pris le virage trop vite.

Aucune des six déductions ne figure explicitement dans l'énoncé. Par exemple, la déduction (b) ne peut être obtenue qu'en faisant les présupposés suivants :

- Le verbe *éviter* suggère l'existence d'au moins deux éléments dont un est mobile. Les deux éléments ont des trajectoires proches ;
- La modalité *pouvoir* suggère une possibilité de contrôle ;
- La forme négative de la phrase dans l'énoncé.

Si le lecteur ne possède pas de connaissances préalables, alors aucune déduction ne peut être faite à partir des textes.

2.3.3 Typologie de textes

L'ordre des connaissances véhiculées dans un texte dépend du genre de texte à analyser. Le typage d'un texte permet de classer ce texte dans une catégorie prédéfinie *a priori*, c'est-à-dire à lui affecter un genre. Par exemple, les travaux de [Biber, 1992; Kessler *et al.*, 1997] sur la détection du genre d'un texte associent un type de texte à la catégorie sociale de l'auteur du texte (tranche d'âge, origine culturelle, profession, etc.). Cela induit des préoccupations que nous avons déjà mentionnées en § 2.3.2.1 pour analyser des données textuelles en traitant les dimensions culturelles véhiculées par un texte. Le typage d'un texte permet également de construire *a posteriori* des classes de textes.

Deux classifications de textes sont présentées dans [Habert, 2000]. Une classification structurelle (dite *a priori*) et une classification fonctionnelle (dite *a posteriori*).

- La classification structurelle consiste à trouver des corrélations entre des types prédéfinis de textes et des marqueurs linguistiques (de temps, d'aspects, de questions, modaux, etc.) contenus dans des textes. Les critères d'une classification structurelle reposent sur les conditions de production des textes (type de l'émetteur, du récepteur, du canal), sur les buts visés par les textes (expliquer, convaincre, raconter, etc.) sur le genre de textes (rapport, article, conférence, conversation, etc.) ou sur une combinaison de ces critères.
- La classification fonctionnelle (*i.e.* le clustering) consiste à faire émerger les types de textes de façon inductive et *a posteriori.* Pour ce faire, la classification fonctionnelle s'appuie sur des classes homogènes de textes. Un regroupement cohérent des textes est réalisé également suivant le repérage d'un ensemble de marqueurs linguistiques présents dans les textes qui les séparent en classes.

Par exemple, l'outil TYPTEX (Typage et Profilage de Textes) [Habert *et al.*, 2000] est issu de la typologie décrite ci-dessus. L'outil TYPTEX permet de classer structurellement et fonctionnellement chaque texte en s'appuyant sur une étude statistique multidimensionnelle de l'ensemble des textes. Une autre étude détaillée de la caractérisation des aspects structurels et fonctionnels d'un texte est présentée dans [Ide, 1994] dans le cadre de la *Text Encoding Initiative* (TEI) : projet

dont le but est d'émettre des recommandations pour l'édition, le codage et l'échange standard de documents électroniques. Il est également possible de choisir de classer l'importance d'un texte selon d'autres critères comme la notoriété des auteurs, le type de la publication (dans un journal, une conférence, etc.).

Nous suggérons une autre typologie de textes dans laquelle nous distinguons deux catégories de textes : (1) les textes scientifiques et techniques et (2) les textes de la langue commune. Dans la première catégorie nous trouvons les textes dont l'univers du discours est limité à ce que nous appelons un domaine de spécialité : les articles scientifiques, les documentations techniques, les bulletins météorologiques, etc. Dans la seconde catégorie, nous trouvons les textes dont l'univers du discours est plus ouvert : les romans, les dictionnaires ou les articles de presse. La distinction entre ces deux catégories ne se fait pas sur la longueur du texte puisqu'une documentation technique peut être longue et un article de presse peut être bref. La distinction se fait donc sur l'étendue de l'univers du discours et non la longueur du texte.

Les textes scientifiques et techniques (*i.e.* la catégorie 1) constituent un cadre adapté à une représentation formelle pour deux raisons :

(1) Du point de vue de la structure, nous prenons l'exemple de la revue scientifique « *Microbiology* ». La rédaction des articles soumis est contrainte par une structure de texte établie par l'éditeur. Les articles pour cette revue sont structurellement codifiés et obéissent à l'ordre suivant : (a) la page de titre, (b) le résumé, (c) l'introduction, (d) les conditions et méthodes de l'expérience, (e) les résultats ; (f) une discussion si nécessaire et la conclusion, (g) les références bibliographiques, (h) les tableaux, (i) les figures, (j) une section sur la théorie si nécessaire et des annexes. De plus, l'article doit tenir dans un nombre de pages minimum et maximum. L'intérêt d'une structure standard de textes est qu'il est possible de cibler la recherche de connaissances dans des parties de l'article. L'article dont le résumé a été donné en exemple en § 2.3.1 ne suit pas strictement cette structure[4]. Lorsque nous avons cherché le nom du gène dans le corps du texte car il n'était pas cité dans le résumé, nous n'étions pas guidé par la structure du texte de l'article. Si tel était le cas, nous l'aurions directement trouvé dans la section (e) les résultats.

(2) Du point de vue du vocabulaire – que nous appelons par la suite terminologie –, un domaine de spécialité est limité. En effet, le domaine du discours (*i.e.* le contexte) des textes scientifiques et techniques permet de désambiguïser un terme *polysémique*, c'est-à-dire le réduire à un seul sens parmi ses sens possibles. Par exemple, lorsque nous rencontrons le terme « avocat » dans un procès-verbal judiciaire, il est probable qu'il s'agisse plus d'une personne que d'un fruit.

La longueur d'un texte n'est pas proportionnelle à la quantité d'information présente dans ce texte. Par exemple, dans un résumé d'article scientifique l'auteur situe les travaux et les avancées scientifiques en peu de mots. La description compacte de l'information contenue dans les résumés est un atout majeur dont nous nous servons pour expérimenter le processus de FdT. Dans les résumés, nous retrouvons une forte densité de termes issus du domaine, un maximum de contenu informationnel et un minimum d'information inutile (que nous appelons par la suite du *bruit*). Les fautes d'orthographe et de frappe sont également très peu présentes dans un résumé car une attention particulière est portée par l'auteur lors de la présentation de son article. De plus, lors

[4]Cet article est court (3 pages). Il est issu du JOURNAL OF BACTERIOLOGY et non de la revue MICROBIOLOGY.

du processus de soumission de l'article, les relecteurs peuvent identifier les fautes d'orthographe subsistantes.

La seconde catégorie de textes est plus difficile à traiter. La une d'un journal par exemple possède des caractéristiques intrinsèques qui justifient son utilité. La une doit présenter, sur une surface de papier limitée ou sur un écran de navigateur dans la version électronique, le maximum du contenu qui est développé dans les pages intérieures. Les introductions aux articles eux-mêmes ne sont pas des résumés. Les introductions sont rédigées pour renvoyer vers l'article en pages intérieures tout en préservant le contenu et en dévoilant une partie qui nous incite à consulter la suite. Pour que l'information soit complète, il faut souvent se reporter à la suite de l'article. Faut-il considérer la partie de l'article présentée dans la une comme une entité textuelle en soi ? ou faut-il lier cette partie à la suite du texte en page intérieure ? Le style journalistique, en soi, pose un problème dans le choix des mots et des expressions, dans l'usage de périphrases (c'est-à-dire un terme substituant un autre terme : « messagère du printemps » pour « hirondelle »), dans l'utilisation d'ellipses (c'est-à-dire l'omission d'un ou plusieurs éléments sous-entendus dans une ou plusieurs phrases : « nous espérons que le lecteur en trouvera peu dans ce manuscrit, autrement certaines parties tomberaient comme un cheveu. »), etc. Par conséquent, comment représenter une information manquante ou volontairement incomplète dans l'introduction présente dans la une d'un journal ?

De ce fait, nous avons choisi de traiter les textes de types scientifique et technique. De plus, nous nous plaçons à un niveau d'analyse intermédiaire entre les niveaux *sémantique* et *pragmatique* de TAL. Au niveau sémantique, le choix de textes scientifiques et techniques nous permet de justifier une représentation particulière des textes (*i.e.* une représentation par des termes). Ce type de textes évite également une dispersion de la signification des sens des textes. Nous nous restreignons au sens relatif au domaine de spécialité des textes. Au niveau pragmatique, l'analyse des textes dépend des connaissances générales sur le monde de référence (*i.e.* le contexte). De la même façon que pour le niveau sémantique, les textes scientifiques et techniques ont un domaine de discours restreint, les influences de ce domaine sur la détermination des significations possibles des termes du texte sont donc minimes.

Le recours à l'analyste pour définir l'objectif de fouille et interpréter la pertinence de connaissances extraites par un processus de FdT est nécessaire. Une fois que les textes sont choisis et l'objectif de fouille défini, la première tâche à effectuer est de choisir la représentation adéquate des données textuelles qui convient à la FdT.

2.3.4 Différentes représentations des textes

Nous décrivons quelque-unes des représentations possibles pour décrire le contenu d'un texte. Nous évoquons d'abord des représentations qui s'attachent à modéliser la forme des textes, puis des techniques qui demandent une analyse grammaticale fine des textes. Nous mettons ensuite l'accent sur la représentation des textes utilisée dans le domaine de l'extraction d'information. Aucune de ces représentations ne constitue une réponse aux questions ouvertes d'une modélisation qui rende compte de façon complète et satisfaisante à l'analyse du contenu d'un texte en vue de sa compréhension. Nous terminons par la représentation utilisée en recherche d'information en expliquant pourquoi elle convient bien à un processus de FdT.

Les travaux de [Tazi et Virbel, 1985] s'intéressent aux connaissances que nous pouvons acquérir à partir d'un texte en considérant la structure graphique du texte. Les connaissances éditoriales

permettent d'isoler et de renforcer des segments de textes que l'auteur a l'intention de mettre en valeur. Le titre, la casse (majuscule/minuscule), le découpage en paragraphes, l'indentation des phrases, les énumérations sont des formes graphiques qui donnent un statut inégal de l'importance des segments de textes. L'exemple qui est traité est celui de la liste énumérative. La liste peut contenir des sous-listes. Une énumération peut être indicée par des tirets, des puces, des chiffres, des lettres, etc. Le but de ces travaux est de spécifier un éditeur de textes capable de proposer des formes graphiques à certaines séquences de textes selon leur importance du point de vue de l'auteur. Des termes déclencheurs comme : « introduction/en introduction », « résumé/pour résumer »,« nous démontrons », etc. sont également associés à cette analyse du format éditorial du texte. De la même façon, la recherche de contextes définitoires [Pearson, 1998] est également une analyse faite à partir de termes déclencheurs de définitions et en repérant des initiales ou des acronymes.

La représentation des textes par des grammaires constructives [Kay et Fillmore, 1999] permet de définir des structures d'arbres de cas ou rôles sémantiques (*i.e.* case frames) et des relations entre les structures à partir de graphes canoniques de prédicats. Le formalisme des graphes conceptuels [Sowa, 2002] est adapté pour la généralisation de constructions grammaticales indépendamment de l'emplacement (agent, action, patient) dans une phrase. Par exemple, nous pouvons réduire une forme interrogative et sa réponse : « Que mange Jean ? Une pomme. » à la forme dite canonique sous forme indicative et affirmative ($< \texttt{Jean} = \texttt{agent}/\texttt{mange} = \texttt{action}/\texttt{pomme} = \texttt{patient} >$).

Nous pouvons représenter un texte par un *objet* structuré complexe (*i.e.* un prédicat) possédant une ou plusieurs propriétés valuées. Grâce à la représentation par objets, nous pouvons appliquer des processus de raisonnement, dont la classification d'objets et la définition récursive de nouveaux objets par composition d'objets existants. L'exemple typique est celui des modèles de bases de données objets du système $\mathcal{CVL}$ [Calvanese *et al.*, 1995] qui permet de raisonner et classifier des objets décrits dans une base de données. L'utilisation des logiques de descriptions [Nebel, 1990; Napoli, 1997], issues de la logique des prédicats, permet également de faire des raisonnements sur des prédicats appelés concepts. Les modèles de représentation par objets ont une grande puissance d'expressivité mais au prix de contraintes fortes de préparation des données pour définir et construire les objets dans une base de connaissances et ainsi effectuer des raisonnements. Les travaux de [Hahn et Reimer, 1998] sur la découverte de connaissances et le résumé de textes en utilisant une logique de descriptions se heurtent à un problème de taille de la base de connaissances.

Représentation des textes pour l'extraction d'information (EI) L'extraction d'information (EI) [Grishman, 1997] consiste à reconnaître dans les textes des arguments correspondant à un ou plusieurs prédicats définis à l'avance par l'analyste comme étant les modèles constituant l'objet du processus d'extraction d'information. Pour ce faire, il faut repérer ces prédicats dans les textes. Les textes doivent être préalablement étiquetés sémantiquement. À chaque mot du texte est associé sa catégorie sémantique. De plus, l'analyste doit construire un dictionnaire de prédicats spécifique au domaine, ce qui peut s'avérer être une tâche coûteuse en temps. L'ensemble de ces prédicats sont structurés dans un objet appelé *patron* d'extraction (ou template). Par exemple, comme nous l'avons évoqué dans l'exemple en fin de § 2.3.2.2 sur les repérage de concepts et d'entités nommées. La structure du patron est celle de TAB. 2.1 et les prédicats sont $\texttt{Type}(< \texttt{x} >)$

$=$ “`inhibition`” ou bien `Cible`$($`Agent` $=$ “`gerE protein`”, $<$ `y` $>)$ $=$ “`SigK gene`”. Dans le système AUTO-SLOG [Riloff et Jones, 1999], les dictionnaires de patrons sont construits semi-automatiquement grâce à des techniques d'apprentissage à partir d'un corpus annoté sémantiquement. Les exemples de patrons trouvés et validés par l'analyste, dans un premier temps, servent à trouver par généralisation de nouveaux patrons par une technique de boostrap [Jones *et al.*, 2003]. De façon analogue, le système CRYSTAL [Soderland *et al.*, 1995] apprend des patrons de concepts et des règles à partir du plein texte et WHISK [Soderland, 1999] à partir de données semi-structurées, par exemple, à partir des métadonnées de pages Web (*i.e.* leurs balises HTML). Les campagnes annuelles MUC (*Message Understanding Conferences*) servent de cadre à l'évaluation de différents systèmes d'EI en s'appuyant sur un corpus de textes et des patrons d'extraction en entrée communs à tous les systèmes.

Le sens d'un mot est aussi défini à travers les autres mots qui l'entourent. Cette idée a permis la naissance de la statistique linguistique (ou distributionnelle) ainsi que les modèles statistiques du langage. Un modèle statistique de langage consiste à prendre une fenêtre de mots (les uni, bi, tri, ...n-grammes) et à calculer le mot le plus probable qui suit la fenêtre de mots. Une phase d'apprentissage permet de prédire un mot sachant la suite courante de mots. Nous nous servons des régularités des cooccurrences des mots dans les textes pour trouver des *motifs* que nous suggérons comme candidats à l'interprétation par l'analyste. En effet, la théorie harissienne [Harris, 1968] définit le contenu d'un texte grâce au principe statistique de cooccurrence des mots. Un mot habituellement présent, *i.e.* en collocation, avec un autre mot ne peut pas qu'être dû au hasard. La cooccurrence de mots reflète des liens sémantiques entre mots d'un texte. Les travaux en sémantique lexicale [Anick et Pustejovsky, 1990] sont fondés sur le principe de la cooccurrence de mots. Par exemple, nous pouvons dire qu'une majorité de textes qui contiennent les mots *vache folle* contiennent aussi les mots *viande, ESB, maladie de K-Jacob,* etc.

Représentation des textes pour la recherche d'information (RI) La recherche d'information (RI) [Van-Rijsbergen, 1979; Salton, 1989] (appelée également recherche documentaire) permet de retrouver une liste de documents (textes, sites Web, images, vidéos, etc.) en réponse à une requête formulée sous la forme d'une expression combinant un ensemble de descripteurs (*i.e.* de mots-clés) et des opérateurs logiques. L'interrogation de bases de données documentaires dans les bibliothèques ainsi que les moteurs de recherche sur le Web s'appuient sur des technique de recherche d'information. Les requêtes sont atomiques lorsque les descripteurs sont connectés par défaut par l'opérateur logique (ET). La requête est complexe lorsqu'une expression combine des requêtes atomiques et avec les opérateurs logiques (OU, NON). La requête est représentée par son vecteur caractéristique qui est constitué d'un ensemble de mots-clés. Les textes sont également représentés par un vecteur caractéristique représentant son contenu (appelé aussi sac de mots). Il s'agit de trouver, parmi les textes, ceux dont le vecteur caractéristique est le plus proche du vecteur de la requête. La liste des textes est ordonnée selon une mesure de similarité entre les deux vecteurs caractéristiques.

Les travaux de [Wilkinson, 1994] prennent en compte une représentation du texte par un vecteur caractéristique pour chacune des sections du texte. Cette approche permet de renvoyer la partie du document qui répond à la requête. Un découpage, au préalable, d'un texte en parties est nécessaire pour appliquer cette approche. Le passage à l'échelle peut s'avérer délicat car il faut rester prudent sur les conclusions de l'expérimentation qui porte sur une base de textes très pe-

tite. Les campagnes annuelles TREC (*Text Retrieval Evaluation Conferences*) servent de cadre à l'évaluation de différents systèmes de RI en s'appuyant sur un corpus de textes et des requêtes en entrée communs à tous les systèmes.

Nous nous inspirons de la représentation des textes par des termes-clés telle qu'elle est utilisée en RI. En effet, l'utilisation des techniques de FdD durant le processus de FdT impose le modèle : une instance = {un ensemble de propriétés}. Une instance est pour nous un texte. Une proporiété constitue un terme-clé. L'incapacité des outils de FdD à traiter des données structurées pour une base de données textuelles de taille réaliste nous impose le choix de la représentation en sacs de termes. De plus, la mise en oeuvre du processus de FdT impose des limitations à l'utilisation des techniques de TAL. Les outils de TAL ne sont pas utilisables dans leur version d'analyse sémantique ou pragmatique pour la FdT. En effet, ces analyses requièrent l'utilisation de modèles du domaine qui ne sont pas toujours disponibles. La richesse sémantique des textes, la très rare disponibilité d'un modèle sémantique du domaine et surtout le nombre de textes à traiter conduisent à des analyses lourdes faisant intervenir des processus informatiques de taille exponentielle [Rajman et Besançon, 1997]. L'utilisation d'outils *surfaciques* de TAL pour prétraiter les textes est une solution pour palier l'absence d'un modèle relatif aux textes. Le prétraitement de surface en TAL peut être une analyse morpho-syntaxique des mots pour réaliser une indexation automatique des textes par des termes du domaine.

Nous faisons donc le choix de la représentation en sacs de termes-index identifiés en utilisant des outils surfaciques de TAL. Nous développons, dans la suite de ce chapitre, la mise en œuvre effective de la modélisation des textes (étape 1 du processus de FdT). Nous donnons, ensuite, des notions de la technique de FdD que nous utilisons (étape 2 du processus de FdT), et ce, afin de pouvoir faire un premier bilan de notre définition de la FdT.

2.4 Notre proposition pour la modélisation des textes

D'après notre processus de FdT, le passage du texte brut à sa modélisation se fait en deux étapes. La première étape, la *sélection* et le *prétraitement,* est chargée d'extraire dans les textes les parties textuelles intéressantes et de les annoter pour que des outils de TAL puissent être mis en œuvre. La seconde étape, l'*indexation,* doit représenter le texte dans un système formel sur lequel les outils de FdD peuvent être appliqués.

2.4.1 Sélection et prétraitement des textes

Nous partons d'une base de textes bruts avec un objectif de fouille fixé par l'analyste. Durant cette phase dans l'étape de modélisation des données textuelles, nous devons (1) sélectionner les textes qui correspondent à l'objectif de fouille. Pour cela, il faut spécifier la requête à faire sur la base de données documentaire pour réunir les textes qui nous intéressent ; puis (2) prétraiter les textes en vue de leur utilisation par la phase suivante d'indexation des textes.

2.4.1.1 Sélection des champs textuels dans les bases de textes

L'ensemble de textes (*i.e.* le corpus) de nos expérimentations est extrait de la base de données documentaire PASCAL-BIOMED de l'INIST-CNRS[5] et de la base de données MEDLINE [MedLine, 2003] constituées de documents, plus précisément des notices bibliographiques, d'environ 12 millions de résumés d'articles scientifiques collectés depuis le milieu des années soixante.

De même qu'aux données peuvent être associées des métadonnées, les textes sont également caractérisés, dans ces bases, par un ensemble de données contextuelles qui sont codées dans des champs XML : titre, auteur(s), date, statut (publié ou non), mots-clés, etc. La figure 2.2 donne une vue partielle d'un texte de notre corpus.

Texte n° 391
Titre : Sequencing of gyrase and topoisomerase IV quinolone-resistance-determining regions of Chlamydia trachomatis and characterization of quinolone-resistant mutants obtained In vitro.
Auteur(s) : Dessus-Babus-S, Bebear-CM, Charron-A, Bebear-C, de-Barbeyrac-B
Résumé : The L2 reference strain of Chlamydia trachomatis was exposed to subinhibitory concentrations of ofloxacin (0.5 microg/ml) and sparfloxacin (0.015 microg/ml) to select fluoroquinolone-resistant mutants. In this study, **two resistant strains were isolated after four rounds of selection [...]** A point mutation was found in the gyrA quinolone-resistance-determining region (QRDR) of both resistant strains, leading to a *Ser83–>Ile* substitution (Escherichia coli numbering) in the corresponding protein. The gyrB, parC, and parE QRDRs of the resistant strains were identical to those of the reference strain. These results suggest that in C. trachomatis, DNA gyrase is the primary target of ofloxacin and sparfloxacin.

FIG. 2.2 – Vue partielle d'une notice bibliographique (texte raccourci).

La première étape du prétraitement porte donc sur l'extraction pour chaque notice des deux champs constitués de textes en langage naturel (*i.e.* non formellement structurés) : le *titre* et le *résumé*. Nous utilisons, pour ce faire, la librairie DILIB [Ducloy, 1999] qui manipule des structures XML et permet d'identifier et de traiter des portions de textes, de créer des index d'éléments en relation et des fichiers inverses de ces relations.

Équation logique de sélection L'équation logique de sélection permet de constituer un corpus homogène de textes traitant d'un sujet particulier. Les textes de notre expérimentation portent sur la biologie moléculaire, plus particulièrement sur le mécanisme de la résistance des bactéries aux antibiotiques. L'équation de sélection est la requête suivante :

(i) Dans la base PASCAL-BIOMED : DEF = (bactérie ET résistance ET (antibiotique OU antibactérien OU anti-infectieux OU antituberculeux OU antilepreux OU antimicrobien)) SAUF phytopathogène ;

(ii) Dans la base MEDLINE : bacteri⋆[6] and MIME[7] = drug resistance, microbial.

Notre corpus est ainsi constitué de 1 361 textes de notices bibliographiques dont les deux tiers proviennent de PASCAL-BIOMED et le dernier tiers de MEDLINE.

[5] INstitut de l'Information Scientifique et Technique, URI : Unité Recherche et Innovation, qui nous a fourni le corpus.

[6] « bacteri⋆ » est une expression régulière qui signifie : tout terme en anglais qui commence par bacteri (*i.e.* bacteria, bacterial, bactericide, bacterin, bacteriology, bacteriophage, bacteriostasis, etc.).

[7] MIME signifie « Minor MeSH descriptors », les termes descripteurs les plus importants de résumés (Subject Headings) de notices bibliographiques de MEDLINE.

2.4.1.2 Étiquetage morpho-syntaxique

L'étiquetage morpho-syntaxique (Part-Of-Speech (POS) tagging) [Church, 1988] correspond à la préparation des textes pour l'application d'outils de TAL dans la phase de modélisation du contenu. L'étiquetage morpho-syntaxique associe à chaque mot d'une phrase sa catégorie morphologique (genre, nombre) et syntaxique (nom, adjectif, verbe, etc.). Plusieurs étiqueteurs, ou *taggers*, existent à l'heure actuelle sur l'anglais et atteignent des performances autour de 99,5% de correction (le quotient du nombre de mots correctement étiquetés sur le nombre total de mots étiqueté), ce qui en fait des outils de traitement automatique fiables. Les étiqueteurs utilisent, à la base, un modèle statistique de langage appris sur un corpus d'entraînement qui peut prédire – calculer la probabilité maximale – l'apparition de la catégorie d'un mot en fonction de la catégorie du mot (ou de la fenêtre de mots) précédemment rencontré(s). Nous utilisons l'étiqueteur de Brill [Brill et Pop, 1999] qui intègre également un lexique, des règles lexicales et contextuelles – appelées patrons lexicaux – qui le rend plus adaptable à un nouveau domaine scientifique pour lequel le vocabulaire ou les tournures langagières sont plus spécifiques. Il suffit d'adapter et de réécrire des règles lexicales spécifiques à notre corpus. Par exemple, les phrases (f1 et e1) étiquetées donnent respectivement les phrases (f2 et e2) :

(f1) Les fractions pectiques contiennent des proportions hautement estérifiées

(f2) Les/DTN :pl fractions/SBC :pl pectiques/ADJ :pl contiennent/VCJ :pl des/PREP :pl proportions/SBC :pl hautement/ADV estérifiées/ADJ2PAR :pl

ou la phrase en gras extraite du texte de la figure 2.2 :

(e1) Two resistant strains were isolated after four rounds of selection.

(e2) Two/CD resistant/JJ strains/NNS :pl were/VBD isolated/VBN after/IN four/CD rounds/NNS :pl of/IN selection/NN ./.

La forme des textes influe sur la qualité de l'étiquetage : les étiqueteurs sont initialement prévus pour fonctionner sur des phrases complètes isolées, syntaxiquement correctes mais pas sur des ensembles de phrases (des paragraphes). Ainsi, nous avons conçu, dans l'équipe ORPAILLEUR, une nouvelle configuration de l'étiqueteur de Brill pour qu'il soit adapté au traitement de séquences nominales isolées c'est-à-dire une suite de noms pour étiqueter des listes de termes issus d'un thésaurus du domaine [Muller *et al.*, 1997].

Une fois que nous avons réalisé le prétraitement des textes (sélection des champs textuels, requête et étiquetage), nous obtenons un corpus constitué du titre et du résumé des `1 361` notices bibliographiques qui se présentent sous la forme de l'exemple donné en FIG. 2.2 de la page 25. Il reste une étape avant de terminer la première partie appelée « modélisation » dans notre processus de FdT de FIG. 2.1, page 10.

2.4.2 Indexation terminologique pour la modélisation du contenu

Dans la mesure où l'analyste fouille dans les textes sans connaissance au préalable nous avons adopté une modélisation du contenu des textes qui repose sur l'ensemble des *termes* que possèdent ces textes.

Définition 2.1 (Terme) *Un terme est un syntagme, c'est-à-dire qu'il est constitué d'un ou plusieurs mots pris ensemble dans une construction syntaxique considérée comme une unité insécable. Ce terme ne prend de sens que par rapport au contexte dans lequel il est utilisé (corps de*

métier, domaine technique, domaine scientifique, etc.). Ce contexte sera appelé « domaine de spécialité ». Le terme ainsi constitué dénote un objet (abstrait ou concret) du domaine de spécialité[8].

Le sens des mots composant un terme ne suivent pas forcément un schéma de composition. Par exemple, une carte de retrait bancaire dite « carte bleue » n'est pas toujours de couleur bleue. En revanche un « lave-linge » est un équipement domestique qui sert à laver le linge.

La modélisation du contenu des textes consiste en une indexation terminologique contrôlée à partir d'une liste de termes attestée. L'indexation par les termes-index permet d'associer un concept à un groupe de mots – une notion qui appartient à une base de connaissances du domaine de spécialité disponible *a priori* ou que nous construisons – et permet ainsi de passer d'un élément de nature linguistique à un élément que nous qualifions de type connaissance. De plus, l'indexation terminologique constitue la première étape de représentation de la sémantique d'un énoncé sans recours à un choix *a priori* sur la nature de la représentation. Cette représentation suppose que la cooccurrence de termes dans un même texte reflète une proximité sémantique entre ces termes [Church et Hanks, 1989]. Les travaux comparatifs de [Rajman et Besançon, 1997] sur une expérimentation de FdT par une analyse des textes en considérant tous les mots (*i.e.* le plein texte) d'un corpus journalistique d'une agence de presse et ceux de [Feldman et Hirsh, 1997] avec le système CART fondé sur une analyse par des termes-clés, montrent que les premiers ne gagnent pas en qualité de connaissances extraites. Au contraire, ils obtiennent plus de bruit en prenant en compte tous les mots présents dans le texte.

Pour indexer les textes de notre corpus, nous avons utilisé la plate-forme de traitement linguistique ILC (Infométrie, Langage, Connaissance) [Toussaint *et al.*, 1998]. La plate-forme ILC est un environnement qui comprend une chaîne de constitution d'un ensemble de termes du domaine et d'indexation de textes que nous décrivons ci-après.

2.4.2.1 Constitution de ressources terminologiques

Avant d'indexer automatiquement le corpus, il convient de constituer une liste contrôlée de termes pertinents du domaine (que nous appelons une *nomenclature*) et de rechercher ces termes dans les textes. Il convient de s'appuyer sur un sous-ensemble terminologique représentatif du domaine pour l'indexation automatique. Une nomenclature résulte de la fusion de plusieurs thésaurus du domaine, appelés ressources terminologiques dans [François *et al.*, 2001]. L'utilisation d'une nomenclature rend le processus d'indexation *supervisé.* L'indexation est alors plus performante et réduit le nombre de termes collectés comparé à une indexation *non supervisée.* Les différentes ressources terminologiques fusionnées ont une cohérence suffisante pour établir des liens de synonymie entre termes et réduire l'ensemble des termes à un sous-ensemble de termes dits *préférentiels.* Les termes regroupés sous leur terme préférentiel sont également gardés dans leur graphie d'origine avec le statut de termes en forme variante.

Il n'existe pas de nomenclature *ad hoc* ayant la couverture nécessaire pour indexer les textes de biologie moléculaire afin de réaliser notre expérimentation. La concaténation de plusieurs nomenclatures existantes est donc nécessaire. Pour ce faire, nous avons pris un ensemble de nomenclatures existantes :

[8] Cette définition s'écarte, quelque peu, de la définition d'un terme par les linguistes du *Cercle de Vienne* : choix entre *signifié* et *concept,* telle qu'elle est présentée dans [Rastier, 1995].

- UMLS : sous-partie de la base de connaissances de la National Library of Medecine (NLM) constituée de 171 039 termes de microbiologie et de biologie moléculaire, de substances chimiques et biochimiques, de fonctions physiologiques, de pathologies et de techniques de laboratoire ;
- MX : vocabulaire multidisciplinaire de l'INIST constitué de 85 149 termes. Le vocabulaire MX est utilisé pour repérer les noms de maladies, de médicaments, de bactéries et du vocabulaire concernant le mécanisme de résistance des bactéries aux antibiotiques ;
- GENE : vocabulaire de 15 348 termes de noms de gènes collectés à partir des bases de données factuelles GeneBank et SwissProt ;
- MÉDIC : vocabulaire de l'INIST constitué de 1 894 termes portant sur les termes désignant des médicaments et qui figurent dans les notices PASCAL-BIOMED mais n'apparaissent pas dans MX et dans l'UMLS ;
- Bactéries : vocabulaire de 1 751 termes de noms de bactéries rassemblés à partir des bases de données factuelles GeneBank et SwissProt ;
- Enzymes : vocabulaire de l'INIST constitué de 749 termes portant sur les enzymes ;
- Toxi : petit vocabulaire de l'INIST constitué de 156 termes portant sur les produits toxiques.

En plus des ressources collectées dans les bases existantes, une liste de supplémentaire de termes appelée ACQ est automatiquement construite à partir des textes du corpus a été faite par l'INIST. 4 005 termes ont été repérés à partir de patrons syntaxiques. Par exemple, des mutations de gènes comme présenté dans FIG. 2.2, page 25 : « *A point mutation was found in (...) leading to a* **Ser83–>Ile** *substitution in the corresponding protein* ». Plus de la moitié (2 721) ont été validés par l'analyste.

Par la suite, si nous rencontrons un terme de notre nomenclature alors ce terme fera partie de l'indexation du texte.

2.4.2.2 Identification des termes et de leurs variantes : travaux en terminologie

Nous prenons en compte différentes variations morphologiques et syntaxiques dans les termes rencontrés (*i.e.* les candidats-termes) qui conservent la sémantique du terme et renvoient vers vers un même concept (un terme préférentiel). C'est un point essentiel dans le processus de préparation de textes. Un texte fait référence à différentes graphies et synonymes pour désigner un même concept. Deux termes-index qui renvoient vers le même concept constituent deux entrées différentes pour notre processus de FdT. Il est donc important de savoir que ces deux termes-index renvoient vers le même concept. La découverte de cooccurrences entre deux concepts (*i.e.* entre deux termes-index issus de concepts différents) dans les textes est plus efficace si nous évitons, en amont du processus de FdT, la dispersion des termes d'indexation entre un terme préférentiel qui sera pris en compte et ses variantes qui seront ramenées au terme préférentiel.
Nous présentons quatre types de variations morphologiques et syntaxiques citées dans [Daille, 2002] :

- Graphique : la variation graphique concerne le changement de graphie (*i.e.* de casse). Nous constatons ces changements surtout dans les titres d'articles en anglais où tous les termes sont mis en majuscules, ou d'autres variations graphiques comme le statut optionnel du trait d'union comme par exemple, le terme « Mot[-]clé » ;
- Flexionnelle : la variation flexionnelle concerne les termes mis aux pluriels (*i.e.* termes fléchis) et qu'il faut rattacher à un même terme. Par exemple, le terme *conservation de produit* au singulier fournit au pluriel la forme fléchie *conservations de produit* ou *conservations de*

produits ;

- Syntaxique : certaines variations sont dites syntaxiques faibles concernent les mots dits « vides », dits *grammaticaux*, comme les prépositions, les déterminants, etc. Par exemple, *chromatographie* **en** *colonne* $\leftrightarrow$ *chromatographie* **sur** *colonne.* Les autres variations syntaxiques concernent l'ajout d'un modifieur (pronom, adjectif, adverbe, etc.) au terme de base non fléchi. Par exemple, *lait pasteurisé* $\rightarrow$ *lait complet pasteurisé* $\rightarrow$ *lait complètement pasteurisé,* ou concernent les énumérations : *analyse de particule* $\rightarrow$ *l'analyse et le tri de particules* ;
- Morpho-syntaxiques : les variations morpho-syntaxiques affectent la structure interne du terme de base. Les mots vides grammaticaux composant le terme subissent des modification de morphologie (*pourrissement après récolte* $\leftrightarrow$ *pourrissement post-récolte*), de morphologie dérivationnelle (*acidité du sang* $\leftrightarrow$ *acidité sanguine*) ou plus complexe (*éco-emballage* $\leftrightarrow$ *emballage écologique*).

Les variations graphiques, flexionnelles, syntaxiques et morpho-syntaxiques s'enchâssent car une variation flexionnelle peut s'appliquer à une variation graphique, une variation syntaxique peut s'appliquer à une variation flexionnelle, etc. Il est important de prendre en considération ces variations si nous voulons indexer automatiquement un texte par des termes attestés et repérer les termes variants qui sont des synonymes.

2.4.2.3 Mise en œuvre de l'indexation terminologique : utilisation de FASTER

Pour l'indexation des textes, nous utilisons FASTER : Filtrage et Acquisition Syntaxique de TERmes [Jacquemin, 1994][9] qui est un outil informatique qui met en œuvre des principes linguistique de traitement des groupes nominaux et d'identification, dans les textes, de termes d'une nomenclature attestée. FASTER est également un analyseur syntaxique fondé sur les grammaires d'unification [Shieber, 1986] et, plus précisément, il s'appuie sur le formalisme des règles de variations syntaxiques PATR-II pour la forme logique des grammaires TAG d'arbres adjoints [Vijay-Shankar, 1992].

L'analyse syntaxique d'un texte est dite de surface (*shallow parsing*) car le but est de se servir de cette analyse pour identifier des éléments d'information contenus dans un texte et pouvoir par la suite appliquer une méthode de fouille de textes. La méthode de fouille de textes s'appuie sur l'extraction des règles d'association entre les éléments d'information contenus dans les textes afin d'aboutir à l'extraction de connaissances à partir de ces textes. Notre objectif n'est donc pas d'effectuer la tâche de la compréhension du texte à la suite de cette analyse.

Nous avons choisi de procéder de cette manière car l'analyse syntaxique profonde (*deep parsing*) d'un texte en vue de sa compréhension est un processus lourd. Un arbre syntaxique est produit en remplaçant chacune des occurrences des mots de la phrase par son symbole représentant les règles de réécriture de la grammaire dite générative. Est-ce pertinent de faire une analyse profonde des textes dans notre expérimentation ? Par exemple, prenons l'arbre syntaxique correspondant à la première phrase simplifiée issue d'un résumé de notre corpus de biologie moléculaire de la FIG. 2.2 : « The strain of Chlamydia was exposed to subinhibitory concentrations ».

Nous voyons que l'analyse en FIG. 2.3 d'une simple phrase simple correspond à un arbre syntaxique large et profond. Si ce processus est reproduit pour chacune des phrases du texte et pour tous les textes du corpus, nous imaginons la complexité qui est générée.

[9] L'acronyme signifie dans sa version en anglais : FAst Syntactic Term Recogniser (FASTR).

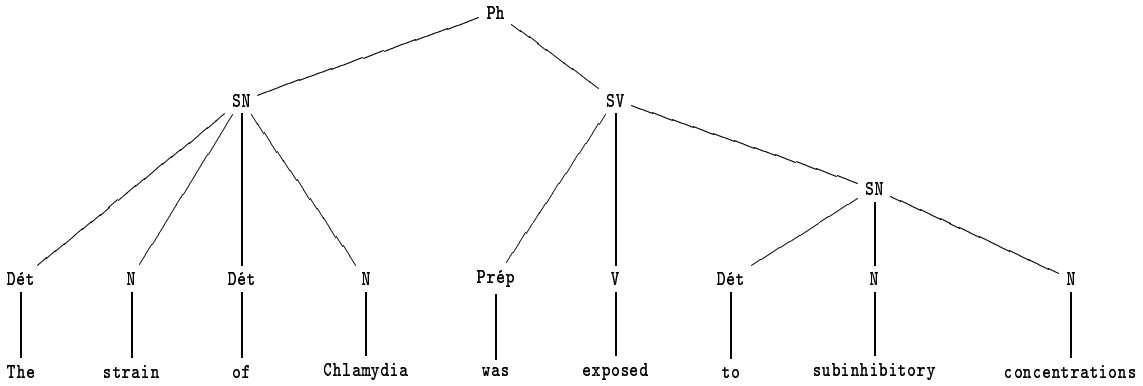


FIG. 2.3 – Arbre issu d'une analyse syntaxique profonde d'une phrase de FIG. 2.2 (étiquettes en français).

La règle PATR-II de la figure 2.4 sert à reconnaître des formes adverbiales `DE NOM EN NOM` (« de minute en minute », « de ville en ville », etc.). La règle stipule qu'un adverbe (complexe) est formé de la suite d'une préposition, d'un nom, d'une deuxième préposition et d'un deuxième nom aux conditions suivantes : la première préposition doit avoir pour lemme « de », la deuxième « en », le premier et le deuxième nom doivent avoir le même trait, c'est-à-dire la même forme graphique.

```
(rule (adv → prep1 nom1 prep2 nom2)
(prep1 lemme) = "de"
(prep2 lemme) = "en"
(nom1 forme) = (nom2 forme))
```

FIG. 2.4 – Exemple d'une règle syntaxique PATR-II.

Le but d'une indexation est de minimiser le silence, c'est-à-dire le fait de ne pas réussir à reconnaître un terme dans un texte. FASTER permet de reconnaître un terme sous des formes variantes. Chaque terme de la nomenclature attestée est caractérisé par sa structure syntaxique (*i.e.* son étiquette morpho-syntaxique). Étant donné une forme variante rencontrée dans un texte, FASTER va considérer cette nouvelle forme comme désignant le même *concept* s'il peut appliquer des méta-règles de transformation de la structure syntaxique (de la forme attestée vers la forme rencontrée). Les méta-règles sont données en plus des règles PATR-II. Par exemple, le terme « *transfer of capsular biosynthesis genes* » doit être considéré comme une forme variante du terme attesté de la nomenclature « *gene transfer* ». Ainsi, la forme attestée « *gene transfer* » peut être reconnue par une opération d'inversion sous la forme « *transfer of genes* » puis par une opération d'insertion sous la forme « *transfer of capsular biosynthesis genes* ».
Tous nos textes sont donc traités, de cette façon, par FASTER. Nous obtenons pour le texte de la figure 2.2, l'ensemble de termes de la figure 2.5.

Certaines variantes ne sont pas acceptables pour l'analyste et l'intérêt de FASTER est de ne garder comme formes variantes que celles qui sont issues d'une transformation linguistique

Terme(s) : "characterization" "determine region" "dna" "escherichia coli" "gyra gene" "gyrase" "gyrb gene" "mutation" "ofloxacin" "parc gene" "pare gene" "point mutation" "protein" "quinolone" "sparfloxacin" "substitution" "topoisomerase"

FIG. 2.5 – Ensemble des termes indexant le texte de la notice bibliographique (figure 2.2, page 25).

permettant de préserver le sens. Dans [Jacquemin, 1997], trois critères sont énoncés et respectés dans FASTER :

(1) Les mots « pleins », dits *lexicaux,* du terme initial sont tous présents, les seuls mots pouvant être élidés sont les mots « vides » – par exemple, *moniteur temps réel* est une variante de *moniteur* **en** *temps réel* ;

(2) Les modifications morphologiques subies par les mots pleins sont des variations flexionnelles ou dérivationnelles – respectivement, *tension***s** *artérielle***s** et *tension des artères* sont des variantes de *tension artérielle* ;

(3) L'ordre des mots peut être modifié et des mots peuvent être insérés à l'intérieur de la variante, mais les relations de dépendance lexicale du terme initial doivent se retrouver à l'identique dans la forme variante du terme – voir l'exemple du paragraphe précédent entre « *transfer of capsular biosynthesis genes* » et « *gene transfer* ».[10]

Cependant, nous sommes conscients qu'une analyse de surface peut induire des erreurs d'indexation et introduire de l'ambiguïté, du bruit. Par exemple, dans l'exemple de texte donné en annexe A.3, page 128, nous relevons la phrase suivante :

« (…) **mice** challenged with a metronidazole-resistant or -sensitive strain isolated from the stomach of a *mouse* **were treated** with metronidazole or amoxicillin. ».

Le terme-index qui est extrait par FASTER est *treat mouse* par une règle PATR-II de type « XX,31,Perm » (*i.e.* une permutation) qui s'applique sur le terme *mouse were treated* reconnu en tant que candidat-terme dans le texte. Pourtant, le sujet de *were treated* est *mice* et non l'occurrence de *mouse* qui est placée juste avant. Ce problème montre les limites de l'approche qui donne le même statut à tous les mots. Ce problème soulève aussi la difficulté de prendre en compte une information lorsqu'elle est présente par parties et « disséminée » dans le texte. La question que nous nous posons est : Comment représenter une information manquante ou volontairement incomplète et présente par parties ? Une analogie intuitive peut être faite avec la une d'un journal décrite en § 2.3.3, page 19. Un résumé d'article scientifique ne peut pas nous dispenser de lire la suite de l'article. Nous ne pouvons pas donner de réponse satisfaisante dans l'absolu à cette question. Nous pensons, néanmoins, qu'il s'agit de faire au mieux pour modéliser une information en faisant un compromis entre l'absence d'information (*i.e.* le silence) et la présence d'erreurs (*i.e.* le bruit).

2.4.3 Représentation des textes par des termes

Nous choisissons une représentation des textes par un ensemble de termes non structurés mais dont la cooccurrence dans les textes révèle des connaissances intéressantes (synonymie, anto-

[10] Une liste des variations dans l'indexation entre la graphie du terme trouvée dans les textes et la graphie d'indexation correspondante est donnée en annexe A.2, page 127.

nymie, méronymie, énumération, proximité conceptuelle, etc.). Nous utilisons un modèle entité-association [Chen, 1976], quelque peu simplifié, présenté dans la FIG. 2.6. Un texte « possède » un ensemble de termes, un terme « caractérise » un ensemble de textes.

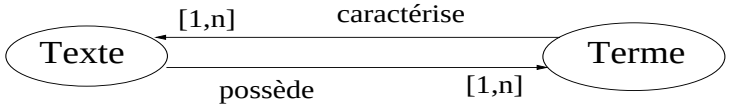


FIG. 2.6 – Un modèle simple d'entité-association utilisé pour représenter les textes en FdT.

Deux raisons nous ont guidé à faire ce choix :

(1) Nous ne disposons pas d'un corpus annoté sémantiquement (*i.e.* où chaque terme a un rôle sémantique de sujet, d'action, d'agent, etc.). L'annotation sémantique du corpus nous permettrait de construire des prédicats pour raisonner sur des objets et inférer des connaissances (*c.f.*, les différentes représentations des textes en § 2.3.4, page 21) ;

(2) La représentation d'un texte par un ensemble de termes est bien adaptée au calcul des règles d'association car l'extraction de règles d'association est fondée sur le calcul des motifs fréquents. Nous considérons l'ensemble des termes qui caractérisent un texte comme un motif. Un motif est fréquent s'il apparaît au moins un certain nombre de fois dans les textes.

La représentation des textes par des termes possède des avantages et des inconvénients. Nous en donnons quelques-uns dans les deux paragraphes suivants.

2.4.3.1 Avantages de la représentation des textes par des termes

La représentation des textes par des termes est communément utilisée dans de nombreuses approches, notamment en RI. Les avantages de cette représentation sont :

(a) Éviter que l'ordre des termes dû à l'agencement des idées n'influe sur la représentation du contenu des textes. En effet, la représentation en « sac de mots » ne tient pas compte de la séquence de termes dans les textes repérés lors de l'indexation ;

(b) Garantir la *robustesse* de la phase de modélisation des textes. En effet, notre but n'est pas une analyse de chacun des textes de notre corpus, mais une analyse de l'ensemble des textes afin de dégager des corrélations entre les concepts du domaine. Nous n'avons pas besoin d'un système qui fasse des vérifications et de la gestion d'erreurs d'analyse pour la modélisation de chaque texte ;

(c) Être capable d'utiliser des techniques de FdD pour traiter des données volumineuses. En effet, les techniques de FdD sont incapables de traiter des données structurées. Les systèmes logiques fondés sur une représentation par des objets structurés sont élégants et les inférences dégagées sont prouvables au sens de la correction et de la complétude. Cependant, ces systèmes se heurtent au problème de la taille des données à traiter. Une base de textes réaliste ne peut être manipulée par ces systèmes ;

(d) S'ancrer dans le domaine. La représentation terminologique utilise une ressource externe constituée d'une nomenclature de termes du domaine. Grâce à la liste restreinte aux termes du domaine, nous évitons la dispersion du processus de FdT vers des termes périphériques au domaine.

2.4.3.2 Limites de la représentation des textes par des termes

Un inconvénient majeur à ce type de représentation est illustré, par exemple, par l'absence de prise en compte de la *négation.* En effet, prenons l'exemple de deux parties de textes extraites d'un corpus de biologie moléculaire sur lequel nous nous sommes appuyés pour réaliser nos expérimentations :

> [n° 000015] When maintained under nonselective conditions, neither the aadA mRNA nor the AadA protein were detected in these subclones. Moreover, since the integrated transforming DNA was not altered or lost expression of the RbcS2 : :aadA : :RbcS2 gene(s) appears to be repressed [...] Identification of an ABC transporter gene that exhibits mRNA level overexpression in fluoroquinolone-resistant Mycobacterium smegmatis.
>
> [n° 000867] However, no resistant strain was detected in the amoxicillin treatment group.

La représentation d'un texte par un ensemble de termes est jusqu'à un certain point pauvre. Ainsi, dans deux phrases comme :

(1) « *When maintained under nonselective conditions, neither the aadA mRNA nor the AadA protein were detected in these subclones* » ;

(2) « *Identification of an ABC transporter gene that exhibits mRNA level over-expression in fluoroquinolone-resistant Mycobacterium smegmatis* ».

Le terme *mRNA* sera identifié comme étant présent dans le texte et donc associé aux gènes *RbcS2* et *aadA* dans la phrase (1) ainsi que dans la phrase (2). Cependant, dans la première phrase, la négation montre bien que *mRNA* n'est pas présent dans cette expérience alors qu'il l'est dans la seconde. La représentation par un ensemble de termes ne permet pas de refléter cette différence qui peut, par la suite, engendrer des erreurs d'interprétation dans les termes mis en relation dans les règles d'association. Il en est de même pour la souche résistante (*resistant strain*) dans l'extrait $\mathtt{n}^\circ$ `000867`. Dans ce cas, seul l'accès aux textes pendant la phase de validation et d'interprétation des connaissances extraites permet de rendre compte de cette différence et de lever l'ambiguïté.

Le deuxième inconvénient vient du fait que nous considérons la présence ou l'absence d'un terme comme principal fondement de la description symbolique d'un texte. Un terme-index est soit présent, soit absent dans un texte. La dualité présence/absence nous oblige à choisir minutieusement le seuil au delà duquel un terme est considéré comme *fréquent* ou pas. Un terme fréquent, par rapport à un seuil fixé par un nombre d'occurrences minimal, participe ou pas à la construction d'un motif fréquent (*cf.* § 2.5 ci-après). La méthode décrite dans [Latiri-Chérif *et al.*, 2002] permet de résoudre le problème de présence/absence. Nous pouvons choisir de fixer un poids $\mathtt{P}$ à chacun des termes indexant un texte. Une valeur réelle variant entre $\mathtt{0}$ et $\mathtt{1}$ est affectée à chaque terme $\mathtt{t}$, selon que ce terme est plus ou moins représentatif du contenu d'un texte $\mathtt{d}$. Le critère de représentativité du terme est un choix difficile. Cela peut être la fréquence d'occurrence du terme $\mathtt{t}$ dans le texte ou dans le corpus entier, sa présence dans le titre, sa position dans le texte (dans le résumé, dans les sections importantes du texte). Nous pouvons par la suite éviter une explosion combinatoire dans la représentation des poids en restreignant le nombre de poids des termes à un sous-ensemble discret tel que le poids du terme $\mathtt{t}$ dans le texte $\mathtt{d}$ soit représenté par la fonction d'appartenance floue $\widetilde{\mathtt{P}}$ avec :

$$\widetilde{\mathtt{P}}_{\mathtt{t}}(\mathtt{d}) \in \{\mathtt{fort},\ \mathtt{moyen},\ \mathtt{faible}\} \tag{2.1}$$

La fonction $\widetilde{\mathtt{P}}$ dans (2.1) permet d'assigner à chaque poids $\mathtt{P} \in [\mathtt{0}, \mathtt{1}]$ une des trois valeurs possibles en découpant l'intervalle $[\mathtt{0}, \mathtt{1}]$ en sous-intervalles flous et en résolvant les conflits d'appartenance

dans les portions d'intervalles floues.

Nous avons décrit la première étape de notre processus de FdT de (*cf.* FIG. 2.1, page 10) par la mise en œuvre de l'étape de modélisation des textes. À présent, nous pouvons appliquer les techniques de FdD. L'algorithme de FdD que nous appliquons est fondé sur l'extraction de motifs fréquents afin de générer un ensemble de règles d'association. Nous donnons, ici, brièvement une idée générale de la technique de FdD que nous utilisons pour la FdT. Nous développons en détail les notions de motifs fréquents et de règles d'association dans le chapitre suivant.

2.5 Notions de motifs fréquents et règles d'association pour la FdT

Un motif est un ensemble de propriétés appartenant à un objet. Un motif fréquent est défini comme un motif présent dans un nombre « plus grand qu'un seuil de fréquence donné » d'objets d'une base de données. Par conséquent, pour qu'un motif soit *fréquent*, il suffit que le nombre de fois où il apparaît soit supérieur ou égal à un seuil σ fixé en paramètre. Le *support* d'un motif est défini comme étant le nombre d'objets possédant ce motif. Le support d'un motif est également exprimable en pourcentage d'objets possédant ce motif par rapport au nombre d'objets total de la base de données.

Dans notre cas, un objet est un texte, une propriété est un terme et un motif est un ensemble de termes. L'ensemble de tous les textes $\mathcal{D}$ et de tous les termes $\mathcal{T}$ sont liés par la relation d'indexation. Cette relation peut être représentée sous la forme d'une matrice $\mathcal{D} \times \mathcal{T}$ de booléens (`1` dénote la *présence* et `0` signifie l'*absence* du terme `t` dans un texte `d`). La matrice de cooccurrence des termes dans les textes constitue la structure en entrée du processus de fouille de textes.

Par exemple, nous reprenons les extraits des textes `n°000015` et `n°000867` donnés en § 2.3.1, page 13. La FIG. 2.7 donne la matrice de cooccurrence ainsi que la matrice d'indexation correspondante. Seul le terme « resistant strain » est commun à ces deux textes. Tous les termes sont ordonnés par ordre alphabétique et sont stockés en minuscules dans des structures XML. La notice bibliographique complète du texte `n°000867` est donnée en annexe § A.3, page 128.

Texte	Concepts
`000015`	"aadA gene" "mRNA gene" "myco. smegmatis" "protein" **[…]** "quinolone" "RbcS2 gene" "*resistant strain*" "subclones"
`000867`	"amoxicillin" "*resistant strain*" "treatment"

↱	aada	amoxicillin	mrna	m. smegmatis	protein	quinolone	rbcs2	r. strain	subclones	treatment
`000015`	1	0	1	1	1	1	1	$1^{\bullet}$	1	0
`000867`	0	1	0	0	0	0	0	$1^{\bullet}$	0	1

FIG. 2.7 – (a) Matrice de cooccurrence des termes pour deux textes – (b) Matrice d'indexation pour ces deux textes.

L'approche naïve pour chercher les motifs fréquents consiste à compter le nombre de fois où chaque ensemble des parties de $\mathcal{T}$ apparaît. Ce qui donne $2^{\mathcal{T}}$ sous-ensembles à tester et conduit à

une combinaison exponentielle pour la recherche des motifs fréquents. Les approches standards de recherche de motifs fréquents s'appuient sur des algorithmes par niveaux (qui en réalité parcourent le treillis des parties $2^{\mathcal{T}}$ en largeur). Pour notre part, nous utilisons l'algorithme « Close » [Pasquier *et al.*, 1999b] qui minimise cet espace de recherche. À partir des motifs fréquents nous calculons les règles d'association entre les termes. Une règle d'association est une règle probabiliste exprimant une corrélation entre la présence de termes dans l'ensemble des textes impliquant la présence d'autres termes dans ces textes. Les notions de motif fréquent, de règles d'association ainsi que l'algorithme *Close* sont présentés dans la section 3.2 consacrée à l'extraction des règles d'association pour la FdT.

2.6 Fouille de textes : bilan

Le processus de FdT est conduit par un analyste qui est expert dans un domaine particulier. Elle donne à celui-ci une vue synthétique du contenu d'un corpus, exhibe des relations entre les différentes notions présentes dans un texte ou des relations entre les textes. Ces relations reflètent des liens de généricité, de similitude, de causalité ou de tendance. L'objectif de la FdT est donc de permettre à l'expert de retrouver, à partir d'un corpus donné, des relations connues dans son domaine, de pouvoir les localiser explicitement dans les textes, de classifier des familles de textes construites à partir d'une ou plusieurs de ces relations. La FdT permet également de découvrir de nouvelles relations. En ce sens, notre définition rejoint celle de [Fayyad *et al.*, 1996b] pour l'ECBD qu'ils qualifient de « processus non trivial d'identification de motifs (d'information) valides, nouveaux, potentiellement utiles et au final compréhensibles à partir d'un ensemble de données. ». De notre point de vue, le principe de la fouille de textes sans connaissance préalable, sans point de vue *a priori* aboutit à des résultats non interprétables et dont la pertinence est difficile à juger.

Le processus de FdT que nous étudions possède des particularités par rapport au processus plus général d'ECBD. Ces particularités viennent du fait qu'il s'applique à des données textuelles.

Nous définissons donc la fouille de textes à travers trois étapes. La modélisation des textes, l'activation d'outils de FdD et l'interprétation des informations extraites.

Nous considérons qu'utiliser les outils de FdD seuls, et non la chaîne complète que nous présentons en FIG. 2.1, revient à faire de l'extraction de connaissances de façon *incomplète*. En effet, le processus de fouille de textes tel que nous le concevons s'appuie sur l'utilisation :

1. d'une méthode opérationnelle d'extraction des règles d'association ;
2. d'un classement des règles suivant des mesures de qualité ;
3. d'un environnement interactif d'accès aux règles et au contenu des textes.

L'extraction des règles d'association (1) se fait en deux étapes. Premièrement, nous calculons les motifs fréquents qui s'appuient sur les motifs fermés fréquents en utilisant l'algorithme *Close* [Pasquier *et al.*, 1999b]. Ces motifs fréquents permettent, deuxièmement, de construire des règles d'association. Nous nous appuyons sur le processus d'extraction de motifs fréquents et de règles d'association pour faire émerger des éléments d'information à partir des textes susceptibles d'être interprétés et devenir des éléments de connaissance utiles et réutilisables. Les mesures de qualité des règles calculées en (2) sont des mesures qui pondèrent chaque règle et permettent donc de les « classer ». Un environnement de navigation (3) aide l'analyste à interpréter les règles d'association obtenues en (1). Il lui permet d'accéder au contenu des textes liés à une règle (*cf.* FIG. 2.2

page 25, complétée par les termes issus de l'indexation du titre et du résumé de FIG. 2.5 page 31). Le chapitre 3 décrit, de façon générale, des méthodes et des mises en œuvre informatiques pour la FdT. En particulier, le chapitre 3 détaille la mise en œuvre de l'étape de FdD correspondant au point (1) tel que nous l'avons défini ci-dessus.

Chapitre 3

Organisation de données textuelles pour la fouille de textes

« Information is only useful when it can be located and synthesized into knowledge. »
Mani Shabrang, business intelligence center, Midland, USA

Sommaire

Introduction

Nous présentons dans ce chapitre différentes méthodes qui opèrent sur le contenu des textes et visent à extraire et structurer des éléments d'information. Ces méthodes sont souvent assimilées à la fouille de textes même si, en réalité, elles portent essentiellement sur l'étape de FdD. Les deux premières sections (§ 3.1 et § 3.2) présentent deux approches très différentes : la première approche repose sur la classification supervisée (en § 3.1.1) ou non supervisée (en § 3.1.2) et la

seconde approche s'appuie sur l'extraction de règles d'association. Nous motivons en § 3.3 notre choix d'utilisation des règles d'association pour l'étape de FdD appliquée à la FdT.

La classification des textes est probablement la méthode de FdD la plus appliquée pour traiter des données textuelles. La raison de cet engouement repose notamment sur les critères suivants :

- La portabilité d'un domaine à un autre des applications issues des méthodes de classification. La classification de textes n'utilise pas de connaissances *a priori* du domaine, elle peut donc être mise en œuvre indifféremment sur des textes de chimie, de littérature, etc. Ces méthodes, souvent issues du domaine de la *recherche d'information*, sont appliquées sur une représentation des textes par mots (sans prendre en compte la notion de *terme* décrite en page 26) et peuvent donc traiter des corpus dans une langue ou dans une autre ;
- Ces différentes méthodes de classification sont, le plus souvent, applicables sur de grandes masses de données (notamment, pour grand nombre de mots-clés) ;
- L'indépendance de ces approches par rapport aux domaines traités par les textes a permis le développement d'outils qui sont très largement diffusés.

En revanche, ces approches se heurtent à des limites qui se font de plus en plus contraignantes :

- Le travail de l'analyste est d'interpréter la signification de ces classes de textes par rapport à ses besoins et ses connaissances. C'est un travail complexe dans la mesure où ce type de méthodes ne le guident pas dans son interprétation ;
- L'interprétation des classes nécessite une double expertise : celle du domaine des données (biologie, chimie, etc.) et une expertise en classification. Il faut, en effet, être capable de détecter des mots de l'indexation qui constituent du bruit, par exemple par leur apparition dans de nombreuses classes qui rend ces classes non interprétables par l'analyste. Il faut noter également que le filtrage de ce bruit dans les textes entraîne un recalcul de la classification ;
- Les classifications obtenues ne sont pas incrémentales au sens où l'ajout d'un texte dans la base de textes peut faire apparaître une nouvelle classe et/ou faire disparaître une autre classe par la fusion de deux classes de l'étape précédente.

De notre point de vue, les techniques que nous présentons pour la classification de textes servent de première étape au processus de FdT. En effet, nous considérons que la classification de textes constitue un prétraitement des données textuelles en vue de classer dans une catégorie tous des textes d'un corpus. Selon nous, pour compléter le processus de FdT, il faut activer un processus de FdD à proprement dit , de plus, il faut se préoccuper de valider et d'interpréter les éléments d'information extraits pour en faire des connaissances.

Le choix de l'utilisation des règles d'association dans notre processus de FdT est lié au fait que nous voulons fouiller dans les textes de façon non supervisée — sans imposer de contraintes *a priori* à notre processus mais en ayant un objectif de fouille défini par l'analyste. L'extraction des règles d'association est une technique de FdD qui a fait ses preuves pour la fouille dans de grandes masses de données. De plus, la facilité d'interprétation des règles d'association par un analyste qui connaît le domaine de fouille motive notre choix d'utiliser cette technique de FdD pour le processus de FdT. De ce fait, nous montrons, en § 3.3, que notre approche de FdT par l'extraction de règles d'association est utile pour cinq applications prenant en entrée des données textuelles : (i) la structuration d'une terminologie pour la construction de thésaurus, (ii) la structuration des connaissances d'un domaine par l'analyse de concepts formels, pour la construction et la main-

tenance d'ontologies du domaine, (iii) l'extraction d'information, (iv) la veille technologique et stratégique et (v) la recherche d'information.

3.1 Classification appliquée aux données textuelles

La classification des textes[11] fait partie du processus d'extraction d'éléments d'information dans des données textuelles. Les processus de classification des textes à mettre en place dépendent du du niveau d'analyse (lexical, syntaxique, sémantique, structurel, etc.) des textes. Par exemple :

- structurer des textes selon des thèmes communs,
- construire des ensembles homogènes de textes selon un ou plusieurs points de vue,
- rechercher des ensembles de paragraphes liés selon une mesure de similarité,
- rechercher des ensembles de termes cooccurrents dans un texte.

L'organisation des données textuelles peut s'appuyer sur une classification hiérarchique comme pour l'utilisation des arbres de décision ou les graphes conceptuels. Les données textuelles peuvent être organisées en classes non hiérarchisées comme lors de l'utilisation de modèles statistiques de langage. A.-H. Tan [Tan, 1999] oppose deux approches :

- Une approche fondée sur une organisation des textes entre eux – classification de documents textuels et leurs visualisations ;
- Une approche s'appuyant sur une organisation des concepts entre eux. Les concepts sont les unités d'information pertinentes décrites par les textes – construction de modèles prédictifs du domaine, découverte d'associations et leurs visualisations.

En revanche, l'examen des travaux de recherche actuels se réclamant de la FdT montre qu'une seule approche est majoritairement traitée. Il s'agit de la classification de documents textuels. L'engouement pour la classification de documents est dû à l'explosion des bases de données textuelles mises en ligne sur le Web.

En classification de documents textuels, l'analyse des textes se fait au niveau du thème général du texte et non au niveau du contenu. L'objectif visé est de répartir un ensemble donné de textes en catégories homogènes auxquelles nous pouvons affecter un thème. Par exemple, classer un ensemble de textes journalistiques dans des rubriques (*ex.*, politique, économie, sports, etc.) ou classer des documents dans les thèmes des annuaires thématiques Web (*ex.*, Yahoo !, Voilà, Infoseek, etc.).

Il existe deux grandes approches pour la classification de textes : l'approche supervisée et non supervisée. Nous présentons dans ce qui suit les deux approches ainsi que les méthodes utilisées pour la classification de textes.

3.1.1 Classification supervisée de textes

Les approches par classification supervisée de documents textuels sont présentées suivant l'importance des activités de recherche et des publications qui en découlent.

[11] En anglais : "Text classification", également désignée par "text categorization" ou par "topic spotting" dans [Sebastiani, 2003].

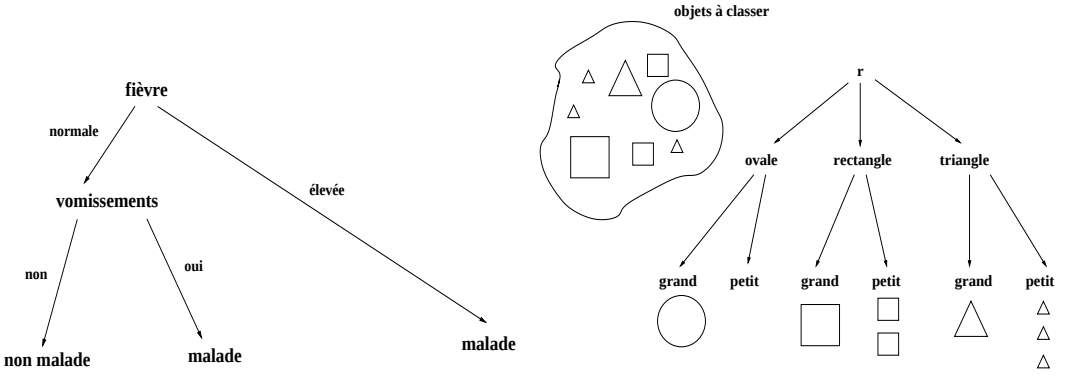


FIG. 3.1 – Exemples de classifications par arbres de décision : (a) binaire (à gauche) et (b) non binaire (à droite).

3.1.1.1 Arbres de décision

Un arbre de décision [Breiman *et al.*, 1984; Quinlan, 1986] est un modèle descriptif et prédictif d'un ensemble de données. Les arbres de décision sont utilisés dans le domaine de l'ECBD dite symbolique. C'est un sous-domaine de l'apprentissage par induction qui consiste à *apprendre* des fonctions de discrétisation, par classification et approximation, sur un ensemble d'objets (*ex.*, des personnes) décrits par des paires de propriétés (*ex.*, taille, poids, etc.) et de valeurs associées (*ex.*, grand, léger, etc.). Les arbres sont dits binaires lorsque chaque propriété prend deux valeurs possibles.

Caractéristiques des données à modéliser par un arbre de décision Soit une population $\mathcal{O} = \{\mathtt{o}_1, \ldots, \mathtt{o}_\mathtt{p}\}$ de $\mathtt{p}$ objets et $\mathcal{Y} = \{\mathtt{y}_1, \ldots, \mathtt{y}_\mathtt{n}\}$ un ensemble de $\mathtt{n}$ classes à expliquer et $\mathcal{X} = \{\mathtt{x}_1, \ldots, \mathtt{x}_\mathtt{m}\}$ un ensemble de $\mathtt{m}$ propriétés explicatives. Il s'agit de classer tous les objets $\mathtt{o}_\mathtt{i} \in \mathcal{O}$ dans l'une ou l'autre des classes de $\mathcal{Y}$, c'est-à-dire que :

$$\forall\, o_i \in \mathcal{O},\ \exists!\ j \in [1, n] \mid o_i \in y_j$$

et par la suite, il faut justifier pourquoi $\mathtt{o}_\mathtt{i} \in \mathtt{y}_\mathtt{j}$ grâce aux propriétés explicatives de $\mathcal{X}$.

Par exemple, on considère une classe booléenne « malade » qu'on veut expliquer, c'est-à-dire trouver un critère de décision d'appartenance à la classe grâce à un ensemble de propriétés telles que des symptômes « fièvre, vomissements, douleurs, etc. » que possèdent un individu. L'arbre de décision permet de déclarer qu'un individu est dans la classe `Malade` ou la classe `Non_Malade` malade. L'arbre de décision représente donc une fonction de classification booléenne dans le cas dit *simple*. La FIG. (3.1-a) représente le cas simple où :

$$\mathcal{X} = \{\mathtt{fièvre}, \mathtt{vomissements}\} \text{ et } \mathcal{Y} = \{\mathtt{malade}\}$$

Les fonctions booléennes peuvent être étendues à un nombre fini de classes > 2 pour traiter des propriétés de type numérique (entiers, réels, etc.) [Mitchell, 1997]. Si $\mathcal{X} = \{\mathtt{ovale}, \mathtt{rectangle}, \mathtt{triangle}\}$ et $\mathcal{Y} = \{\mathtt{petit}, \mathtt{grand}\}$, alors on obtient la classification présentée en FIG. (3.1-b). En outre, les arbres de décision sont adaptés aux données présentant les caractéristiques suivantes :

– les domaines des valeurs pour une propriété donnée sont considérés comme des disjonctions de valeurs. Si on veut obtenir une classification utile, il faut considérer qu'un objet, selon un point de vue donné, est dans une et une seule classe.
– les données peuvent contenir des propriétés à valeurs manquantes ou inconnues.

Classification dans un arbre de décision Un arbre de décision prend en entrée un *objet* décrit par son ensemble de *propriétés.* L'*intension* est l'ensemble des propriétés d'un objet. En sortie, l'algorithme de classification affecte cet objet à une et une seule classe d'objets. Les classes représentent les nœuds de l'arbre de décision. La classification est dite *incertaine* lorsque les valeurs de certaines propriétés sont non spécifiées aux nœuds-feuille de l'arbre de décision.

Les algorithmes de construction d'arbres de décision, dont les plus connus et utilisés sont `ID3` et `C4.5` [Quinlan, 1993], sont globalement décrits par l'Algorithme 1. Nous soulignons les techniques sous-jacentes qu'il faut mettre en œuvre. L'opération la plus importante et complexe est de se donner un critère de choix pour la formation des nœuds de l'arbre. Pour cela, il faut disposer d'une mesure permettant de calculer l'homogénéité des nœuds et le degré de séparation entre nœuds pour former des classes d'objets (*ex.*, le gain d'information [Salton, 1989]).

Algorithme 1: Algorithme de conctruction d'un arbre de décision

Entrée :
- un nœud racine `r`;
- un ensemble de nœuds à partir desquels sont issues au moins deux branches;
- un ensemble de nœuds terminaux qui permettent de classer les objets par rapport à la variable à expliquer `y`;

Sortie : segmenter la population d'objets en des classes les plus homogènes possibles et avec une séparation maximale entre les classes;

```
1 : pour chaque nœud courant n faire
    2 : /*À partir du nœud racine r*/;
    3 : établir l'ensemble des divisions admissibles;
    /* cet ensemble dépend de la propriété choisie et des modalités (i.e., des valeurs possibles)
    associées à cette propriété */;
    4 : si n n'est pas un nœud terminal alors
        */ pour se donner un critère de choix, il faut disposer d'une mesure permettant de calculer
        l'homogénéité des classes et le degré de séparation des classes. Nous utilisons, lors des
        divisions, des mesures telles que le gain d'information [Salton, 1989] */;
        5 : /*développer l'arbre*/;
        6 : construire les branches partant du nœud courant ;
        7 : choisir la « meilleure » division admissible;
        8 : pour chaque modalité de la variable faire
            9 : créer une branche;
    10 : sinon
        11 : arrêter le développement de cette branche et aller au nœud voisin de n;
        /* le nœud voisin est déterminé par une stratégie de parcours d'arbre (en largeur, en pro-
        fondeur, mixte, etc.) */
```

Représentation des arbres de décision Les arbres de décision sont représentés soit sous forme graphique (*cf.* FIG. 3.1), soit par les fonctions de classification décrites par des règles de décision

sous la forme de tests. Une règle de décision se présente sous la forme d'un test : « `Si Conditions alors Conclusions` ». La classification consiste à appliquer une règle de décision pour chaque étape de parcours de l'arbre. L'arbre en entier rassemble tous les chemins possibles. Les tests représentent, formellement, une disjonction de conjonctions de contraintes sur les valeurs des propriétés. Par exemple dans la figure (FIG. 3.1-a), on obtient la règle suivante pour un individu :

$$((\texttt{température} = \texttt{élevée}) \vee (\texttt{température} = \texttt{normale} \wedge \texttt{vomissements} = \texttt{oui}) \Longrightarrow \texttt{malade})$$

Il n'y a pas de possibilité de connaître *a priori* la taille de l'arbre (*i.e.* le nombre de nœuds, le nombre de niveaux et la largeur de l'arbre). Cependant, plus les propriétés explicatives sont indépendantes entre elles, plus la taille et la complexité de l'arbre de décision augmentent [Zhang et Zhang, 2002].

L'utilisation des arbres de décision pour le traitement des données textuelles se fait dans le domaine de la classification de textes. [Apté *et al.*, 1998] et [Johnson *et al.*, 2002] mènent, en parallèle, des travaux en classification de textes que ces auteurs assimilent à la fouille de textes, ce qui est un point de vue à considérer dans notre étude de la définition de FdT (*cf.* § (3.1.4)).

Nous pouvons imaginer les objets à classer de la figure (FIG. 3.1-b) comme des textes qui ont les caractéristiques de décrire ces objets. Un texte est : « je décris un objet en forme de cercle de grande taille ». En pratique, chaque texte est un objet représenté par un vecteur caractéristique dont les composantes sont des couples de termes-clés du texte (*i.e.* les propriétés de l'objet) associés à leurs fréquences $\langle(\texttt{cercle}, 1), (\texttt{grand}, 1)\rangle$. Comme l'ensemble des techniques de classification supervisée, il faut disposer de données d'apprentissage et de catégories prédéfinies par un analyste. Les catégories prédéfinies correspondent aux feuilles de l'arbre de décision. Le traitement consiste à apprendre des règles qui respectent cette classification. Tout nouveau texte est, par la suite, assigné à une seule des catégories prédéfinies. La catégorie *cible* d'un nouveau texte est celle dont la règle de décision est la plus proche du vecteur caractéristique du texte. La pertinence d'une telle classification est validée par les mesures de précision et de rappel (*cf.* § (3.1.3)).

3.1.1.2 Classification bayésienne naïve

La classification bayésienne simple ou naïve (Naive-Bayes) est une méthode de classification supervisée numérique dont le but est de rechercher la dépendance entre les propriétés par un calcul de probabilité conditionnelle. L'hypothèse naïve *a priori* du classifieur est que toutes les propriétés sont conditionnellement indépendantes. Le nom de classification bayésienne vient de l'utilisation du théorème fondamental naïf énoncé par T. Bayes :

Pour tout k :

$$P(B_k|A) = \frac{P(A|B_k)\ P(B_k)}{\sum_{n\geqslant 1} P(A|B_n)\ P(B_n)}$$

En supposant le principe des probabilités totales :

$$P(A) = \sum_{n\geqslant 1} P(A|B_n)\ P(B_n)$$

L'avantage de la formule de Bayes est que la difficulté de calculer des probabilités *a posteriori* de A sachant des événements $B_1, \ldots, B_n$ revient à des calculs de probabilités conditionnelles fixées *a priori* $A|B_1, \ldots, A|B_n$.

La classification bayésienne naïve est bien adaptée pour modéliser des relations *simples* entre propriétés des objets. [McCallum et Nigam, 1998; Mladenić, 1999] utilisent la classification bayésienne pour catégoriser des documents textuels.

Soient $\mathcal{D} = \{\mathtt{d_i},\ \mathtt{i} = 1 \ldots |\mathcal{D}|\}$ et $\mathcal{C} = \{\mathtt{C_k},\ \mathtt{k} = 1 \ldots |\mathcal{C}|\}$ deux ensembles, respectivement, de textes et de classes prédéfinies. Soit $\mathcal{V}$ un ensemble de vecteurs caractéristiques associés à chaque texte de $\mathcal{D}$. Soit $\mathtt{d_i}$ un texte dont $\mathtt{v_i} =< \mathtt{w_{i1}}, \ldots, \mathtt{w_{i|v_i|}} >$ est le vecteur caractéristique associé dans le modèle $\mathcal{M} = \langle \mathcal{D}, \mathcal{C}, \mathcal{V} \rangle$. L'utilisation de la classification bayésienne naïve consiste à associer à tout texte $\mathtt{d_i} \in \mathcal{D}$, une des classes $\mathtt{C_k} \in \mathcal{C}$. Le texte $\mathtt{d_i}$ est affecté à la classe $\mathtt{C_k}$ dont la probabilité conditionnelle par rapport à $\mathtt{v_i}$ est la plus élevée.

$$\mathtt{P}(\mathtt{C_k}|\mathtt{v_i}) = \mathtt{P}(\mathtt{C_k}) \times \frac{\mathtt{P}(\mathtt{v_i}|\mathtt{C_k})}{\mathtt{P}(\mathtt{v_i})} \tag{3.1}$$

L'hypothèse naïve du classifieur est fondée sur l'indépendance des occurrences des termes dans un texte :

$$\mathtt{P}(\mathtt{v_i}|\mathtt{C_k}) = \prod_{\mathtt{j}=1}^{|\mathtt{v_i}|} \mathtt{P}(\mathtt{w_{ij}}|\mathtt{C_k}) \tag{3.2}$$

Nous obtenons la formule 3.3 après avoir simplifié la formule 3.1 naïve de Bayes en ignorant $\mathtt{P}(\mathtt{v_i})$ qui ne dépend pas de la classe $\mathtt{C_k}$. En effet, $\mathtt{P}(\mathtt{v_i})$ ne sert qu'à normaliser le résultat puisque $\sum_{\mathtt{i}=1}^{|\mathcal{D}|} \mathtt{P}(\mathtt{v_i}) = \sum_{\mathtt{i}=1}^{|\mathcal{D}|} \sum_{\mathtt{k}=1}^{|\mathcal{C}|} \mathtt{P}(\mathtt{C_k}|\mathtt{v_i}) = 1.$

$$\mathtt{P}(\mathtt{C_k}|\mathtt{v_i}) \approx \mathtt{P}(\mathtt{C_k}) \times \prod_{\mathtt{j}=1}^{|\mathtt{v_i}|} \mathtt{P}(\mathtt{w_{ij}}|\mathtt{C_k}) \tag{3.3}$$

Il faut donc affecter $\mathtt{d_i}$ à la classe $\mathtt{C_{cible}}$ la plus probable :

$$\mathtt{C_{cible}} = \arg\max_{\mathtt{C_k}}\ \mathtt{P}(\mathtt{C_k}|\mathtt{v_i}),\ \mathtt{k} = 1, \ldots, |\mathcal{C}| \tag{3.4}$$

3.1.1.3 Modèles statistiques du langage

Nous présentons deux modèles fondés sur une analyse statistique du langage pour la classification de textes.

Classifieur TF/IDF Le classifieur TF/IDF, proposé par Salton [Salton, 1989], représente chacune des $\mathtt{n}$ classes prédéfinies $\mathcal{C} = \{\mathtt{C_1}, \ldots, \mathtt{C_n}\}$ par un vecteur caractéristique constitué d'un ensemble de termes-clés prédéfinis. L'ensemble des termes-clés constitue un vocabulaire $\mathtt{V}$. La dimension du vecteur caractéristique est fixée à $|\mathtt{V}|$, le cardinal du vocabulaire. Un terme-clé peut appartenir à une ou plusieurs classes $\mathcal{C}$. Une classe peut avoir une composante nulle si le terme correspondant est absent.

Un texte est également représenté par un vecteur caractéristique de même dimension. Les composantes des vecteurs caractéristiques combinent deux valeurs :

- le nombre d'occurrences du terme $\mathtt{t_i}$ dans la classe $\mathtt{C_j}$, noté $\mathtt{TF}(\mathtt{t_i}, \mathtt{C_j})$,
- le nombre de classes où le terme $\mathtt{t_i}$ apparaît au moins une fois, noté $\mathtt{DF}(\mathtt{t_i})$.

Dans la pratique l'inverse de cette valeur est utilisée :

$$\mathtt{IDF}(\mathtt{t_i}) = \mathtt{log_2}\left(\frac{|\mathtt{n}|}{\mathtt{DF}(\mathtt{t_i})}\right)$$

$\mathtt{IDF}(\mathtt{t_i})$ aura une valeur élevée pour un terme très caractéristique de peu de classes. Inversement, cette valeur est faible si le terme est présent dans beaucoup de classes, par exemple, si le terme est uniformément distribué dans les classes. Le terme répandu dans les classes n'est pas typique d'une classe et n'est pas discriminant entre les classes. La valeur $\mathtt{d_{ij}}$ du produit $(\mathtt{TF}(\mathtt{t_i}, \mathtt{C_j}) \times \mathtt{IDF}(\mathtt{t_i}))$[12] correspond au poids du terme $\mathtt{t_i}$ dans la classe $\mathtt{C_j}$. De même, le poids du terme $\mathtt{t_i}$ pour un texte $\mathtt{d_k}$ est mesurée par TF. Le pouvoir de discrimination de ce terme est mesuré par IDF. Ainsi, un terme qui a une valeur de TF/IDF élevée doit être à la fois important dans ce texte, et doit apparaître peu dans les autres textes.

Une mesure de similarité est ensuite calculée entre un texte $\mathtt{d_k}$ et une classe $\mathtt{C_j}$ à travers les composantes de leurs vecteurs caractéristiques respectifs :

$$\mathtt{sim}(\mathtt{d_k}, \mathtt{C_j}) = \frac{\sum_{\mathtt{i}=1}^{|\mathtt{V}|} \mathtt{d_{ki}} \mathtt{d_{ij}}}{\sqrt{\sum_{\mathtt{i}=1}^{|\mathtt{V}|} (\mathtt{d_{ki}})^2 \sum_{\mathtt{i}=1}^{|\mathtt{V}|} (\mathtt{d_{ij}})^2}} \tag{3.5}$$

L'équation (3.5) est appelée mesure `cosinus` car elle mesure le cosinus de l'angle formé par le vecteur du texte (*i.e.* les composantes $\mathtt{d_{ki}}$) et le vecteur de la classe (*i.e.* les composantes $\mathtt{d_{ij}}$) du numérateur. Ces deux vecteurs sont normalisés par le dénominateur de cette équation pour ne pas favoriser la classe dont le vecteur caractéristique est générique et aura tendance à attirer tous les textes à classer.

Nous affectons le texte $\mathtt{d_k}$ à la classe $\mathtt{C_i} \in \mathcal{C}$ qui maximise la mesure de similarité de la formule (3.5). Une valeur de `1` signifie que les deux vecteurs sont identiques (le texte appartient fortement à cette classe), une valeur de `0` signifie que la classe et le texte ne sont pas liés. Des classes de documents utilisant une mesure sur l'équation (3.5) pour la détection de thèmes sont également créées par le système TOPCAT (Topic Categorization) de [Clifton et Cooley, 1999].

Support Vector Machine (SVM**)** Les Support Vector Machines (SVMs[13]) ont été introduites par V. Vapnik [Vapnik, 1995]. Cette méthode consiste à trouver un hyperplan séparateur entre des ensembles de textes en deux classes. Les textes sont représentés dans le cas d'un plan euclidien par un point $\mathtt{p} = (\mathtt{x}, \mathtt{y})$. Les coordonnées d'un texte sont celles du modèle vectoriel de Salton [Salton, 1989] formé de couples de mots et leurs fréquences associées.

La figure (3.2-a) représente le cas simple de deux classes dans un espace de dimension `2`. Pour reprendre l'exemple figure (FIG. 3.1-b), les abcisses représentent la fréquence d'apparition du terme « petit » et les ordonnées celle du terme « cercle ». le texte `d` a pour coordonnées $\langle 1, 1 \rangle$, les termes petit et cercle apparaissent une seule fois. En phase d'apprentissage, les textes en points blancs sont classés dans la $\mathtt{catégorie_1}$, les textes en points noirs dans la $\mathtt{catégorie_2}$; il s'agit de trouver une droite séparatrice qui minimise les exemples qu'on ne sait pas classer (*cf.* les points en gris dans la figure (3.2-b)). Il existe une infinité de droites pour séparer les deux catégories.

[12] La pondération des termes TF/IDF est utilisée également en recherche d'information. TF signifie « Term Frequency » et IDF signifie « Inverted Document Frequency ».

[13] L'une des traductions convenables en français que nous avons trouvée pour les SVMs est « machines à vecteurs supports » dans [Canu, 2002].

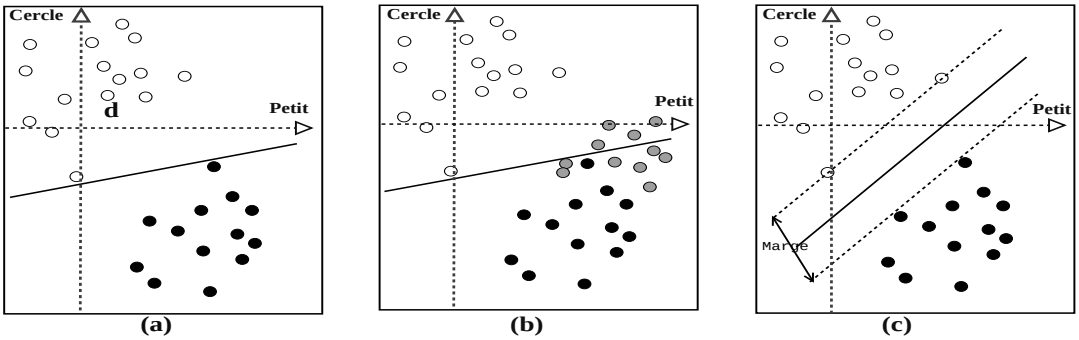


FIG. 3.2 – Exemples de classifications par une SVM.

[Joachims, 1998] propose de chercher une droite séparatrice et de prendre en compte une marge sur la droite séparatrice qui minimise les textes mal classés (*cf.* la figure (3.2-c)). Un nouveau texte est classé dans une des deux catégories selon sa situation par rapport à la droite séparatrice trouvée en phase d'apprentissage par la SVM. Selon [Dumais *et al.*, 1998], pour la classification de textes, cette méthode s'est avérée efficace dans son cas.

3.1.2 Classification non supervisée de textes

La classification non supervisée de textes est caractérisée par l'absence de catégories prédéfinies et de l'utilisation d'un corpus d'apprentissage. Nous calculons sur tout le corpus une classification sans savoir *a priori* le nombre de catégories produites. Nous présentons la classification non supervisée de textes par deux méthodes : les réseaux bayésiens et l'utilisation des graphes conceptuels.

3.1.2.1 Réseaux bayésiens

Une variante de la classification bayésienne naïve est la définition d'un réseau bayésien [Pearl, 1988; Jensen, 1996]. Un réseau bayésien est un graphe probabiliste acyclique orienté. Dans le modèle des réseaux bayésiens, on ne suppose pas que les propriétés sont toutes indépendantes, certaines sont donc liées. Un texte est représenté par un réseau bayésien qui reflète la structure du texte (introduction, section, paragraphe, conclusion, etc.). Un nœud du réseau est donc une partie d'un texte. La transition d'un nœud du réseau vers ses nœuds voisins montre le lien entre les parties d'un texte exprimée par une probabilité conditionnelle. Le corpus entier constitue une forêt de réseaux et reflète la structure du corpus. Un mécanisme d'unification de graphes permet de faire de la recherche documentaire. Il s'agit de trouver les réseaux bayésiens (textes) les plus proches d'une requête modélisée également par un réseau bayésien [Piwowarski *et al.*, 2002]. La technique d'unification permet également de classer des textes en catégories, de la même manière que pour la recherche documentaire, en s'appuyant sur la proximité structurelle entre textes. Cependant, l'expérience montre que la classification en utilisant les réseaux bayésiens alourdit considérablement les calculs et les résultats n'augmentent pas de façon significative [Denoue, 2000]. Le moteur de recherche INQUERY [Callan *et al.*, 1992] utilise la classification en construisant un réseau bayésien pour la recherche documentaire sur le Web.

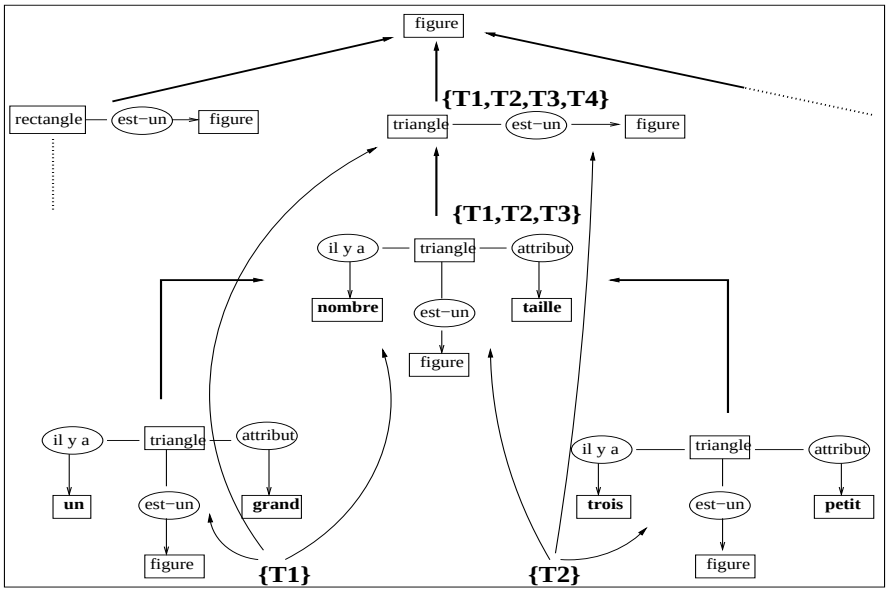


FIG. 3.3 – Exemple de classification de textes par une hiérarchie de graphes conceptuels.

3.1.2.2 Formalisme des graphes conceptuels

L'utilisation des graphes conceptuels [Sowa, 2002] est le seul formalisme qui permet de faire de la classification de textes à partir d'une représentation du texte plus complexe que les méthodes précédentes. La représentation de chaque phrase d'un texte est une structure de graphe de concepts reliés entre eux. Cette représentation est différente des représentations du texte, dans les méthodes précédentes, par un ensemble de termes-clés présents dans les textes et leurs fréquences d'apparition.

Le travail de Montes-y-Gómez [Montes_y_Gómez *et al.*, 2002] sur la classification de textes décrit chaque texte par un ensemble de graphes de concepts construits après l'analyse des phrases, des paragraphes ou du texte entier. Les graphes conceptuels issus des textes sont groupés selon les similarités des concepts ou les similarités des relations entre les concepts. Les travaux antérieurs de [Mechkour *et al.*, 1995], similaires à celui de Montes-y-Gómez, ont été développés pour la classification d'images, décrites sous forme textuelles, en utilisant le formalisme des graphes conceptuels.

Reprenons l'exemple d'un ensemble de textes décrivant une partie des objets de la figure (FIG. 3.1-b) (*i.e.* les triangles) :

– T1 : il y a un grand triangle. Un triangle est une figure géométrique,
– T2 : il y a trois petits triangles. Un triangle est une figure géométrique,

Nous pouvons généraliser ces 2 textes en :

– T3 : il y a un nombre de triangles. Un triangle a une taille. Un triangle est une figure géométrique.

De même, on peut déduire un texte décrivant le sous-graphe :

– T4 : Un triangle est une figure géométrique.

Montes-y-Gómez, dans [Montes_y_Gómez *et al.*, 2001], présente un algorithme d'unification de graphes et des mesures de similarité entre graphes (pour la généralisation de graphes et la re-

cherche de sous-graphes communs). Le traitement consiste à construire une hiérarchie de graphes conceptuels s'appuyant sur des similarités et des régularités de structures des graphes conceptuels issus de l'analyse des textes. Les textes sont rattachés à certains concepts du graphe. Ce qui permet de structurer le corpus en un nombre de classes de textes. Une classe de la hiérarchie hérite des textes de ses sous-classes (*cf.* FIG. 3.3). Néanmoins, les résultats effectifs d'une telle démarche de classification ne sont pas donnés dans les travaux de Montes-y-Gómez. La méthode de prétraitement pour construire de façon automatique des graphes conceptuels à partir du contenu des textes n'est, également, pas explicite. Nous pensons qu'une méthode orientée TAL fondée sur le repérage de marqueurs linguistiques – ceux qui nomment les relations : « il y a », « est-un », etc. – est nécessaire. Dans ce cas, un ensemble exhaustif de marqueurs lié au domaine des textes doit être établi. Il n'est pas reproductible sur un autre corpus.

3.1.3 Mesures de qualité d'une classification de textes

La classification supervisée mesure l'importance de chaque terme pour classer de nouveaux textes. Par exemple, une mesure venant de la théorie de l'information de C. Shannon (« information gain ») fondée sur un calcul d'entropie mesure la typicité d'un terme. Plus un mot est lié à une catégorie et pas aux autres, plus il est important : si un nouveau texte le contient, ce terme sera plus discriminant. D'autres mesures semblables ont été mises au point. Parmi les plus utilisées en recherche d'information en général et en classification de textes en particulier, nous présentons les mesures de *précision* et de *rappel*[14].

Soit $\mathtt{S}$ l'ensemble des textes classés comme ayant la propriété recherchée (*i.e.* textes dans la catégorie considérée) ; soit $\mathtt{V}$ l'ensemble de tous les textes qui possèdent effectivement cette propriété (*i.e.* textes pertinents). $|\mathtt{X}|$ est le cardinal d'un ensemble $\mathtt{X}$.

Mesure de précision La précision $\mathtt{P}$ est une mesure de pertinence de la classification. La précision est le rapport du nombre de textes pertinents et bien classés dans une catégorie au nombre total de textes classés dans cette catégorie. Autrement dit, le nombre de réponses correctes sur le nombre de réponses fournies.

$$\mathtt{P} = \frac{|\mathtt{S} \cap \mathtt{V}|}{|\mathtt{S}|} \tag{3.6}$$

Mesure de rappel Le rappel $\mathtt{R}$ est une mesure de couverture de la classification. Le rappel est le rapport du nombre de textes pertinents et bien classés dans une catégorie au nombre total de textes pertinents de cette catégorie. Autrement dit, le nombre de réponses correctes sur le nombre de réponses attendues.

$$\mathtt{R} = \frac{|\mathtt{S} \cap \mathtt{V}|}{|\mathtt{V}|} \tag{3.7}$$

Une précision de 100% signifie donc que tous les textes rapportés sont pertinents, un rappel de 100% que tous les textes pertinents ont été trouvés — l'ensemble vide a une précision de 100%, l'ensemble de tous les textes a un rappel de 100%.

Il est plus simple d'évaluer les résultats d'une classification de textes supervisée que non supervisée. $\mathtt{V}$ n'est connu en extension qu'en classification supervisée. En effet, parmi les $\mathtt{N}$ exemples de

[14] Ces deux mesures sont désignées en anglais par : « precision » et « recall ». Ces mesures sont également utilisées en recherche d'information [Salton, 1989].

textes classés, on utilise une partie des textes pour l'entraînement, et le reste pour le test. Pendant la phase de test, on soumet chaque texte à l'algorithme de classification supervisée et on vérifie qu'on trouve la bonne classe.

3.1.4 Bilan de la classification appliquée aux données textuelles

Les méthodes présentées ci-dessus, qu'elles soient supervisées ou non, ont pour but d'apprendre une structuration du corpus de textes en classes. En classification supervisée, on cherche à apprendre un critère de classification à partir d'exemples positifs et négatifs et on évalue la qualité du classifieur avec un ensemble de nouveaux textes à classer. En classification non supervisée, il s'agit plutôt de justifier une classification fondée sur des similarités apprises en analysant l'ensemble des données textuelles. Selon [Dumais *et al.*, 1998], pour la classification de textes, les études comparatives faites sur les classifieurs issus des différentes méthodes, les SVMs apparaissent comme les plus performantes. Lorsqu'on a des textes à valeurs de propriétés manquantes ou incertaines, les arbres de décision paraissent utiles. La méthode s'appuyant sur l'utilisation des graphes conceptuels est l'unique formalisme qui représente les textes par une structure autre qu'un ensemble de termes-clés.

Nous présentons, en § 3.2, les règles d'association définies de façon formelle. L'extraction de règles d'association constituent l'étape d'activation d'une technique de FdD dans notre processus de FdT.

3.2 Extraction de règles d'association pour la FdT

L'extraction de règles d'association est une méthode assez répandue en fouille de données, même si elle est peut-être moins courante que les méthodes de classification. Une telle méthode vise à extraire d'un corpus de textes des liens entre les termes caractérisant les textes. Ces liens sont exprimés à travers des règles du type $\mathtt{A} \Longrightarrow \mathtt{B}$ et ne sont donc pas précisément identifiés (ils ne correspondent pas à une relation sémantique du domaine). L'extraction de règles d'association est la méthode de fouille de données sur laquelle nous nous appuyons durant toute la suite de notre mémoire. Même si les avantages et inconvénients de cette méthode apparaîtront assez clairement dans les chapitres suivants, nous pouvons lister les points qui nous ont motivé pour choisir cette méthode :

(1) - Une règle d'association est facile à lire. La lecture intuitive d'une telle règle est : « quand un texte possède A, il y a de grandes chances qu'il possède B » ;

(2) - Une règle d'association est généralement composée de peu de termes. Cela facilite le travail d'interprétation de l'analyste qui, par rapport à ses connaissances, cherche à relier seulement quelques notions pour chaque règle (c'est-à-dire, par exemple, identifier la relation entre les termes présents par rapport à son domaine de spécialité). Ceci est à opposer à des classes volumineuses issues d'une classification de textes contenant des relations diverses entre les termes ;

(3) - Les règles d'association peuvent être pondérées par une mesure de validité appelée *confiance*. Si la règle n'est pas valide, elle sera pondérée par la confiance et interprétée par : « Dans $\mathtt{x}\%$ des cas, les textes qui possèdent A possèdent B ». Cependant chaque règle traite de propriétés symboliques booléennes. Cela s'oppose à certaines méthodes de classification où les classes

sont constituées de propriétés pondérées par une valeur qu'il est parfois difficile d'interpréter. L'utilisation de propriétés symboliques booléennes constitue pour nous un avantage puisqu'il est possible de confronter ces règles à un modèle de connaissances. C'est un point fort de nos travaux de thèse que nous proposons dans le chapitre § 5.

Dans un premier temps, nous proposons de définir le cadre formel d'extraction des règles d'association et des méthodes qui en découlent pour les extraire. Nous définissons une règle d'association ainsi que les éléments constitutifs d'une règle : les *motifs fréquents* obtenus grâce aux *motifs fermés* et aux *motifs générateurs*. Enfin, nous présentons l'algorithme d'extraction des motifs fréquents – Close [Pasquier *et al.*, 1999b] – et l'algorithme de construction des règles d'association informatives [Bastide, 2000].

3.2.1 Définition d'une règle d'association

Une règle d'association (RA) est une règle d'implication conditionnelle permettant de trouver des corrélations entre des éléments qui sont liés par une relation $\mathcal{R}$. Dans le contexte de la FdT, les règles d'association sont interprétées comme une cooccurrence de termes impliquant la présence d'autres termes dans les textes en accord avec la définition usuelle en sémantique lexicale [Anick et Pustejovsky, 1990] (*cf.* § 2.3.4).

Les règles d'association ont été initialement étudiées en analyse de données [Guigues et Duquenne, 1986; Luxenburger, 1991], puis en fouille de données afin de trouver des régularités, des corrélations dans des bases de données de grandes tailles. Par la suite, les règles d'association ont été appliquées à la fouille de textes afin d'apprendre des relations de corrélations entre des éléments textuels, par exemple des termes constituant des *mots-clés* d'un texte [Feldman *et al.*, 1998; Kodratoff, 1999] ou les travaux plus récents de [Delgado *et al.*, 2002].

Définition 3.1 (Règle d'association) *Une règle d'association* $\mathtt{R} : \mathtt{B} \Longrightarrow \mathtt{H}$ *est constituée d'un ensemble de termes* B *(prémisse) impliquant un ensemble de termes* H *(conséquent)*[15]*. Une règle d'association est notée :*

$$\mathtt{R} : \mathtt{t}_1 \sqcap \ldots \sqcap \mathtt{t}_k \Longrightarrow \mathtt{t}_{k+1} \sqcap \ldots \sqcap \mathtt{t}_n$$

où $\{\mathtt{t}_1, \ldots, \mathtt{t}_k\}$ *et* $\{\mathtt{t}_{k+1}, \ldots, \mathtt{t}_n\}$ *sont deux ensembles non vides de* termes *et l'opérateur* $\sqcap$ *exprime la présence simultanée des ensembles de termes de la règle.*

Les règles d'association sont définies dans une forme particulière dans [Agrawal *et al.*, 1993] –une seule propriété en conséquent de la règle. La forme généralisée de la définition 3.1 est introduite dans [Agrawal et Srikant, 1994]. L'interprétation intuitive de la règle R en FdT est : si des textes contiennent les termes $\mathtt{t}_1$ et $\mathtt{t}_2$ ...et $\mathtt{t}_k$ alors ces textes ont tendance à contenir également, avec une probabilité P, les termes $\mathtt{t}_{k+1}$ et $\mathtt{t}_{k+2}$ ...et $\mathtt{t}_n$. L'utilisation du signe « $\Longrightarrow$ » est un abus de notation puisqu'il ne s'agit pas de l'implication logique classique (vraie/fausse) mais d'une implication particulière qui est vraie avec une probabilité P. Nous pouvons donc noter une règle par : « $\mathtt{R} : \mathtt{B} \overset{\mathtt{P}}{\Longrightarrow} \mathtt{H}$ ». $\mathtt{B} \sqcap \mathtt{H}$ est appelé *motif*, et dénote l'ensemble des termes $\mathtt{t}_i$ $(\mathtt{pour\ i} \in \{1, \ldots, n\})$.

[15] $\mathtt{B} = \mathtt{Body}$ et $\mathtt{H} = \mathtt{Head}$. Ces dénominations font référence à la méthode de la *résolution* de clauses en programmation logique (H :- B).

Une règle d'association se construit à partir du motif $\mathtt{B} \sqcap \mathtt{H}$ à condition que ce motif soit *fréquent.* La recherche de motifs fréquents est une étape préalable sur laquelle s'appuie la construction de règles d'association.

3.2.2 Définition d'un motif fréquent

Nous reprenons certaines définitions et notations utilisées dans [Bastide, 2000] et dans [Pasquier *et al.*, 1999b] pour la définition des motifs fréquents.

Soit un ensemble fini $\mathcal{T} = \{\mathtt{t}_1, \mathtt{t}_2, \ldots, \mathtt{t}_n\}$ de termes caractérisant un ensemble fini de textes $\mathcal{D} = \{\mathtt{d}_1, \mathtt{d}_2, \ldots, \mathtt{d}_m\}$[16]. $\mathcal{T}$ et $\mathcal{D}$ sont liés par une relation binaire $\mathcal{R} \subseteq \mathcal{T} \times \mathcal{D}$ dans un contexte $\mathcal{C} = < \mathcal{T}, \mathcal{D}, \mathcal{R} >$.

Exemple : Soit $\mathcal{T}$ et $\mathcal{D}$ deux ensembles.

Ensemble de termes :
$\mathcal{T} = \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}, \mathtt{e}\}$

Ensemble de textes :
$\mathcal{D} = \{\mathtt{d}_1, \mathtt{d}_2, \mathtt{d}_3, \mathtt{d}_4, \mathtt{d}_5, \mathtt{d}_6\}$

$\mathcal{R}$	$\mathtt{d}_1$	$\mathtt{d}_2$	$\mathtt{d}_3$	$\mathtt{d}_4$	$\mathtt{d}_5$	$\mathtt{d}_6$
a	×		×·		×·	
b		×	×·	×	×·	×
c	×	×	×·		×·	×
d	×					
e		×	×·	×	×·	×

Définition 3.2 (Motif) *Un motif est un sous-ensemble de $\mathcal{T}$. On dit qu'un motif (ou un* itemset T) est inclus dans un texte $\mathtt{d}_i$ (ou que $\mathtt{d}_i$ contient T) si T et $\mathtt{d}_i$ sont en relation par $\mathcal{R}$, c'est-à-dire que $\forall \mathtt{t} \in \mathtt{T}, (\mathtt{t}, \mathtt{d}_i) \in \mathcal{R}$. Un motif de taille k est appelé *k-motif.* Par exemple, $\mathtt{d}_3$ et $\mathtt{d}_5$ contiennent le 4-motif $\{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{e}\}$[17].

Définition 3.3 (Image d'un motif) *Soit* f *la fonction qui fait correspondre à un motif* T *l'ensemble des textes (objets) qui le contiennent :* $\mathtt{f}(\mathtt{T}) = \{\mathtt{d}_i \in \mathcal{D} \mid \mathtt{d}_i \ \mathtt{contient}\ \mathtt{T}\}$. *L'image d'un motif est appelée l'*extension *du motif. Par exemple, l'extension du 4-motif* $\{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{e}\}$ *est l'ensemble de textes* $\mathtt{f}(\{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{e}\}) = \{\mathtt{d}_3, \mathtt{d}_5\}$.

Définition 3.4 (Support d'un motif) *Le support d'un motif* T *est défini par la fréquence d'apparition du motif dans l'ensemble $\mathcal{D}$. Le support est égal au cardinal de l'image de* T, i.e. $\mathtt{support}(\mathtt{T}) = |\mathtt{f}(\mathtt{T})|$, *lorsqu'il est exprimé en absolu. Sinon, le support est égal à la proportion d'apparition du motif dans l'ensemble de textes $\mathcal{D}$,* i.e. $\mathtt{support}(\mathtt{T}) = \dfrac{|\mathtt{f}(\mathtt{T})|}{|\mathcal{D}|}$.

Définition 3.5 (Motif fréquent) *Un motif* T *est dit fréquent s'il apparaît un nombre de fois supérieur à un seuil de support dans l'ensemble de textes $\mathcal{D}$,* i.e. $\mathtt{support}(\mathtt{T}) \geqslant \mathtt{minsup}$ *où* minsup *est un* seuil *(ou support minimal). Par exemple, si* $\mathtt{minsup} = 3$[18] *alors le motif* $\{\mathtt{a}, \mathtt{c}, \mathtt{e}\}$ *n'est pas fréquent (*i.e. *non fréquent car* $|\{\mathtt{d}_3, \mathtt{d}_5\}| = 2$*). Nous notons l'ensemble des motifs fréquents* $\mathcal{L} = \{\mathtt{T} \subseteq \mathcal{T} \mid \mathtt{support}(\mathtt{T}) \geqslant \mathtt{minsup}\}$

[16] Les noms items et objets sont classiquement utilisés en fouille de données pour désigner respectivement les termes et les textes dans notre contexte de FdT.

[17] Le 4-motif $\{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{e}\}$ est différencié par $(\times\cdot)$ dans les cases correspondantes du tableau de la relation $\mathcal{R}$.

[18] $\mathtt{minsup} = \frac{3}{6} = 0,5$, nous parlons alors de support minimal à 50%.

3.2.3 Extraction de règles d'association

Pour extraire une règle d'association, nous nous appuyons sur deux mesures : le *support* et la *confiance*. Les deux mesures permettent à l'analyste de définir deux seuils pour la génération des règles d'association « valides ». Le seuil `minsup` est le support minimum que nous utilisons également pour déterminer qu'un motif est fréquent (*cf.* la définition 3.5). La valeur `minconf` permet de choisir le seuil minimal de confiance accordé à la règle d'association pour être considérée comme valide[19].

Mesures de support et de confiance. La mesure de **support** de $\mathtt{B} \Longrightarrow \mathtt{H}$ dénote le cardinal de l'intersection des images des motifs B et H. Nous définissons donc le support d'une règle comme le support du motif $\mathtt{B} \sqcap \mathtt{H}$.

$$|\mathtt{f}(\mathtt{B} \sqcap \mathtt{H})| = |\mathtt{f}(\mathtt{B}) \cap \mathtt{f}(\mathtt{H})| \tag{3.8}$$

En termes de motifs, $\mathtt{B} \sqcap \mathtt{H}$ représente le motif concaténant l'ensemble des termes de B et de H, c'est-à-dire $\{\mathtt{t}_1, \ldots, \mathtt{t}_n\}$ qui est l'ensemble des termes $\mathtt{t}_i$, pour $\mathtt{i} \in \{1, 2, \ldots, n\}$ qui doivent apparaître simultanément (*i.e.* cooccurrer) dans les textes. Le nombre de ces textes sert de support à la règle $\mathtt{B} \Longrightarrow \mathtt{H}$. Par conséquent, $\mathtt{support}(\mathtt{B} \sqcap \mathtt{H}) = \mathtt{support}(\mathtt{B\ et\ H})$, c'est-à-dire le nombre de textes du corpus qui ont contribué à l'extraction de la règle et qui contiennent tous les termes $\{\mathtt{t}_1, \ldots, \mathtt{t}_k\}$ de B « et » tous les termes $\{\mathtt{t}_{k+1}, \ldots, \mathtt{t}_n\}$ de H conformément à l'équation (3.8).

En pratique, le support d'une règle représente le rapport entre le cardinal de l'ensemble des textes décrits par le motif $\mathtt{B} \sqcap \mathtt{H}$ et le cardinal de l'ensemble des textes du corpus $\mathcal{D}$.

$$\mathtt{support}(\mathtt{R}) = \mathtt{support}(\mathtt{B} \sqcap \mathtt{H}) = \frac{|\mathtt{f}(\mathtt{B} \sqcap \mathtt{H})|}{|\mathcal{D}|} \tag{3.9}$$

Dans les travaux [Kuntz *et al.*, 2000; Guillet, 2004], le support est défini en termes de fréquence : $\mathtt{support}(\mathtt{B} \Longrightarrow \mathtt{H}) = \mathtt{freq}(\mathtt{B} \cup \mathtt{H})$. En ce qui nous concerne, $\mathtt{freq}(\mathtt{B} \cup \mathtt{H})$ représente ce que nous définissons par $|\mathtt{f}(\mathtt{B} \sqcap \mathtt{H})|$.
La mesure de support est symétrique, c'est-à-dire que $\mathtt{support}(\mathtt{B} \Longrightarrow \mathtt{H}) = \mathtt{support}(\mathtt{H} \Longrightarrow \mathtt{B})$. La mesure de support est décroissante par rapport à la taille du motif, *i.e.* $\mathtt{support}(\mathtt{B} \sqcap \mathtt{H}) \leqslant \mathtt{support}(\mathtt{B})$ et $\mathtt{support}(\mathtt{B} \sqcap \mathtt{H}) \leqslant \mathtt{support}(\mathtt{H})$.

La mesure de **confiance** de R est définie par :

$$\mathtt{confiance}(\mathtt{R}) = \frac{\mathtt{support}(\mathtt{B} \sqcap \mathtt{H})}{\mathtt{support}(\mathtt{B})} \in [0, 1] \tag{3.10}$$

Si nous considérons le motif $\mathtt{B} \sqcap \mathtt{H}$ comme un *événement* ayant une certaine probabilité, alors ce motif sera considéré comme l'occurrence simultanée de tous les événements élémentaires de chacun des termes composant le motif B et le motif H. En termes de probabilités, le motif $\mathtt{B} \sqcap \mathtt{H}$ sera dénoté par l'événement $\mathtt{B} \cap \mathtt{H}$. La confiance est donc représentée par la probabilité conditionnelle[20] :

$$\mathtt{confiance}(\mathtt{R}) = \frac{\mathtt{support}(\mathtt{B} \sqcap \mathtt{H})}{\mathtt{support}(\mathtt{B})} = \frac{\mathtt{P}(\mathtt{B} \cap \mathtt{H})}{\mathtt{P}(\mathtt{B})} = \mathtt{P}(\mathtt{H}|\mathtt{B}) \tag{3.11}$$

[19] Les valeurs `minsup` et `minconf` sont, respectivement, désignées en fouille de données par σ_s et σ_c.
[20] *Cf.* la justification présentée en annexe § B.2.

La confiance donne une mesure du pourcentage d'exemples de la règle. Le complémentaire à 1 mesure le pourcentage de contre-exemples. Un contre-exemple pour une règle d'association signifie qu'il y a des textes qui possèdent les termes de B mais pas nécessairement tous les termes de H. Lorsque la confiance vaut 1, la règle est *exacte* (ou totale). Elle s'exprime sous la forme d'une condition : « S'il pleut dehors, alors le sol sera mouillé ». Sinon la règle est dite *approximative* (ou partielle) et se voit attribuer une confiance variant entre 0 et 1. Par exemple : « Dans 85% des cas (*i.e.* avec une mesure de confiance de 0,85), les textes journalistiques qui parlent de grève de trains parlent également par relation causale de {bouchon, voiture, avion} ». Il est également intéressant d'observer que « Dans 15% des autres cas, les textes parlent de grève de trains mais pas de {bouchon, voiture, avion} en même temps ».

Étapes de construction des règles d'association La construction de règles d'association se décompose en deux étapes :

(a) déterminer l'ensemble des motifs fréquents ($\geqslant$ minsup) ainsi que leurs supports.

(b) générer, pour chaque motif fréquent, toutes les règles d'association dont la confiance est supérieure ou égale à minconf [Guillaume, 2000].

3.2.4 Formalisation mathématique

L'objet de cette section est de définir formellement les parties B et H d'une règle d'association. L'étape préalable à la construction de règles d'association est le calcul des motifs fréquents (*cf.* § (3.2.2)) du contexte formel $\mathcal{C} = \langle \mathcal{T}, \mathcal{D}, \mathcal{R} \rangle$. La recherche des motifs fréquents peut se faire par le calcul des motifs *fermés* (B $\cup$ H) fréquents et des motifs *générateurs* (B). Dans le contexte de FdT, la définition des motifs fermés et des motifs générateurs nous amène à définir deux fonctions f et g qui établissent une correspondance de Galois entre les textes (les objets) et les termes qui indexent les textes (leurs propriétés).

3.2.4.1 Correspondance de Galois

Nous définissons, de façon duale, deux relations binaires f et g, respectivement entre $\mathcal{T}$ et $\mathcal{D}$ dans le contexte $\mathcal{C}$ et entre $\mathcal{D}$ et $\mathcal{T}$, définis § 3.2.2, telles que :

$$\mathtt{f} : 2^{\mathcal{T}} \longrightarrow 2^{\mathcal{D}}$$
$$T \longmapsto \mathtt{f}(T) = \{d \in \mathcal{D} \mid \forall t \in T \ \ (t \mathcal{R} d)\}$$

$$\mathtt{g} : 2^{\mathcal{D}} \longrightarrow 2^{\mathcal{T}}$$
$$D \longmapsto \mathtt{g}(D) = \{t \in \mathcal{T} \mid \forall d \in D \ \ (t \mathcal{R} d)\}$$

$\mathtt{f}(T)$ sera aussi noté $Extension(T)$. $\mathtt{g}(D)$ sera également noté $Intension(D)$. Un élément $T \in 2^{\mathcal{T}}$ est un motif du contexte $\mathcal{C}$ (*cf.* FIG. 3.4).

Nous pouvons montrer que le couple de fonctions (f, g) constitue une correspondance de Galois entre les deux ordres partiels (*i.e.* les ensembles ordonnés partiellement par rapport à l'inclusion) $\left(2^{\mathcal{T}}, \subseteq\right)$ et $\left(2^{\mathcal{D}}, \subseteq\right)$[21], c'est-à-dire que :

[21] La notation $2^{\mathcal{T}}$ signifie $\mathcal{P}(\mathcal{T})$ l'ensemble des parties de $\mathcal{T}$; ce qui nous permet de noter que $T \in 2^{\mathcal{T}}$.

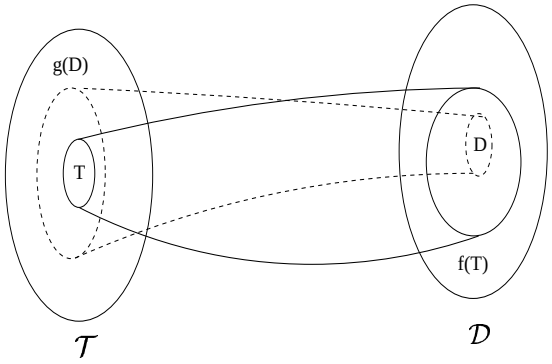


FIG. 3.4 – Illustration de la correspondance de Galois dans le contexte $\mathcal{C} = \langle \mathcal{T}, \mathcal{D}, \mathcal{R} \rangle$ de FdT.

$$\forall T \in 2^{\mathcal{T}}, \forall D \in 2^{\mathcal{D}} \quad T \subseteq \mathtt{g}(D) \Longleftrightarrow D \subseteq \mathtt{f}(T)$$

Propriétés de la correspondance de Galois

- `f` et `g` sont décroissantes :
  - $\forall T_1, T_2 \in 2^{\mathcal{T}}$
  $$T_1 \subseteq T_2 \Rightarrow f(T_1) \supseteq f(T_2)$$
  - $\forall D_1, D_2 \in 2^{\mathcal{D}}$
  $$D_1 \subseteq D_2 \Rightarrow g(D_1) \supseteq g(D_2)$$

Explication intuitive : Les textes de T_2 possèdent en commun tous les termes de $f(T_2)$. Lorsque nous retirons un ou plusieurs textes à T_2 pour obtenir T_1, les textes de T_1 auront toujours en commun, dans $f(T_1)$, les termes de $f(T_2)$, plus éventuellement d'autres (*i.e.* $f(T_2) \subseteq f(T_1)$). La fonction `f` est donc décroissante. De façon duale, la fonction `g` est également décroissante.

- leurs composées $h = \mathtt{g} \circ \mathtt{f}$ et $h' = \mathtt{f} \circ \mathtt{g}$ sont des opérateurs de fermeture, c'est-à-dire que :
  - $\forall T \in 2^{\mathcal{T}},\ T \subseteq h(T)$ et $\forall D \in 2^{\mathcal{D}},\ D \subseteq h'(D)$ (extensivité),
  - $\forall T \in 2^{\mathcal{T}},\ h \circ h(T) = h(T)$ et $\forall D \in 2^{\mathcal{D}},\ h' \circ h'(D) = h'(D)$ (idempotence).

3.2.4.2 Définitions d'un motif fermé fréquent et d'un motif générateur

Les propriétés de la correspondance de Galois ci-dessus nous permettent de définir un motif *fermé,* un motif *fermé fréquent* et un motif *générateur* :

Définition 3.6 (Motif fermé) `F` *est un motif « fermé » si et seulement si* $\mathtt{h}(\mathtt{F}) = \mathtt{F}$. *Pour calculer le motif fermé* `F`*, il faut calculer l'intersection des « intensions » de tous les textes qui possèdent ce motif en utilisant la formule :*

$$\mathtt{h}(\mathtt{F}) = \bigcap \mathtt{g}(\mathtt{d}) \ \mathtt{avec\ d\ tel\ que\ F} \subseteq \mathtt{g}(\mathtt{d}) \tag{3.12}$$

Dans le contexte $\mathcal{C}$ *(de l'exemple 3.2.2),* $\{\mathtt{b}, \mathtt{c}, \mathtt{e}\}$ *est un motif fermé, car c'est le motif maximal commun aux textes* $\{\mathtt{d}_2, \mathtt{d}_3, \mathtt{d}_5, \mathtt{d}_6\}$ *(*i.e. *avec* $\{\mathtt{b}, \mathtt{c}, \mathtt{e}\} \subseteq \mathtt{g}(\mathtt{d_i})$ `pour i` $\in \{2, 3, 5, 6\}$*).*

$$\mathtt{h}(\{\mathtt{b}, \mathtt{c}, \mathtt{e}\}) = \bigcap_{\mathtt{i} \in \{2,3,5,6\}} \mathtt{g}(\mathtt{d_i}) = \{\mathtt{b}, \mathtt{c}, \mathtt{e}\} \cap \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{e}\} \cap \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{e}\} \cap \{\mathtt{b}, \mathtt{c}, \mathtt{e}\} \cap \{\mathtt{b}, \mathtt{c}, \mathtt{e}\} = \{\mathtt{b}, \mathtt{c}, \mathtt{e}\}$$

Définition 3.7 (Motif fermé fréquent) FF *est un motif fermé* fréquent *si et seulement si* FF *est un motif fermé et* FF *est un motif fréquent pour un seuil donné* minsup*.*

Propriétés 3.1

(a) Soit un couple de fonctions (f, g) constituant une correspondance de Galois, alors $f \circ g \circ f = f \circ h = f$ et $g \circ f \circ g = g \circ h' = g$;

(b) La fermeture d'un motif est le plus petit fermé contenant ce motif ;

(c) Le support d'un motif est égal au support de sa fermeture ;

(d) Tout sous-motif *d'un motif fermé fréquent est un motif fréquent ; mais n'est pas forcément un motif fermé ;*

(e) Tout super-motif *d'un motif fermé non fréquent est un motif non fréquent.*

Démonstrations 3.1

(a) Soit T un motif,
- *$f \circ g \circ f(T) \subseteq f(T)$?*
  $T \subseteq h(T)$ car h est extensive
  $f(T) \supseteq f \circ h(T)$ car f est décroissante
  $f \circ g \circ f(T) \subseteq f(T)$ ■
- *$f \circ g \circ f(T) \supseteq f(T)$?*
  $f(T) \subseteq h' \circ f(T)$ car h' est extensive
  $f(T) \subseteq f \circ g \circ f(T)$ ■

Idem pour un ensemble de textes D, $g \circ f \circ g(D) = g(D)$.

(b) Soit T un motif quelconque et T' un motif fermé tel que : $T \subseteq T' \subseteq h(T)$, on veut montrer que $h(T) = T'$?
Comme T' est un motif fermé, on a : $T \subseteq h(T') \subseteq h(T)$ alors :
$f(T) \supseteq f \circ h(T')$; car f *est décroissante*
$f(T) \supseteq f(T')$; par la Propriété 3.1 (a), i.e *$f \circ h = f$*
$g \circ f(T) \subseteq g \circ f(T')$; car g *est décroissante*
$h(T) \subseteq T'$
Or, nous avons supposé que : $T' \subseteq h(T)$
d'où $h(T) = h(T')$ ■

(c) par la Propriété 3.1 (a), i.e. *$f \circ g \circ f = f$, nous avons :*

$$support(h(T)) = \frac{|f \circ h(T)|}{|\mathcal{D}|} = \frac{|f \circ g \circ f(T)|}{|\mathcal{D}|} = \frac{|f(T)|}{|\mathcal{D}|} = support(T)$$ ■

(d) Soit $\mathcal{L}_{\mathtt{FF}}$, l'ensemble des fermés fréquents et $\mathcal{L}$, l'ensemble des motifs fréquents défini § 3.2.2.
Soit T, T′ *deux motifs tels que* $\mathtt{T} \in \mathcal{L}_{\mathtt{FF}}$ *et* $\mathtt{T}' \subseteq \mathtt{T}$
$\mathtt{T}' \subseteq \mathtt{T} \Longrightarrow \mathtt{f(T')} \supseteq \mathtt{f(T)} \Longrightarrow \mathtt{support(T')} \geqslant \mathtt{support(T)} \geqslant \mathtt{minsup}$
par conséquent, $\mathtt{T}' \in \mathcal{L}$*, mais on ne peut pas conclure l'appartenance ou non du motif* T′ *à* $\mathcal{L}_{\mathtt{FF}}$ ■

(e) De façon analogue, soit T et T' deux motifs tels que $\mathtt{T} \notin \mathcal{L}_{\mathtt{FF}}$, $\mathtt{T}' \in \mathcal{L}$ *et* $\mathtt{T} \subseteq \mathtt{T}'$,
$\mathtt{M}' \supseteq \mathtt{T} \Longrightarrow \mathtt{f(T')} \subseteq \mathtt{f(T)} \Longrightarrow \mathtt{support(T')} \leqslant \mathtt{support(T)} \leqslant \mathtt{minsup}$
par conséquent, $\mathtt{T}' \notin \mathcal{L}$ ■

Définition 3.8 (Motif générateur) *Un motif* générateur G_k *d'un motif fermé* $\mathtt{F_k}$ *de taille* $\mathtt{k}$ *est un motif minimal (*i.e. *le plus petit motif par rapport à l'inclusion) dont la fermeture est égale à* $\mathtt{F_k}$*, Plusieurs motifs peuvent générer le même motif fermé. Ces motifs font partie d'un classe d'équivalence (*cf. *définition 3.9,* FIG. *3.5 et* FIG. *3.6 en § 3.2.4.4). Autrement dit,* G_k *est un motif* générateur *d'un fermé* $\mathtt{F_k}$*, si et seulement si :* $\nexists\, k' < k,\ G_{k'} \subsetneq G_k\ tels\ que\ h(G_{k'}) = \mathtt{F_k}$.

Propriétés 3.2

(a) Tout motif générateur $\mathtt{T} \in \mathtt{G_k}$ *ne peut être inclus dans les motifs fermés de ses sous-motifs.* $\{T = T_1 \otimes T_2 \in G_{k-1}\ tel\ que\ T \not\subseteq h(T_1) et T \not\subseteq h(T_2)\}$. *L'opérateur* $\otimes$ *correspond à effectuer une jointure telle qu'elle est définie pour les bases de données ;*

(b) Tout sous-motif d'un motif générateur fréquent est un motif générateur fréquent ;

(c) Tout super-motif d'un motif générateur non fréquent est un motif générateur non fréquent.

Démonstrations 3.2

(a) Soit C_k *un motif générateur candidat d'un fermé* $\mathtt{F_k}$ *et* G'_k*, avec* $k' \leqslant k$*, un sous-motif générateur.*
$G_{k'} \subseteq C_k \subseteq h(G_{k'})$*, par application de* h *et son idempotence, on a* $h(G_{k'}) \subseteq h(C_k) \subseteq h(G_{k'})$
$h(G_{k'}) = h(C_k)$*,* C_k *n'est pas un motif générateur car par définition* C_k *ne génère pas un nouveau fermé et n'est pas minimal* ■

(b) les deux propriétés 3.2-(b et c) des motifs générateurs sont démontrées dans [Bastide, 2000] (p. 51).

L'algorithme original de recherche des motifs fréquents s'appelle *Apriori* [Agrawal et Srikant, 1994], il a été utilisé pour la première fois sur des données textuelles dans [Feldman et Dagan, 1995]. Cet algorithme se révèle inefficace pour traiter des données issues d'une expérimentation de taille réelle car il génère un trop grand nombre de règles d'association à analyser. En revanche, nous détaillons l'algorithme *Close* [Pasquier *et al.*, 1999b] qui recherche un sous-ensemble de motifs fréquents particuliers : les motifs *fermés* fréquents. Les motifs fréquents, et leurs supports, sont déduits du sous-ensemble des motifs fermés fréquents. L'idée générale d'extraction des motifs fermés fréquents et la même que celle de l'algorithme *Apriori* à la différence que les motifs fermés fréquents sont moins nombreux et plus rapides à trouver — en nombre d'accès à la base de données —- que les motifs fréquents. L'algorithme *Close* est donc optimal et plus approprié pour le nombre de textes que nous traitons, car l'espace de recherche des motifs est réduit. En effet, dans le pire cas le nombre de motifs fréquents est de $(2^{|\mathcal{T}|} = 2^{\mathtt{n}})$ (*i.e.* l'ensemble des parties de $\mathcal{T}$). Pour les motifs fermés fréquents, une étude de complexité présentée dans [Godin, 1989] montre que ce nombre est inférieur à $2^{|\mathtt{f(T)}|} \times \mathtt{n}$. $|\mathtt{f(T)}|$ est une borne supérieure correspondant au motif fréquent $\mathtt{T}$ ayant le plus grand cardinal de l'extension. De plus, des résultats expérimentaux dans cette même étude montrent que ce nombre est en pratique en $\mathcal{O}(\mathtt{n})$, plus précisément que le nombre de motifs fermés dans le pire cas est égal à la $\mathtt{moyenne}(|\mathtt{f(T)}|) \times \mathtt{n}$.

Nous adaptons la présentation de l'algorithme *Close* pour générer des règles d'association à partir de données textuelles.

3.2.4.3 Présentation de l'algorithme Close

L'algorithme *Close* [Pasquier *et al.*, 1999b] est inspiré de l'algorithme *Apriori* [Agrawal et Srikant, 1994] pour la recherche de motifs fréquents par lecture et comptage des données en entrée. Il existe d'autres algorithmes qui sont des variantes de *Apriori*. Par exemple, les algorithmes *CLOSETS* [Pei *et al.*, 2000] et CHARM [Zaki et Hsiao, 1999]. Des études comparatives des performances de ces algorithmes sont présentées dans [Bastide, 2000; Taouil, 2000; Zheng *et al.*, 2001].

L'algorithme *Close* est composé de quatre étapes. Les trois premières étapes concernent la recherche de tous les motifs fermés fréquents. Ces trois étapes ont la plus grande complexité calculatoire (en mémoire et en accès à la base de données). La quatrième étape est un calcul simple, sans accéder à la base de données, qui découle des calculs faits durant les trois premières étapes. Le nombre de motifs fermés fréquent est très inférieur au nombre de motifs fréquents, même dans le pire cas, lorsque les données sont fortement corrélées. Ce qui rend la recherche de motifs fermés fréquents (*Close*) moins coûteuse que la recherche de motifs fréquents (*Apriori*). L'idée de *Close* est de calculer les motifs fermés fréquents puis de trouver l'ensemble des motifs fréquents sans recours à la lecture et au comptage des données. De plus, *Close* utilise une technique itérative, dite *par niveaux*, dans la prise en compte des motifs à traiter en s'appuyant sur la propriété 3.1 (d) de § 3.2.4.2 stipulant qu'un motif fermé fréquent ne peut contenir que des sous-motifs fréquents. Afin de respecter cette propriété, l'itération démarre à partir de chaque terme de $\mathcal{T}$ (*i.e.* les 1-motifs). De façon itérative, la construction des k-motifs se fait par jointure, deux à deux, des $(k-1)$-motifs. À la $k^{\text{ème}}$ itération, les k-motifs sont appelés motifs candidats C_k et les $(k-1)$-motifs sont appelés motifs clés ou générateurs G_{k-1}. *Close* construit les k-motifs par jointure, deux à deux, des $(k-1)$-motifs. La prise en compte d'un motif m se fait par le calcul : $C_k = G_{k-1} \otimes m$. Le choix de m suit une stratégie de parcours en profondeur d'un arbre dont les nœuds sont mis dans un ordre lexicographique [Ganter et Wille, 1999]. L'ordre lexicographique est l'ordre des entrées des mots dans un dictionnaire. Par exemple, $(\texttt{a}, \texttt{aa}, \texttt{aab}, \texttt{ab}, \texttt{aba}, \texttt{abb}, \texttt{abc}, \ldots)$ est une liste triée selon l'ordre lexicographique dont le parcours définit à chaque étape un niveau de construction des motifs candidats C_k. Pour récapituler, l'algorithme *Close* procède donc par niveaux :

- Au niveau 1 : calcul du support de chaque 1-motif (*i.e.* la fréquence d'apparition de chaque terme de $\mathcal{T}$ dans le contexte $\mathcal{C}$) ; suppression des termes non fréquents (*i.e.* dont le support est strictement inférieur à `minsup`), calcul de leurs fermés par la formule (3.12) ;
- Au niveau k : calcul des k-motifs générateurs candidats ; calcul de leurs fermés et leurs supports, suppression des k-motifs non fréquents et des k-motifs non générateurs (*cf.* propriétés 3.1-(d) et 3.2-(a)) ;
- Au niveau $k+1$: les k-motifs générateurs sont utilisés pour générer les $(k+1)$-motifs candidats ; puis le traitement fait au niveau k est renouvelé.

Nous donnons ci-après l'algorithme formel, puis nous commentons les quatre étapes qui constituent *Close* :

L'algorithme *Close* calcule, dans un premier temps, tous les motifs fermés fréquents ainsi que les motifs générateurs (étape 1 à 3). Les motifs générateurs de fermés servent pour le calcul des règles d'association. Par la suite, l'ensemble des motifs fréquents et leurs supports sont calculés à partir des motifs fermés fréquents (étape 4). Nous donnons les commentaires pour les quatre étapes.

(Étape 1 : candidats) calcul des ensembles des motifs potentiellement générateurs C_k (*i.e.*

Algorithme 2: Algorithme *Close* de construction des motifs fermés fréquents

Entrée : n_{max} est le niveau maximal correspondant au plus grand motif de la matrice d'entrée ;
Sortie : l'ensemble des triplets (générateur, support, fermés).
pour chaque k *de* 2 *à* n_{max} **faire**
  /* un traitement particulier est effectué pour les 1-motif par simple comptage de leurs support */
  /* **(Étape 1 : candidats)** : génération de l'ensemble des motifs candidats selon l'ordre lexicographique*/
  $C_k = G_{k-1} \otimes m$;
  /* **(Étape 2 : fermés)** : calcul des fermés*/
  pour chaque $c_k \in C_k$ **faire**
    $f_k \leftarrow h(c_k)$;
    si $f_k \neq c_k$ **alors**
      si $f_k \neq \emptyset$ *et* $\texttt{support}(f_k) \geqslant \texttt{minsup}$ **alors**
        $F_k^C = F_k^C \cup \{f_k\}$;
    /* sinon, ce motif est déjà dans l'ensemble des fermés*/

    /* **(Étape 3 : élagage)** : appliquer deux stratégies*/
    pour chaque $c_{k-1} \subsetneq c_k$ **faire**
      /* $1^{\texttt{ère}}$ stratégie */
      si $c_{k-1} \in G_{k-1}$ **alors**
        /* $2^{\texttt{ème}}$ stratégie */
        si $c_k \not\subseteq h(G_{k-1})$ **alors**
          $G_k = G_k \cup \{c_k\}$;
  retourner $\left(\cup_k G_k^C\right)$, leurs supports et $\left(\cup_k F_k^C\right)$;
/* **(Étape 4 : fréquents)** : inférence des motifs fréquents (non utilisée)*/

les candidats générateurs) par jointure de motifs générateurs fréquents G_{k-1} de niveau inférieur (suivant l'ordre lexicographique). C_k est initialisé à $\emptyset$;

(Étape 2 : fermés) calcul des fermetures (*i.e.* les fermés[22]) de ces candidats par application de h, nous déterminons les supports (*cf.* propriété 3.1-(c)) et nous les ajoutons à l'ensemble F_k^C des fermés s'ils sont non égaux à leurs fermés, non vides et fréquents ($\geqslant minsup$). F_k^C est initialisé à $\emptyset$;

(Étape 3 : élagage) suppression des candidats de C_k non fréquents. Cette suppression est faite suivant deux stratégies d'élagage : les motifs générateurs candidats dans C_k sont gardés si et seulement si :

- Aucun de ses sous-motifs n'est un motif générateur non fréquent (*cf.* propriété 3.2-(b)) ;
- S'il n'est pas inclus dans la fermeture d'un de ses sous-motifs générateurs G_{k-1} ;

– Retourner l'ensemble restant des motifs générateurs $\left(\cup_k G_k^C\right)$, leurs supports ainsi que l'ensemble des fermés $\left(\cup_k F_k^C\right)$ que nous utilisons lors de la génération des règles d'association ;

(Étape 4 : fréquents) grâce à la propriété 3.1-(b), nous savons que nous pouvons calculer l'ensemble des motifs fréquents à partir des motifs fermés fréquents. Ce calcul consiste à trier les couples de motifs fermés fréquents et leurs supports par cardinalité décroissante.

[22] Car la fermeture d'un motif est elle-même un motif fermé.

Pour un k-motif, l'ensemble de ses $(k-1)$-motifs non fermés est ajouté à la liste (*cf.* propriété 3.1-(d)). Les nouveaux $(k-1)$-motifs ont le même support que le k-motif fermé qui permet de les retrouver (*cf.* propriété 3.1-(c)). Le processus est renouvelé jusqu'à atteindre le niveau des 1-motifs.

En sortie de l'algorithme *Close*, nous avons besoin de garder une trace des motifs fermés fréquents et des générateurs. Le calcul des motifs fréquents à partir des motifs fermés fréquents est une étape que nous n'exploitons pas pour construire les règles d'association informatives que nous définissons § 3.2.4.4.

Exemple Nous présentons un exemple dans les tableaux (TAB. 3.2) du déroulement de l'algorithme *Close*, correspondant au tableau (TAB. 3.1) du même exemple donné en § 3.2.2. À la première passe, le motif `d`, ainsi que tous ses super-motifs (`ad`, `bd`, `abd`, ...), sont élagués car le support de $\texttt{Support}(\texttt{d}) = \frac{1}{6}$ est inférieur à $\texttt{minsup} = \frac{2}{6}$. Dans la deuxième passe, le motif-candidat `ac` est élagué de l'ensemble des motifs générateur car un de ses sous-motifs (*i.e.* `a`) possède comme fermeture le motif-candidat `ac`, etc.

TAB. 3.1 – Représentation sous forme tabulaire de la matrice d'entrée de l'exemple § (3.2.2)

Texte	Termes
d_1	acd
d_2	bce
d_3	abce
d_4	be
d_5	abce
d_6	bce

TAB. 3.2 – Déroulement de l'algorithme Close pour l'exemple de la table 3.1 avec $\texttt{minsup} = 2/6$

Générateur	Support	Fermeture
a	3	ac
b	5	be
c	5	c
~~d~~	~~1~~	~~acd~~
e	5	be

1ère passe →

Générateur	Support	Fermeture
ae	2	abce
ab	2	abce
bc	4	bce
ce	4	bce

2ème passe →

Fermé	Support
c	5
ac	3
be	5
bce	4
abce	2

fermés fréquents →

3.2.4.4 Présentation de l'algorithme de génération des règles d'association informatives

Tout motif fermé fréquent `m` est susceptible d'engendrer $(2^{|\texttt{m}|} - 2)$ règles d'association possibles. Le nombre maximal de règles généré par l'ensemble `M` de tous les motifs est : $3^{|\texttt{M}|} - 2^{|\texttt{M}|+1}+1$ (*cf.* la démonstration présentée en annexe § B.1). Parmi ces règles, il y en a qui sont *redondantes.* Une règle est dite redondante par rapport à d'autres règles d'association, si l'information qu'elle apporte est, par ailleurs, présente dans d'autres règles. Une règle d'association redondante est donc inutile ou moins informative qu'une ou plusieurs autres règles. Le problème de

recherche d'une famille minimale de règles à partir de laquelle nous pouvons retrouver toutes les autres règles est appelé la recherche d'une « base » de règles. Les bases de règles d'implication de Duquenne-Guigues [Guigues et Duquenne, 1986] et de règles d'implications partielles de Luxenburger [Luxenburger, 1991] ont été adaptées aux règles d'association exactes et approximatives dans [Zaki, 2000; Pasquier *et al.*, 1999a]. Nous donnons les définitions des règles d'association informatives, puis nous décrirons les règles exactes et approximatives.

Règle d'association informative. Une règle d'association informative est une règle telle que, par rapport à l'ordre d'inclusion :

- le motif `B` est minimal,
- le motif `H` est maximal.

Explication intuitive : Nous avons besoin de savoir quelles cooccurrences de termes au minimum impliquent quels autres termes. Le tableau (TAB. 3.3) montre qu'une règle $\texttt{ab} \Longrightarrow \texttt{cd}$, si elle est valide, peut engendrer d'autres règles qui sont redondantes car elles peuvent se calculer à partir de $\texttt{ab} \Longrightarrow \texttt{cd}$. À valeur de support identique (*i.e.* $\texttt{support}(\texttt{abcd})$) et à valeurs de confiance près, il suffit d'avoir `ab` et *a fortiori* un de ses sur-ensembles ($\{\texttt{abc}\}, \{\texttt{abd}\}, \{\texttt{abcd}\}, \ldots$) pour avoir « `cd` » voire un de ses sous-ensembles non vides ($\{\texttt{c}\}$ et $\{\texttt{d}\}$). La justification mathématique, d'un point de vue logique, de la redondance des quatre règles de TAB. 3.3 pour $\texttt{ab} \Longrightarrow \texttt{cd}$ est donnée en annexe B.3.

TAB. 3.3 – Ensemble de règles d'association redondantes engendrées par une règle valide

Règle de départ	Règle redondante
$\texttt{ab} \Longrightarrow \texttt{cd}$	$\texttt{ab} \Longrightarrow \texttt{c}$
	$\texttt{ab} \Longrightarrow \texttt{d}$
	$\texttt{abc} \Longrightarrow \texttt{d}$
	$\texttt{abd} \Longrightarrow \texttt{c}$

Définition 3.9 (Classe d'équivalence de motifs fréquents) *Soit deux motifs* $\texttt{T}, \texttt{T}' \in \mathcal{T}$. *La relation* θ *est définie par :*

$$T \; \theta \; T' \Longleftrightarrow f(T) = f(T').$$

La classe d'équivalence est donnée par :

$$[T] = \{T' \in \mathcal{T} \mid T \; \theta \; T'\}.$$

Propriétés 3.3

(a) θ *est une relation d'équivalence ;*

(b) Deux motifs appartenant à la même classe d'équivalence ont le même support et la même fermeture

$$T \; \theta \; T' \Longrightarrow support(T) = support(T') \; et \; h(T) = h(T')$$

(c) Deux motifs de support égal et dont l'un est sous-motif de l'autre appartiennent à la même classe

$$T \subseteq T' \; et \; support(T) = support(T') \Longrightarrow T \; \theta \; T'$$

(d) Les motifs minimaux d'une classe d'équivalence sont les motifs générateurs fréquents

$$T \; est\; générateur \iff T \in \min[T]$$

(e) Le motif maximal d'une classe d'équivalence est le motif fermé

$$T \; est \; fermé \iff T = \max[T]$$

Démonstrations 3.3

(a) $\mathtt{T}\ \theta\ \mathtt{T}' \iff \mathtt{f}(\mathtt{T}) = \mathtt{f}(\mathtt{T}') \iff \frac{|\mathtt{f}(\mathtt{T})|}{|\mathcal{D}|} = \frac{|\mathtt{f}(\mathtt{T}')|}{|\mathcal{D}|} \iff \mathtt{support}(\mathtt{T}) = \mathtt{support}(\mathtt{T}')$.
De même $\mathtt{T}\ \theta\ \mathtt{T}' \iff \mathtt{f}(\mathtt{T}) = \mathtt{f}(\mathtt{T}') \Longrightarrow \mathtt{g} \circ \mathtt{f}(\mathtt{T}) = \mathtt{g} \circ \mathtt{f}(\mathtt{T}') \Longrightarrow \mathtt{h}(\mathtt{T}) = \mathtt{h}(\mathtt{T}')$ *; et réciproquement puisque* $\mathtt{f} \circ \mathtt{h} = \mathtt{f}$ *(*cf. *propriété 3.1-(b)),* $\mathtt{h}(\mathtt{T}) = \mathtt{h}(\mathtt{T}') \Longrightarrow \mathtt{f} \circ \mathtt{h}(\mathtt{T}) = \mathtt{f} \circ \mathtt{h}(\mathtt{T}')$ $\iff \mathtt{f}(\mathtt{T}) = \mathtt{f}(\mathtt{T}') \iff \mathtt{T}\ \theta\ \mathtt{T}'$ ■

(b) $\mathtt{T} \subseteq \mathtt{T}' \Longrightarrow \mathtt{f}(\mathtt{T}) \supseteq \mathtt{f}(\mathtt{T}')$*, de plus* $\mathtt{support}(\mathtt{T}) = \mathtt{support}(\mathtt{T}') \iff |\mathtt{f}(\mathtt{T})| = |\mathtt{f}(\mathtt{T}')|$*. Nous avons donc,* $\mathtt{f}(\mathtt{T}) = \mathtt{f}(\mathtt{T}')$ *d'où* $\mathtt{T}\ \theta\ \mathtt{T}'$ ■

(c) par définition d'un motif générateur,

(d) par la propriété d'extensivité de $\mathtt{h}$ *(*$\forall \mathtt{T}, \mathtt{T} \subseteq \mathtt{h}(\mathtt{T})$*),* $\mathtt{h}(\mathtt{T})$ *est donc le motif fréquent maximal de* $[\mathtt{T}]$ ■

Les définitions de motifs générateurs, de motifs fermés fréquents et de classes d'équivalence de motifs fréquents nous permettent de définir une règle d'association informative :

Définition 3.10 (Règle d'association informative) *Une règle d'association* $\mathtt{R} : \mathtt{B} \Longrightarrow \mathtt{H}$ *est informative, s'il n'existe pas de règle* $\mathtt{R}' : \mathtt{B}' \Longrightarrow \mathtt{H}'$ *telle que :*

- $\mathtt{support}(\mathtt{R}) = \mathtt{support}(\mathtt{R}')$,
- $\mathtt{confiance}(\mathtt{R}) = \mathtt{confiance}(\mathtt{R}')$,
- $\mathtt{B} \supseteq \mathtt{B}'$ *et* $\mathtt{H} \subseteq \mathtt{H}'$.

Soit $\mathtt{B}$ et $\mathtt{H}$ deux motifs particuliers de $\mathcal{T}$ pris tels que :

- $\mathtt{B} \cup \mathtt{H} = \{\mathtt{fermeture}\} = \mathtt{h}(\mathtt{B}) = \cap_{\mathtt{B} \subseteq \mathtt{g}(\mathtt{d})} \mathtt{g}(\mathtt{d})$,
- $\mathtt{B} = \{\mathtt{générateur}\}$ l'ensemble des motifs générateurs,
- $\mathtt{H} = \{\mathtt{fermeture}\} \setminus \{\mathtt{générateur}\}$ l'ensemble des fermetures privées de leurs générateurs.

Une règle d'association informative $\mathtt{R} : \mathtt{B} \Longrightarrow \mathtt{H}$ est donc, en pratique, calculée par :

$$B \Longrightarrow ((B \cup H) \setminus B) \; \mathtt{avec} \begin{cases} B \in \mathtt{min}[B \cup H] & \mathtt{est\ un\ motif\ générateur\ fréquent} \\ (B \cup H) = \mathtt{max}[B \cup H] & \mathtt{est\ un\ motif\ fermé\ fréquent} \\ B \subsetneq B \cup H & \mathtt{inclusion\ stricte\ pour\ H} \neq \emptyset \end{cases}$$

Par exemple, le tableau TAB. 3.1, correspondant à la matrice en § 3.2.2 page 50, nous donne pour **minsup=2/6 et minconf=2/5** par application de l'algorithme 3 l'ensemble des classes, des générateurs, des fermés et des règles de FIG. 3.5 lorsque les liens de redondance sont gardés ou lorsque ces liens sont supprimés dans FIG. 3.6.

Dans le 1$^{\mathtt{er}}$ tableau à gauche de TAB. 3.4, par exemple pour la règle soulignée ($\mathtt{R} : \mathtt{a} \Rightarrow \mathtt{bce}$), le motif générateur est "a", le motif fermé est "abce", et $\mathtt{confiance}(\mathtt{R}) = \frac{\mathtt{support}(\mathtt{abce})}{\mathtt{support}(\mathtt{a})} = \frac{2}{3}$.

Algorithme 3: Algorithme de génération de règles d'association informatives

Entrée :
- K : ensemble de triplets (G : motif générateur, support, F_G : motif fermé) issus de l'exécution de *Close* ;
- minsup : support minimum supérieur à celui de génération des motifs fermés fréquents de *Close*, sinon appliqué par défaut ;
- minconf : confiance minimale ;

Sortie :
- R_A : ensemble des règles d'association valides.

pour chaque *motif générateur* $G \in K$ **faire**
  trouver l'ensemble des motifs fermés candidats $C_G = \{F_G / G \subsetneq F_G\}$;
  trier l'ensemble C_G par cardinalité croissante ;
  pour chaque *motif fermé* $F_G \in C_G$ **faire**
    si *conf*$(r : G \Longrightarrow F_G \setminus G) \geqslant$ minconf **alors**
      /* Optionnel : Élagage des sur-motifs fermés candidats suivants d'une classe d'équivalence non directe pour le générateur courant */
      si *conf*$(r : G \Longrightarrow F_G \setminus G) \neq 1$ **alors**
        ôter de C_G tout F tel que $F_G \subseteq F$;
      $R_A \longleftarrow R_A \cup \{r\}$;
      afficher cette règle ;

TAB. 3.4 – Déroulement de l'algorithme d'extraction de règles informatives pour l'exemple 3.1 avec minconf $= 2/5$ et règles redondantes gardées

Gén.	Ferm.	Règle	Conf.
a	ac	a ⇒ c	3/3
	abce	a ⇒ bce	2/3

Gén.	Ferm.	Règle	Conf.
b	be	b ⇒ e	5/5
	bce	b ⇒ ce	4/5
	abce	b ⇒ ace	2/5

Gén.	Ferm.	Règle	Conf.
c	c		
	ac	c ⇒ a	3/5
	bce	c ⇒ be	4/5
	abce	c ⇒ abe	2/5

Gén.	Ferm.	Règle	Conf.
e	be	e ⇒ b	5/5
	bce	e ⇒ bc	4/5
	abce	e ⇒ abc	2/5

Gén.	Ferm.	Règle	Conf.
ae	abce	ae ⇒ bc	2/2

Gén.	Ferm.	Règle	Conf.
ab	abce	ab ⇒ ce	2/2

Gén.	Ferm.	Règle	Conf.
bc	bce	bc ⇒ e	4/4
	abce	bc ⇒ ae	2/4

Gén.	Ferm.	Règle	Conf.
ce	bce	ce ⇒ b	4/4
	abce	ce ⇒ ab	2/4

Les supports de "a" et "abce" sont données dans le déroulement de l'algorithme *Close* en TAB. 3.2, page 58. Nous pouvons calculer le support d'une règle approximative par rapport aux textes vérifiant les parties B et H de la règle en accord avec la formule (3.9) donnée § 3.2.3, page 51. Ce qui nous donne, par exemple, pour la même règle (R : a ⇒ bce) du tableau (TAB. 3.4) le $\texttt{support}_1(R) = \frac{|f(abce)|}{|\mathcal{D}|} = \frac{2}{6}$. Nous pouvons également choisir de redéfinir le support d'une règle d'association en prenant en compte les contre-exemples $\texttt{support}_2(R) = \frac{|f(a)|}{|\mathcal{D}|} = \frac{3}{6}$ (ici, 2 exemples et 1 contre-exemple). Nous avons choisi d'implanter la deuxième définition qui est sans incidence sur l'algorithme d'extraction des règles car $\texttt{support}_2(R) \geqslant \texttt{support}_1(R) \geqslant \texttt{minsup}$. Les textes qui constituent des contre-exemples participent à l'extraction de la règle via la confiance. Les contre-exemples à une règle sont des indicateurs intéressants lors de la phase d'interprétation des règles d'association.

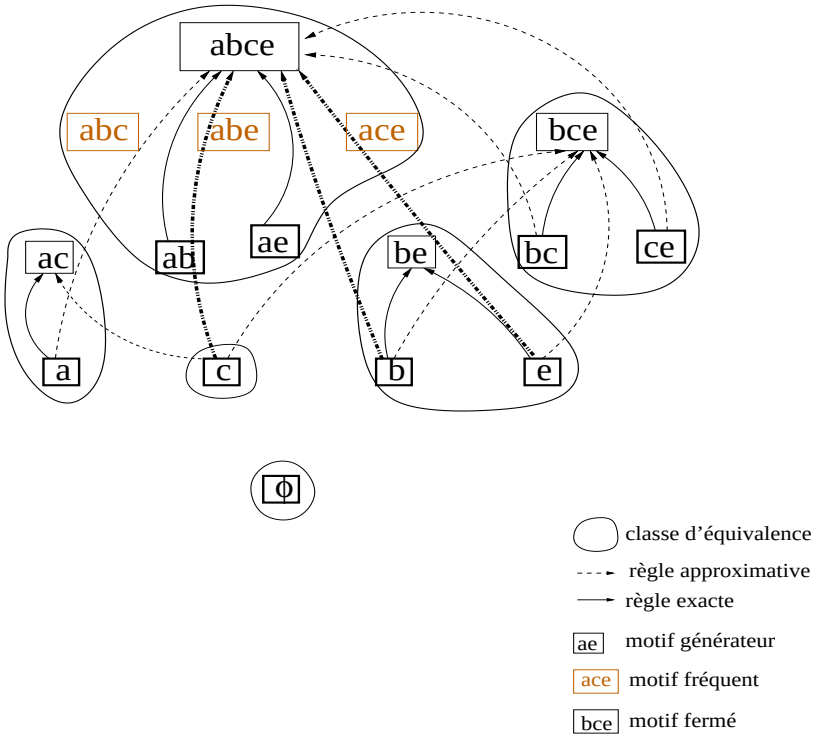


FIG. 3.5 – Calcul des règles pour **minsup=2/6 et minconf=2/5** illustrant les ensembles `B` et `H` – liens redondants gardés en pointillés gras.

TAB. 3.5 – Déroulement de l'algorithme d'extraction de règles informatives pour l'exemple 3.1 avec `minconf` $= 2/5$ et règles redondantes barrées

Gén.	Ferm.	Règle	Conf.
a	ac	a ⇒ c	3/3
	abce	a ⇒ bce	2/3

Gén.	Ferm.	Règle	Conf.
b	be	b ⇒ e	5/5
	bce	b ⇒ ce	4/5
	~~abce~~	~~b ⇒ ace~~	~~2/5~~

Gén.	Ferm.	Règle	Conf.
c	c		
	ac	c ⇒ a	3/5
	bce	c ⇒ be	4/5
	~~abce~~	~~c ⇒ abe~~	~~2/5~~

Gén.	Ferm.	Règle	Conf.
e	be	e ⇒ b	5/5
	bce	e ⇒ bc	4/5
	~~abce~~	~~e ⇒ abc~~	~~2/5~~

Gén.	Ferm.	Règle	Conf.
ae	abce	ae ⇒ bc	2/2

Gén.	Ferm.	Règle	Conf.
ab	abce	ab ⇒ ce	2/2

Gén.	Ferm.	Règle	Conf.
bc	bce	bc ⇒ e	4/4
	abce	bc ⇒ ae	2/4

Gén.	Ferm.	Règle	Conf.
ce	bce	ce ⇒ b	4/4
	abce	ce ⇒ ab	2/4

Les trois règles que nous éliminons dans les tableaux (TAB. 3.5) dénotent des liens redondants et non informatifs. La règle approximative `b` $\Longrightarrow$ `ace` se retrouve dans la règle exacte `ab` $\Longrightarrow$ `ce` et la règle de confiance supérieure `bc` $\Longrightarrow$ `ae`. La règle approximative `c` $\Longrightarrow$ `abe` se retrouve dans les deux règles `bc` $\Longrightarrow$ `ae` et `ce` $\Longrightarrow$ `ab`. De même pour la règle `e` $\Longrightarrow$ `abc` qui est retrouvée par les règles `ce` $\Longrightarrow$ `ab` et `ae` $\Longrightarrow$ `bc`. Nous pouvons réduire l'ensemble des règles sans perte d'information en utilisant la partie optionnelle de l'algorithme 3.

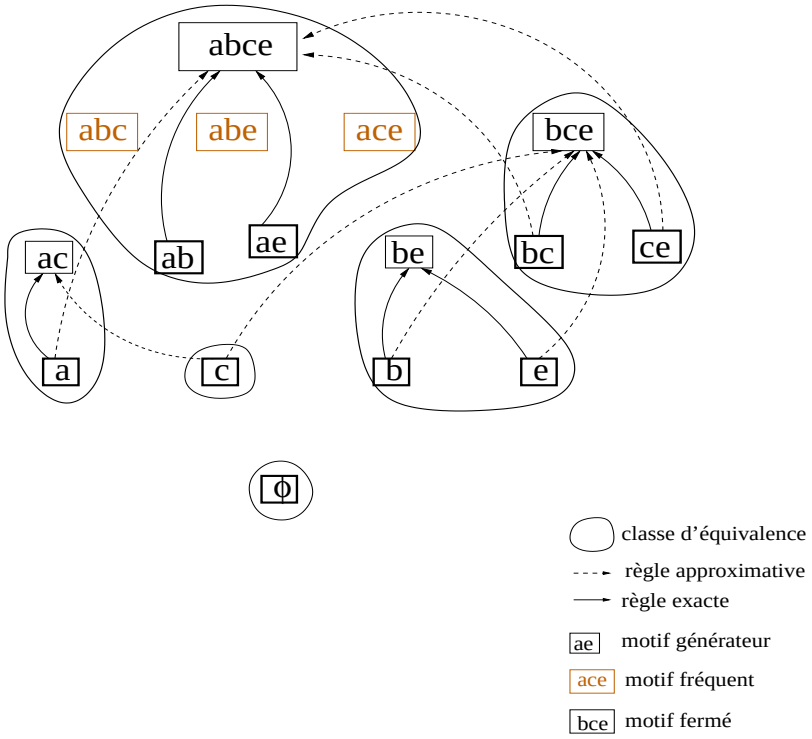


FIG. 3.6 – Calcul des règles pour **minsup=2/6 et minconf=2/5** illustrant les ensembles `B` et `H` – liens redondants supprimés.

3.3 Intérêt des motifs et des règles d'association pour des applications sur les textes

Nous présentons, dans cette section, comment le processus d'extraction des motifs fermés fréquents et des règles d'association peut servir différents types d'applications. Cet intérêt constitue un retour d'expériences de l'utilisation de la technique d'extraction de règles d'association sur les données textuelles. Nous montrons l'intérêt du processus de fouille de textes pour la structuration d'une terminologie et la structuration de connaissances relatives à un domaine. La structuration de connaissances d'un domaine est développée en deux parties : l'analyse formelle de concepts et la structuration d'ontologies. Nous poursuivons la présentation de l'intérêt des règles d'association pour la FdT, en montrant comment les règles d'association peuvent aider à réaliser des tâches de recherche d'information, d'extraction d'information et de veille technologique et stratégique.

3.3.1 Filtrage d'une terminologie pour la constitution d'un thésaurus

Le processus de fouille de textes est sensible à la phase d'indexation. Si un terme est absent de l'indexation d'un seul texte, cela peut entraîner la disparition d'une règle du fait du seuil *minsup* choisi. Nous constatons que le processus d'indexation terminologique décrit § 2.4.2.3 29 permet de révéler la qualité de l'indexation. À l'issue de la phase d'indexation, nous observons une grande disparité des termes bien que le corpus soit spécialisé (*i.e.* résistance des bactéries aux antibiotiques). Nous retrouvons ce phénomène régulièrement en analyse automatique de corpus. Un texte

fait référence à des termes périphériques au domaine qui introduisent du *bruit*. Certaines règles d'association révèlent ces termes périphériques au domaine. Ces termes constituent pour nous du bruit et n'ont pas été repérés par l'analyste lors du nettoyage manuel des termes-index des textes. Par exemple, dans la règle « mycobacterium tuberculosis » $\Longrightarrow$ « tuberculosis », la maladie « tuberculosis » n'est pas pertinente dans le domaine de discours du corpus. Le terme « tuberculosis » résulte du repérage de la bactérie « mycobacterium tuberculosis ». C'est un sous-terme présent dans le vocabulaire médical utilisé pour la phase d'indexation.

Comme le montre FIG. 2.1, le processus de fouille de texte comprend une boucle de réutilisation des connaissances extraites par un retour à l'étape d'indexation. Par conséquent, il est possible, par un processus itératif de :

(1) Filtrer un terme-bruit repéré dans une règle d'association en éliminant toutes les occurrences de ce terme dans l'indexation des textes ;

(2) Extraire les règles s'appuyant sur cette nouvelle indexation ;

(3) Retourner au point (1).

Il faut, néanmoins, être prudent lors de l'élimination des termes considérés comme un terme-bruit issu de l'indexation automatique. Une élimination du terme-bruit dans tous les textes où il apparaît peut avoir une incidence d'un texte à un autre. L'incidence sur la caractérisation de son contenu peut être plus ou moins importante selon le texte. Nous savons exactement quels textes ont permis de générer la règle. L'analyste doit, d'abord, consulter ces textes pour s'assurer du statut de terme-bruit avant de l'éliminer de l'indexation.

L'ensemble des termes présents indifféremment en partie gauche ou droite des règles d'association constitue un thésaurus du domaine. Le thésaurus peut être affiné en éliminant les termes-bruit. Les règles d'association peuvent donc servir pour la construction de ressources terminologiques d'un domaine particulier. Lorsque nous ne disposons pas de vocabulaire du domaine, nous pouvons imaginer d'utiliser tous les termes des règles d'association comme un premier thésaurus du domaine et l'affiner au fur et à mesure que l'analyste rencontre des termes bruit jusqu'à obtenir un thésaurus satisfaisant. Ce processus itératif est généralement appelé : bootstrapping. Cette façon de faire rejoint le but des travaux de Condamines [Condamines, 2002] ou de [Séguéla et Aussenac-Gilles, 1999], en TAL, fondés sur le repérage de marqueurs linguistiques dans les textes pour l'extraction de schémas de relations de dépendance entre concepts (hyponymie, hypéronymie et synonymie) à partir de corpus.

3.3.2 Structuration de connaissances d'un domaine

L'interêt de l'extraction des règles d'association pour la structuration de connaissances d'un domaine est développée en deux parties : d'abord, nous présentons comment les règles d'association sont utilisées en analyse formelle de concepts ; puis nous évoquons leur utilisation dans le cadre de la structuration d'ontologies.

3.3.2.1 Analyse de concepts formels : construction d'un treillis de Galois

L'analyse de concepts formels (ACF) [Ganter et Wille, 1999] est une méthode formelle d'apprentissage de concepts à partir de données. l'ACF) est fondée sur la construction d'un treillis

de Galois (ou treillis de concepts). La construction d'un *treillis de Galois* [Duquenne, 1999] — ou la classification par treillis de Galois — répond aux objectifs de l'ECBD. En particulier, elle permet de construire des concepts à partir d'un ensemble d'individus munis de leurs propriétés, puis d'organiser ces concepts dans une structure hiérarchique à partir de laquelle il est possible d'observer des corrélations entre les individus et leurs propriétés communes. La construction d'un treillis de Galois permet de se donner une structure mathématique pour l'analyse de concepts issus du domaine.

Nous avons introduit § 3.2.4.1 la connexion de Galois pour une relation $\mathtt{g} = \mathtt{Intension}$ définie entre un ensemble de textes $\mathcal{D}$ et l'ensemble des termes-clés qui indexent ces textes $\mathcal{T}$, et sa relation duale $\mathtt{f} = \mathtt{Extension}$ définie sur $(\mathcal{T} \times \mathcal{D})$ entre un ensemble de termes et les textes qui les possèdent. Nous avons également introduit l'opérateur de fermeture $\mathtt{h} = \mathtt{g} \circ \mathtt{f}$. Les relations $\mathtt{f}$ et $\mathtt{g}$ ainsi que leur composée $\mathtt{h}$ nous permettent de construire des motifs fermés fréquents. La construction des motifs fermés correspond à la recherche des rectangles maximaux dans une matrice booléenne [Norris, 1978].

Si $\mathtt{T}$ est un motif fermé fréquent (*i.e.* $\mathtt{h}(\mathtt{T}) = \mathtt{T}$), un concept fréquent (*i.e.* dit concept fort) est un couple $(\mathtt{T}, \mathtt{f}(\mathtt{T}))$ dont l'extension est supérieure à un seuil donné. L'ensemble des concepts constitue un treillis de Galois. Les concepts sont organisés selon une structure d'ordre partiel par une relation de *subsomption* notée $\prec$.

Propriétés 3.4

Deux concepts quelconques du treillis possèdent :

- *Un unique plus petit concept du treillis qui les subsume (*i.e. *qui les généralise), constituant leur borne supérieure ;*
- *Un unique plus grand concept qu'ils subsument (*i.e. *qui les spécialise), constituant leur borne inférieure.*

Soit $\mathtt{C}_1 = (\mathtt{T}, \mathtt{f}(\mathtt{T}))$ et $\mathtt{C}_2 = (\mathtt{T}', \mathtt{f}(\mathtt{T}'))$ deux concepts du treillis, $\mathtt{C}_2$ subsume $\mathtt{C}_1$ si et seulement si :

$$C_1 \prec C_2 \Longleftrightarrow T' \subsetneq T \Longleftrightarrow f(T') \supsetneq f(T)$$

La sémantique de la relation de subsomption (généralisation et spécialisation) est l'inclusion ensembliste des intensions et des extensions. (FIG. 3.7) illustre un treillis de Galois correspondant à l'exemple du tableau (TAB. 3.1) avec un support minimum de $2/6$. Ce treillis est appelé treillis des *Icebergs* dans [Stumme *et al.*, 2002]. Le terme *iceberg* illustre le principe de la poussée d'Archimède exercée sur un objet plongé dans l'eau. En diminuant le seuil de support, nous obtenons plus de concepts et nous alourdissons le treillis et par conséquent nous faisons émerger des concepts. Nous remarquons que le treillis de Galois (*i.e.* sans support minimum) à droite de (FIG. 3.7) possède un concept supplémentaire $\langle \mathtt{acd}, \mathtt{d}_1 \rangle$ de support $1/6$ et un concept racine $\langle \mathcal{T}, \emptyset \rangle$ pour respecter la propriété (3.4) des treillis.
La construction d'un treillis de Galois (resp. des Icebergs) à partir des motifs fermés fréquents consiste à :

- calculer les extensions des motifs fermés (resp. fréquents) générés par *Close*,
- calculer les liens de subsomption entre les concepts.

L'intérêt de l'ACF pour la FdT est décrite à travers notre première expérience dans [Toussaint *et al.*, 2000] et dont nous donnons le principe d'extraction des règles d'association en § 4.1.1 page 72.

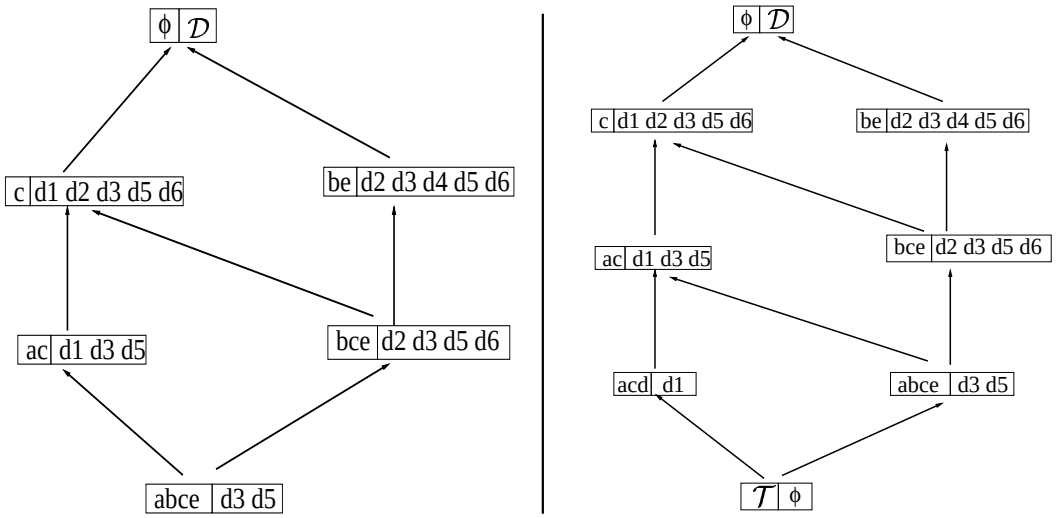


FIG. 3.7 – Treillis des Icebergs avec $\texttt{minsup} = 2/6$ (à gauche) et treillis de Galois (à droite) de l'exemple du tableau (TAB. 3.1).

3.3.2.2 Construction d'ontologies

Une ontologie est une structure hiérarchique de concepts reliés par des liens de généricité, de causalité, de typicité, etc. La construction d'ontologies est une tâche essentielle pour l'acquisition et l'organisation de connaissances dans un domaine pour une application donnée. Les ontologies d'un domaine sont construites avec l'aide d'experts du domaine. Pour être consensuelle, il est difficilement imaginable de créer et maintenir une ontologie universelle couvrant tous les domaines. En effet, l'échec partiel du projet CYC [Lenat et Guha, 1990] est essentiellement dû à une ambition d'universalité et de couverture des connaissances dans tous les domaines.

Les entités composant une ontologie sont abordées selon deux points de vue : linguistique et conceptuel. Le point de vue linguistique privilégie l'approche par extraction de termes et de relations sémantiques entre les termes à partir de textes. Le point de vue conceptuel, issu de l'ingénierie des connaissances, place le concept au centre de l'ontologie et vise à construire un modèle à partir des connaissances, pratiques et avérées, issues de l'expertise humaine dans un domaine particulier.

Une ontologie se construit selon un cycle de construction qui rejoint celui du processus de FdT (identifier les sources d'informations, extraire les concepts, trouver les relations entre les concepts, les valider). Le cycle de vie d'une ontologie est donc présent dans notre processus de FdT puisqu'il nous faut maintenir et enrichir une ontologie avec de nouvelles connaissances et réutiliser une partie de ces connaissances pour d'autres applications. Nous voyons en chapitre 5 que nous procédons à une mise-à-jour d'un modèle de connaissances terminologique proche d'une ontologie.

Un environnement de construction semi-automatique et de maintien d'une ontologie du domaine, appelé TEXT-TO-ONTO, est décrit dans [Maedche et Staab, 2003]. Dans ce système, la recherche de concepts se fait par une méthode statistique du type TF/IDF pour construire des classes de concepts. Les classes construites sont utilisées pour guider l'analyste dans le choix des liens qu'il pourra mettre entre les concepts présents dans les classes. Un calcul de règles d'asso-

ciation sert également d'aide pour la construction effective de l'ontologie. Cet environnement est fortement interactif et représente une aide à la construction manuelle d'une ontologie du domaine. Des heuristiques permettent de comparer une ontologie construite avec l'aide de TEXT-TO-ONTO et une ontologie de référence construite à la main [Maedche et Staab, 2001], ce qui permet de mieux cibler les propositions de mise à jour à faire éventuellement par l'expert dans l'ontologie de référence. Une application pour la construction d'ontologies s'appuyant sur l'analyse de concepts formels, appelée FCA-MERGE, est proposée dans [Stumme et Maedche, 2001].

De plus en plus de travaux concernant la construction automatique d'ontologies s'appliquent à une tâche de structuration des documents Web. C'est un thème de recherche émergent appelé *Web sémantique*, qui a pour but de structurer des documents Web pour les rendre compréhensibles pour un agent « intelligent » qui parcourt les documents et traite leur contenu pour donner des réponses à des requêtes des utilisateurs plus élaborées qu'une liste de documents que renvoie les moteurs de recherche du Web actuel.

3.3.3 Extraction d'information (EI)

L'extraction de règles d'association permet de réaliser des tâches d'extraction d'information pour remplir des patrons. À ce titre, le système TEXTRISE [Nahm et Mooney, 2001] illustre l'application d'un processus de FdT pour l'EI. TEXTRISE apprend à remplir certains attributs de patrons pour de nouveaux textes à partir de règles d'association apprises sur d'autres patrons. Dans une notice bibliographique par exemple, un patron possède un attribut auteur inconnu $\texttt{aut}_?$ mais un attribut mots-clés complet $\{\texttt{mc}_1, \texttt{mc}_2, \texttt{mc}_3\}$. Si durant la phase d'apprentissage nous avons une règle d'association ($\texttt{mc}_1, \texttt{mc}_2 \Longrightarrow \texttt{aut}_1$) appelée *soft-matching rule*, ce texte est attribué, à un degré de confiance près, à $\texttt{aut}_1$.

De notre point de vue, nous présentons les schémas d'interprétation des règles d'association, plus précisément des motifs fréquents sous-jacents à la règle, comme des patrons tels qu'ils sont définis en extraction d'information. Notre processus de FdT apprend donc des patrons et un lexique sémantique spécifique au domaine à travers l'extraction des règles d'association. Par exemple la règle :

"determine region" $\sqcap$ "gyrA gene" $\sqcap$ "gyrase" $\sqcap$ "mutation" $\Longrightarrow$ "quinolone"

signifie, dans le domaine de biologie moléculaire, que la "mutation" du gène "gyrA gene" dans une "région determinée" de la "gyrase" provoque une résistance à la famille des antibiotiques de la "quinolone".

3.3.4 Veille technologique et stratégique

La veille stratégique (appelée également *business intelligence*) est une tâche particulière d'extraction d'information dans le domaine industriel, des innovations technologiques, des avancées scientifiques et techniques, des normes et des brevets industriels. La veille est un processus de mise à jour périodique d'informations. La veille consiste à recueillir l'information, à la synthétiser et à tirer des conclusions pouvant réorienter les choix d'une entreprise vis-à-vis de ses concurrents dans le domaine industriel. L'information est collectée, par des analystes, sur les site Web des concurrents, dans les banques de données ou par des consommateurs.

Les règles d'association révèlent des implications entre termes et permettent de faire de la veille scientifique. Par exemple, [Nauer, 2002] utilise le champs « auteurs » d'articles scientifiques pour savoir quels sont les auteurs qui publient ensemble ? avec qui publient-ils systématiquement,

souvent, peu ou jamais ? Dans [Feldman *et al.*, 1998], l'utilisation des règles d'association permet de chercher des noms de compagnies qui ont fait alliance ou qui ont fusionné. Par exemple :

$$\begin{array}{rcl} \text{"intuit corp"} \sqcap \text{"novell corp"} & \Longrightarrow & \text{"merger"} \\ \text{"apple computer inc"} \sqcap \text{"sun microsystems inc"} & \Longrightarrow & \text{"merger talk"} \\ \text{"america online inc"} \sqcap \text{"bertelsmann ag"} & \Longrightarrow & \text{"joint venture"} \end{array}$$

Les systèmes qui s'inscrivent dans le cadre de la FdT pour la veille stratégique sont de plus en plus nombreux. Un des premiers systèmes qui utilise les données textuelles a été développé pour une tâche de recherche d'information. Il s'agit du système IOTA [Chiaramella *et al.*, 1986]. Il existe d'autres systèmes, issus du monde industriel, qui se définissent dans le cadre de le FdT. Il s'agit, notamment de *IBM Intelligent Miner for texts* [IBM, 1998] qui applique la mesure d'*intérêt* pour la classification de règles d'association (qu'ils appellent la mesure de *lift*). L'outil d'IBM intègre cette classification dans un environnement plus global de classification de documents par sujet et la détection de thèmes et la tâche de recherche documentaire. Plus récemment un outil similaire est développé par la société SAS qui travaille depuis longtemps déjà dans le domaine de la fouille de données. L'outil *SAS Text Miner* est une adaptation à des données textuelles prétraitées de l'outil classique de fouille de données symbolique de SAS (*cf.* `www.sas.com/technologies/analytics/datamining/textminer/`).

3.3.5 Recherche d'information (RI)

Le lien avec les motifs fréquents de termes cooccurrents dans les textes, tel que nous l'utilisons en FdT, se retrouve en recherche documentaire. La réponse à une requête de l'utilisateur (*i.e.* la liste de documents pertinents) est fondée sur le lien de cooccurrence entre les termes de la requête et leur fréquence d'apparition ensemble dans les textes. Dans [Carpineto et Romano, 1996; Carpineto et Romano, 2000], l'utilisation des motifs fermés fréquents permet, par navigation dans le treillis de Galois correspondant, de répondre à une requête par les documents constituant l'extension d'un concept. La requête est simplement considérée comme un nouveau texte à classifier dans le treillis.

Nous présentons dans le prochain chapitre notre méthodologie de sélection de règle d'association selon un critère de présence forte/rare des termes, constituant une règle d'association, dans le corpus. Nous présentons pour ce faire l'outil TAMIS qui réalise la tâche de FdT d'un point de vue *syntaxique.*

Chapitre 4

Description de l'outil TAMIS

« Never use statistics when you know
what you are talking about ».
G. Piatetsky-Shapiro

Sommaire

Introduction

Le présent chapitre présente l'outil ORPAILLEUR de fouille de textes, que nous avons développé, appelé TAMIS (*Text Analysis by Mining Interesting_ruleS*) fondé sur un tri des règles d'un point de vue syntaxique, c'est-à-dire s'appuyant sur la cooccurrence des termes présents dans les règles d'association et leurs distributions dans les textes.

L'outil TAMIS tient compte des spécifications présentées à la fin du chapitre 3 et automatise le processus d'extraction de connaissances à partir de textes présenté en FIG. 2.1. Dans un premier temps, nous présentons le problème d'analyse des résultats d'un processus fondé sur l'extraction des règles d'association informatives. Notre contribution à la résolution du problème de lecture

et d'analyse des règles consiste à utiliser des mesures de qualité pour ordonner les règles extraites [Cherfi *et al.*, 2004b; Cherfi *et al.*, 2003a]. Les différents ordres obtenus représentent des points de vue différents pour l'interprétation des règles par l'analyste. Une expérimentation portant sur un corpus de biologie moléculaire montre l'adéquation des ordres calculés pour l'aide à l'interprétation des règles extraites. Nous montrons également l'utilité du processus de FdT que nous proposons pour : (i) l'extraction de connaissances à partir de textes, (ii) l'amélioration de l'indexation des textes. Ce chapitre concerne la première partie de notre étude et réalisation d'un outil pour faire de la *fouille de textes* avec un TAMIS dit *syntaxique* qui prend en compte la base de données textuelle, sans utiliser les connaissances du domaine.

4.1 Gestion du nombre de règles d'association

Le nombre de règles extrait croît, dans le pire cas, de manière exponentielle par rapport au nombre de termes du corpus. La borne supérieure du nombre de motifs fréquents qui permettent de générer les règles d'association, pour une matrice de $\mathtt{n}$ termes fortement corrélées [23], est de $2^{\mathtt{n}}$. De plus, chaque motif fréquent $\mathtt{m}$ est susceptible de générer $(2^{|\mathtt{m}|} - 2)$ règles d'association possibles. Le nombre maximal de règles que nous pouvons extraire à partir des données est égal, dans le pire cas, à $3^{\mathtt{n}} - 2^{\mathtt{n}+1} + 1$ (*cf.* le détail du calcul présenté en annexe § B.1). L'interprétation des règles par l'analyste devient alors une tâche très fastidieuse, voire impossible.

Le grand nombre de règles d'association générées est traité par différentes approches que nous proposons de classer en quatre catégories :

1. Dans la première catégorie, l'idée consiste à *réduire le nombre de règles* en cherchant une base minimale de règles d'association. À partir d'une base minimale de règles, il est possible de déduire la totalité des règles. La réduction du nombre de règles s'opère soit :
   – Durant le processus de fouille [Guigues et Duquenne, 1986; Luxenburger, 1991; Diatta, 2003] ;
   – Après avoir organisé les données par une structure hiérarchique, par exemple un treillis de Galois [Ganter, 1999; Toussaint *et al.*, 2000; Stumme *et al.*, 2001] ou un espace de généralisation [Bournaud et Courtine, 2001]. Un espace de généralisation est un treillis d'héritage qui ne contient pas les concepts d'intensions vides. Nous revenons sur cette approche à travers deux exemples en § 4.1.1 ;
2. La deuxième catégorie d'approches consiste à *utiliser les connaissances de l'analyste* pour filtrer ces règles et de ne chercher que celles dont la prémisse (B) et/ou la conclusion (H) sont des termes d'un « type » particulier définis par l'analyste : *user-defined rule template* [Klemettinen *et al.*, 1994; Feldman *et al.*, 1998] ou DS : *direction setting* [Liu *et al.*, 1999a]. Ce typage est vu comme une contrainte sur les termes pour la génération des règles d'association qui permet de réduire l'ensemble des règles extraites. Par le même principe, dans [Li *et al.*, 1999], l'espace de recherche des textes est partitionné en deux sous-ensembles. Un sous-ensemble ($\mathtt{D}_{\mathtt{cible}}$) où apparaît un motif choisi comme `cible` et un autre où il n'apparaît pas ($\mathtt{D}_{\overline{\mathtt{cible}}}$). Puis, il s'agit de trouver tous les motifs `X` qui apparaissent au moins une fois dans ($\mathtt{D}_{\mathtt{cible}}$) et aucune fois dans ($\mathtt{D}_{\overline{\mathtt{cible}}}$). Ce qui permet de générer toutes les

[23] Une matrice de données booléenne fortement corrélées est une matrice très peu creuse (*i.e.* peu d'éléments nuls).

règles *exactes* telles que : $\mathtt{X} \Longrightarrow \mathtt{cible}$. Cette simplification de l'espace de recherche permet de s'affranchir du seuil `minsup` qu'il est difficile de choisir *a priori* sans prendre l'avis de l'analyste ou sans connaître la nature des données que nous fouillons. Cependant, nous pensons que cette approche peut introduire un biais sur les règles extraites à cause du paradoxe de *Simpson* [Simpson, 1951]. Le paradoxe de Simpson est un problème classique et identifié comme une source d'erreurs en apprentissage. Par exemple, une population est partitionnée en $\mathtt{Pop}_1$ et $\mathtt{Pop}_2$ suivant une première propriété $\mathtt{W}$ donnée. Si nous observons une certaine caractéristique $\mathtt{C}$ sur ces deux populations, nous trouvons par exemple que $\mathtt{P}_{\mathtt{Pop}_1}(\mathtt{C}) \geqslant \mathtt{P}_{\mathtt{Pop}_2}(\mathtt{C})$. Si nous subdivisons ces deux populations selon une seconde propriété $\mathtt{W}'$ en $\mathtt{Pop}_{1,\mathtt{W}'}$, $\mathtt{Pop}_{1,\overline{\mathtt{W}'}}$, $\mathtt{Pop}_{2,\mathtt{W}'}$ et $\mathtt{Pop}_{2,\overline{\mathtt{W}'}}$. Le paradoxe tient au fait que l'ordre des valeurs des probabilités pour ces deux sous-populations peut être inversé. En effet, nous obtenons que $\mathtt{P}_{\mathtt{Pop}_{1,\mathtt{W}'}}(\mathtt{C}) \leqslant \mathtt{P}_{\mathtt{Pop}_{2,\mathtt{W}'}}(\mathtt{C})$ et que $\mathtt{P}_{\mathtt{Pop}_{1,\overline{\mathtt{W}'}}}(\mathtt{C}) \leqslant \mathtt{P}_{\mathtt{Pop}_{2,\overline{\mathtt{W}'}}}(\mathtt{C})$. Un exemple de ce paradoxe est que si les taux de mortalité chez les femmes célibataires et les femmes mariées dans une ville $\mathtt{V}_1$ sont respectivement inférieurs à ceux d'une ville $\mathtt{V}_2$, il est néanmoins possible que le taux de mortalité des femmes dans la ville $\mathtt{V}_1$ soit supérieur à celui de la ville $\mathtt{V}_2$. Le paradoxe de Simpson rend délicat le choix de la partition des données en données d'apprentissage et en données d'expérience. Freitas [Freitas, 1998] utilise ce paradoxe comme un avantage et cherche les règles « surprenantes » en proposant un algorithme qui détecte les propriétés vérifiant le paradoxe de Simpson ;

3. La troisième catégorie concerne les approches par *utilisation de mesures de qualité.* Par exemple, dans l'approche présentée dans [Bayardo et Agrawal, 1999], le filtrage des règles porte sur l'utilisation de mesures de qualité associées aux règles. Cette approche de filtrage des règles est particulière au sens où le filtrage ne repose pas sur la structure de la règle, comme pour les travaux de *Klemettinen* et *Liu* cités dans la première catégorie. En effet, [Bayardo et Agrawal, 1999] définissent deux ordres partiels ($\leqslant_{\mathtt{sc}}$ et $\leqslant_{\mathtt{s\neg c}}$) sur les règles d'association en combinant les mesures de *support* et de *confiance* (*cf.* définitions du support et de la confiance d'une règle en § 3.2.3 page 51).

$$\mathtt{r}_1 \leqslant_{\mathtt{sc}} \mathtt{r}_2 \begin{cases} \mathtt{support}(\mathtt{r}_1) & \leqslant \mathtt{support}(\mathtt{r}_2) \\ \mathtt{confiance}(\mathtt{r}_1) & < \mathtt{confiance}(\mathtt{r}_2) \end{cases} \quad \text{ou} \quad \begin{cases} \mathtt{support}(\mathtt{r}_1) & < \mathtt{support}(\mathtt{r}_2) \\ \mathtt{confiance}(\mathtt{r}_1) & \leqslant \mathtt{confiance}(\mathtt{r}_2) \end{cases}$$

$$\mathtt{r}_1 \leqslant_{\mathtt{s\neg c}} \mathtt{r}_2 \begin{cases} \mathtt{support}(\mathtt{r}_1) & \leqslant \mathtt{support}(\mathtt{r}_2) \\ \mathtt{confiance}(\mathtt{r}_1) & > \mathtt{confiance}(\mathtt{r}_2) \end{cases} \quad \text{ou} \quad \begin{cases} \mathtt{support}(\mathtt{r}_1) & < \mathtt{support}(\mathtt{r}_2) \\ \mathtt{confiance}(\mathtt{r}_1) & \geqslant \mathtt{confiance}(\mathtt{r}_2) \end{cases}$$

Les règles qui satisfont ces deux ordres sont présentées comme les plus pertinentes parmi toutes les règles pour différentes mesures de qualité utilisées en FdD. Notre approche pour la gestion du nombre de règles d'association se place dans cette troisième catégorie. Une autre approche présentée dans [Lehn *et al.*, 2004] est fondée sur la détection des dépendances fonctionnelles que l'analyste n'a pas besoin de voir apparaître dans l'ensemble des règles. L'idée consiste à éliminer les règles d'association qui dénotent une dépendance fonctionnelle à l'image de la dépendance entre attributs dans une base de données. L'analyste peut inférer les règles éliminées – qui reflètent une dépendance fonctionnelle – en utilisant des propriétés de base de la logique propositionnelle ;

4. La quatrième catégorie rassemble des *techniques incrémentales* [Cheung *et al.*, 1996] permettant de générer les règles en ajoutant, un à un, les documents dans le corpus. Un critère de maintenance permet alors de délimiter les étapes où des modifications importantes dans

les règles sont apparues. À chaque étape, nous pouvons observer les nouvelles règles apparues par rapport à celles générées à l'étape précédente. Un algorithme incrémental de génération des règles d'association est, également, proposé dans [Godin et Missaoui, 1994]. Une technique que nous qualifions de « décrémentale » est présentée dans [Kuntz *et al.*, 2000]. L'analyste a la possibilité de sélectionner une sous-partie des règles grâce à un outil de visualisation de graphes d'implications entre les motifs. Lorsqu'un motif `M` particulier est sélectionné par l'analyste, un processus d'extraction de la sous-partie de règles d'implication dites « locales » est activé s'appuyant sur la prémisse `M`. Les autres motifs en `B` sont ignorés. Les techniques incrémentales n'ont pas retenu notre attention en raison du volume de données que nous avons manipulées dans nos expérimentations. De plus, l'analyse du contenu d'un corpus de textes ne nous semble pas liée à un ordre particulier dans l'ajout des textes dans la base de textes.

Les quatre sous-sections suivantes présentent des exemples de travaux qui introduisent des notions (extraction réduite des règles, connaissances de l'analyste, mesures de qualité) que nous manipulons dans la suite du mémoire. Les exemples de § 4.1.1 illustrent la catégorie 1. La section § 4.1.2 illustre la catégorie 2. La section § 4.1.3 illustre la catégorie 3 et nous finissons en § 4.1.4 par positionner notre approche dans la catégorie 3.

4.1.1 Approche par réduction du nombre de règles : deux exemples

Nous avons utilisé, dans une expérience sur des données textuelles agricoles dans [Toussaint *et al.*, 2000], l'approche proposée dans [Simon et Napoli, 1999] montrant qu'un sous-ensemble de règles d'association est obtenu en suivant les liens d'héritages directs entre concepts dans un treillis de Galois.

Nous distinguons pour chaque concept $\mathtt{C_i}$:

- L'ensemble $\mathtt{P_i}$ des *termes propres* du concept $\mathtt{C_i}$. $\mathtt{P_i} = \{\mathtt{t_i} \in \mathtt{C_i} \mid \forall \mathtt{C_j}\ ;\ \mathtt{C_i} \prec \mathtt{C_j}$ `alors` $\mathtt{t_i} \notin \mathtt{C_j}\}$ est l'ensemble des termes qui n'appartiennent pas aux intensions des concepts $\mathtt{C_j}$ qui subsument (*i.e.* sont plus généraux que) $\mathtt{C_i}$;
- L'ensemble $\mathtt{H_i}$ des *termes hérités* du concept $\mathtt{C_i}$. $\mathtt{H_i} = \{\mathtt{t_i} \in \mathtt{C_i} \mid \exists \mathtt{j}$ `tel que` $\mathtt{t_i} \in \mathtt{C_j}$ `et` $\mathtt{C_i} \prec \mathtt{C_j}\}$ est l'ensemble des termes qui appartiennent aux intensions des concepts $\mathtt{C_j}$ qui subsument $\mathtt{C_i}$.

Si $\mathtt{P_i} \neq \emptyset$, alors tout concept correspondant $\mathtt{C_i}$ du treillis de Galois permet de générer des règles d'association :

$$\begin{cases} \textit{exactes} \ : \ \mathtt{t_i} \Longrightarrow \mathtt{H_i}\ \mathtt{où}\ \mathtt{t_i} \in \mathtt{P_i} \\ \textit{approximatives} \ : \ \mathtt{t_i} \Longrightarrow \mathtt{P_i}\ \mathtt{où}\ \mathtt{t_i} \in \mathtt{H_i}\ \mathtt{avec}\ \mathtt{confiance} = \frac{|\mathtt{f(C_i)}|}{|\mathtt{f(C_j)}|} \end{cases}$$

`f` est la fonction `Extension` de la correspondance de Galois (*cf.* paragraphe 3.2.4.1 page 52).

Par exemple, le concept $\mathtt{C_6}$ du treillis de Galois représenté en figure (FIG. 4.1) possède un terme propre $\mathtt{P_6} = \{\mathtt{d}\}$ et deux termes hérités $\mathtt{H_6} = \{\mathtt{a}, \mathtt{c}\}$. De même, le concept $\mathtt{C_4}$ avec $\mathtt{P_4} = \{\mathtt{a}\}$ et $\mathtt{H_4} = \{\mathtt{c}\}$. Ainsi, $\mathtt{C_6}$ et $\mathtt{C_4}$ permettent de générer les règles suivantes :

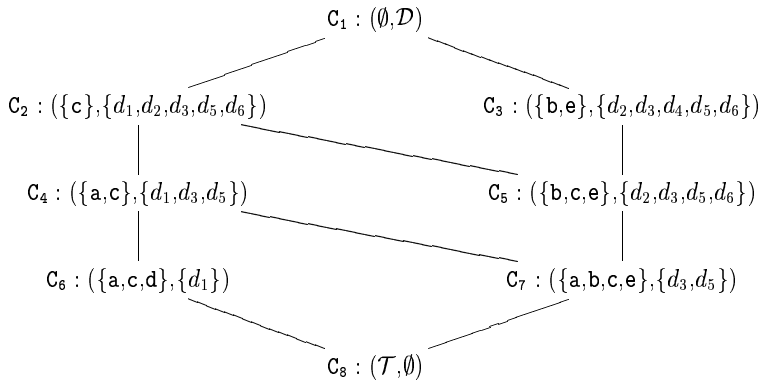


FIG. 4.1 – Treillis de Galois du tableau (TAB. 3.1).

$\mathtt{R_1} : (\mathtt{d} \Longrightarrow \mathtt{a}, \mathtt{c})$ règle exacte.
$\mathtt{R_2} : (\mathtt{a}, \mathtt{c} \Longrightarrow \mathtt{d})$ avec $\mathtt{confiance}(\mathtt{R_2}) = \frac{|\{d_1\}|}{|\{d_1,d_3,d_5\}|} = \frac{1}{3}$.
$\mathtt{R_3} : (\mathtt{a} \Longrightarrow \mathtt{c})$ règle exacte.
$\mathtt{R_4} : (\mathtt{c} \Longrightarrow \mathtt{a})$ avec $\mathtt{confiance}(\mathtt{R_4}) = \frac{|\{d_1,d_3,d_5\}|}{|\{d_1,d_2,d_3,d_5,d_6\}|} = \frac{3}{5}$.

Il est intéressant de noter que les règles exactes extraites par cette méthode sont un très petit sous-ensemble des règles d'association informatives exactes extraites en utilisant l'algorithme *Close*. Nous réduisons les 7 règles exactes extraites au $\mathtt{minsup} = \frac{2}{6}$ dans le tableau (TAB. 3.4) à 2 règles seulement (*i.e.* $\mathtt{R_1}$ et $\mathtt{R_3}$) sans support minimum.

En revanche, les règles approximatives reliant les motifs fermés entre eux ne sont pas présentes dans les règles informatives réduites que nous construisons (*cf.* figure (FIG. 3.5)), sauf lorsque le motif est, à la fois générateur et fermé, comme le motif ($\{\mathtt{c}\}$) correspondant au concept $\mathtt{C_2}$ du treillis de Galois qui permet de générer la règle $\mathtt{R_4} : \mathtt{c} \Longrightarrow \mathtt{a}$. Par exemple, la règle ($\mathtt{R_2} : \mathtt{a}, \mathtt{c} \Longrightarrow \mathtt{d}$) ne figure pas dans la liste des règles extraites même au seuil $\mathtt{minsup} = \frac{1}{6}$. Cette règle ne respecte pas le critère de minimalité de l'antécédent. Nous réduisons les 10 règles approximatives extraites au $\mathtt{minsup} = \frac{2}{6}$ dans le tableau (TAB. 3.4) à 2 règles seulement (*i.e.* $\mathtt{R_2}$ et $\mathtt{R_4}$) sans support minimum.

Bournaud et Courtine [Bournaud et Courtine, 2001] utilisent un treillis d'héritage « particulier » duquel sont ôtés les concepts d'*intensions* vides pour ne générer que les règles d'association exactes. Un treillis d'héritage est un treillis de Galois dans lequel n'apparaissent que les concepts propres, c'est-à-dire les couples d'intension et d'extension propres. La relation de subsomption permet de revenir au treillis de Galois initial. En enlevant au treillis d'héritage (*cf.* figure (FIG. 4.2 à gauche)) les concepts $\mathtt{C_5}$ et $\mathtt{C_7}$ (*cf.* figure (FIG. 4.2 à droite)), nous pouvons générer, en s'appuyant sur le même principe que [Toussaint *et al.*, 2000] pour les propriétés propres et héritées : ($\mathtt{t_i} \Longrightarrow \mathtt{H_i}$ où $\mathtt{t_i} \in \mathtt{P_i}$), les deux mêmes règles exactes $\mathtt{R_1}$ et $\mathtt{R_3}$ suivantes :

$\mathtt{R_1} : \mathtt{d} \Longrightarrow \mathtt{a}, \mathtt{c}$
$\mathtt{R_3} : \mathtt{a} \Longrightarrow \mathtt{c}$

Nous devons donc faire un compromis entre le nombre de règles à générer et l'interprétabilité des règles extraites.

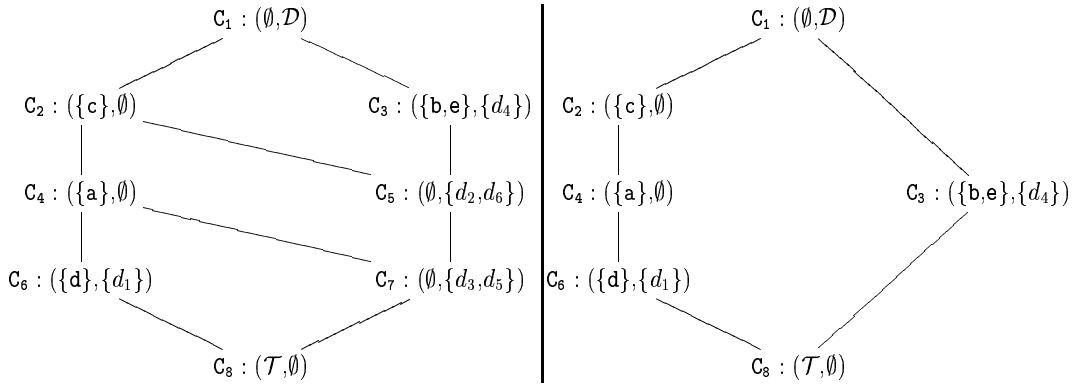


FIG. 4.2 – Treillis d'héritage (à gauche) et Espace de généralisation (à droite) du tableau (TAB. 3.1).

4.1.2 Approche par utilisation des connaissances de l'analyste

Les principaux travaux pour illustrer cette approche pour l'analyse des règles d'association sont décrits par S. Sahar et par B. Liu.

Dans [Sahar, 1999], les règles d'association minimales en B et en H sont générées de façon incrémentale. L'analyste est mis à contribution pour classer les règles d'association selon quatre points de vue :

- Les règles vraies et non intéressantes : ce sont des règles dont les parties B et H ont une signification triviale. Par exemple "époux" $\Longrightarrow$ "marié". Lorsqu'une règle de ce type est rencontrée par l'analyste, il faut la garder et ne pas générer la famille de règles qui possèdent B' et H' tels que B' est un sous-motif de B (par rapport à l'ordre d'inclusion $B' \subsetneq B$) et H' est un sous-motif de H ;
- Les règles fausses et intéressantes : ce sont des règles dont la signification est fausse ("homme" $\Longrightarrow$ "marié"), mais qui risquent de servir de sous-règle intéressante. Une sous-règle d'une règle est définie selon l'ordre d'inclusion des deux motifs : $B' \sqcap H' \subsetneq B \sqcap H$; par exemple, ("homme", "possède véhicule 4x4" $\Longrightarrow$ marié). Les règles de cette catégorie sont toutes présentées à l'analyste.
- Les règles fausses et non intéressantes : ce sont des règles dont la signification est fausse ("salaire élevé" $\Longrightarrow$ "marié") mais dont la connaissance lorsqu'elle est augmentée d'autres termes en parties B et H n'intéresse pas l'analyste ;
- Les règles vraies et intéressantes : ce sont les règles idéales du point de vue de l'analyste. Toutes les sur-règles $B' \Longrightarrow H'$ (telles que $B' \sqcap H' \supsetneq B \sqcap H$) de cette règle sont générées.

L'algorithme propose la règle candidate à l'analyste qui la classe dans une des quatre catégories et la famille de règles proches est validée ou rejetée automatiquement de la base de connaissances.

La critique que nous formulons, à propos de cette méthode, est la perte potentielle de sur-règles intéressantes sans prendre de précautions particulières. Par exemple si nous ignorons toute une famille de sous-règles d'une règle car elle est triviale, nous pouvons perdre une association potentiellement intéressante entre d'une condition minimale (B) impliquant la conclusion jugée triviale dans le point de vue : règles vraies et non intéressantes. Les règles d'association sont

générées en utilisant l'algorithme *Apriori* [Agrawal et Srikant, 1994] — toutes les combinaisons d'associations possibles sont alors générées — puis ces règles et leurs familles correspondantes sont classées dans les quatre catégories *a posteriori.* Il est vrai que c'est un moyen de réduire le trop grand nombre de règles. D'ailleurs, nous utilisons une approche similaire dans notre processus de FdT.

Dans [Liu *et al.*, 1997; Liu *et al.*, 1999b], la recherche de motifs qui décrivent la structure de règles intéressantes s'appuie sur les connaissances générales de l'analyste. Ces connaissances appelées *general impressions* sont vagues et imprécises mais permettent de décrire la structure (*i.e.* le patron) des règles intéressantes suivant les indications de l'analyste. Les règles sont classées selon leur divergence par rapport aux connaissances de l'analyste selon quatre critères :

- Les règles d'association conformes aux connaissances du domaine (`confirm`) ;
- Les règles d'association dont la partie `B` est conforme et la partie `H` non conforme, c'est-à-dire, qui sont surprenantes (`unexpConseq`) ;
- Les règles d'association dont la partie `B` est surprenante et la partie `H` conforme (`unexpCond`) ;
- Les règles d'association dont les deux parties `B` et `H` sont surprenantes (`bothsideUnexp`).

Une mesure de divergence quantifiant le degré de termes différents entre la règle à analyser `i` et la règle de référence `j` est associée à chaque critère. La divergence est mesurée en partie gauche par $\mathtt{L_{ij}}$ et en partie droite par $\mathtt{R_{ij}}$. Les quatre mesures heuristiques suivantes sont utilisées :

$$\mathtt{m_1{=}confirm_{ij}} = \mathtt{L_{ij}} \times \mathtt{R_{ij}} \qquad \mathtt{m_2{=}unexpConseq_{ij}}{=}\begin{cases} 0 & \mathtt{L_{ij}} - \mathtt{R_{ij}} \leqslant 0 \\ \mathtt{L_{ij}} - \mathtt{R_{ij}} & \mathtt{L_{ij}} - \mathtt{R_{ij}} > 0 \end{cases}$$

$$\mathtt{m_3{=}unexpCond_{ij}}{=}\begin{cases} 0 & \mathtt{R_{ij}} - \mathtt{L_{ij}} \leqslant 0 \\ \mathtt{R_{ij}} - \mathtt{L_{ij}} & \mathtt{R_{ij}} - \mathtt{L_{ij}} > 0 \end{cases} \qquad \mathtt{bothsideUnexp_{ij}} = 1 - \max(\mathtt{m_1}, \mathtt{m_2}, \mathtt{m_3})$$

Le procédé est identique dans [Liu *et al.*, 1999a]. La recherche des règles particulières dont la partie `H` est un motif à un seul terme `y`. C'est une règle de référence notée ($\emptyset \Longrightarrow \mathtt{y}$). Le test du χ^2 affecte une des valeurs de direction parmi $\{-1, 0, 1\}$ à la règle $\mathtt{x_1} \Longrightarrow \mathtt{y}$, et, de façon incrémentale, nous pouvons calculer les directions de ($\mathtt{x_1}, \mathtt{x_2} \Longrightarrow \mathtt{y}$), ..., ($\mathtt{x_1}, \ldots, \mathtt{x_k} \Longrightarrow \mathtt{y}$) sachant les directions des ($\mathtt{k} - 1$) sous-règles précédentes. Un changement de direction (de -1 vers 0) dénote un apport de connaissance de cette règle d'association. Dans [Subramonian, 1998], une expérience est menée jusqu'à construire une base de connaissances en classant les règles extraites une à une. Aucune opérationnalisation du processus n'est proposée dans ce travail.

Une autre approche utilisant des connaissances du domaine organisées dans une structure de treillis de Galois est présentée dans [Ganter, 1999]. Les connaissances sont utilisées pour supprimer les règles d'association qui violent certaines contraintes. Pour cela, des opérateurs de la logique propositionnelle sont définis. Par exemple, un opérateur de disjonction exclusive : si les concepts `homme` et `femme` sont *subsumés* par le concept `individu`, alors les règles `homme` $\Longrightarrow$ `femme` et `femme` $\Longrightarrow$ `homme` seront rejetées.

4.1.3 Approche par utilisation de mesures de qualité

Le processus de FdD génère des combinaisons de règles telles que la présence de certaines règles peut en rendre d'autres redondantes. Une des principales difficultés de l'analyse des résultats d'un processus de fouille de textes est de trouver de « bonnes » mesures de qualité des règles d'association extraites. Une mesure de qualité doit être indépendante du domaine de fouille décrit

par les données. Une mesure de qualité permet d'ordonner les règles selon un critère et de présenter, en premier lieu, les règles les plus pertinentes, c'est-à-dire des règles qui, potentiellement, présenteraient un intérêt pour l'analyste.

Les mesures de qualité sont de deux types : les mesures objectives et subjectives.

- Les mesures objectives sont dites *dirigées par les données* car ces mesures concernent la structure du motif correspondant et la nature des données à traiter ;
- Les mesures subjectives ont dites *dirigées par l'analyse* car ces mesures concernent l'objectif de fouille, c'est-à-dire que ces mesures caractérisent la classe des motifs que recherche l'analyste [Silberschatz et Tuzhilin, 1996] ;
- Les mesures heuristiques combinées entre objectivité et subjectivité [Shah *et al.*, 1999; Hussain *et al.*, 2000; Padmanabhan et Tuzhilin, 2000].

Propriétés requises pour une mesure de qualité Une mesure de qualité $\mathtt{M}$ d'une règle d'association doit, idéalement, posséder les quatre propriétés objectives suivantes :

Propriétés 4.1 (Propriétés objectives)

(O1) $\mathtt{M} = \mathtt{0}$ *ou égale à toute autre situation de référence. La situation de référence est une valeur particulière d'une mesure de qualité, par exemple, la situation d'*indépendance *;*

(O2) Si $\mathtt{f}(\mathtt{B} \sqcap \mathtt{H})$ *croît en même temps que* $\mathtt{f}(\mathtt{B})$ *et que* $\mathtt{f}(\mathtt{H})$ *est constant, alors* $\mathtt{M}$ *est croissante ;*

(O3) Si $\mathtt{f}(\mathtt{B})$ *décroît en même temps que* $\mathtt{f}(\mathtt{H})$ *et que* $\mathtt{f}(\mathtt{B} \sqcap \mathtt{H})$ *est constant, alors* $\mathtt{M}$ *est décroissante ;*

(O4) Si $\mathtt{f}(\mathtt{H})$ *décroît en même temps que* $\mathtt{f}(\mathtt{B})$ *et que* $\mathtt{f}(\mathtt{B} \sqcap \mathtt{H})$ *est constant, alors* $\mathtt{M}$ *est décroissante.*

D'autres caractéristiques souhaitables d'une mesure de qualité pour les règles d'association (incompatibilité, répulsion, indépendance, attraction, implication) sont décrites dans ([Guillaume, 2000], chapitre 2).

Dans [Silberschatz et Tuzhilin, 1996] des mesures « subjectives », du point de vue de l'analyste, permettent de mesurer la qualité des règles d'association extraites par le processus de fouille que des mesures objectives ne trouvent pas pertinentes. L'idée est de s'appuyer sur le jugement de l'analyste pour préférer une bonne règle mesurée de façon approximative à une mauvaise règle mesurée de façon exacte. Deux mesures subjectives sont proposées à l'analyste :

Propriétés 4.2 (Propriétés subjectives)

(S1) Utilité (actionability) : Une règle est utile ($\mathtt{i}$*) si elle est constitue un cas particulier par rapport à un ensemble d'autres règles proches et (*$\mathtt{ii}$*) si elle peut être transformée par l'analyste en une autre règle plus intéressante et donner lieu à une prise une décision de l'analyste dans sa tâche ou pour son domaine [Piatetsky-Shapiro et Matheus, 1994] ;*

(S2) Surprise : si la règle surprend l'analyste en contredisant les connaissances du domaine et qu'il ne peut pas la rattacher à une règle plausible.

Les mesures (S1) et (S2) sont liées. La recherche des règles surprenantes est subordonnée au non rattachement à une règle plausible. Un ensemble de règles attestées représente les croyances de

l'analyste et constituent les règles de référence. Une règle paraît surprenante si la fréquence d'apparition de ses termes dans le corpus dévie sensiblement de la fréquence d'apparition des termes d'une règle, proche, dite de référence. La règle de référence est une règle connue de l'analyste et attestée dans le domaine de spécialité. La règle de référence est considérée comme proche au sens où son motif diffère de quelques termes par rapport au motif de la règle surprenante extraite. Cependant, les mesures subjectives requièrent l'avis de l'analyste pour chaque règle à rattacher. Une mesure subjective demeure intéressante pour définir une méthodologie d'analyse manuelle de l'ensemble des règles d'association.

Pour notre part, nous nous intéressons aux mesures de qualité dites *objectives* associées aux règles d'association. Notre application entre dans le cadre de l'apprentissage non supervisé. L'intervention de l'expert consiste à vérifier l'adéquation, sans autre préalable, des règles que nous lui présentons par rapport à ses connaissances. l'approche que nous choisissons pour traiter le problème détaille le processus de calcul des mesures de qualité que nous utilisons pour classer les règles d'association.

4.1.4 Notre approche de l'utilisation de mesures de qualité

Les approches que nous avons présentées, hormis celle de [Bayardo et Agrawal, 1999], ont un point commun : la réduction du nombre de règles dites moins informatives ou d'autres qui sont jugées redondantes. L'approche que nous proposons consiste à conserver les règles informatives réduites (*cf.* § 3.2.4.4 page 58). En effet, dans une approche d'apprentissage non supervisée de règles d'association, il n'est pas possible de préjuger de celles qui seront, au final, retenues par l'analyste. Nous cherchons à aider l'analyste dans la lecture et l'interprétation de ces règles en les triant suivant des valeurs données par des mesures de qualité. Nous utilisons ces mesures pour construire des « points de vue » complémentaires sur l'ensemble des règles. Nous suggérons un moyen de sélectionner, parmi ces règles, celles qui présentent un intérêt particulier pour l'analyste. Pour cela, nous procédons en deux étapes :

1. Nous calculons des mesures de qualité associées à chacune des règles qui proposent à l'analyste une classification des règles et une sélection de celles qui semblent pertinentes selon les valeurs données par les mesures de qualité ;

2. L'analyste identifie un sous-ensemble de règles qui présente un intérêt particulier par rapport à ses besoins et à ses connaissances en visualisant le classement que nous lui fournissons. L'analyste valide certaines règles et en rejette d'autres. Le but de notre processus de sélection est que le sous-ensemble de règles d'association identifié par l'analyste soit présent en tête de liste.

Nous n'exigeons pas un ordre strict entre les règles identifiées par l'analyste comme étant intéressantes et celles qui ne le sont pas. L'analyste peut trouver, parmi les « pépites » de connaissances potentielles que nous lui proposons, certaines qui ne sont pas intéressantes. Néanmoins, le sous-ensembe qu'il identifie doit être *présent*[24] dans le classement, par les mesures de qualité, que nous proposons pour l'ensemble des règles d'association extraites.

[24] Le terme approprié (mais flou) serait *majoritairement présent.*

4.2 Mesures de qualité des règles d'association

Soit $\mathcal{D}(\mathtt{B})$, $\mathcal{D}(\mathtt{H})$ et $(\mathcal{D}(\mathtt{B} \sqcap \mathtt{H}) = \mathcal{D}(\mathtt{B}) \cap \mathcal{D}(\mathtt{H}))$ les ensembles de textes de $\mathcal{D}$ qui possèdent respectivement tous les termes des motifs $\mathtt{B}$, $\mathtt{H}$ et $\mathtt{B} \sqcap \mathtt{H}$ d'une règle d'association (*cf.* figure (FIG. 4.3)).

L'ensemble $\mathcal{D}(\mathtt{X})$ correspond à $\mathtt{f}(\mathtt{T})$ de la correspondance de Galois définie en § 3.2.4.1. $|\mathcal{D}(\mathtt{X})|$ dénote le cardinal de l'intersection des images des motifs $\mathtt{B}$ et $\mathtt{H}$ (*cf.* § 3.2.3).

Soit $\mathtt{P}(\mathtt{X})$ la probabilité de l'événement : « observer le motif $\mathtt{X}$ dans l'ensemble des textes ». Cette interprétation rejoint le formalisme développé dans [Guillaume, 2000]. La probabilité $\mathtt{P}(\mathtt{X})$ est définie par :

$$0 \leqslant \left(\mathtt{P}(\mathtt{X}) = \frac{|\mathcal{D}(\mathtt{X})|}{|\mathcal{D}|}\right) \leqslant 1 \tag{4.1}$$

Trois valeurs de probabilités des motifs $\mathtt{B}$, $\mathtt{H}$ et $\mathtt{B} \sqcap \mathtt{H}$ ont un impact sur la valeur des mesures que nous utilisons. Il s'agit respectivement de : $\mathtt{P}(\mathtt{B})$, $\mathtt{P}(\mathtt{H})$ et $\mathtt{P}(\mathtt{B} \cap \mathtt{H})$ [Cherfi et Toussaint, 2002a; Cherfi et Toussaint, 2002b].

Nous rappelons que la probabilité du motif $\mathtt{B} \sqcap \mathtt{H}$ est égale à $\mathtt{P}(\mathtt{B} \cap \mathtt{H})$ (*cf.* § 3.2.3). Cette probabilité signifie la probabilité d'avoir simultanément tous les termes de la partie $\mathtt{B}$ et de la partie $\mathtt{H}$.

$$\mathtt{P}(\mathtt{B} \cap \mathtt{H}) = \frac{|\mathcal{D}(\mathtt{B}) \cap \mathcal{D}(\mathtt{H})|}{|\mathcal{D}|}$$

4.2.1 Situation de référence

Une situation de référence est une valeur fixée d'une mesure de qualité qui permet de caractériser une situation particulière pour les données mesurées en termes d'événements : « cooccurrence des motifs $\mathtt{B}$ et $\mathtt{H}$ dans les textes du corpus ». Deux situations de référence sont mesurables : les cas d'indépendance et d'incompatibilité.

Définition 4.1 (Événements indépendants) *Les motifs* $\mathtt{B}$ *et* $\mathtt{H}$ *sont dits indépendants si la fréquence d'apparition du motif* $\mathtt{H}$ *dans un texte ne dépend pas de la présence ou de l'absence du motif* $\mathtt{B}$ *dans ce texte. Si les motifs* $\mathtt{B}$ *et* $\mathtt{H}$ *sont considérés comme des événements, alors le cas d'indépendance de deux événements* $\mathtt{B}$ *et* $\mathtt{H}$ *dénote que la probabilité d'avoir l'événement* $\mathtt{H}$ *n'est pas influencée par le fait d'avoir préalablement l'événement* $\mathtt{B}$.

Nous considérons que deux motifs $\mathtt{B}$ *et* $\mathtt{H}$ *sont indépendants si et seulement si :*

$$\mathtt{P}(\mathtt{H}|\mathtt{B}) = \tfrac{\mathtt{P}(\mathtt{B} \cap \mathtt{H})}{\mathtt{P}(\mathtt{B})} = \mathtt{P}(\mathtt{H}), \quad \textit{c'est-à-dire } \mathtt{P}(\mathtt{B} \cap \mathtt{H}) = \mathtt{P}(\mathtt{B}) \times \mathtt{P}(\mathtt{H})$$

Définition 4.2 (Événements incompatibles) *Le cas d'incompatibilité de deux événements* $\mathtt{B}$ *et* $\mathtt{H}$ *signifie que nous ne trouvons jamais (ou très rarement) ces deux motifs présents simultanément dans les textes.*

Nous considérons que deux motifs $\mathtt{B}$ *et* $\mathtt{H}$ *sont incompatibles si et seulement si :*

$$\mathtt{P}(\mathtt{B} \cap \mathtt{H}) \approx 0$$

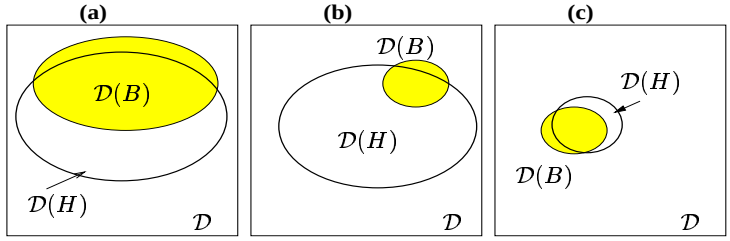


FIG. 4.3 – Principaux cas illustrant les variations de $\mathcal{D}(\mathtt{B})$ et $\mathcal{D}(\mathtt{H})$ – $\mathcal{D}$ est l'espace représentant l'ensemble des textes du corpus.

4.2.2 Cas de distribution des termes dans les textes

Plus $\mathcal{D}(\mathtt{X})$ est grand (*i.e.* plus $\mathcal{D}(\mathtt{X})$ couvre l'espace $\mathcal{D}$), plus $\mathtt{P}(\mathtt{X})$ est proche de $\mathtt{1}$. Le motif $\mathtt{X}$ est très fréquent et décrit presque tous les textes. Par conséquent, les connaissances potentiellement apportées par ce motif et par la règle sous-jacente, du point de vue de l'extraction de connaissances par l'analyste, sont considérées comme des connaissances non informatives ou triviales. La figure (FIG.4.3) représente, en particulier, trois principales distributions des termes $\mathtt{B}$ et $\mathtt{H}$ qui nous intéressent :

Cas (a) – $\mathtt{P}(\mathtt{B})$ et $\mathtt{P}(\mathtt{H})$ sont toutes deux proches de $\mathtt{1}$ dans ce cas. Les règles du cas (a) sont considérées comme les moins informatives. Un ensemble de termes présent dans presque tous les textes impliquera, très probablement, un autre ensemble présent dans tous les textes. Il y a de grandes chances que ces termes désignent des *concepts* génériques du domaine. Par exemple, deux termes très répandus qui ont permis de sélectionner les textes du corpus d'expérience comme « mutation » et « résistance » ne donnent aucune information s'ils constituent la règle (« mutation » $\Longrightarrow$ « résistance ») ;

Cas (b) – comme $\mathtt{P}(\mathtt{B})$ est plus faible, le cas (b) paraît, en ce sens, plus intéressant. L'inconvénient est que tout texte qui possède $\mathtt{B}$ aura tendance à posséder $\mathtt{H}$;

Cas (c) – ce cas est le plus intéressant. Les termes y sont rares et apparaissent presque à chaque fois ensemble (*i.e.* $\mathtt{P}(\mathtt{B} \cap \mathtt{H}) \simeq \mathtt{P}(\mathtt{B}) \simeq \mathtt{P}(\mathtt{H})$). Ces termes sont donc vraisemblablement reliés dans un contexte du domaine ;

Les algorithmes de fouille de données favorisent le cas de la figure (FIG. 4.3-(a)) car, à l'origine, ces algorithmes ont pour but de rechercher de motifs fréquents dans des bases de données d'articles du *panier de la ménagère* afin de trouver quelles marchandises sont achetées conjointement dans les supermarchés. Par exemple $\mathtt{bière}, \mathtt{saucisses} \Longrightarrow \mathtt{chips}, \mathtt{moutarde}$. Nous considérons que la classe de motifs à chercher est dirigée par la tâche fixée par l'analyste pour la fouille dans les textes techniques et que ce n'est pas forcément les motifs trouvés par les algorithmes classiques (ceux correspondant au cas (a)) qu'il faut chercher. En ce sens, nous rejoignons la réflexion de [Freitas, 1998].

Nous nous intéressons aux règles qui reflètent les motifs de la figure (FIG. 4.3-(c)) car ils sont porteurs d'une connaissance très peu présente dans les textes et potentiellement utiles pour l'analyste. Cette connaissance est appelée « pépite de connaissance » dans [Azé, 2003].

4.2.3 Mesures de support et de confiance

Nous réécrivons les mesures de *support* et de *confiance* des règles d'association en utilisant la formule (4.1) par :

- La mesure de support d'une règle d'association est la probabilité conjointe $\mathtt{P}(\mathtt{B} \cap \mathtt{H})$. Le support est la probabilité que $\mathtt{B}$ et $\mathtt{H}$ soient à *vrais* en même temps. Dans la pratique, le support mesure l'intersection des deux ensembles $\mathcal{D}(\mathtt{B})$ et $\mathcal{D}(\mathtt{H})$;
- La mesure de confiance d'une règle d'association est la probabilité conditionnelle de trouver $\mathtt{H}$ sachant $\mathtt{B}$, soit $\mathtt{P}(\mathtt{H}|\mathtt{B}) = \frac{\mathtt{P}(\mathtt{B} \cap \mathtt{H})}{\mathtt{P}(\mathtt{B})}$. Dans la pratique, la confiance mesure la surface d'inclusion de l'ensemble $\mathcal{D}(\mathtt{B})$ dans $\mathcal{D}(\mathtt{H})$.

Critique des mesures de support et de confiance Les mesures de support et de confiance ne différencient pas les cas (a), (b) et (c) de la figure (FIG. 4.3). Le support représente l'intersection $\mathcal{D}(\mathtt{B}) \cap \mathcal{D}(\mathtt{H})$, il peut alors distinguer (a) de (b) et de (c) mais ne peut pas distinguer (b) de (c). La confiance représente la surface d'inclusion de $\mathcal{D}(\mathtt{B})$ dans $\mathcal{D}(\mathtt{H})$ et n'est pas un facteur discriminant de ces trois cas.

Pour les raisons ci-dessus, les mesures de support et de confiance ne sont pas suffisantes, à elles seules, pour identifier les cas du plus significatif (c) vers le moins significatif (a). Leurs caractéristiques statistiques ne reflètent pas la significativité de la règle. Le paragraphe suivant montre que d'autres *mesures de qualité* sont capables de différencier les trois cas possibles de la figure (FIG. 4.3).

4.2.4 Autres mesures de qualité des règles

Nous présentons d'autres mesures de qualité qui permettent différents classements des règles d'association. Les mesures que nous présentons constituent des mesures classiques en fouille de données et sont synthétisées dans [Lavrač *et al.*, 1999; Guillaume, 2000; Guillet, 2004]. Ces mesures de qualité sont des transformations de la mesure de confiance $\mathtt{P}(\mathtt{H}|\mathtt{B})$ qui permettent de la comparer à $\mathtt{P}(\mathtt{H})$ [Tan *et al.*, 2002; Lenca *et al.*, 2003]. La comparaison se fait en centrant la confiance sur $\mathtt{P}(\mathtt{H})$ avec différents coefficients d'échelle ou bien en divisant par $\mathtt{P}(\mathtt{H})$.

4.2.4.1 L'intérêt

L'**intérêt** [IBM, 1998] (ou *lift*) mesure la déviation du support de la règle par rapport au cas d'indépendance. La valeur de l'intérêt est donnée par :

$$\mathtt{int}\,[\mathtt{B} \implies \mathtt{H}] = \frac{\mathtt{P}(\mathtt{B} \cap \mathtt{H})}{\mathtt{P}(\mathtt{B}) \times \mathtt{P}(\mathtt{H})} \tag{4.2}$$

L'intérêt varie dans l'intervalle $[0, +\infty[$. Cette mesure dénote une indépendance de $\mathtt{B}$ et $\mathtt{H}$ si l'intérêt vaut 1. Plus $\mathtt{B}$ et $\mathtt{H}$ sont incompatibles, plus $\mathtt{P}(\mathtt{B} \cap \mathtt{H})$ tend vers 0, et donc plus l'intérêt est proche de 0. Plus $\mathtt{B}$ et $\mathtt{H}$ sont dépendants, plus l'*intérêt* est supérieur à 1.

Par définition, on a $(\mathcal{D}(\mathtt{B}) \cap \mathcal{D}(\mathtt{H}) \subseteq \mathcal{D}(\mathtt{B})$ et $(\mathcal{D}(\mathtt{B}) \cap \mathcal{D}(\mathtt{H})) \subseteq \mathcal{D}(\mathtt{H})$. Plus $\mathcal{D}(\mathtt{B})$ et $\mathcal{D}(\mathtt{H})$ sont petits dans $\mathcal{D}$ et donc sont proches de leur intersection, plus la valeur de l'intérêt augmente. Si $\mathtt{P}(\mathtt{B} \cap \mathtt{H}) \simeq \mathtt{P}(\mathtt{B})$ alors $\mathtt{int}\,[\mathtt{B} \implies \mathtt{H}] \simeq \frac{\mathtt{P}(\mathtt{B})}{\mathtt{P}(\mathtt{B}) \times \mathtt{P}(\mathtt{H})} = \frac{1}{\mathtt{P}(\mathtt{H})}$, de la même manière lorsque

$\mathtt{P}(\mathtt{B} \cap \mathtt{H}) \simeq \mathtt{P}(\mathtt{H})$, nous avons $\mathtt{int}\,[\mathtt{B} \Longrightarrow \mathtt{H}] = \frac{1}{\mathtt{P}(\mathtt{B})}$. Quand $\mathtt{P}(\mathtt{B})$ ou $\mathtt{P}(\mathtt{H})$ tendent vers 0, l'intérêt augmente. Par conséquent, les règles qui se trouvent dans le « bon » cas (c) sont classées en premier. Enfin, l'intérêt est une mesure symétrique $\mathtt{int}\,[\mathtt{B} \Longrightarrow \mathtt{H}] = \mathtt{int}\,[\mathtt{H} \Longrightarrow \mathtt{B}]$.

4.2.4.2 La conviction

La **conviction** [Brin *et al.*, 1997] privilégie les contre-exemples à une règle. Elle mesure la déviation du support du contre-exemple à la règle dû au motif $\mathtt{B} \sqcap \neg\mathtt{H}$ par rapport à l'indépendance de $\mathtt{B}$ et $\neg\mathtt{H}$. Dans notre contexte, $\neg\mathtt{H}$ signifie l'absence d'au moins un terme du motif dans au moins un texte de $\mathcal{D}(H)$. $|\mathcal{D}(\neg\mathtt{H})| = |\mathcal{D}| - |\mathcal{D}(\mathtt{H})|$ et donc $\mathtt{P}(\neg\mathtt{H}) = 1 - \mathtt{P}(\mathtt{H})$.

$$\mathtt{conv}\,[\mathtt{B} \implies \mathtt{H}] = \frac{\mathtt{P}(\mathtt{B}) \times \mathtt{P}(\neg\mathtt{H})}{\mathtt{P}(\mathtt{B} \cap \neg\mathtt{H})} \tag{4.3}$$

La conviction vaut $\left(\frac{1}{\mathtt{int}[\mathtt{B} \Longrightarrow \neg\mathtt{H}]}\right)$, n'est pas symétrique, et mesure la validité de la direction de l'implication (de $\mathtt{B}$ vers $\mathtt{H}$) pour les contre-exemples.

La valeur de conviction augmente lorsque $1 - \mathtt{P}(\mathtt{H})$ est faible (*i.e.* $\mathtt{P}(\mathtt{H})$ élevé), $\mathtt{P}(\mathtt{B})$ est élevé et lorsque $\mathtt{P}(\mathtt{B} \cap \mathtt{H}) \simeq \mathtt{P}(\mathtt{B})$ car $\mathtt{P}(\mathtt{B}) = \mathtt{P}(\mathtt{B} \cap \mathtt{H}) + \mathtt{P}(\mathtt{B} \cap \neg\mathtt{H})$. Ce qui classe les règles du cas (c) en premier.

Comme l'intérêt, cette mesure varie dans l'intervalle $[0, +\infty[$. La mesure de conviction dénote une dépendance entre $\mathtt{B}$ et $\mathtt{H}$ si sa valeur est > 1, une indépendance si elle est $= 1$ et pas de dépendance si sa valeur est comprise dans $[0, 1[$. La mesure de conviction n'est pas calculable pour les règles exactes puisque nous ne pouvons diviser par $\mathtt{P}(\mathtt{B} \cap \neg\mathtt{H})$ qui vaut 0, car il n'y a aucun contre-exemples à la règle.

4.2.4.3 La dépendance

La mesure de **dépendance** est utilisée pour mesurer une distance de la confiance de la règle par rapport au cas d'indépendance de $\mathtt{B}$ et $\mathtt{H}$.

$$\mathtt{dep}\,[\mathtt{B} \implies \mathtt{H}] = |\mathtt{P}(\mathtt{H}|\mathtt{B}) - \mathtt{P}(\mathtt{H})| \tag{4.4}$$

Cette mesure varie dans l'intervalle $[0, 1[$ car c'est une valeur positive. Plus la valeur de la dépendance est proche de 0 (resp.1) plus $\mathtt{B}$ et $\mathtt{H}$ sont indépendants (resp. dépendants). Ce qui augmente le plus sa valeur est la taille de $\mathcal{D}(\mathtt{H})$. Nous obtenons alors des valeurs sensiblement égales pour les cas (a) et (b). C'est particulièrement notable pour les règles exactes où la confiance $\mathtt{P}(\mathtt{H}|\mathtt{B})$ vaut 1 et donc $\mathtt{dep}\,[\mathtt{B} \implies \mathtt{H}] = 1 - \mathtt{P}(\mathtt{H})$ ne dépend pas de $\mathtt{P}(\mathtt{B})$. Par conséquent, la dépendance permet de séparer les règles du cas (c) des règles du cas (a) et du cas (b).

Pour cette raison, les deux mesures suivantes qui représentent également des dépendances sont définies.

4.2.4.4 La nouveauté et la satisfaction

La mesure de **nouveauté** [Piatetsky-Shapiro, 1991] est définie par :

$$\mathtt{nov}\,[\mathtt{B} \Longrightarrow \mathtt{H}] = \mathtt{P}(\mathtt{B} \cap \mathtt{H}) - \mathtt{P}(\mathtt{B}) \times \mathtt{P}(\mathtt{H}) \tag{4.5}$$

La valeur absolue de cette mesure vaut ($\mathtt{dep}\,[\mathtt{B} \implies \mathtt{H}] \times \mathtt{P}(\mathtt{B})$). Plus $\mathtt{P}(\mathtt{B})$ est faible, plus la valeur de la nouveauté est faible. Ainsi, les règles du cas (b) sont rejetées en fin de classement et sont différenciées du cas (a), alors que la dépendance ne le fait pas.

Nous nous intéressons aux valeurs faibles de de cette mesure (*i.e.* autour de la valeur d'indépendance 0). La nouveauté varie entre $]-1, 1[$ et prend une valeur négative lorsque $\mathtt{P}(\mathtt{B} \cap \mathtt{H}) < \mathtt{P}(\mathtt{B}) \times \mathtt{P}(\mathtt{H})$. La nouveauté s'approche de -1 pour des règles de faibles supports $\mathtt{P}(\mathtt{B} \cap \mathtt{H}) \simeq 0$.

Il faut souligner un paradoxe pour la mesure de nouveauté par rapport à la dépendance. La mesure de nouveauté est pour certains cas contre-intuitive. Par exemple, dans le cas (a), la dépendance entre les événements B et H est forte puisque l'ensemble des textes ayant B permet de déduire l'ensemble des textes ayant H. Cependant nous avons bien $\mathtt{P}(\mathtt{B}) \times \mathtt{P}(\mathtt{H}) \approx \mathtt{P}(\mathtt{B} \cap \mathtt{H}) \approx 1$. Une règle du cas (a) semble apporter une nouveauté même si les termes qui y apparaissent sont dépendants dans le corpus.
La nouveauté est symétrique (c'est-à-dire, $\mathtt{nov}(\mathtt{B} \implies \mathtt{H}) = \mathtt{nov}(\mathtt{H} \implies \mathtt{B})$) alors que l'une peut avoir plus de contre-exemples que l'autre. Pour cette raison, nous utilisons la mesure suivante appelée **satisfaction** :

$$\mathtt{sat}\,[\mathtt{B} \implies \mathtt{H}] = \frac{(\mathtt{P}(\neg\mathtt{H}) - \mathtt{P}(\neg\mathtt{H} \mid \mathtt{B}))}{\mathtt{P}(\neg\mathtt{H})} \tag{4.6}$$

qui s'écrit également : $|\mathtt{sat}\,[\mathtt{B} \implies \mathtt{H}]| = \frac{\mathtt{P}(\mathtt{H}|\mathtt{B}) - \mathtt{P}(\mathtt{H})}{1 - \mathtt{P}(\mathtt{H})} = \frac{\mathtt{dep}\,[\mathtt{B} \implies \mathtt{H}]}{\mathtt{P}(\neg\mathtt{H})}$ car $\mathtt{P}(\neg\mathtt{H}) - \mathtt{P}(\neg\mathtt{H}|\mathtt{B}) = (1 - \mathtt{P}(\mathtt{H})) - (1 - \mathtt{P}(\mathtt{H}|\mathtt{B})) = \mathtt{P}(\mathtt{H}|\mathtt{B}) - \mathtt{P}(\mathtt{H})$, avec $\mathtt{P}(\mathtt{H}|\mathtt{B}) + \mathtt{P}(\mathtt{H}|\neg\mathtt{B}) = 1$.

En multipliant par $\mathtt{P}(\mathtt{B})$ on a $\mathtt{sat}\,[\mathtt{B} \implies \mathtt{H}] = \frac{\mathtt{P}(\mathtt{B}) \times \mathtt{P}(\neg\mathtt{H}) - \mathtt{P}(\mathtt{B} \cap \neg\mathtt{H})}{\mathtt{P}(\mathtt{B}) \times \mathtt{P}(\neg\mathtt{H})}$ en divisant par $\mathtt{P}(\mathtt{B} \cap \neg\mathtt{H})$ on a $\mathtt{sat}\,[\mathtt{B} \implies \mathtt{H}] = \frac{\mathtt{conv}[\mathtt{B} \implies \mathtt{H}] - 1}{\mathtt{conv}[\mathtt{B} \implies \mathtt{H}]}$.

Cette mesure varie dans l'intervalle $]-\infty, 1]$ et vaut 0 en cas d'indépendance de B et H. La satisfaction n'est pas utile pour classer les règles exactes car sa valeur est 1 (puisque les règles exactes ont une confiance $\mathtt{P}(\mathtt{H} \mid \mathtt{B}) = 1$).

Pour cette mesure, $\mathtt{P}(\mathtt{H})$ apparaît au numérateur et au dénominateur, donc la variation de cette mesure dépend de $\mathtt{P}(\mathtt{B})$. Plus $\mathtt{P}(\mathtt{B})$ est faible, plus cette mesure est élevée. Par l'intermédiaire de cette mesure, les règles du cas (a) sont rejetées en fin de classement et sont différenciées du cas (b). Nous nous intéressons sommes aux fortes valeurs de satisfaction (*i.e.* proche de la valeur 1).

En somme, ces deux mesures peuvent être consultées simultanément lorsqu'on se trouve dans les cas (a) ou (b) (*i.e.* pour des règles à faible dépendance). Plus la nouveauté est faible et la satisfaction forte, plus la règle est considérée comme significative. L'utilisation conjointe de la *nouveauté* et de la *satisfaction* est illustrée, par un exemple, à la fin du paragraphe 4.3.2.3.

4.2.5 Combinaison des mesures de qualité

Nous présentons un algorithme permettant de combiner les différentes mesures de qualité que nous avons présentées [Cherfi *et al.*, 2003a]. De par leurs caractéristiques, nous proposons de classer les règles d'association selon les valeurs croissantes des mesures de qualité (sauf pour la mesure de *nouveauté*).

Notations pour l'algorithme 4 : $\mathtt{rang}(\mathtt{X})$ renvoie la valeur de la mesure de qualité « X » associée à la règle « r ». Nous dirons que $\mathtt{rang}(\mathtt{X})$ est *élevé* pour la règle r si la position de cette règle est *en haut* de classement pour une mesure de qualité (respectivement *faible* si la règle est *en bas* de classement). Le classement est une liste de l'ensemble des règles extraites ordonnée décroissante pour les valeurs de la mesure X.

Algorithme 4: Combinaison des mesures de qualité

Entrée : $\mathtt{E} := \emptyset$: l'ensemble initialement vide de règles intéressantes;
Sortie : $\mathtt{E}$ l'ensemble final de règles intéressantes;
pour chaque *règle extraite* $\mathtt{r}$ **faire**
  si *rang(*$\mathtt{int}$*) est élevé* **alors**
    $\mathtt{E} := \mathtt{E} \bigcup \{\mathtt{r}\}$
  sinon
    si *rang(*$\mathtt{conv}$*) est élevé* **alors**
      $\mathtt{E} := \mathtt{E} \bigcup \{\mathtt{r}\}$
    sinon
      si *rang(*$\mathtt{dep}$*) est élevé* **alors**
        $\mathtt{E} := \mathtt{E} \bigcup \{\mathtt{r}\}$
      sinon
        si *rang(*$\mathtt{nov}$*) est faible* **alors**
          si *rang(*$\mathtt{sat}$*) est élevé* **alors**
            $\mathtt{E} := \mathtt{E} \bigcup \{\mathtt{r}\}$
renvoyer $\mathtt{E}$

Nous résumons dans le tableau (TAB. 4.1) les caractéristiques des différentes mesures de qualité que nous avons utilisées. Pour chaque mesure, nous rappelons les intervalles de définitions, les valeurs particulière mesurant les cas d'indépendance statistique des termes dans le corpus. Nous donnons également les valeurs des situations de référence et nous indiquons lorsque la mesure est symétrique.

TAB. 4.1 – Caractéristiques des mesures de qualité utilisées

Mesure	Formule	Intervalle	Valeur d'indépendance	Situation de référence	Symétrie
$\mathtt{int}\,[\mathtt{B} \Longrightarrow \mathtt{H}]$	$\frac{\mathtt{P(B \cap H)}}{\mathtt{P(B) \times P(H)}}$	$[0, +\infty[$	1	$= 0$, incompatibles	Oui
$\mathtt{conv}\,[\mathtt{B} \Longrightarrow \mathtt{H}]$	$\frac{\mathtt{P(B) \times P(\neg H)}}{\mathtt{P(B \cap \neg H)}}$	$[0, +\infty[$	1	> 1, dépendants $[0, 1[$, non dép.	Non
$\mathtt{dep}\,[\mathtt{B} \Longrightarrow \mathtt{H}]$	$\lvert\mathtt{P(H \mid B)} - \mathtt{P(H)}\rvert$	$[0, 1[$	0	$\simeq 1$, dépendants	Non
$\mathtt{nov}\,[\mathtt{B} \Longrightarrow \mathtt{H}]$	$\mathtt{P(B \cap H)} - \mathtt{P(B)} \times \mathtt{P(H)}$	$]-1, 1[$	0	$\simeq -1$, support faible	Oui
$\mathtt{sat}\,[\mathtt{B} \Longrightarrow \mathtt{H}]$	$\frac{\mathtt{(P(\neg H) - P(\neg H \mid B))}}{\mathtt{P(\neg H)}}$	$]-\infty, 1]$	0	$= 1$, règle exacte	Non

4.3 Application au corpus de biologie moléculaire

Afin de valider notre méthodologie, nous appliquons l'algorithme de combinaison des mesures de qualité des règles sur un corpus de textes de taille réelle. Cette étape de validation et d'interprétation des règles d'association constitue la dernière étape de notre processus de FdT. Nous

commençons par décrire les données textuelles en entrée de processus en § 4.3.1, nous proposons ensuite § 4.3.2 notre méthode d'interprétation des résultats du processus de FdT. Nous décrivons globalement les résultats du processus § 4.3.2.1 puis en confrontant les règles extraites à l'avis de l'analyste en § 4.3.2.2. Nous montrons et discutons de l'adéquation des connaissances extraites par les règles par rapport au domaine de fouille en § 4.3.2.3 et § 4.3.2.4.

4.3.1 Description des données

Notre corpus est composé de 1 361 documents d'environ 240 000 mots, pour un volume de 1,6 Mø. Un *document* est constitué d'un *identifiant* unique (*i.e.* un numéro), d'un titre, d'une liste d'auteurs, du résumé sous forme textuelle et d'une liste de termes caractérisant ce résumé. Les textes sont en anglais et traitent de biologie moléculaire, plus particulièrement des mutations génétiques en lien avec une résistance aux antibiotiques.

Deux indexations ont été mises en œuvre avec ce corpus sur la biologie moléculaire. La première indexation est entièrement automatique et a été réalisée en utilisant FASTER. L'ensemble des textes a été indexé par un total de 22 885 termes qui correspondent à 3 337 termes différents. Parmi ces termes, 1 762 (soit 52,8 %) sont des termes n'apparaissant qu'une seule fois (*i.e.* des termes *hapax*). Cette distribution des termes dans le corpus est bien connue en analyse de l'information textuelle. Elle est due, notamment, aux termes périphériques du domaine présents dans la description des textes en langage naturel.

Une seconde indexation a eu lieu avec les 22 885 termes filtrés manuellement par les documentalistes de l'INIST. Ce filtrage manuel permet d'éliminer une grande partie considérée comme du bruit – près de la moitié. Il résulte que l'ensemble des textes a été indexé par un total de 14 374 termes dont 632 différents (soit 18,94 % du nombre de termes différents par rapport à la première expérience).

4.3.2 Expérimentations et interprétation

Cette partie caractérise, d'un point de vue qualitatif, les règles d'association extraites par le processus de fouille ainsi que leur interprétation par un analyste. Il faut noter que les mesures présentées en § 4.2 ne couvrent qu'une partie de leurs valeurs possibles. Par exemple, nous n'observons pas de cas d'indépendance entre B et H pour les règles extraites. De même, nous n'avons pas de valeurs négatives pour la *nouveauté.*

4.3.2.1 Description des résultats

Nous avons appliqué le processus de fouille sur le corpus avec les deux indexations introduites en § 4.3.1.

Pour la première expérience portant sur les 3 337 termes issus de l'indexation automatique de FASTER, nous avons fixé `minsup` à 0,7% (correspondant à un seuil minimum de support de 10 textes pour les règles extraites). Le seuil de support a été fixé à cette valeur car 49 % des termes apparaissent entre 5 et 15 fois dans les textes. Nous avons donc choisi de prendre la valeur moyenne de 10. La valeur `minconf` est fixée à 100% (*i.e.* règles exactes). Nous avons obtenu 1 202 règles. Les règles sont trop nombreuses pour être analysées une par une. Comme le soulignent [Gras *et al.*, 2001] : « *... le nombre de règles calculé peut être très élevé et les tâches de dépouillement, d'interprétation et de synthèse des résultats peuvent alors devenir extrêmement complexes, voire*

inextricables, pour l'utilisateur ». Dans la seconde expérience portant sur les 632 termes filtrés, nous avons fixé `minconf` à 80% — nous acceptons au plus 20% de contre-exemples à une règle — et nous avons conservé `minsup` égal à 10 textes. Nous avons obtenu 347 règles, dont 128 règles exactes. Nous diminuons le nombre de règles extrait de plus d'un tiers. C'est un nombre de règles interprétable en quelques heures par l'analyste. Le choix des seuils `minsup` et `minconf` est délicat et généralement proposé par l'analyste, du fait de son expertise dans le domaine de fouille. L'usage de seuils comme critère de décision en fouille de données est une étape importante pour mener une expérimentation en FdT. Il est courant de fixer plusieurs seuils et de faire un compromis entre le nombre de règles d'association extraites et la capacité ainsi que le temps que peut consacrer l'analyste durant la phase d'interprétation des règles.

TAB. 4.2 – Pourcentage de règles obtenues par cas de distribution des termes

Cas	% de règles
(a)	10
(b)	28.5
(c)	**61.5**

Plus de 60% des règles relèvent du cas (c) de FIG. 4.3, le plus intéressant de notre point de vue.

4.3.2.2 Méthode d'interprétation et confrontation aux commentaires de l'analyste

Nous avons réalisé l'outil TAMIS en deux versions : une version avec une interface de visualisation en Java et une version avec une interface Web. Un ensemble de termes représentant les textes est donné en entrée. Le traitement consiste à extraire pour des seuils de support et de confiance donnés en paramétres, l'ensemble des règles d'association informatives valides, puis à calculer pour chaque règle les mesures de qualité correspondante, et enfin, de proposer à l'analyste de voir les règles suivant le tri par les différentes mesures que nous lui suggérons en suivant l'algorithme 4 proposé en § 4.2.5[25].

Nous avons alors proposé les 347 règles obtenues lors de la seconde expérience à un analyste, expert du domaine. Les règles n'ont pas été classées afin de lui laisser une libre appréciation. La confrontation à l'avis de l'analyste a montré que les règles qu'il retenait se trouvaient dans les cas de FIG. 4.3 (c) et (b). Il est important de repérer les règles qui lui paraissent « interprétables » pour l'analyste.

Définition 4.3 (Règle interprétable) *Une règle est* interprétable *pour l'analyste s'il peut relier tous les termes apparaissant dans* `B` *et* `H`*. Ces termes dénotent une relation sémantique dans le domaine (généricité, lien de composition, causalité, synonymie, hypéronymie, etc.). Le travail de l'analyste consiste à expliquer pourquoi il est sémantiquement valide que les termes apparaissent ensemble dans cette règle.*

Analyse par l'expert. Les textes décrivent le phénomène de la mutation des gènes dans les bactéries provoquant une résistance aux antibiotiques. Cela suit le principe biologique de : « Ce qui ne tue pas une bactérie, la rend plus forte ».

[25] Une décription de l'outil TAMIS en versions Java et pour le Web en annexe C

L'information génétique a pour support l'ADN présent dans chacune des cellules qui composent tout organisme vivant. L'ADN détermine les caractéristiques d'une cellule en interaction avec l'environnement. Cette information est présente sous forme de séquences nucléotidiques (*i.e.* de gènes) pouvant correspondre à des séquences protéiques de la cellule [Zaccai et Garrec, 1998]. Certains antibiotiques permettent d'inhiber la synthèse protéique en empêchant, par exemple la production d'enzymes par la bactérie. Cette bactérie ne se reproduit plus et/ou meurt. Au sein du génome de la bactérie, certaines mutations peuvent provoquer une résistance aux antibiotiques qui ne pourront donc plus se fixer sur la bactérie. C'est un schéma du phénomène de résistance des bactéries aux antibiotiques.

Voici huit règles extraites avec l'outil TAMIS et les explications associées données par l'analyste :

Numéro : **120**
Règle : **"determine region"** $\sqcap$ **"gyrA gene"** $\sqcap$ **"gyrase"** $\sqcap$ **"mutation"** $\Longrightarrow$ **"quinolone"**
pB : "0,008" *pH* : "0,059" *pBH* : "0,008" *Support* : "11" *Confiance* : "1,000" *Intérêt* : "17,012"
Conviction : "indéfinie" *Dépendance* : "0,941" *Nouveauté* : "0,008" *Satisfaction* : "1,000"

La règle 120 reflète le phénomène de résistance. Elle indique que les textes cités décrivent la mutation du gène "gyrA" qui contrôle le comportement de l'enzyme "gyrase" dans une zone précise de l'ADN. Cet enzyme est responsable de la résistance aux antibiotiques de la famille des "Quinolones". Pour avoir le schéma complet du mécanisme de résistance, il manque dans la règle le nom de la bactérie, qui n'intervient pas car ce n'est pas la même bactérie pour les 11 textes.

Numéro : **279**
Règle : **"mutation"** $\sqcap$ **"parC gene"** $\sqcap$ **"quinolone"** $\Longrightarrow$ **"gyrA gene"**
pB : "0,015" *pH* : "0,046" *pBH* : "0,014" *Support* : "21" *Confiance* : "0,952" *Intérêt* : "20,574"
Conviction : "20,028" *Dépendance* : "0,906" *Nouveauté* : "0,014" *Satisfaction* : "0,950"

Le commentaire sur la règle 279 fait ressortir le fait que le gène "parC" a été découvert plus récemment que le gène "gyrA". Ces deux gènes sont liés par mutation combinée et les bactéries résistent alors aux Quinolones. Chaque fois qu'un texte cite le gène "parC", ce texte fait référence à "gyrA".

Numéro : **202**
Règle : **"grlA gene"** $\Longrightarrow$ **"mutation"** $\sqcap$ **"staphylococcus Aureus"**
pB : "0,009" *pH* : "0,023" *pBH* : "0,008" *Support* : "12" *Confiance* : "0,917" *Intérêt* : "40,245" *Conviction* : "11,727" *Dépendance* : "0,894" *Nouveauté* : "0,008" *Satisfaction* : "0,915"

Numéro : **270**
Règle : **"mecA"** $\sqcap$ **"meticillin"** $\Longrightarrow$ **"mecA gene"** $\sqcap$ **"staphylococcus Aureus"**
pB : "0,009" *pH* : "0,012" *pBH* : "0,009" *Support* : "12" *Confiance* : "1,000" *Intérêt* : "80,059"
Conviction : "indéfinie" *Dépendance* : "0,988" *Nouveauté* : "0,009" *Satisfaction* : "1,000"

Les commentaires sur les deux règles 202 et 270 indiquent que la "meticillin" inhibe le gène "mecA" des bactéries et permet de guérir des infections dues à la mutation du gène "grlA" causé par la bactérie "Staphylococcus Aureus".

Numéro : **293**
Règle : **"mycobacterium tuberculosis"** $\Longrightarrow$ **"tuberculosis"**
pB : "0,053" *pH* : "0,067" *pBH* : "0,053" *Support* : "72" *Confiance* : "1,000" *Intérêt* : "14,956" *Conviction* : "indéfinie" *Dépendence* : "0,933" *Nouveauté* : "0,049" *Satisfaction* : "1,000"

Numéro : **335**
Règle : **"restriction enzyme"** $\Longrightarrow$ **"enzyme"**
pB : "0,008" *pH* : "0,112" *pBH* : "0,008" *Support* : "11" *Confiance* : "1,000" *Intérêt* : "8,954" *Conviction* : "indéfinie" *Dépendance* : "0,888" *Nouveauté* : "0,007" *Satisfaction* : "1,000"

Certaines règles comme celles ci-dessus sont inintéressantes. La plupart sont dues à un effet de bord crée par l'outil d'indexation qui, dans son processus d'extraction de termes, procède par reconnaissance de termes les plus longs puis par découpage en sous-termes. (*ex.* "mycobacterium tuberculosis" dans la règle 293 et "restriction enzyme" dans la règle 335).

Numéro : **183**
Règle : **"epidemic strain"** $\Longrightarrow$ **"outbreak"**
pB : "0,012" *pH* : "0,057" *pBH* : "0,012" *Support* : "16" *Confiance* : "1,000" *Intérêt* : "17,449" *Conviction* : "indéfinie" *Dépendance* : "0,943" *Nouveauté* : "0,011" *Satisfaction* : "1,000"

Numéro : **2**
Règle : **"agar dilution"** $\Longrightarrow$ **"dilution method"**
pB : "0,019" *pH* : "0,025" *pBH* : "0,019" *Support* : "26" *Confiance* : "1,000" *Intérêt* : "40,029" *Conviction* : "indéfinie" *Dépendance* : "0,975" *Nouveauté* : "0,019" *Satisfaction* : "1,000"

D'autres règles relient des termes à leurs synonymes. Les auteurs emploient indifféremment des termes et leurs synonymes pour décrire un même concept (*ex.* dans la règle 183). Enfin, des liens d'hypéronymie sont observés sur plusieurs règles, comme dans la règle 2 où la dilution de l'"agar" est une méthode de dilution couramment utilisée dans le domaine.

4.3.2.3 Adéquation des mesures de qualité à l'analyse de l'expert

Nous supposons que le corpus de biologie moléculaire de nos expérimentations reflète les connaissances du domaine, et que l'indexation reflète le contenu des textes de ce corpus. L'analyste a globalement réussi à interpréter, par rapport au domaine, les règles que nous lui avons présentées. Les règles citées ci-dessous sont présentées, accompagnées des mesures de qualité correspondantes, en annexe D.

La mesure d'*intérêt*, par définition, classe en premier les règles ayant des termes rares en B et en H de figure (FIG. 4.3-(c)). On s'attend à ce que l'analyste préfère ce genre de règle. L'expérience montre bien que les deux règles 270 et 202, présentées au paragraphe précédent, ont des valeurs très supérieures à la valeur en cas d'indépendance $= 1$ pour cette mesure. Ces deux règles ont respectivement comme valeur d'intérêt $80,059$ et $40,245$. Ces règles sont porteuses de connaissances pour l'analyste puisqu'il a réussi à les expliquer facilement (voir en début du paragraphe 4.3.2.2). Par ailleurs, la règle 159 ("dna" $\sqcap$ "gyrA gene" $\Longrightarrow$ "mutation") qui illustre FIG. 4.3-(b)

ainsi que la règle 228 ("Gyrase" ⊓ "protein" ⟹ "mutation") pour FIG. 4.3-(a) sont moins informatives. Leur intérêt et leur conviction sont plus proches de 1 (respectivement $4,929$ et $5,086$). Le comportement symétrique de la mesure d'intérêt peut se révéler intéressant. Par exemple, la règle 108 ("dalfopristin" ⟹ "quinupristin") et la règle 332 ("quinupristin" ⟹ "dalfopristin") ont la même forte valeur d'intérêt (de $75,611$). Cette mesure a permis de les rapprocher dans le classement. Nous avons mis en valeur des similitudes de comportement de populations de bactéries en résistance à deux antibiotiques ("quinupristine" et dalfopristine). Ce qui est confirmé par l'analyste.

En confrontant plusieurs règles à fortes valeurs de *conviction,* nous avons retrouvé dans les textes une antériorité dans la découverte du gène GyrA par rapport à ParC. Nous avons vérifié sur nos données que cette mesure renforce la direction de l'implication de B vers H. Dans notre exemple, l'analyste a souligné que ParC et GyrA sont deux gènes régulièrement présents ensemble dans les règles et il le justifiait par leurs comportements comparables du point de vue de la mutation. Pourtant, le sens de l'implication $\ldots \sqcap \texttt{ParC} \sqcap \ldots \Longrightarrow \ldots \sqcap \texttt{GyrA} \sqcap \ldots$ dans des règles de fortes valeurs de *conviction* contribue à les différencier. Par exemple la règle 279, déjà présentée, a une valeur de conviction largement supérieure à 1 ($20,028$). En revanche, la règle 215 ("gyra gene" ⊓ "pare gene" ⟹ "parc gene" ⊓ "quinolone") dans le sens "gyrA gene" vers "parC gene" est moins bien classée ($11,735$). La conviction peut ainsi faire une distinction entre les règles 279 et 215, alors que l'intérêt les classe de façon proche car toutes les deux illustrent le cas (c) de la figure 4.3. Finalement, l'explication réside dans le fait que les textes les plus anciens de notre corpus ne traitent que de GyrA alors que les textes plus récents traitent de GyrA et de ParC.

La *dépendance* est forte pour de faibles valeurs de $\texttt{P}(\texttt{H})$, ce qui nous place également dans la figure (FIG. 4.3-(c)). Les règles exactes 270 et 120, ou à valeur de confiance proche de 1 (*ex.* règle 279) sont celles qui sont les plus dépendantes (car $\texttt{dep}[\texttt{B} \Longrightarrow \texttt{H}] = 1 - \texttt{P}(\texttt{H})$).

Enfin, les deux règles suivantes illustrent le comportement de la *nouveauté* et de la *satisfaction.* La règle non informative 273 ("meticillin" ⟹ "staphylococcus Aureus") correspond à la figure (FIG. 4.3-(a)), alors que celle qui est mieux interprétée 265 ("mecA gene" ⊓ "meticillin" ⟹ "Staphylococcus Aureus") — à cause de la présence du gène — correspond à la figure (FIG. 4.3-(b)). Ces deux règles ont des valeurs de dépendance faible (resp. $0,733$ et $0,790$) . Néanmoins, le classement par nouveauté place 273 devant 265, alors que la satisfaction les classe inversement. Ces deux mesures peuvent donc distinguer le cas moins informatif (b) du cas non informatif (a), là où la dépendance ne peut aider à les différencier.

La règle 329 ("quinolone" ⊓ "substitution" ⟹ "gyrA gene") est confirmée dans tous les textes ex. n°000311 :« Mutants with the single Ser-91 to Phe *substitution* in *GyrA gene* were ... less susceptible to norfloxacin and ciprofloxacin than the wild type. ». norfloxacin et ciprofloxacin sont tous les deux des quinolones.

4.3.2.4 Éléments de discussion

En extraction de connaissances, on aurait tendance à chercher les règles les plus génériques vérifiées sur un grand nombre d'exemples (*i.e.* ayant des supports élevés). Néanmoins, l'analyste juge, par exemple, que la règle : ("aztreonam" "clavulanic acid" "enzyme "⟹ "β-lactamase") est plus interprétable que : ("aztreonam" "enzyme" ⟹ "β-lactamase"), qui se trouve être plus générique et couvre plus d'exemples (16 textes contre 11).

Les deux règles exactes 219 ("gyrA gene" "resistance mechanism" ⟹ "quinolone") et 326 ("quinolone" "resistance mechanism" ⟹ "gyrA gene") portent exactement sur les mêmes textes.

Comme le mécanisme de résistance porte sur les quinolones, l'analyste préfère la seconde règle. Toutes les mesures discriminantes (ici, *intérêt* et *satisfaction*) les classent dans le bon ordre. 10 textes sur 11 confirment un phénomène de résistance dû à la mutation sur le gène *GyrA* mais un seul texte (n°1032) apporte une contradiction à l'interprétation des deux règles par la phrase : « No changes in the quinolone-resistant determing regions of parC, parE, gyrA, or gyrB were found in this mutant. ». Ce qui montre que la *négation* dans les textes, si elle n'est pas prise en compte dans le processus de fouille et pour le formalisme des règles d'association en particulier, reste un problème entier qu'il nous reste à étudier.

4.4 Approches comparables

Les travaux de [Azé et Roche, 2003; Roche *et al.*, 2003] se différencient par l'extraction de règles vérifiant certaines contraintes (au maximum K termes en B et un seul terme en H). Cette contrainte permet de ne pas utiliser de seuil de support (difficile à fixer a priori). Cette stratégie de diminution de l'espace de recherche ainsi que l'utilisation d'une mesure dite de « moindre contradiction », qui favorise l'extraction des règles ayant un minimum de contre-exemples, permet de réduire le nombre de règles extraites par un algorithme itératif qui affine l'espace de recherche.

Dans les travaux de [Faure *et al.*, 1998], nous partons de schémas de sous-catégorisation pour « apprendre » une hiérarchie de concepts (*i.e.* ontologie) par une classification hiérarchique ascendante (CHA) et par l'utilisation de relations grammaticales dans les textes, par exemple :

$$[Secher]\ COD\ < aliment >$$

$$[Secher]\ CC\ < air >$$

Ces schémas de sous-catégorisation sont appris à partir d'exemples contenus dans un corpus étiqueté sur les recettes de cuisine. Toutes les occurrences du verbe "sécher" font apparaître un aliment en complément d'objet direct (COD) et un terme comme "air" en complément circonstanciel (CC). [Suzuki et Kodratoff, 1998] reprend le corpus étiqueté par les schémas de sous-catégorisation et cherche à trouver les dépendances les plus *pertinentes* entre des concepts et des ensembles de documents en donnant une mesure d'intensité aux règles d'association générées. L'intensité dans les règles d'association est également utilisée dans [Gras *et al.*, 2001] par le calcul d'une pondération des règles avec une fonction entropique tenant compte, à la fois des contre-exemples à la règle et à sa contraposée $\neg H \Longrightarrow \neg B$.

Enfin, dans [Feldman *et al.*, 1998], l'exploitation des règles se fait par la sélection de celles pour lesquelles les termes dans B et H sont d'un certain type. Cela permet de descendre jusqu'à des indices de *confiance* très faibles (de l'ordre de 0,1). Par exemple, chercher toutes les compagnies qui ont fait alliance ou qui ont fusionné : "intuit corp" "novell corp" $\Longrightarrow$ "merger".

4.5 Conclusion

Nous avons décrit l'outil orpailleur de fouille de textes appelé TAMIS : Text Analysis by Mining Interesting_ruleS. C'est la partie de notre étude et réalisation d'un outil pour faire de la *fouille de textes* dite *syntaxique.* L'outil TAMIS automatise le processus d'extraction de connaissances à partir de textes de *Fayyad et al.*. Nous avons présenté le problème d'analyse des résultats d'un

processus fondé sur l'extraction des règles d'association informatives. Nous avons montré l'utilité du processus de FdT que nous proposons pour : (i) l'extraction de connaissances à partir de textes, (ii) l'identification du bruit pour améliorer l'indexation des textes. Notre contribution à la résolution du problème de sélection pour l'analyse des règles extraites consiste à utiliser des mesures de qualité pour ordonner les règles extraites. Les classements par mesure de qualité des règles donnent des points de vue différents pour l'interprétation des règles par l'analyste. Une expérimentation portant sur un corpus de biologie moléculaire a montré l'adéquation des ordres calculés pour l'aide à l'interprétation des règles extraites. Les critères de sélection des règles par les mesures de qualité portent sur les textes eux-mêmes. Nous n'utilisons pas les connaissances du domaine décrites *a priori* dans un modèle existant du domaine, par exemple une *ontologie* du domaine. Pour ce faire, nous décrivons au chapitre 5 une sélection des règles qui utilise un modèle de connaissances.

Chapitre 5

Description de l'outil Sem-TAMIS : utilisation d'un modèle de connaissances

« It is impossible for a man to learn
what he thinks he already knows. »
Anonyme

Sommaire

Introduction

Nous présentons dans ce chapitre le deuxième module de l'outil ORPAILLEUR de fouille de textes, que nous avons développé, que nous appelons Sem-TAMIS (*Semantic Text Analysis by Mining Interesting_ruleS*). Sem-TAMIS s'appuie sur un classement des règles d'un point de vue sémantique, c'est-à-dire que l'outil Sem-TAMIS est fondé sur une approche exploitant un modèle de connaissances pour la sélection de règles d'association extraites à partir de bases de données textuelles [Cherfi *et al.*, 2004a; Janetzko *et al.*, 2004]. Nous montrons l'utilité du processus de FdT

que nous proposons pour : (i) la sélection des règles d'association apportant des connaissances nouvelles. Nous appelons connaissances nouvelles les connaissances qui ne sont pas présentes dans un modèle de connaissances *a priori* ; (ii) l'enrichissement de ce modèle de connaissances à partir des règles d'association extraites des textes. Nous définissons le modèle de connaissances comme une source d'information disponible dans un domaine particulier à la manière d'un réseau sémantique tel qu'utilisé dans les systèmes de représentation de connaissances [Sowa, 1992].

Sem-TAMIS est donc un module complémentaire de l'outil TAMIS pour l'étude et réalisation d'un outil pour faire de la *fouille de textes* dite *sémantique.* Dans le chapitre 4, nous avons abordé la sélection des règles d'association par des mesures de qualité s'appuyant uniquement sur les données (les textes) en entrée. Chaque mesure permet de mettre en valeur un certain type de règles d'association : celles qui portent sur des signaux d'information faibles (les pépites potentielles de connaissances). Ces règles sont celles qui ont le moins de contre-exemples et qui sont stables en présence de bruit dans les données [Azé, 2003]. Nous avons proposé une combinaison de ces mesures de qualité [Cherfi *et al.*, 2003b].

À partir de l'expérience acquise lors du développement de l'outil TAMIS, nous constatons les limites de la méthodologie d'utilisation des mesures de qualité syntaxiques car le classement que nous proposons s'appuie sur les données elles-mêmes et se fait sans prendre en compte les connaissances du domaine. Il est donc difficile de développer une approche statistique indépendante des données pour mesurer la qualité des règles extraites puisque ces mesures s'appuient sur le même ensemble de données que le processus d'extraction de règles d'association.

Nous décrivons, en premier lieu, le processus de fouille de textes qui utilise un modèle de connaissances tel que nous le concevons. Ce modèle est appelé « modèle terminologique » car il porte sur des termes du domaine, de façon analogue à un thésaurus pour une documentation technique. Nous décrivons, par la suite, la mise en œuvre de ce processus dans un cadre probabiliste. Nous proposons, pour ce faire, une mesure de qualité des règles par rapport au modèle terminologique que nous appelons la *vraisemblance.* Nous évaluons le comportement de la mesure de vraisemblance sur un modèle formel, puis sur une expérimentation portant sur le même corpus de biologie moléculaire que celui que nous utilisons pour valider l'outil TAMIS du chapitre 4. Enfin, nous proposons une stratégie pour l'enrichissement du modèle terminologique avec les connaissances correspondant aux règles d'association extraites et validées par l'analyste.

5.1 Fouille de textes avec un modèle de connaissances

Dans le schéma général de référence d'ECBD introduit dans [Fayyad *et al.*, 1996a] et que nous adaptons — pour l'appliquer aux textes — en FIG. 2.1 (*cf.* § 2.2.1), le processus de fouille de données est suivi d'une phase d'interprétation. Au cours de cette phase d'interprétation, les règles d'association extraites sont évaluées par un analyste pour déterminer si une nouvelle connaissance, présente dans les règles et inconnue de cet analyste, peut être validée.

Nous souhaitons classer les règles d'association qui sont présentées à l'analyste en les classant par qualité décroissante en fonction des connaissances disponibles sur le domaine des données. Nous considérons qu'une règle d'association est de *bonne qualité* si cette règle contient des informations susceptibles d'enrichir le modèle de connaissances. Les autres règles sont qualifiées de « triviales » puisqu'elles reflètent une connaissance déjà présente dans le modèle. Notre objectif est de faciliter la tâche de l'analyste en définissant une méthodologie de sélection des règles

de qualité et de rejet des règles triviales qui exploite un modèle de connaissances. Notre modèle de connaissances est dans le cas présent un modèle terminologique caractérisé par un réseau de termes structurés hiérarchiquement par une relation de généralisation appelée EST-UN. La relation EST-UN a la sémantique classique de l'intension des ensembles d'objets dénotés par les termes. Par exemple, dans FIG. 5.1, une « tarte aux pommes » EST-UNE « tarte » qui elle-même EST-UN « dessert »[26], etc. La règle d'association ("tarte aux pommes" $\Longrightarrow$ "tarte") est considérée comme triviale. En revanche, la règle ("tarte aux pommes" $\sqcap$ "chocolat" $\Longrightarrow$ "tarte") ou ("tarte aux pommes" $\Longrightarrow$ "tarte" $\sqcap$ "chocolat") ne sont pas triviales car le terme « chocolat » n'est pas relié, par la relation EST-UN, aux autres termes de la règle dans le modèle connaissances.

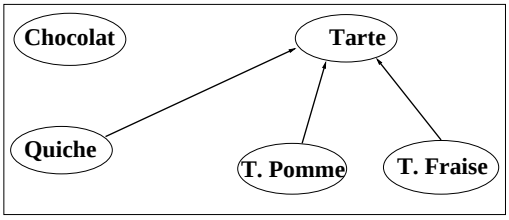


FIG. 5.1 – Exemple d'un modèle terminologique (les liens entre les termes représentent la relation EST-UN).

La relation EST-UN est antisymétrique, c'est-à-dire que le sens du lien dans le modèle de connaissances est important. Ainsi, nous considérons que la règle ("tarte" $\Longrightarrow$ "quiche") n'est pas *triviale* par rapport au modèle terminologique de FIG. 5.1. Par conséquent, cette règle ne doit pas être rejetée.

Intérêt de l'utilisation d'un modèle de connaissance. Peu d'approches pour la sélection des règles d'association exploitent un modèle de connaissances [Janetzko *et al.*, 2004]. Or, la qualité d'une règle ne doit pas être définie à l'aide de mesures de qualité statistiques issues uniquement des données, mais en évaluant l'apport de la règle par rapport aux connaissances du domaine déjà acquises : il existe souvent des sources de connaissances disponibles (par exemple, des ontologies) qui peuvent être exploitées. Le modèle de connaissances doit donc aider à interpréter les résultats du processus de fouille et réciproquement, les résultats de la fouille doivent contribuer à la mise à jour des connaissances. Nous considérons ainsi que la construction et l'enrichissement d'un modèle de connaissances sont des processus itératifs d'extraction de connaissances. À partir de données stables au cours des itérations, l'ensemble des règles d'associations extraites et leurs mesures de qualité statistiques restent identiques alors que le classement des règles suivant notre mesure de qualité évolue en fonction des mises à jour successives du modèle de connaissances.

5.1.1 Modèle terminologique

Soit $\mathcal{T} = \{\mathtt{t}_1, \ldots, \mathtt{t}_n\}$ un ensemble de termes $\mathtt{t}_i$, $\mathtt{i} \in \{1, \ldots, n\}$ qui sert de vocabulaire de référence pour indexer un ensemble de textes $\mathcal{D} = \{\mathtt{d}_1, \ldots, \mathtt{d}_m\}$ – notés $\mathtt{d}_j$ pour *document* –, $\mathtt{j} \in \{1, \ldots, m\}$ (*cf.* FIG. 5.2).

Définition 5.1 (Relation EST-UN**)** *La relation* EST-UN *est définie sur* $\mathcal{H} \times \mathcal{H}$*. Cette relation est*

[26] Ce lien n'est pas représenté en FIG. 5.1.

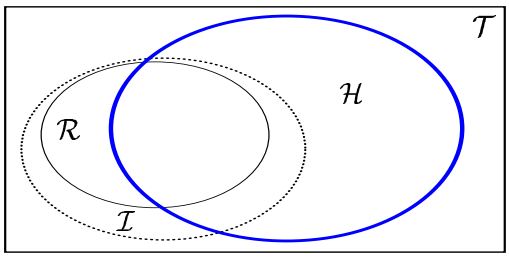


FIG. 5.2 – Les différents ensembles de termes : $\mathcal{T}$: ensemble des termes, $\mathcal{I}$: ensemble des termes d'indexation des textes, $\mathcal{R}$: sous-ensemble des termes d'indexation $\mathcal{I}$ apparaissant dans les règles d'association, et $\mathcal{H}$: ensemble des termes du modèle M.

réflexive, antisymétrique et transitive. Elle constitue un ordre partiel. Le graphe de la relation EST-UN*, non nécessairement connexe, matérialise le modèle terminologique.*

Définition 5.2 (Modèle terminologique) *Nous définissons le modèle de connaissances comme un modèle terminologique caractérisé par un réseau de termes $\mathcal{H} \subseteq \mathcal{T}$ structurés hiérarchiquement par une relation de généralisation* EST-UN*. Le réseau de termes est un graphe qui n'est pas nécessairement connexe. Un terme appartenant à ce modèle peut être seul ou être relié par la relation de généralisation* EST-UN *à un ou plusieurs autres termes.*

Nous introduisons les caractéristiques des règles d'association par rapport à un modèle de connaissances du domaine en commençant par les rappels suivants :

Rappel de la définition d'une règle d'association

Chaque texte est représenté par un ensemble de termes qui indexent son contenu. $\mathcal{I}$ désigne le sous-ensemble de termes qui indexent au moins un texte.
Une règle d'association est une implication de la forme $\mathtt{B} \overset{\mathtt{P}}{\Longrightarrow} \mathtt{H}$ avec $\mathtt{B} \subseteq \mathcal{I}$, $\mathtt{H} \subseteq \mathcal{I}$ et $\mathtt{B} \cap \mathtt{H} = \emptyset$. Soit $\mathtt{B} = \{\mathtt{t}_1, \ldots, \mathtt{t}_p\}$ l'ensemble des termes de la partie gauche d'une règle d'association r et $\mathtt{H} = \{\mathtt{t}_{p+1}, \ldots, \mathtt{t}_q\}$ l'ensemble des termes de sa partie droite. r : $\mathtt{t}_1 \sqcap \ldots \sqcap \mathtt{t}_p \overset{\mathtt{P}}{\Longrightarrow} \mathtt{t}_{p+1} \sqcap \ldots \sqcap \mathtt{t}_q$ signifie que tous les textes de $\mathcal{D}$ contenant les termes $\mathtt{t}_1$ et $\mathtt{t}_2 \ldots$ et $\mathtt{t}_p$ ont tendance à contenir aussi les termes $\mathtt{t}_{p+1}$ et $\mathtt{t}_{p+2} \ldots$ et $\mathtt{t}_q$ avec une probabilité P.

Rappel de la définition des mesures de support et de confiance d'une règle

La mesure de support de r est le nombre de textes contenant les termes du motif $\mathtt{B} \sqcap \mathtt{H} = \{\mathtt{t}_1, \ldots, \mathtt{t}_p, \ldots, \mathtt{t}_q\}$. La mesure de confiance de r est le rapport entre le nombre de textes contenant l'ensemble des termes $\mathtt{B} \sqcap \mathtt{H}$ et le nombre de textes contenant $\mathtt{B}$ $(\mathtt{t}_1, \ldots, \mathtt{t}_p)$. Ce rapport définit la probabilité conditionnelle $\mathtt{P}(\mathtt{H}|\mathtt{B})$. Le support et la confiance sont deux mesures associées aux règles d'association et exploitées par les algorithmes d'extraction de règles pour en réduire la complexité (*cf.* § 4.2.3). Deux valeurs de seuil sont définies : `minsup` pour le support minimal et `minconf` pour la confiance minimale.

Du fait des seuils `minsup` et `minconf`, tous les termes de $\mathcal{I}$ ne sont pas présents dans les règles d'association. Nous disposons d'un ensemble de règles d'association $\mathtt{r}_i : \mathtt{B}_i \Longrightarrow \mathtt{H}_i \mid \mathtt{i} \in \{1, \ldots, n\}$,

préalablement extraites par notre processus de FdT. Nous désignons par $\mathcal{R}$, le sous-ensemble des termes présents, au moins une fois, en $\mathtt{B_i}$ ou en $\mathtt{H_i}$, de l'ensemble des règles d'association, *i.e.* $\mathcal{R} = \bigcup_{\mathtt{r_i}} (\mathtt{B_i} \sqcap \mathtt{H_i})$. FIG. 5.2 montre les intersections et inclusions possibles entre $\mathcal{I}$, $\mathcal{R}$ et $\mathcal{H}$. L'ensemble des termes $\mathcal{H}$ du modèle est à dissocier de l'ensemble des termes indexant les textes $\mathcal{I}$. En effet, nous ne pouvons pas garantir une parfaite adéquation entre le modèle de connaissances initial et le contenu des textes. Notamment, il nous semble intéressant de considérer qu'un modèle n'est pas complet au sens où il ne contient pas de façon exhaustive tous les termes du domaine. Plus $\mathcal{I} \cap \mathcal{H}$ est grand, plus le modèle est complet par rapport à l'ensemble des textes, plus $\mathcal{I} \cap \mathcal{R}$ est grand, meilleure est la couverture du modèle par rapport à l'ensemble des règles extraites.

5.1.2 Définition d'une règle triviale

Nous cherchons à supprimer les règles d'association expriment une relation EST-UN, c'est-à-dire pour lesquelles tous les termes de B sont liés par une relation de généralisation avec les termes de H dans le modèle de connaissances. Ces règles sont appelées *règles d'association triviales.*

Comment définir ce qu'est une règle triviale par rapport au modèle de connaissances introduit ? La construction et la mise à jour d'un modèle consistent à ajouter de nouveaux termes et des relations de généralisation entre les termes.

Définition 5.3 (Règle d'association triviale) *Une règle d'association entre deux termes* a *et* b*, notée* $\mathtt{a} \Longrightarrow \mathtt{b}$*, est triviale si la relation de généralisation que l'analyste peut ajouter au modèle à partir de cette règle,* i.e. a EST-UN b*, est déjà présente dans le modèle.*

Par exemple, le terme « tarte » est plus général que le terme « quiche » et la règle ("quiche" $\Longrightarrow$ "tarte") est considérée comme triviale et donc à rejeter puisqu'elle exprime une relation de généralisation connue dans le modèle terminologique de FIG. 5.1. En revanche, la règle "tarte aux fraises" $\Longrightarrow$ "chocolat" $\sqcap$ "quiche" doit être conservée. Cette règle exprime potentiellement une relation intéressante entre « tarte aux fraises », « chocolat » et « quiche » pour lesquels il n'existe pas de lien hiérarchique EST-UN.

Il nous faut préciser que dans le cadre de la FdT, une règle $\mathtt{a} \Longrightarrow \mathtt{b}$ signifie que chaque fois qu'il y a une occurrence de a dans un texte, il y a (avec une certaine confiance) aussi une occurrence de b. Cela ne signifie pas formellement que, dans le modèle terminologique, on peut systématiquement ajouter que (a EST-UN b). L'ajout d'une telle relation de généralisation est à l'initiative et sous le contrôle de l'analyste.

Notre approche vise à opérer une sélection dans l'ensemble des règles extraites en rejetant les règles qui sont triviales par rapport au modèle de connaissances donné. Les règles triviales sont aussi appelées règles *taxinomiques.*

5.1.3 Modèle de connaissances probabiliste

Le modèle de connaissances pour la sélection des règles d'association présenté ci-après et exploité dans un modèle probabiliste, construit sur $(\mathcal{H} \times \mathcal{H}$, EST-UN$)$ et noté M. L'objectif est de définir la vraisemblance d'une règle $r : a \Longrightarrow b$ comme la probabilité de trouver un chemin, on réflexif, de "a" vers "b" dans le modèle de connaissances (*i.e.* la probabilité de transition de a vers b). nous ne considérons pas la transition réflexive "$\mathtt{a} - \mathtt{a}$", "$\mathtt{b} - \mathtt{b}$", etc.

Étapes de définition d'un modèle Nous définissons un modèle de connaissances probabiliste par les trois étapes suivantes :

I. Calculer les chemins de longueur minimale La distribution de probabilités est calculée en utilisant le chemin de longueur minimale qui relie, dans le modèle M, un terme $\mathtt{t_i}$ à un terme $\mathtt{t_j}$. Les chemins que nous considérons sont non réflexifs. La longueur du chemin minimal entre $\mathtt{t_i}$ et $\mathtt{t_j}$ dans M est notée $\ell(\mathtt{t_i}, \mathtt{t_j})$.

II. Calcul d'une fonction décroissante δ pour les chemins minimaux Considérons la situation où il existe un chemin (minimal) entre $\mathtt{t_i}$ et $\mathtt{t_j}$ dans le modèle M. Selon la théorie de l'activation en expansion [Collins et Loftus, 1975], plus ce chemin est long, plus la valeur de vraisemblance doit être faible. Nous introduisons donc la fonction strictement décroissante δ. Cette fonction est définie par :

$$\begin{aligned} \delta \ &: \mathcal{H} \times \mathcal{H} \longrightarrow]0, 1] \\ (\mathtt{t_i}, \mathtt{t_j}) &\longmapsto \frac{1}{\ell(\mathtt{t_i}, \mathtt{t_j})} \end{aligned}$$

III. Définition d'une distribution de probabilités sur le modèle Pour être conforme à la notion de distribution de probabilités, il est impératif que la somme des probabilités associées aux divers chemins minimaux entre $\mathtt{t_i}$ et tous les autres termes de $\mathcal{H}$ soit égale à $\mathtt{1}$. Soit $\mathcal{B}_\mathtt{i} = \{\mathtt{x} \in \mathcal{H} \mid \mathtt{t_i Est - un^+ x}\}$ l'ensemble des termes $\mathtt{x}$ de M reliés à $\mathtt{t_i}$ par un chemin de longueur $\geqslant \mathtt{1}$. La cardinalité $|\mathcal{B}_\mathtt{i}|$ est appelée, par la suite, le *facteur de branchement* de $\mathtt{t_i}$. Il s'agit donc de normaliser la fonction δ par la fonction ξ définie par :

$$\begin{aligned} \xi \ &: \mathcal{H} \longrightarrow]0, 1] \\ \mathtt{t_i} &\longmapsto \left(\sum_{\mathcal{B}_\mathtt{i}} \delta(\mathtt{t_i}, \mathtt{x}) \right)^{-1} \end{aligned}$$

L'effet de cette normalisation par ξ, intuitivement, est que plus il existe de transitions entre le terme $\mathtt{t_i}$ et d'autres termes du modèle M, plus $\xi(\mathtt{t_i})$ est faible. À l'inverse, moins il y a de transitions, plus $\xi(\mathtt{t_i})$ est fort. Ce qui équivaut à pondérer la longueur des chemins par le *facteur de branchement* du nœud de départ $\mathtt{t_i}$.

En réunissant les trois fonctions ℓ, δ et ξ précédentes, nous définissons la distribution de probabilités P_M ($\in [\mathtt{0}, \mathtt{1}]$) sur M par :

$$P_M(\mathtt{t_i}, \mathtt{t_j}) = \xi(\mathtt{t_i}) \ \cdot \ \delta(\mathtt{t_i}, \mathtt{t_j}) \tag{5.1}$$

Ce qui définit la P_M probabilité d'une transition entre deux termes dans le modèle M. La probabilité P_M est conforme à la théorie de l'activation de propagation, « spreading activation theory » [Collins et Loftus, 1975], selon laquelle un marqueur d'information part d'un nœud du réseau et se propage à travers ce réseau. La force de ce marqueur est fonction du nombre de relations existant entre le terme de départ et les termes d'arrivée du marqueur. De plus cette force s'affaiblit de façon proportionnelle à la distance parcourue par le marqueur. La force du marqueur partant de $\mathtt{a}$ pour

aller à b dans le modèle est définie par la probabilité de transition d'un terme "a" vers un terme "b" La distribution de probabilités attribue à chaque couple de termes $(\mathtt{t_i}, \mathtt{t_j})$ du modèle M une probabilité de transition de $\mathtt{t_i}$ vers $\mathtt{t_j}$.
Pour le calcul, nous réécrivons la fonction P_M de (5.1), définie par ℓ, δ et ξ, en une seule fonction de ℓ telle que :

$$P_M \ : \mathcal{H} \times \mathcal{H} \longrightarrow]0, 1]$$

$$(\mathtt{t_i}, \mathtt{t_j}) \longmapsto \left[\ell(\mathtt{t_i}, \mathtt{t_j}) \ \times \left(\sum_{\mathcal{B}_\mathtt{i} = \{\mathtt{x} \in \mathcal{H} \ | \ \mathtt{t_i Est-un^+ x}\}} \frac{1}{\ell(\mathtt{t_i}, \mathtt{x})} \right) \right]^{-1} \tag{5.2}$$

Pour la définition de P_M dans (5.2), le premier facteur assure que plus le chemin entre $\mathtt{t_i}$ et $\mathtt{t_j}$ est court, plus la probabilité est forte. Le second facteur assure que la somme de toutes les probabilités de transitions de $\mathtt{t_i}$ vers les autres termes auxquels il est relié dans le modèle est égale à 1. Si le facteur de branchement $\mathcal{B}_\mathtt{i}$ est grand, c'est-à-dire que $\mathtt{t_i}$ est un terme relié par EST-UN à beaucoup d'autres termes du modèle, alors les probabilités de transitions à partir de $\mathtt{t_i}$ seront plus faibles que si $\mathcal{B}_\mathtt{i}$ est plus petit, c'est-à-dire que le terme $\mathtt{t_i}$ est relié à peu d'autres termes du modèle M.

5.2 Définition de la vraisemblance d'une règle

Nous définissons la *vraisemblance* d'une règle d'association *simple* $a \Longrightarrow b$, où a et b sont deux termes, comme une mesure associée à chaque règle d'association extraite. La vraisemblance est définie par la probabilité $P_M(a, b)$. Plus le lien hiérarchique entre a et b est fort dans le modèle, plus cette vraisemblance est forte, et donc plus la règle est considérée comme triviale pour un analyste. Nous observons cependant que cette définition ne permet de calculer que la vraisemblance de règles simples pour lesquelles on suppose que les termes indexant les textes et présents dans les règles sont également décrits dans le modèle, c'est-à-dire que $a, b \in \mathcal{R} \cap \mathcal{H}$.

Prenons l'exemple du modèle M introduit en FIG.5.3(a), la distribution de probabilités est représentée par la matrice donnée FIG. 5.3(b). Cette table nous donne pour tout couple (t_i, t_j) la valeur de $P_M(t_i, t_j)$. Pour un chemin court entre deux termes – par exemple (b,c) – la distribution donne une forte probabilité $(0, 30)$ alors que pour un chemin long – par exemple (b,d) – elle donne une valeur faible $(0, 05)$ – en italique dans FIG. 5.3(b) –. Le calcul de vraisemblance pour une règle simple se fait par un simple accès à cette matrice de probabilités. Par exemple, la règle $b \Longrightarrow e$ a pour vraisemblance dans le modèle M : $P_M(b \Longrightarrow e) = P_M(b, e) = 0, 30$.

Nous généralisons dans § 5.2.2 le calcul de vraisemblance des règles d'association par rapport à un modèle M. En effet, pour que le processus de sélection des règles puisse être utilisé sur des données réelles, nous devons résoudre deux problèmes :

- Les règles simples où $|\mathtt{B}| = |\mathtt{H}| = 1$ ne représentent qu'un sous-ensemble réduit par rapport à l'ensemble des règles extraites. Il est donc nécessaire de généraliser la définition de la vraisemblance pour traiter les règles *complexes* (où $|\mathtt{B} \times \mathtt{H}| > 1$) ;
- Nous considérons qu'un modèle de connaissances peut toujours être enrichi. Nous devons pouvoir calculer la vraisemblance d'une règle même si certains termes de la règle n'appartiennent pas au modèle de connaissances. Nous devons donc étendre notre distribution de probabilités pour prendre en compte le modèle M (dont les termes sont dans $\mathcal{H}$) et l'en-

semble des termes présents dans les règles (que nous avons noté $\mathcal{R}$). La distribution de probabilités doit donc être étendue à l'ensemble $\mathcal{R} \cup \mathcal{H}$.

Nous dissocions la présentation de ces deux points. La section 5.2.1 propose trois possibilités pour étendre la distribution de probabilités et discute de l'impact de ces choix sur la vraisemblance des règles complexes. La section 5.2.2 suppose que la distribution de probabilités est étendue à $\mathcal{R} \cup \mathcal{H}$ par l'une des trois possibilités et définit la vraisemblance pour les règles complexes.

5.2.1 Extension de la distribution de probabilités

La distribution de probabilités qui a été introduite en paragraphe 5.1.3 doit être étendue pour traiter deux cas de figure :

1 – L'équation 5.2 en § 5.1.3 qui définit la distribution de probabilités permet de calculer la probabilité d'une transition entre deux termes du modèle de connaissances M qui sont reliés par au moins un chemin. Lorsqu'il n'existe pas de chemin entre deux termes t_k et t_l, la probabilité n'est pas calculable et doit être fixée arbitrairement ;

2 – Il existe des termes présents dans les règles d'association qui ne font pas (encore) partie du modèle de connaissances M. Ce sont les termes $t \in \mathcal{R} \setminus \mathcal{H}$. Pour les prendre en compte, il est possible d'étendre la distribution de probabilités à l'ensemble contenant à la fois les termes du modèle et les termes des règles, c'est-à-dire, à l'ensemble $\mathcal{R} \cup \mathcal{H}$. Tout terme $t \in \mathcal{R} \setminus \mathcal{H}$ se trouve ainsi inclus dans le modèle de connaissances en tant que terme isolé : aucune relation n'a été définie dans le modèle pour ce terme. Ce point nous ramène donc au problème évoqué au point 1 ci-dessus.

Nous considérons à présent que le modèle de connaissances est étendu de $\mathcal{H}$ à $\mathcal{R} \cup \mathcal{H}$. Trois stratégies peuvent être mises en œuvre pour traiter les cas où il n'existe pas de chemin entre deux termes.

Probabilité nulle : La première solution consiste à associer une probabilité de transition nulle pour tout couple de termes entre lesquels il n'existe pas de chemin dans le modèle de connaissances. Cette approche est intéressante lorsqu'il s'agit de règles simples. En effet, la valeur 0 permet d'identifier facilement les règles simples non taxinomiques à partir de la matrice des probabilités de transitions. L'analyste peut alors chercher à interpréter une règle taxinomique simple et, éventuellement, en déduire qu'il faut mettre à jour le modèle, c'est-à-dire, introduire un lien taxinomique entre les deux termes. En revanche, cette méthode défavorise les règles complexes. Il suffit qu'il existe un couple de termes sans transition pour que la vraisemblance de la règle soit nulle. Il n'y a donc pas de continuité dans la vraisemblance. Dès qu'il y a un couple de termes non taxinomique, la vraisemblance est nulle, inversement, lorsque la vraisemblance est non nulle, tous les couples de termes sont taxinomiques. Afin d'éviter ce problème, nous pouvons utiliser une valeur fixe arbitraire que nous définissons ci-dessous.

Valeur fixe arbitraire : Afin de réduire le nombre de règles dont la vraisemblance serait nulle, une seconde stratégie consiste à attribuer une valeur fixée arbitrairement aux couples de termes pour lesquels il n'existe pas de transition. Cette valeur est la probabilité minimale pour M : $\mathtt{P_M(t_k, t_l)} = \frac{1}{n+1}$ avec $\mathtt{n}$ le nombre de termes de $\mathcal{H}$. Ainsi, dans FIG. 5.3 (b),
$P_M(\mathtt{b}, \mathtt{c}) = \frac{1}{\mathtt{n}+1} = \frac{1}{6}$ car $\mathtt{n} = 5$.

Par exemple pour $(\mathtt{b}, \mathtt{c})$, nous avons : $P_M(\mathtt{b}, \mathtt{c}) = 1 \times \left(\left(3 \times \frac{1}{1}\right) + \left(2 \times \frac{1}{6}\right)\right)^{-1} = 0,3$ et pour $(\mathtt{b}, \mathtt{d})$: $P_M(\mathtt{b}, \mathtt{d}) = \frac{1}{6} \times \left(\left(3 \times \frac{1}{1}\right) + \left(2 \times \frac{1}{6}\right)\right)^{-1} = 0,05$.

Stratégie mixte : Une troisième stratégie — celle que nous adoptons pour notre expérimentation — consiste à associer une probabilité nulle dans le cas de règles simples et à appliquer une valeur fixe arbitraire dans le cas de règles complexes.

5.2.2 Vraisemblance des règles complexes

À présent, la distribution de probabilités est définie pour tous les couples de termes $(t_k, t_l) \in (\mathcal{R} \cup \mathcal{H}) \times (\mathcal{R} \cup \mathcal{H})$. Le calcul de la vraisemblance pour une règle complexe est fondé sur le calcul de la probabilité dans la théorie de l'activation de propagation pour chaque couple de termes issu du produit cartésien de la partie droite avec la partie gauche de la règle. Étant donnée une règle complexe $r : t_1 ... t_i \rightarrow t_{i+1} ... t_p$ (où $\mathtt{B} = \{t_1, ..., t_i\}$ et $\mathtt{H} = \{t_{i+1}, ..., t_p\}$), la probabilité du produit cartésien est :

$$P_M(\mathtt{B} \times \mathtt{H}) = \prod_{(t_k, t_l) \in \mathtt{B} \times \mathtt{H}} P_M(t_k, t_l) \tag{5.3}$$

qui s'écrit en extension :

$$P_M(\mathtt{B} \times \mathtt{H}) = \prod (P_M(t_1, t_{i+1}) ... P_M(t_1, t_p) ... P_M(t_i, t_{i+1}) ... P_M(t_i, t_p))$$

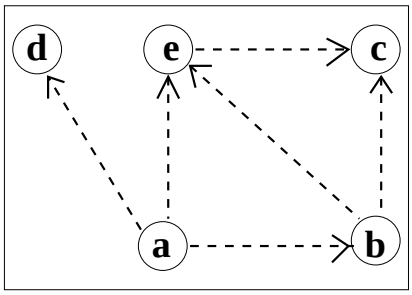
Nous observons cependant que plus le nombre de termes présents dans une règle est important, plus la probabilité $P_M(\mathtt{B} \times \mathtt{H})$ est faible. Or le nombre de termes présents dans une règle ne doit pas affecter la vraisemblance d'une règle. L'équation 5.4 généralise l'équation 5.3 en prenant la moyenne géométrique de la probabilité du produit cartésien. Nous définissons ainsi la vraisemblance d'une règle :

$$P_M(r_i) = \sqrt[|\mathtt{B_i}| \times |\mathtt{H_i}|]{P_M(\mathtt{B} \times \mathtt{H})} = \sqrt[|\mathtt{B_i}| \times |\mathtt{H_i}|]{\prod_{(t_k, t_l) \in \mathtt{B_i} \times \mathtt{H_i}} P_M(t_k, t_l)} \tag{5.4}$$

Nous soulignons que l'équation 5.4 pour les règles complexes est également applicable aux règles simples puisque $|\mathtt{B}| = |\mathtt{H}| = 1$.

5.3 Exemple formel

Prenons un exemple formel repris de [Pasquier *et al.*, 1999b] afin d'étudier le comportement de l'équation (5.4) pour identifier les règles d'association triviales. Soit le modèle de connaissances décrit par la FIG. 5.3(a) qui doit être interprété de la façon suivante : "a" EST-UN "b", "e" EST-UN "c", etc. Sur ce modèle, nous calculons la distribution de probabilités dont la matrice est donnée par FIG. 5.3(b). Nous allons étudier un ensemble de textes $\{d_1, \ldots, d_6\}$ décrits par un ensemble de termes d'indexation $\{a, \ldots, e\}$ par rapport à ce modèle (*cf.* FIG. 5.4(1)).



↱	a	b	c	d	e
a	0,22	0,22	0,11	0,22	0,22
b	0,05	0,30	*0,30*	*0,05*	0,30
c	0,10	0,10	0,60	0,10	0,10
d	0,10	0,10	0,10	0,60	0,10
e	0,06	0,06	0,40	0,06	0,40

FIG. 5.3 – (a) Le modèle de connaissances M– (b) Probabilités de transition pour M.

5.3.1 Comportement de la vraisemblance par rapport au modèle

Le but de cet exemple est de pouvoir illustrer le comportement de la vraisemblance sur un ensemble réduit de règles et un modèle de connaissances de petite taille. Vingt règles d'association, numérotées $r_1, \ldots, r_{20}$, sont extraites avec un support minimal $\texttt{minsup} = 1$ et une confiance minimale $\texttt{minconf} = 0,1$ et leurs valeurs de vraisemblance sont calculées à la FIG. 5.4 (2). Par exemple, pour la règle r_6, nous avons :
$P_M(\texttt{b} \Longrightarrow \texttt{a,c,e}) = (P_M(b,a) \times P_M(b,c) \times P_M(b,e))^{1/3} = (0,05 \times 0,3 \times 0,3)^{1/3} = 0,165$.

Nous classons ces règles en 8 classes. La colonne de gauche de FIG. 5.4 (2) contient des règles taxinomiques dites T-règles et la colonne de droite des règles non taxinomiques dites ¬T-règles, c'est-à-dire des règles non triviales qui relient des termes entre lesquels il n'existe pas de lien taxinomique. Les lignes de cette table regroupent les règles en fonction de leur structure, i.e., le nombre de termes présents dans B et H : en ligne 1 se trouvent les règles simples T-règles $(1,1)$ taxinomiques et ¬T-règles $(1,1)$ non taxinomiques, en ligne 2 les T-règles $(1,n)$ et ¬T-règles $(1,n)$, puis $(n,1)$ et (n,m) avec $n,\ m > 1$.

Texte	Termes
d_1	{acd}
d_2	{bce}
d_3	{abce}
d_4	{be}
d_5	{abce}
d_6	{bce}

n°	T	n/d	$P_M(r)$	n°	¬T	n/d	$P_M(r)$
r_1	b⇒e	0/1	0,300	r_{19}	e⇒b	1/0	0,000
r_{11}	a⇒c	0/0	0,111	r_{20}	c⇒a	1/0	0,000
r_2	b⇒c, e	0/2	0,300				
r_4	a⇒b, c, e	0/2	0,176	r_{13}	d⇒a, c	2/0	0,100
r_6	b⇒a, c, e	1/2	0,165	r_{14}	c⇒b, e	2/0	0,100
r_7	e⇒b, c	1/1	0,163	r_{15}	c⇒a, d	2/0	0,100
r_9	a⇒c, d	0/1	0,157	r_{16}	c⇒a, b, e	3/0	0,100
r_{10}	e⇒a, b, c	2/1	0,121				
r_5	b, c⇒e	1/1	0,173	r_{17}	c, e⇒b	2/0	0,081
r_3	a, b⇒c, e	0/3	0,217	r_{12}	b, c⇒a, e	3/1	0,110
r_8	a, e⇒b, c	1/2	0,160	r_{18}	c, e⇒a, b	4/0	0,081

FIG. 5.4 – (1) La base de données textuelles – (2) Mesure de vraisemblance pour les règles de l'exemple FIG.5.3(a) et le modèle M.

L'analyse de cet exemple formel montre que la vraisemblance permet d'attribuer une valeur forte aux règles triviales par rapport au modèle M et une valeur faible aux règles qui sont faiblement taxinomiques.

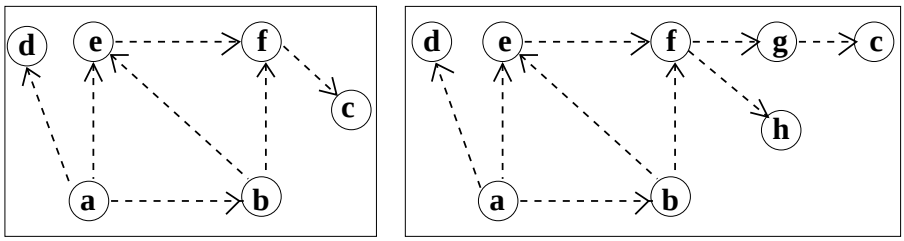


FIG. 5.5 – Les variantes M_1 et M_2 du modèle de connaissances M de FIG. 5.3 (a).

Nous constatons, de façon empirique, un seuil $s = 0,111$ qui sépare les T-règles ($P_M(r_i) \geqslant 0,111$) des ¬T-règles ($P_M(r_i) < 0,111$). Ce point sera discuté en section 5.3.2.

Les T-règles $(1,1)$ sont purement taxinomiques. En accord avec la définition 5.4 de la vraisemblance, plus la longueur du lien taxinomique est importante (la longueur est 1 pour r_1 et 2 pour r_{11}), plus la valeur de vraisemblance est faible ($P_M(r_1) > P_M(r_{11})$). Ainsi, r_{11} est moins triviale que r_1 (selon la propriété en § 5.1.3). À l'inverse, pour les ¬T-règles $(1,1)$, il n'y a pas de chemin de "e" vers "b" (règle r_{19}) ni de "c" vers "e" (règle r_{20}). Nous avons donc $P_M(r_{19}) = P_M(r_{20}) = 0$. Notons que la direction des relations taxinomiques est respectée.

Les T-règles $(1,n)$, $(n,1)$ et (n,m) de FIG. 5.4 (2), nous pouvons vérifier deux principes découlant des propriétés attendues de la vraisemblance que nous avons définie : (i) moins il y a de liens non taxinomiques entre les termes de B et de H, plus la valeur de la vraisemblance est élevée. (ii) plus les liens taxinomiques sont directs, plus la valeur de vraisemblance est élevée.

La colonne n/d de FIG. 5.4 (2) donne le nombre de couples de termes de $B \times H$ qui ne sont pas des relations taxinomiques (noté n) et le nombre de relations taxinomiques directes avec un chemin de longueur 1 (noté d). Par exemple, pour la règle r_8, 1/2 pour n/d signifie que, parmi les $|B \times H| = 2 \times 2 = 4$ couples de termes, un couple est non taxinomique, *i.e.* (e, b) et que deux couples sont taxinomiques directs, *i.e.* (a, b) et (e, c). Il y a donc un couple taxinomique indirect, *i.e.* (a, c).

5.3.2 Discussion

Les deux colonnes de FIG. 5.4 (2) séparent les règles taxinomiques des règles non taxinomiques. Cependant, la question de l'existence d'un seuil s pour la valeur de vraisemblance se pose. Nous montrons, dans cette section, que ce seuil ne peut être défini formellement et dépend du modèle choisi. Nous souhaitons également caractériser le comportement de notre méthodologie lorsque le modèle évolue. Pour cela, nous prenons le même ensemble de règles $\{r_1, \ldots, r_{20}\}$.

Si nous opérons sur le modèle des modifications majeures, par exemple, en créant un lien taxinomique entre deux termes (t_u, t_v) intervenant dans le calcul de la vraisemblance d'une règle r, alors l'analyse faite en section 5.3.1 montre que le calcul de vraisemblance sur le nouveau modèle donne une valeur plus forte pour r.

L'impact de modifications mineures du modèle engendre des changements pour la vraisemblance d'une règle qui sont plus subtils. Nous définissons une modification mineure comme suit : prenons les couples de termes (t_u, t_v) du modèle M qui interviennent dans le calcul de la vraisemblance des différentes règles. S'il existe un chemin entre t_u et t_v, alors le nouveau modèle que nous définissons préserve l'existence d'un chemin (éventuellement différent du chemin dans M).

TAB. 5.1 – Mesures P_{M_1} (à gauche) et P_{M_2} (à droite) pour les 20 règles de TAB. 5.4

n° T		P_{M_1}	n° ¬T		P_{M_1}	n° T		P_{M_2}	n° ¬T		P_{M_2}
r_1	b⇒e	0,286	**r_{12}**	**b,c⇒a,e**	**0,091**	r_1	b⇒e	0,231	**r_{12}**	**b,c⇒a,e**	**0,105**
r_2	b⇒c,e	0,187	r_{13}	d⇒a,c	0,083	r_2	b⇒c,e	0,127	r_{13}	d⇒a,c	0,062
r_3	a,b⇒c,e	0,149	r_{14}	c⇒b,e	0,083	r_5	b,c⇒e	0,117	r_{14}	c⇒b,e	0,062
r_5	b,c⇒e	0,148	r_{15}	c⇒a,d	0,083	r_4	a⇒b,c,e	0,116	r_{15}	c⇒a,d	0,062
r_4	a⇒b,c,e	0,143	r_{16}	c⇒a,b,e	0,083	r_3	a,b⇒c,e	0,108	r_{16}	c⇒a,b,e	0,062
r_9	a⇒c,d	0,119	**r_{10}**	**e⇒a,b,c**	0,074	r_9	a⇒c,d	0,092	**r_7**	**e⇒b,c**	0,052
r_6	b⇒a,c,e	0,109	**r_{11}**	**a⇒c**	0,069	r_6	b⇒a,c,e	0,073	**r_{11}**	**a⇒c**	0,046
r_8	a,e⇒b,c	0,104	r_{17}	c,e⇒b	0,063	r_8	a,e⇒b,c	0,069	**r_{10}**	**e⇒a,b,c**	0,044
r_7	e⇒b,c	0,092	r_{18}	c,e⇒a,b	0,063				r_{17}	c,e⇒b	0,043
			r_{19}	e⇒b	0,000				r_{18}	c,e⇒a,b	0,043
			r_{20}	c⇒a	0,000				r_{19}	e⇒b	0,000
									r_{20}	c⇒a	0,000

S'il n'existe pas de chemin entre t_u et t_v, alors le nouveau modèle préserve également le fait que ce chemin n'existe pas.

Nous souhaitons montrer avec l'ensemble $\{r_1, \ldots, r_{20}\}$ que :

1. Ces modifications mineures ont une incidence sur la valeur du seuil ;
2. Une règle classée dans M comme taxinomique peut se trouver classée parmi les règles non taxinomiques.

Nous introduisons deux modèles M_1 et M_2 (cf. FIG. 5.5) légèrement différents de M. Pour assurer que les modifications sur le modèle M sont mineures, ces modifications portent sur les termes *puits* (en théorie des graphes [Berge, 1985]), *i. e.*, des termes qui ne sont à l'origine d'aucune relation avec un autre terme. "c" et "d" vérifient cette propriété. Dans M_1, l'introduction du terme "f" rallonge tous les chemins entre un terme quelconque (différent de "c") et le terme "c". Le fait de n'introduire qu'un seul nouveau terme, augmente faiblement le facteur de branchement. Dans M_2, la longueur des chemins et le facteur de branchement sont augmentés par rapport à M_1.

Nous observons l'évolution des valeurs de vraisemblance affectées aux règles d'association $\{r_1, \ldots, r_{20}\}$. Dans M, les règles $\{r_1, \ldots, r_{11}\}$ étaient classées comme T-règles et les règles $\{r_{12}, \ldots, r_{20}\}$ comme ¬T-règles. Dans la mesure où la nature des liens entre termes dans M_1 et M_2 reste inchangée, nous considérons que la règle r_{12} reste la règle seuil séparant les T-règles et les ¬T-règles. Dans ces conditions, on observe un abaissement du seuil de $s = 0,111$ pour M, à $s_1 = 0,091$ pour M_1 et $s_2 = 0,105$ pour M_2.

Nous observons particulièrement les règles où le terme "c" est présent et nous remarquons que :

- La règle r_{10} est considérée comme taxinomique dans M. r_{10} a deux liens non taxinomiques ((e,a),(e,b)) contre un lien taxinomique direct (e,c). De ce fait, cette règle devrait être non taxinomique. L'affaiblissement du lien taxinomique (e,c) dans M_1 suffit à faire passer la règle r_{10} parmi les règles non taxinomiques. *A fortiori,* dans M_2 ;
- La règle r_7 a une valeur de vraisemblance légèrement supérieure à r_{10}, *i.e.* plus taxinomique, dans M que r_{10} puisqu'elle implique un lien non taxinomique (e, b) pour un lien taxinomique direct (e,c). Elle reste classée taxinomique dans M_1 mais devient non taxinomique

dans M_2 ;
- Pour M_1 et M_2 , la règle r_{11} purement taxinomique indirecte passe également parmi les règles non taxinomiques ;
- Seules les règles d'association ayant le terme "c" en partie droite H changent de statut, ce qui est conforme à nos attentes compte tenu des modifications choisies pour définir M_1 et M_2.

Ces résultats s'analysent comme suit :

1 – La mesure de vraisemblance que nous proposons se comporte de façon cohérente par rapport à sa définition lorsque nous l'appliquons sur les données et que la hiérarchie du modèle subit des modifications « mineures ». Le score de vraisemblance ne modifie pas le classement des règles d'association purement taxinomiques ou purement non taxinomiques ;

2 – Les valeurs pour le seuil des règles taxinomiques et non taxinomiques ne sont pas indépendantes du modèle de connaissances choisi. Par conséquent, les valeurs de seuils ne peuvent être fixées *a priori* ;

3 – Si les règles présentent des connaissances nouvelles, alors le modèle de connaissances peut être enrichi de façon incrémentale. Et nous avons le moyen de compléter, après validation par l'analyste, ce modèle avec de nouveaux termes identifiés grâce aux règles d'association.

5.4 Expérimentation sur des données textuelles

L'ensemble des textes de notre corpus a été indexé par le même ensemble des termes d'indexation $\mathcal{I}$, avec $|\mathcal{I}| = 632$ termes, que nous avons décrit en § 4.3.1. Nous avons obtenu 347 règles d'association avec les seuils $\texttt{minsup} = 10$ et $\texttt{minconf} = 0,8$ (*cf.* l'expérimentation décrite en § 4.3.2.1). Le modèle de connaissances utilisé pour la sélection des règles est issu du métathésaurus UMLS [UMLS, 2000]. Il contient quelques 125 000 termes venant d'environ 100 thésaurus médicaux et biologiques. Alors que le métathésaurus contient 11 relations différentes, nous nous sommes limités aux relations de type EST-UN ("PAR" : parent). Ce modèle ne couvre $\mathcal{I}$ que partiellement. Au total, 438 termes de $\mathcal{H}$ sont identiques à ceux de $\mathcal{I}$ ($\approx 70\%$ des termes). Le modèle est donc incomplet. Parmi les 347 règles, en fixant un seuil $\texttt{s} = 0$, 136 d'entre elles (soit $\approx 40\%$) ont une vraisemblance $\neq 0$ et ce sont toutes des règles complexes. Nous nous sommes focalisés sur des règles de vraisemblance nulle soit 211 règles dont 46 simples et 165 complexes.

Nous avons réalisé une classification des règles d'association en règles triviales / non triviales. Certaines règles non rejetées sont triviales mais l'incomplétude du modèle n'a pas permis de les identifier. De même, certaines règles rejetées ne sont pas triviales en raison des probabilités de transition très élevées de certains liens taxinomiques par rapport à d'autres liens non taxinomiques.

Pour évaluer la qualité de cette classification, nous déterminons 4 classes de règles : les vraies-positives (non taxinomiques $\neg$T et évaluées comme non triviales), les fausses positives ($\neg$T, mais qui sont triviales), les vraies-négatives (taxinomiques et triviales) et les fausses-négatives (taxinomiques, mais non triviales).

L'évaluation des règles est réalisée par l'analyste afin de leur assigner une de ces 4 classes. Parmi les 136 règles qui ont été rejetées par notre processus de sélection des règles, 122 (soit 90%) sont vraies-négatives et 14 (soit 10%) sont fausses-négatives. Le faible pourcentage de règles fausses-négatives montre que la mesure de vraisemblance que nous proposons est capable d'identifier les règles taxinomiques.

Parmi les 211 règles d'association qui sont non rejetées ($\neg$T), il y a 115 (soit 55%) qui sont vraies-positives et 96 (soit 45%) qui sont fausses-positives, *i.e.* déjà connues de l'analyste. Le fort pourcentage de règles vraies-positives s'explique par l'incomplétude du modèle disponible dans le domaine traité par les textes. En revanche, le fort pourcentage de fausses-positives nous permet de souligner les termes absents du modèle et de pouvoir l'enrichir de ces nouveaux termes.

TAB. 5.2 résume les résultats obtenus sur nos textes de biologie moléculaire.

TAB. 5.2 – Confrontation : Connaissances de l'analyste / Calcul de la vraisemblance selon M

	Évaluation par rapport aux connaissances de l'analyste	
Vraisemblance calculée sur M	Taxinomique	Non taxinomique
Taxinomique	90% (vraies négatives)	10% (fausses négatives)
Non taxinomique	45% (fausses positives)	55% (vraies positives)

Sur nos données de biologie moléculaire, nous voyons que la mesure de vraisemblance se comporte en accord avec sa définition 5.4. La distributivité du calcul de vraisemblance pour une règle, c'est-à-dire les probabilités de chaque terme de B avec chacun des termes H, montre que certaines configurations des termes en B et H donnent la même valeur de vraisemblance pour une règle. Par exemple, la permutation d'un terme de B et d'un terme de H. Ce qui est illustré par les valeurs de vraisemblance pour les règles taxinomiques $\{r_1, r_2, r_3, r_4\}$ du tableau ci-dessous. Nous voyons aussi qu'une sous-règle $\{r_6\}$ d'une règle non taxinomique $\{r_5\}$ est non taxinomique.

N°	Règle	Vraisemblance	Taxinomique
r_1	0,00104	"aztreonam" $\sqcap$ "lactams"$\Longrightarrow$"lactam"	Oui
r_2	0,00104	"aztreonam" $\sqcap$ "lactam"$\Longrightarrow$"lactams"	Oui
r_3	0,00131	"aztreonam" $\sqcap$ "lactamase" $\sqcap$ "lactams" $\Longrightarrow$ "lactam"	Oui
r_4	0,00131	"aztreonam" $\sqcap$ "lactam" $\sqcap$ "lactamase" $\Longrightarrow$ "lactams"	Oui
r_5	0,00000	"aztreonam" $\sqcap$ "clavulanic acid" $\sqcap$ "enzyme"$\Longrightarrow$"lactamase"	Non
r_6	0,00000	"aztreonam" $\sqcap$ "clavulanic acid" $\Longrightarrow$"lactamase"	Non

5.5 Enrichissement incrémental du modèle terminologique

Nous proposons une discussion sur l'enrichissement du modèle de connaissance et sur l'impact de cet enrichissement. Afin de minimiser l'intervention de l'analyste et d'automatiser le processus d'enrichissement du modèle terminologique, nous proposons une stratégie de placement des termes dans le modèle terminologique. Nous illustrons, en FIG. 5.6 et FIG. 5.7, l'insertion de termes dans le modèle grâce à des règles simples, par exemple $\{\mathtt{R}_1, \ldots, \mathtt{R}_4\}$, et à des règles complexes, par exemple $\{\mathtt{R}_5, \mathtt{R}_6, \mathtt{R}_7\}$, décrites ci-dessous.

La possibilité d'enrichir notre modèle de connaissances (*i.e.* l'UMLS) à partir des règles classées comme étant non taxinomiques est un atout majeur de l'approche que nous proposons. En effet, supposons qu'une règle R soit classée comme non taxinomique dans notre modèle de connaissances qui nous sert initialement à classer les règles du corpus de biologie moléculaire La règle sera identifiée parmi l'ensemble des règles car $\mathtt{P}_{\mathtt{UMLS}}(\mathtt{R}) = 0$. Supposons que l'analyste la juge

triviale. L'enrichissement consiste à ajouter les termes de la règle qui en sont absents. Le nouveau terme sera placé dans le modèle en liens taxinomiques avec les termes donnés par la règle. Nous illustrons notre discussion avec des exemples de règles qui ne demandent pas d'expertise au lecteur. Nous distinguons deux cas, les règles simples et complexes.

R_1 : "outer membrane protein" $\Longrightarrow$ "membrane protein"
R_2 : "polymyxin b" $\Longrightarrow$ "polymyxin"
R_3 : "dna topoisomerase atp hydrolysing" $\Longrightarrow$ "gyrase"
R_4 : "disk diffusion" $\Longrightarrow$ "diffusion test"

Règles simples : Le cas d'une règle simple produit une modification mineure dans le modèle UMLS. Si le terme absent t_{new} est en partie B de la règle, alors nous pouvons ajouter un lien taxinomique EST-UN entre t_{new} et t_H. Par exemple, la règle R_1 est interprétée dans le contexte de biologie moléculaire comme la « membrane externe d'une protéine » EST-UNE « membrane d'une protéine ». Nous plaçons le premier terme en dessous du second.

Si le terme absent t_{new} est en partie H, alors nous pouvons ajouter un lien taxinomique EST-UN entre t_B et t_{new}, puis nous pouvons ajouter un lien taxinomique entre t_{new} et tous les termes auxquels t_B est auparavant relié. Par exemple, pour la règle R_2, la « polymyxin b » EST-UNE « polymyxin ». De plus, l'analyste sait que « polymyxin b » EST-UN médicament antibiotique : « drug. antibiotic ». Nous pouvons donc insérer « polymyxin » entre « polymyxin b » et « drug. antibiotic ».

Si les termes t_B et t_H sont tous les deux absents de l'UMLS, alors : – (i) si l'analyste découvre que ce sont deux termes synonymes, alors nous pouvons les relier entre eux et nous les plaçons sous le terme racine « NO_PARENT ». Par exemple, dans la règle R_3, « dna topoisomerase atp hydrolysing » et « gyrase » sont deux termes qui désignent la même enzyme. Nous pouvons donc les relier mutuellement et les placer sous « NO_PARENT ». – (ii) si l'analyste découvre que ce sont deux termes taxinomiques alors t_B est relié à t_H, qui est, lui-même, relié à « NO_PARENT ». Par exemple, dans la règle R_4, la « diffusion par disque » EST-UN « test de diffusion » des bactéries dans une souche contaminée. De plus, « test de diffusion » est placé sous« NO_PARENT ».

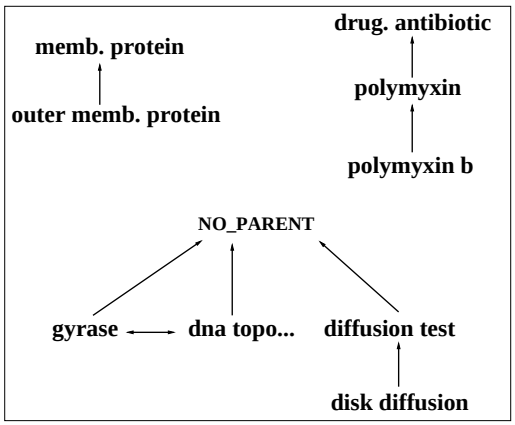


FIG. 5.6 – Schéma de placement pour les règles simples.

Règles complexes : Dans le cas d'une règle complexe, si le terme absent t_{new} est en partie B de la règle, alors nous pouvons ajouter un lien taxinomique EST-UN entre t_{new} et le parent de t_B. Les termes t_B et t_{new} sont placés comme des frères dans l'UMLS. Par exemple, la règle R_5 est

interprétée comme l'« infection » (terme absent du modèle) de l'« appareil urinaire » (présent dans l'UMLS) EST-UNE « infection urinaire » (absent de l'UMLS). Nous plaçons « infection » en tant que frère de « appareil urinaire » qui est, à l'origine, placé sous « NO_PARENT » en insérant « infection urinaire » entre « appareil urinaire » et « NO_PARENT ».

$\mathtt{R}_5$: "infection" $\sqcap$ "urinary tract" $\Longrightarrow$ "urinary infection"
$\mathtt{R}_6$: "clavulanic acid" $\Longrightarrow$ "enzyme" $\sqcap$ "β-lactamase"
$\mathtt{R}_7$: "agar dilution" $\sqcap$ "microdilution" $\Longrightarrow$ "dilution method"

Si le terme absent $\mathtt{t}_{\mathtt{new}}$ est en partie $\mathtt{H}$, alors $\mathtt{t}_{\mathtt{new}}$ est placé comme frère de $\mathtt{t}_{\mathtt{H}}$ et une relation taxinomique entre les termes de $\mathtt{t}_{\mathtt{B}}$ et $\mathtt{t}_{\mathtt{new}}$ est établie. Par exemple, dans la règle $\mathtt{R}_6$, l'« acide clavulanique » inhibe l'« enzyme » appelée « β-lactamase » (absent de l'UMLS). Nous plaçons « β-lactamase » comme terme frère de « enzyme » et nous ajoutons un lien taxinomique entre « acide clavulanique » et « β-lactamase ». Le terme « β-lactamase » devrait être placé, par la suite, sous le terme « enzyme » car la *β-lactamase* est une *enzyme*, mais seules les connaissances de l'expert ou la présence d'une autre règle, comme par exemple « β-lactamase » $\Longrightarrow$ « enzyme », pourrait enrichir le modèle de cette nouvelle relation taxinomique.

Si tous les termes de la règle sont absents, alors nous ajoutons un sous-arbre à « NO_PARENT » dans lequel tous les termes $\mathtt{t}_{\mathtt{B}}$ sont reliés à tous les termes $\mathtt{t}_{\mathtt{H}}$ qui, à leur tour, sont reliés à « NO_PARENT ». Par exemple, dans la règle $\mathtt{R}_7$, la « dilution agar » et la « microdilution » SONT-DES « méthodes de dilution ». Nous plaçons « agar dilution » et « microdilution » sous le terme « dilution method » que nous relions à « NO_PARENT ».

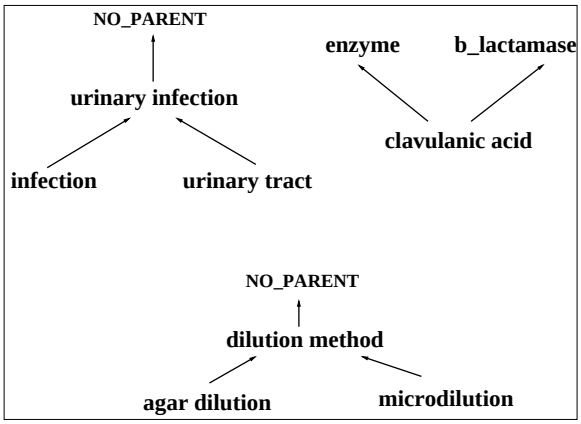


FIG. 5.7 – Schéma de placement pour les règles complexes.

Le processus d'enrichissement du modèle de connaissances est incrémental. Nous pouvons envisager de réutiliser le modèle enrichi pour classer à nouveau les règles selon la mesure de vraisemblance. L'implication dans une règle d'association ne signifie pas, par défaut, la relation EST-UN. Par exemple, nous ne pouvons pas dire pour « urinary infection » $\Longrightarrow$ « urinary tract » qu'un « appareil urinaire » EST-UNE « infection urinaire ». Cette règle simple est classée non taxinomique. La validation par l'analyste est, donc, obligatoire avant toute modification du modèle de connaissances.

5.6 Approches comparables

De nombreux travaux de recherche en fouille de textes se sont concentrés sur la façon de gérer le très grand nombre de règles d'association extraites à partir de corpus de textes. Cependant, la plupart de ces travaux ont abordé le problème du point de vue statistique, sans chercher à y introduire des connaissances.

Les travaux de [Basu *et al.*, 2001] se distinguent et constituent une exception par rapport à ce point puisqu'ils proposent une approche exploitant une base de connaissances pour réduire l'ensemble des règles. Au lieu de s'ancrer dans une approche probabiliste comme la nôtre, ils introduisent une mesure de similarité sémantique entre mots.

Les règles d'association généralisées [Srikant et Agrawal, 1995; Han, 1995; Hipp *et al.*, 2002] constituent une approche différente puisque l'extraction des règles exploite le fait que les termes appartiennent à différents niveaux d'une ontologie. Si l'on connaît les ancêtres d'un terme, alors un critère est appliqué afin de contraindre le processus d'extraction (bloquer les règles qui introduisent à la fois un terme et son ancêtre, par exemple). Ce processus reste d'une grande complexité calculatoire et le nombre de règles généré est au final encore plus élevé. Enfin, un travail similaire exploitant un modèle de connaissances pour la classification de termes est proposé par [Resnik, 1999]. La similarité est fondée sur l'information mutuelle. L'information mutuelle entre deux termes $\mathtt{t}_1$ et $\mathtt{t}_2$ est définie par : $\mathtt{I}(\mathtt{t}_1, \mathtt{t}_2) = \mathtt{P}(\mathtt{t}_1) - \mathtt{P}(\mathtt{t}_1|\mathtt{t}_2)$, c'est-à-dire la différence entre la probabilité d'avoir le terme $\mathtt{t}_1$ et la probabilité conditionnelle d'avoir ce même terme sachant le terme $\mathtt{t}_2$. L'information mutuelle entre $\mathtt{t}_1$ et $\mathtt{t}_2$ vaut zéro si et seulement si $\mathtt{t}_1$ et $\mathtt{t}_2$ sont statistiquement indépendants, sans faire aucune hypothèse sur la relation *a priori* entre $\mathtt{t}_1$ et $\mathtt{t}_2$. La mesure de similarité définie dans ces travaux sert à affecter un seul sens à des termes ambigus selon la proximité sémantique qu'ils ont avec leurs voisins dans un thésaurus (*i.e.* WORDNET).

Enfin, un certain parallèle peut être fait avec l'approche de la gestion du nombre de règles extraites par l'utilisation des connaissances de l'analyste, présentée en § 4.1.2 (page 74). Le processus de FdT que nous définissons est fondé sur une gestion *a posteriori* des règles extraites par le classement des règles par rapport aux connaissances du domaine. Les approches dans [Sahar, 1999; Liu *et al.*, 1999b; Ganter, 1999] font appel aux connaissances de l'analyste dans le processus d'élagage d'un certain nombre de règles qu'ils jugent inintéressantes avant (parfois pendant) l'activation des techniques de FdD. Ces approches présentées en § 4.1.2 ne sont pas reproductibles lorsqu'on change de domaine de fouille. En effet, il faut redéfinir, avec l'analyste, les connaissances qu'il ne souhaite pas extraire pour chaque domaine de fouille, et ensuite, solliciter à nouveau son avis pour l'interprétation des résultats. Notre approche par l'utilisation de la mesure de vraisemblance permet de trier les règles d'association extraites et place l'intervention de l'analyste en bout de chaîne de FdT uniquement pour la prise de décision finale.

5.7 Conclusion

Dans ce chapitre, nous avons proposé une méthodologie de sélection des règles d'association par l'utilisation d'un modèle de connaissances. Cette méthodologie applique successivement un processus symbolique d'extraction de règles d'association et un processus probabiliste de calcul d'une mesure de vraisemblance entre une règle et un modèle de connaissances.

Nous avons appliqué cette mesure pour la sélection des règles non taxinomiques en rejetant celles qui sont taxinomiques car elles sont triviales dans un domaine de spécialité. Nous avons

étudié et montré que les propriétés de la mesure de vraisemblance est cohérente avec les attentes d'un analyste en fouille de textes. Cette mesure garde une cohérence suite à des variations légères du modèle. Enfin, la méthodologie que nous avons présentée permet une démarche incrémentale en fouille de textes car le modèle est progressivement enrichi et la mesure de vraisemblance d'une règle est reflétée par cet enrichissement.

Cette approche peut être étendue suivant plusieurs directions. Tout d'abord, nous souhaitons généraliser la mesure de vraisemblance pour prendre en compte d'autres relations que EST-UN. La relation de causalité entre termes est transitive et peut, à ce titre, être appliquée de la même manière que la relation taxinomique EST-UN.

Si nous disposons d'un modèle d'acceptation (par opposition à rejet) de règles, alors nous pouvons envisager une approche soulignant la conformité d'une règle à un modèle. Par exemple, un système logique de preuves. une expérimentation dans ce domaine nous paraît intéressante. La méthodologie que nous avons présentée ne considère pas les liens existant à l'intérieur de `B` ou à l'intérieur de `H`. Il nous semble intéressant de proposer une variante de cette mesure qui prend en compte les liens entre termes apparaissant du même côté d'une règle. Le choix d'un seuil empirique demeure délicat. Il est fixé par jugement de l'analyste. Nous envisageons de définir un moyen pour apprendre ce seuil à partir de la topologie du modèle choisi. Par exemple, la probabilité qu'un terme du modèle apparaisse dans une règle, le nombre de termes dans le modèle et le nombre de termes présents dans les règles peuvent constituer des paramètres pour l'apprentissage du seuil de vraisemblance. L'enrichissement du modèle peut également être automatisé, nous explorons quelque pistes pour les termes synonymes qui apparaissent dans les règles non rejetées.

Chapitre 6

Conclusion et perspectives

6.1 Conclusion

Ce mémoire de thèse présente une expérience complète combinant une méthode de traitement automatique de corpus de textes et un processus de fouille de données qui prend en compte une évaluation des résultats par l'analyste. L'ensemble de ces processus constitue une approche de la « fouille de textes ». Nous soulignons l'exigence d'avoir une bonne indexation de départ pour extraire des règles informatives. En revanche, cette indexation peut être améliorée en filtrant les termes périphériques au domaine et enrichie par de nouveaux concepts trouvés dans les règles d'association. Nous suggérons une classification des règles selon différents « points de vue » grâce à l'utilisation de différentes mesures de qualités des règles. Bien que cela induit une subjectivité liée à toute expertise humaine, nous avons confronté de façon opérationnelle la valeur des mesures de qualité des règles présentées à l'analyste. Nous avons trouvé qu'une combinaison des mesures d'*intérêt* et de *conviction* permettent de classer les règles qui sont les plus significatives et qui illustrent la le cas d'une distribution *rare* des termes constituant des *pépites* de connaissances potentielles. Nous avons identifié ce cas comme étant le plus informatif du point de vue de l'analyste. Les mesures de qualité de *nouveauté* et de *satisfaction* permettent de distinguer des règles à faible mesure de *dépendance.* Le but de cette méthodologie est de s'assurer de l'apport de mesures qualitatives pour classer en premier, parmi les nombreuses règles extraites, celles qui sont les plus informatives et qui conviennent le mieux pour un analyste du domaine. La connaissance qualitative se définit parfois comme une connaissance non susceptible de se prêter à l'attribution de valeurs numériques ou de mesures. Cette vision peut être considérée de façon plus souple. La complémentarité entre connaissances qualitatives et quantitatives est définie au cours de notre travail. Nous avons défini une mesure qualitative à partir de mesures quantitatives : une mesure de vraisemblance pour une règle extraite par un processus symbolique dépendant d'un modèle donné par une distribution de probabilités des termes dans un modèle terminologique.

Nous avons réalisé le système TAMIS qui permet de classer les règles selon différentes mesures de qualité dites *syntaxiques* car les mesures sont calculées à partir des données (de la distribution des termes dans les textes). Nous proposons un algorithme pour combiner les différents classements des règles (*cf.* algorithme 4, § 4.2.5). D'autre part, nous avons proposé une mesure de vraisemblance dite *sémantique* qui s'appuie sur des connaissances du domaine. La vraisemblance permet de classer les règles, en éliminant les règles déjà présentes dans un modèle de connaissances du domaine.

Spécifications pour un système de fouille de textes
Avec le développement de la veille stratégique (appelée également *business intelligence*), les grandes compagnies industrielles ont développé des outils pour la fouille de textes en s'appuyant sur des techniques numériques et symboliques de recherche documentaire, de catégorisation de documents, etc.
Un sondage de KDNUGGETS[27] réalisé en septembre 2003 auprès de 111 chercheurs ou organisations différentes pose les deux questions suivantes :

Q_1 : quelle est votre expérience en fouille de textes ?

Pourcentage	Réponse
18%	A utilisé des outils libres ou de recherche
16%	Utilise actuellement les outils libres ou de recherche
12%	A utilisé les outils commerciaux
14%	Utilise actuellement les outils commerciaux
12%	Aucun, mais envisagé dans les six prochains mois
27%	Aucun, non envisagé dans les six prochains mois

Ce sondage n'est pas fait sur un échantillon statistique représentatif de la communauté de FdT. Cependant, comme le détaille le tableau ci-dessus, il révèle que 73% des sondés s'intéressent à la FdT, ce qui traduit un fort besoin et un grand intérêt pour la fouille de texte. Nous observons également que l'utilisation, passée ou actuelle, des outils libres ou de recherche est supérieure à l'utilisation des outils commerciaux.

Q_2 : quel pourcentage de temps dans votre projet d'extraction de données est passée sur la suppression des erreurs et la préparation ?

Parmi 187 réponses à cette deuxième question, plus de la moitié déclare que la préparation des données demande 60% de l'effort en FdT.

D'après l'expérience qui fait suite à notre travail, nous estimons que le besoin exprimé par les scientifiques et les industriels de faire de la FdT passe par le développement de systèmes capables de traiter de grandes masses de données. De plus, l'utilisateur doit pouvoir passer d'une base de textes de petite échelle (un compte rendu de réunion, une brochure ou un article scientifique) vers une grande échelle (une documentation technique, des brevets, des textes de loi du journal officiel, le Web) de façon transparente. Le processus doit être adaptable pour des données structurées (les bases de données), les données semi-structurées (les documents du Web) ou les données peu structurées (les textes en langage naturel). Le processus de fouille de textes ne doit pas comporter de modules *ad hoc* à la base de textes qui sert d'application durant la phase de développement et de tests. Un système de fouille doit, par conséquent être générique et indépendante du domaine des textes.

6.2 Perspectives

À court terme, nous pensons que le processus de fouille de textes gagne en précision et en qualité des connaissances extraites lorsque les termes qui indexent les textes n'ont pas tous un statut identique. Les travaux de [Cai *et al.*, 1998] qui définissent une fonction de pondération sur les propriétés des individus d'une base de données sont une piste intéressante. Nous pensons appliquer

[27] KDNUGGETS est le site Web de référence pour la communauté de fouille de données.

cette pondération aux termes pour représenter le contenu des textes. L'étude du seuil $\mathtt{s}$ de vraisemblance qui sépare les règles d'association taxinomiques de celles qui apportent des connaissances nouvelles est à mener très précisément. Soit $\mathtt{n} = |\mathcal{H}|$ le nombre de termes du modèle terminologique $\mathcal{H}$ du domaine. Nous pensons, dans une première étape, que le seuil empirique pourrait être $\mathtt{s} = \frac{1}{\mathtt{n}+1}$ pour les règles dites *simples*. Par la suite, il nous faut affiner ce critère selon le type de textes et les valeurs de *support* et de *confiance* des règles que l'on veut extraire. L'apprentissage automatique de seuils numériques comme critère servant à prendre des décision ultérieures est classique non seulement en apprentissage statistique mais également dans le domaine de l'apprentissage symbolique en général et plus particulièrement en fouille de données. Par exemple, [Claveau *et al.*, 2003] combine les mesures de *rappel* et de *précision* pour l'extraction et la découverte des patrons nom-verbe (*qualia*) par une technique symbolique de programmation logique inductive (PLI) pour l'extension des requêtes en recherche d'information.

À moyen terme, la reproduction de l'expérimentation avec un analyste différent ou sur un corpus différent nous permettra d'avoir une validation systématique des deux approches, syntaxique et sémantique, que nous proposons. La combinaison de relations, autres que la relation EST-UN, qui possèdent les propriétés d'ordre partiel nous paraît un problème difficile pour lequel il nous faut définir une stratégie de combinaison. Nous souhaitons généraliser la mesure de vraisemblance pour prendre en compte d'autres relations que EST-UN. La relation de causalité, par exemple, repose sur une transitivité des liens entre les termes et peut, à ce titre, être appliquée de la même manière que la relation taxinomique EST-UN. La seule chose dont nous avons besoin dans cette relation est la fermeture transitive pour faire un pré-ordre (la réflexivité existe par défaut). Nous n'exploitons pas l'antisymétrie dans cet ordre. Le pré-ordre suffit et afin de construire un ordre partiel, nous pouvons partir d'une relation quelconque, la rendre transitive et réflexive ; encore que la réflexivité n'intervient pas dans ce pré-ordre. L'enrichissement du modèle de connaissances à partir des règles d'association extraites des textes se fait par l'ajout de nouveaux termes dans la hiérarchie du modèle. La stratégie de placement des termes contenus dans les règles n'est pas triviale. Une étude de la sémantique du type flèche — utilisé dans les règles d'inférence et de typage notamment pour la construction de grammaires de dépendances [Dikovsky, 2001] pour le langage naturel — s'avère nécessaire pour l'ajout de liens entre termes de la hiérarchie. Enfin, une projection du treillis des concepts, extrait à partir des textes, sur le modèle de connaissances du domaine est une tâche qui nous paraît difficile mais intéressante puisqu'il nous faut passer du niveau terminologique vers le niveau de concepts dont la sémantique d'intension / extension est différente de celle que nous manipulons durant nos travaux de thèse.

À long terme, nous comptons réaliser un système d'extraction d'information, en amont de TAMIS, pour travailler sur des données de type objets. Pour cela, il nous faut (i) identifier les fragments de textes pertinents, *i.e.* contenant une information, (ii) définir une structure de représentation de l'information, (iii) développer des règles d'inférence permettant d'identifier l'information, (iv) remplir la structure proposée.

Bibliographie

[Agrawal *et al.*, 1993] R. Agrawal, T. Imielinski et A. Swami. Mining Associations Rules between Sets of Items in Large Databases. Dans *Proc. of the ACM-SIGMOD Int'l Conf. on Management of Data*, pages 207–216, Washington, USA, 1993.

[Agrawal et Srikant, 1994] R. Agrawal et R. Srikant. Fast algorithms for mining association rules in large databases. Dans *Proc. of the 20^{th} Int'l Conf. on Very Large Databases (VLDB'94)*, pages 478–499, Santiago, Chile, 1994. Extended version : IBM Research Report RJ 9839.

[Anick et Pustejovsky, 1990] P. Anick et J. Pustejovsky. An Application of lexical Semantics to Knowledge Acquisition from Corpora. Dans *Proc. of the 30^{th} Int'l Conf. on Computational Linguistics (COLING'90)*, volume 3, pages 7–12, Helsinki, 1990.

[Apté *et al.*, 1998] C. V. Apté, F. Damerau et S.M. Weiss. Text Mining with Decision Trees and Decision Rules. Dans *Proc. of the Conf. on Automated Learning and Discovery (CONALD'98)*, Pittsburgh, USA, 1998. Carnegie-Mellon University. 4 pages.

[Azé et Roche, 2003] J. Azé et M. Roche. Une application de la fouille de textes : l'extraction des règles d'association à partir d'un corpus spécialisé. Dans *Actes de Extraction et Gestion des Connaissances (EGC'03)*, rédacteur D. Boulanger M.S. Hacid, Y. Kodratoff, volume 17 de *RSTI/RIA-ECA*, pages 283–294, Lyon, 2003. Hermès Éditions.

[Azé, 2003] J. Azé. Une nouvelle mesure de qualité pour l'extraction de pépites de connaissances. *Extraction des connaissances et apprentissage (ECA)*, 17(1) :171–182, 2003.

[Barbut et Monjardet, 1970] M. Barbut et B. Monjardet. *Ordre et Classification : Algèbre et Combinatoire*, volume I & II. Classique Hachette, Paris, 1970.

[Bastide, 2000] Y. Bastide. *Data mining : algorithmes par niveau, techniques d'implantation et applications*. Thèse de doctorat, Université Blaise Pascal, Clermont-Ferrand II, 2000.

[Basu *et al.*, 2001] S. Basu, R. J. Mooney, K. V. Pasupuleti et J. Ghosh. Evaluating the Novelty of Text-Mined Rules using Lexical Knowledge. Dans *Proc. of the 7^{th} ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD'01)*, pages 233–238, San Francisco, USA, 2001. ACM Press.

[Bayardo et Agrawal, 1999] R. J. Bayardo et R. Agrawal. Mining the Most Interesting Rules. Dans *Proc. of the 5^{th} ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD'99)*, pages 145–154, 1999.

[Berge, 1985] C. Berge. *Graphs (2^{nd} revised edition)*, volume 6 de *North-Holland Mathematical Library*. North-Holland, Amsterdam, The Netherlands, 1985. 413 pages.

[Biber, 1992] D. Biber. The multidimensional approach to linguistic analyses of genre variation : An overview of methodology and finding. *Computers in the Humanities*, 26(5–6) :331–347, 1992.

[Bournaud et Courtine, 2001] I. Bournaud et M. Courtine. Un Espace de Généralisation pour l'Extraction de Règles d'Association. Dans *Actes EGC'01 : Journées Extraction et Gestion des Connaissances*, rédacteurs H. Briand et F. Guillet, volume 1 de *1-2*, pages 129–140, Nantes, France, 2001. Éditions Hermès.

[Breiman *et al.*, 1984] L. Breiman, J.H. Friedman, R.A. Ohlsen et C.J. Stone. *Classification and Regression Trees*. Wadsworth and Brooks Group, 1984.

[Brill et Pop, 1999] E. Brill et M. Pop. Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging. Dans *Natural Language Processing Using Very Large Corpora*, rédacteurs N. Ide et J. Véronis, volume 11 de *Text, Speech and Language Technology*, page 13 pages. Kluwer Academic Publishers, 1999.

[Brin *et al.*, 1997] S. Brin, R. Motwani, J. Ullman et S. Tsur. Dynamic Itemset Counting and Implication Rules for Market Basket Data. Dans *Proceedings of the ACM SIGMOD'97 Conference on Management of Data*, volume 36, pages 255–264, Tucson, USA, 1997.

[Cai *et al.*, 1998] C.H. Cai, A.W.C. Fu, C.H. Cheng et W.W. Kwong. Mining Association Rules with Weighted Items. Dans *Proc. of the Int'l Database Engineering and Applications Symposium (IDEAS'98)*, pages 68–77, Cardiff, UK, 1998. IEEE Computer Society.

[Callan *et al.*, 1992] J P. Callan, W. B. Croft et S. M. Harding. The INQUERY Retrieval System. Dans *Proc. of the 3^{rd} Int'l Conf. on Database and Expert Systems Applications (DEXA'92)*, rédacteurs A. M. Tjoa et I. Ramos, pages 78–83, Valence, Spain, 1992. Springer-Verlag.

[Calvanese *et al.*, 1995] D. Calvanese, G. De Giacomo et M. Lenzerini. Structured Objects : Modeling and Reasoning. Dans *Proc. of the 4^{th} Int. Conf. on Deductive and Object-Oriented Databases (DOOD'95)*, volume 1013 de *Lecture Notes in Computer Science – LNCS*, pages 229–246. Springer-Verlag, 1995.

[Canu, 2002] S. Canu. *Décision et reconnaissance des formes en Signal*, chapitre Modèles connexionnistes et machines à vecteurs supports pour la décision. Hermès, 2002.

[Carpineto et Romano, 1996] C. Carpineto et G. Romano. Information Retrieval through Hybrid Navigation of Lattice Representations. *Int'l Journal of Human-Computer Studies*, 45 :553–578, 1996.

[Carpineto et Romano, 2000] C. Carpineto et G. Romano. Order-Theoretical Ranking. *Journal of the American for Information Science (JASIS'00)*, 51(7) :587–601, 2000.

[Chen, 1976] P. P. Chen. The Entity-Relationship Model – Toward a unified view of data. *ACM Transactions in Database Systems*, 1(1) :9–36, 1976.

[Cherfi *et al.*, 2003a] H. Cherfi, A. Napoli et Y. Toussaint. Towards a Text Mining Methodology Using Frequent Itemsets and Association Rule Extraction. Dans *Journées d'informatique Messine (JIM'03)*, pages 285–294, Metz, France, 2003. E. SanJuan, INRIA Lorraine.

[Cherfi *et al.*, 2003b] H. Cherfi, A. Napoli et Y. Toussaint. Vers une méthodologie de fouille de textes s'appuyant sur l'extraction de motifs fréquents et de règles d'association. Dans *Actes de la Conférence d'Apprentissage (CAp'03)*, rédacteur R. Gilleron, pages 61–76, Laval, 2003. dans le cadre de la plateforme (AFIA'03), Presses universitaires de Grenoble.

[Cherfi *et al.*, 2004a] H. Cherfi, D. Janetzko, A. Napoli et Y. Toussaint. Sélection de règles d'association par un modèle de connaissances pour la fouille de textes. Dans *Actes de CAp'04 : Conférence d'Apprentissage*, rédacteurs M. Liquière et M. Sebban, pages 191–206, Montpellier, 2004. Presses Universitaires de Grenoble. Voir aussi [Janetzko et al., 2004].

[Cherfi *et al.*, 2004b] H. Cherfi, A. Napoli et Y. Toussaint. Towards a Text Mining Methodology Using Frequent Itemsets and Association Rules. *Dans Soft Computing Journal*, 2004. 11 pages. Special Issue on "Recent Advances in Knowledge and Discovery". Springer. À paraître.

[Cherfi et Toussaint, 2002a] H. Cherfi et Y. Toussaint. Adéquation d'indices statistiques avec l'interprétation de règles d'association. Dans *Actes des 6^{mes} Journées internationales d'Analyse statistique des Données Textuelles (JADT'02)*, rédacteurs A. Morin et P. Sébillot, volume 1, pages 233–244, Saint-Malo, 2002. INRIA.

[Cherfi et Toussaint, 2002b] H. Cherfi et Y. Toussaint. How Far Association Rules and Statistical Indices Help Structure Terminology ? Dans *Proc. of the workshop on Natural Language Processing and Machine Learning for Ontology Engineering (OLT'02)*, pages 5–9, Lyon, 2002. In conjunction with the 15^{th} Eur. Conf. on Artificial Intelligence (ECAI'02).

[Cheung *et al.*, 1996] D.W. Cheung, J. Han, V. Ng et C.Y. Wong. Maintenance of discovered association rules in large databases : An incremental updating technique. Dans *Proc. of the 12^{th} IEEE Int'l Conf. on Data Engineering (ICDE'96)*, pages 106–114, New Orleans, USA, 1996.

[Chiaramella *et al.*, 1986] Y. Chiaramella, B. Defude, M.-F. Bruandet et D. Kerkouba. IOTA : A Full Text Information Retrieval System. Dans *Proc. of the 9^{th} ACM Annual Int'l Conf. on Research and Development in Information Retrieval (SIGIR'86)*, pages 207–213, Pisa, Italy, 1986. ACM Press.

[Chomsky, 1957] N. Chomsky. *Syntactic Structures*. Mouton, La Haye, The Netherlands, 1957.

[Church et Hanks, 1989] K. Church et P. Hanks. Word association norms, mutual information, and Lexicography. Dans *Proc. of the 27^{th} Annual Meeting of the Assoc. for Computational Linguistics (ACL'89)*, pages 76–83, Vancouver, 1989. ACL Press.

[Church, 1988] K. Church. A stochastic parts program and noun phrase parser for unrestricted text. Dans *Proc. of the 2^{nd} Conf. on Applied Natural Language Processing (ANLP'88)*, pages 136–143, Austin, USA, 1988.

[Claveau *et al.*, 2003] V. Claveau, P. Sébillot, C. Fabre et P. Bouillon. Learning Semantic Lexicons from a Part-of-Speech and Semantically Tagged Corpus using Inductive Logic Programming. *Journal of Machine Learning Research (JMLR), special issue on ILP*, 4(1) :493–525, 2003.

[Clifton et Cooley, 1999] C. Clifton et R. Cooley. TOPCAT : Data Mining for Topic Identification in a Text Corpus. Dans *Proc. of the 3^{rd} Eur. Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD'99)*, pages 174–183, Prague, 1999.

[Collins et Loftus, 1975] A. Collins et E. Loftus. A Spreading-Activation Theory of Semantic Processing. *Psychological Review*, 82(6) :407–428, 1975.

[Condamines, 2002] A. Condamines. Corpus analysis and conceptual relation patterns. *Terminology : Int'l Journal of theoretical and applied issues in specialized communication*, 8(1) :141–162, 2002. John Benjamins Publishing.

[Daille, 2002] B. Daille. Découvertes linguistiques en corpus. Mémoire d'Habilitation à Diriger des Recherches (HDR), 2002. 55 pages. IRIN – Université de Nantes.

[Davey et Priestley, 1994] B. A. Davey et H. A. Priestley. *Introduction to Lattices and Order*. Cambridge University Press, 4^{th} édition, 1994.

[Delgado *et al.*, 2002] M. Delgado, M. J. Martin-Bautista, D. Sanchez et M.A. Vila. Mining Text Data : Special Features and Patterns. Dans *Pattern Detection and Discovery : Proc. of ESF Exploratory Workshop*, rédacteurs D.J. Hand, N.M. Adams et R.J. Bolton, volume 2447 de *Lecture Notes in Artificial Intelligence – LNAI*, pages 140–153, London, 2002. Springer-Verlag.

[Denoue, 2000] L. Denoue. Cours de classification supervisée de documents, 2000. Université de Savoie.

[Diatta, 2003] J. Diatta. Génération de la base de Guigues-Duquenne-Luxenburger pour les règles d'association par une approche utilisant des mesures de similiarité multivoies. Dans *Actes de CAp'03 : Conférence d'Apprentissage*, rédacteur R. Gilleron, pages 281–298, Laval, 2003. Dans le cadre de la plate-forme AFIA, Presses Universitaires de Grenoble.

[Dikovsky, 2001] A. Dikovsky. Grammars for local and long dependencies. Dans *Proc. of the 39^{th} Int'l Conf. of the Association for Computational Linguistics (ACL'01)*, pages 156–163, Toulouse, 2001. ACL & Morgan Kaufman.

[Ducloy, 1999] J. Ducloy. DILIB, une plate-forme XML pour la génération de serveurs WWW et la veille scientifique et technique. Dans *Micro-Bulletin*. CNRS, 1999.

[Dumais *et al.*, 1998] S. Dumais, J. Platt, D. Heckermann et M. Sahami. Inductive Learning Algorithms and Representations for Text Categorization. Dans *Proc. of the 7^{th} Int'l Conf. on Information and Knowledge Management (CIKM'98)*, pages 148–155, Washington, USA, 1998. ACM Press.

[Duquenne, 1999] V. Duquenne. Latticial structures in data analysis. *Theoretical Computer Science*, 217 :407–436, 1999.

[Faure *et al.*, 1998] D. Faure, C. Nédellec et C. Rouveirol. Acquisition of Semantic Knowledge using Machine learning methods : The System ASIUM. Rapport Technique ICS-TR-88-16, LRI Université Paris-Sud, 1998.

[Fayyad *et al.*, 1996a] U. Fayyad, G. Piatetsky-Shapiro et P. Smyth. From Data Mining to Knowledge Discovery. *AI Magazine*, 17(3) :37–54, 1996.

[Fayyad *et al.*, 1996b] U. Fayyad, G. Piatetsky-Shapiro et P. Smyth. Knowledge Discovery and Data Mining : Towards a Unifying Framework. Dans *Proc. of the 2^{nd} Int'l Conf. on Knowledge Discovery from Databases (KDD'96)*, rédacteurs E. Simoudis, J. Han et U. Fayyad, pages 82–88, Portland, USA, 1996.

[Feldman *et al.*, 1998] R. Feldman, M. Fresko, Y. Kinar, Y. Lindell, O. Liphstat, M. Rajman, Y. Schler et O. Zamir. Text Mining at the Term Level. Dans *Proc. of the 2^{nd} Eur. Symp. on Principles of Data Mining and Knowledge Discovery (PKDD'98)*, rédacteurs J. M. Zytkow et M. Quafafou, volume 1510 de *Lecture Notes in Artificial Intelligence – LNAI*, pages 65–73, Nantes, 1998.

[Feldman et Dagan, 1995] R. Feldman et I. Dagan. Knowledge Discovery in Textual Databases (KDT). Dans *Proc. of the 1^{st} Int'l Conf. on Data Mining and Knowledge Discovery*, rédacteurs U. M. Fayyad et R. Uthurusamy, pages 112–117, Montréal, Canada, 1995. AAAI Press.

[Feldman et Hirsh, 1997] R. Feldman et H. Hirsh. Exploiting Background Information in Knowledge Discovery from Text. *Journal of Intelligent Information Systems*, 9(1) :83–97, 1997.

[François *et al.*, 2001] C. François, J. Royauté et D. Besagni. Apport d'une méthodologie de recherche de termes en corpus dans un processus de KDD : application de veille en biologie moléculaire. Dans *Actes de VSST'01 : Veille Stratégique, Scientifique et technologique*, pages 49–62, Barcelone, 2001. FPC/UPC – SFBA – IRIT.

[Freitas, 1998] A. A. Freitas. On Objective Measures of Rule Surprisingness. Dans *Proc. of the 2^{nd} Eur. Symp. on Principles of Data Mining and Knowledge Discovery (PKDD'98)*, volume 1510 de *Lecture Notes in Artificial Intelligence – LNAI*, pages 1–9, Nantes, France, 1998. Springer-Verlag.

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

[Ganter, 1999] B. Ganter. Attribute exploration with background knowledge. *Theoretical Computer Science*, 217(2) :215–233, 1999.

[Gayral *et al.*, 1994] F. Gayral, P. Grandemange, D. Kayser et F. Lévy. Interprétation des constats d'accidents : représenter le réel et le potentiel. *Revue Traitement Automatique des Langues (TAL) : Approches sémantiques*, 35(1) :65–81, 1994. Rédacteur F. Lévy.

[Godin et Missaoui, 1994] R. Godin et R. Missaoui. An incremental concept formation approach for learning from databases. *Theoretical Computer Science*, 133(2) :387–419, 1994.

[Godin, 1989] R. Godin. Complexité de structures de treillis. *Annales des Sciences Mathématiques du Québec*, 13(1) :19–38, 1989.

[Gras *et al.*, 2001] R. Gras, P. Kuntz, R. Couturier et F. Guillet. Une version entropique de l'intensité d'implication pour les corpus volumineux. Dans *Actes EGC'01 : Journées Extraction et Gestion des Connaissances*, rédacteurs H. Briand et F. Guillet, volume 1 de *1-2*, pages 69–80, Nantes, France, 2001. Éditions Hermès.

[Grishman, 1997] R. Grishman. Information Extraction : Techniques and Challenges. Dans *Proc. of the Int'l Summer School on Information Extraction (SCIE'97)*, rédacteur M. T. Pazienza, volume 1299 de *Lecture Notes in Computer Science – LNCS*, Frascati, Italy, 1997. Springer-Verlag.

[Grobelnik *et al.*, 2000] M. Grobelnik, D. Mladenic et N. Milic-Frayling. Introduction. Dans *Proc. of the workshop on Text Mining*, Boston, 2000. In conjunction with the 6^{th} Int'l Conf. on Knowledge Discovery and Data Mining (KDD'00).

[Guigues et Duquenne, 1986] J.L. Guigues et V. Duquenne. Familles minimales d'implication informatives résultant d'un tableau de données binaires. *Mathématiques, Informatique et Sciences Humaines*, 95(24) :5–18, 1986.

[Guillaume, 2000] S. Guillaume. *Traitement des données volumineuses : Mesures et algorithmes d'extraction de règles d'association et règles ordinales*. Thèse de doctorat, Université de Nantes, 2000.

[Guillet, 2004] F. Guillet. Mesures de qualité des connaissances en ECD. Lecture at Conf. Extraction et Gestion des Connaissances (EGC'04), 2004. Clermont-Ferrand, France.

[Habert *et al.*, 2000] B. Habert, G. Illouz, P. Lafon, S. Fleury, H. Folch, S. Heiden et S. Prévost. Profilage de textes : cadre de travail et expérience. Dans *Journées d'Analyse des Données Textuelles (JADT'00)*, rédacteur M. Rajman, Lausanne, 2000. 8 pages.

[Habert, 2000] B. Habert. Création de dictionnaires sémantiques et typologie des textes. Dans *Actes des journées scientifiques L'Imparfait – Philologie électronique et assistance à l'interprétation des textes*, rédacteur J.-E. Tyvaert, volume 15, pages 171–188, Reims, 2000. Presses Universitaires de Reims.

[Hahn et Reimer, 1998] U. Hahn et U. Reimer. Text Summarization Based on Terminological Logics. Dans *Proc. of the 13^{th} Europ. Conf. on Artificial Intelligence (ECAI'98)*, rédacteur H. Prade, pages 165–169. John Wiley & Sons Ltd, 1998.

[Han, 1995] J. Han. Mining Knowledge at Multiple Concept Levels. Dans *Proc. of the 4^{th} Int'l Conf. on Information and Knowledge Management (CIKM'95)*, pages 19–24, Baltimore, USA, 1995. ACM Press. Invited talk.

[Harris, 1968] Z. Harris. *Mathematical Structures of Languages*. Wiley-Interscience, New-York, 1968.

[Hearst, 1999] M. Hearst. Untangling Text Data Mining. Dans *Proc.of the 37^{th} Annual Meeting of the Association for Computational Linguistics (ACL'99)*. University of Maryland, 1999.

[Hipp *et al.*, 2002] J. Hipp, U. Güntzer et G. Nakhaeizadeh. Data mining of association rules and the process of knowledge discovery in databases. Dans *Data Mining in E-Commerce, Medicine, and Knowledge Management*, pages 15–36. Springer, Heidelberg, 2002.

[Hussain *et al.*, 2000] F. Hussain, H. Liu et H. Lu. Relative Measure for Mining Interesting Rules. Dans *Proc. of Knowledge Discovery and Data Mining, Current Issues and New Applications, the 4^{th} Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD'00)*, rédacteurs T. Terano, H. Liu et A. L. P. Chen, volume 1805 de *Lecture Notes in Computer Science – LNCS*, pages 86–97, Kyoto, Japan, 2000. Springer.

[IBM, 1998] IBM. *Intelligent Miner for Text*. International Business Machine (IBM), 1998.

[Ide, 1994] N. Ide. Encoding standards for large text resources : The Text Encoding Initiative. Dans *Proc. of the 15^{th} Int'l Conf. on Computational Linguistics (COLING'94)*, volume 1, pages 574–578, Kyoto, Japan, 1994.

[Jacquemin, 1994] C. Jacquemin. FASTR : A Unification-Based Front-End to Automatic Indexing. Dans *Proc. of Information Multimedia Information Retrieval Systems and Management*, pages 34–47, New-York, 1994. Rockfeller University.

[Jacquemin, 1997] C. Jacquemin. Variation terminologique : Reconnaissance et acquisition automatique de termes et de leurs variantes. Mémoire d'Habilitation à Diriger des Recherches (HDR), 1997. IRIN – Université de Nantes.

[Janetzko *et al.*, 2004] D. Janetzko, H. Cherfi, R. Kennke, A. Napoli et Y. Toussaint. Knowledge-based selection of association rules for text mining. Dans *Proc. of the 16^{th} European Conference on Artificial Intelligence (ECAI'04)*, rédacteurs R. López de Mántaras et L. Saitta, pages 485–489, Valencia, Spain, 2004. IOS Press.

[Jensen, 1996] F. V. Jensen. *An introduction to Bayesian networks*. UCL Press, London, 1996.

[Joachims, 1998] T. Joachims. Text Categorization with Support Vector Machines : Learning with Many Relevant Features. Dans *Machine Learning : Proc. of the 10^{th} Europ. Conf. on Machine Learning (ECML'98)*, rédacteurs C. Nédellec et C. Rouveirol, volume 1398 de *Lecture Notes in Computer Science – LNCS*, pages 137–142, Chemnitz, Germany, 1998. Springer.

[Johnson *et al.*, 2002] D. E. Johnson, F. J. Oles, T. Zhang et T. Goetz. A decision-tree-based symbolic rule induction system for text categorization. *IBM Systems Journal : Artificial Intelligence*, 41(3) :428–436, 2002.

[Jones *et al.*, 2003] R. Jones, R. Ghani, T. Mitchell et E. Riloff. Active Learning for Information Extraction with Multiple View Feature Sets. ECML-03 Workshop on Adaptive Text Extraction and Mining, 2003. Cavtat-Dubrovnik, Croatie.

[Kay et Fillmore, 1999] P. Kay et C.J. Fillmore. Grammatical constructions and linguistic generalizations : the What's X doing Y ? construction. *Language*, 75 :1–33, 1999.

[Kayser, 1988] D. Kayser. What Kind of Thing is a Concept ? *Computational Intelligence*, 4 :158–165, 1988.

[Kessler *et al.*, 1997] B. Kessler, G. Nunberg et H. Schütze. Automatic Detection of Text Genre. Dans *Proc. of the 35^{th} Annual Meeting of the Association for Computational Linguistic and 8^{th} Conf. of the Eur. Chapter of the Association for Computational Linguistics (ACL'97)*, rédacteurs P. R. Cohen et W. Wahlster, pages 32–38, Somerset (NJ), USA, 1997. ACL Press.

[Klemettinen *et al.*, 1994] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen et A. I. Verkamo. Finding Interesting Rules from Large Sets of Discovered Association Rules. Dans *Proc. of the 3^{rd} Int'l Conf. on Information and Knowledge Management (CIKM'94)*, rédacteurs B. K. Bhargava N. R. Adam et Y. Yesha, pages 401–407, Gaithersburg, USA, 1994. ACM Press.

[Kodratoff, 1999] Y. Kodratoff. Knowledge Discovery in Texts : A Definition, and Applications. Dans *Foundations of Intelligent Systems, Proc. of the 11^{th} Int'l Symposium, ISMS'99*, rédacteurs Z. W. Ras et A. Skowron, volume 1609 de *Lecture Notes in Artificial Intelligence – LNAI*, pages 16–29, Warsaw, Pol., 1999. Springer.

[Kodratoff, 2000a] Y. Kodratoff. Datamining and Textmining. Dans *Proc. of ECD'00*, pages 6–9, Tunis, 2000.

[Kodratoff, 2000b] Y. Kodratoff. Quelques contraintes symboliques sur le numérique en ECD et en ECT, 2000.

[Kuntz *et al.*, 2000] P. Kuntz, F. Guillet, R. Lehn et H. Briand. A User-Driven Process for Mining Association Rules. Dans *Proc. of the 4^{th} Eur. Conf. on Principles of Data Mining and Knowledge Discovery (PKDD'00)*, rédacteurs D.A. Zighed, H.J Komorowski et J.M. Zytkow, volume 1910 de *Lecture Notes in Artificial Intelligence – LNAI*, pages 483–489, Lyon, 2000. Springer-Verlag.

[Latiri-Chérif *et al.*, 2002] C. Latiri-Chérif, S. Elloumi, S. Ben Yahia et A. Jaoua. Textmining : Extension de la connexion de Galois floue. Dans *Actes de EGC'02 : Extraction et Gestion des Connaissances*, rédacteurs D. Hérin et D. A. Zighed, volume 1 de *1-4*, pages 375–386, Montpellier, 2002. Hermès Sciences.

[Lavrač *et al.*, 1999] N. Lavrač, P. Flach et B. Zupan. Rule Evaluation Measures : A Unifying View. Dans *Proc. of the 9^{th} Int'l Workshop on Inductive Logic Programming (ILP'99)*, volume 1634 de *Lecture Notes in Artificial Intelligence – LNAI*, pages 174–185, Bled, Slovenia, 1999. Springer-Verlag, Heidelberg. Co-located with ICML'99.

[Lehn *et al.*, 2004] R. Lehn, F. Guillet et H. Briand. Qualité d'un ensemble de règles : élimination des règles redondantes. Dans *Mesures de Qualité pour la Fouille de Données*, rédacteurs H. Briand, M. Sebag et F. Guillet, Revue des Nouvelles Technologies de l'Information (RNTI), pages 141–167. Cépaduès Éditions, Toulouse, 2004.

[Lenat et Guha, 1990] D. B. Lenat et R. Guha. *Building Large Knowledge Bases*. Addison Wesley, 1990.

[Lenca *et al.*, 2003] P. Lenca, P. Meyer, P. Picouet, B. Vaillant et S. Lallich. Critères d'évaluation des mesures de qualité en ECD. *Revue des Nouvelles Technologies de l'Information*, 1 :123–134, 2003.

[Li *et al.*, 1999] J. Li, X. Zhang, G. Dong, K. Ramamohanarao et Q. Sun. Efficient Mining of High Confidence Association Rules without Support Thresholds. Dans *Proc. of the 3^{rd} Eur. Conf. on Principles of Knowledge Discovery in Databases (PKDD'99)*, rédacteurs J. M. Zytkow et J. Rauch, volume 1704 de *Lecture Notes in Computer Science – LNCS*, pages 406–411, Prague, 1999. Springer.

[Ligozat, 1996] G. Ligozat. Representations of Space and Time. Dans *Survey of the State of the Art in Human Language Technology*, rédacteurs R.A. Cole, J. Mariani, H. Uszkoreit, G.B. Varile, A. Zaenen, A. Zampolli et V. Zue, pages 343–347. Kluwer Academic Publishers, 1996.

[Liu *et al.*, 1997] B. Liu, W. Hsu et S. Chen. Using General Impressions to Analyze Discovered Association Rules. Dans *Proc. of the 3rd Int'l Conf. on Knowledge Discovery and Data Mining (KDD'97)*, pages 31–36, Newport Beach, USA, 1997. AAAI Press.

[Liu *et al.*, 1999a] B. Liu, W. Hsu et Y. Ma. Pruning and Summarizing the Discovered Associations. Dans *Proc. of the 5th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD'99)*, rédacteurs S. Chaudhuri et D. Madigan, pages 125–134, San Diego, USA, 1999. ACM Press.

[Liu *et al.*, 1999b] B. Liu, W. Hsu, K. Wang et S. Chen. Visually Aided Exploration of Interesting Association Rules. Dans *Proc. of Research and Development in Knowledge Discovery and Data Mining : the 3rd Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD'99)*, rédacteurs N. Zhong et L. Zhou, volume 1574 de *Lecture Notes in Computer Science – LNCS*, pages 380–389. Springer, 1999.

[Luxenburger, 1991] M. Luxenburger. Implications partielles dans un contexte. *Mathématiques, Informatique et Sciences Humaines*, 113(29) :35–55, 1991.

[Lévy, 1994] F. Lévy. Introduction aux approche sémantiques. *Revue Traitement Automatique des Langues (TAL) : Approches sémantiques*, 35(1) :3–18, 1994.

[Maedche et Staab, 2000] A. Maedche et S. Staab. Mining Ontologies from Text. Dans *Proc. of the 12th Int'l Conf. on Knowledge Engineering and Knowledge Management (EKAW'00*, rédacteurs R. Dieng et O. Corby, volume 1937 de *Lecture Notes in Artificial Intelligence – LNAI*, pages 189–202, Juan-les-Pins, 2000. Springer-Verlag.

[Maedche et Staab, 2001] A. Maedche et S. Staab. Comparing Ontologies - Similarity Measures and a Comparison Study. Internal Report 408, Institute AIFB, University of Karlsruhe, 2001.

[Maedche et Staab, 2003] A. Maedche et S. Staab. *Handbook on Ontologies in Information Systems Learning*, chapitre Ontology Learning. Springer, S. Staab et R. Studer édition, 2003.

[Marcus *et al.*, 1994] M.P. Marcus, B. Santorini et M.A. Marcinkiewicz. Building a large annotated corpus of English : the Penn Treebank. Dans *Using Large Corpora*, rédacteur S. Armstrong, pages 1–22. MIT Press, Boston, USA, 1994. Republication of Journal of Computational Linguistics, 19(1), 1993.

[McCallum et Nigam, 1998] A. McCallum et K. Nigam. A Comparison of Event Models for Naive Bayes Text Classification. Dans *Proc. of AAAI Workshop on Learning for Text Categorization*, pages 41–48, Madison, USA, 1998. AAAI Press. In conjunction with the 15th Int'l Conf. on Machine Learning (ICML'98).

[Mechkour *et al.*, 1995] M. Mechkour, C. Berrut et Y. Chiaramella. Using Conceptual Graph Frame work for Image Retrieval. Dans *Proc. of the 2nd Int'l Conf. on MultiMedia Modelling (MMM'95)*, pages 127–142, Singapore, 1995. IEEE Press.

[MedLine, 2003] MedLine. Base de données de résumés d'articles scientifiques en médecine du portail Web PubMed, 2003. National Library of Medicine (NLM) accessible par : www.ncbi.nlm.nih.gov/entrez/. Site visité en juin 2003.

[Mitchell, 1997] T. Mitchell. *Machine Learning*, chapitre Decision Tree Learning, pages 52–78. McGraw-Hill, 1997.

[Mladenić, 1999] D. Mladenić. Text-Learning and Related Intelligent Agents : A Survey. *IEEE Intelligent Systems*, 14(4) :44–54, 1999.

[Montague, 1974] R. Montague. *Formal Philosophy*. Yale University Press, 1974.

[Montes_y_Gómez *et al.*, 2001] M. Montes_y_Gómez, A. Gelbukh, A. López-López et R. Baeza-Yates. Text Mining With Conceptual Graphs. Dans *Proc. of Symp. on Natural Language Processing and Knowledge Engineering (NLPKE'01)*, Tucson, USA, 2001. In conjunction with IEEE Conf. on Systems, Man, And Cybernetics (SMC'01). 6 pages.

[Montes_y_Gómez *et al.*, 2002] M. Montes_y_Gómez, A. Gelbukh et A. López-López. Text mining at Detail Level using Conceptual Graphs. Dans *Lecture Notes in Artificial Intelligence – LNAI*, volume 2393, pages 122–136. Springer, 2002.

[Muller *et al.*, 1997] C. Muller, X. Polanco, J. Royauté et Y. Toussaint. Acquisition et structuration des connaissances en corpus : éléments méthodologiques. Rapport Technique RR-3198, INRIA Lorraine, Nancy, 1997. 45 pages.

[Nahm et Mooney, 2001] U. Y. Nahm et R. J. Mooney. Mining Soft-Matching Rules from Textual Data. Dans *Proc. of the 17^{th} Int'l Joint Conf. on Artificial Intelligence (IJCAI'01)*, pages 979–984, Seattle, USA, 2001.

[Napoli, 1997] A. Napoli. Une introduction aux logiques de descriptions. Rapport Technique 3314, INRIA, 1997.

[Nauer, 2002] E. Nauer. Complémentarité entre fouille de données et recherche d'information dans le cadre d'analyses bibliométriques. Dans *Actes du $13^{ème}$ Congrès Francophone de Reconnaissance des Formes et Intelligence Artificielle : RFIA'02*, volume 3, pages 965–974, Angers, 2002.

[Nédellec *et al.*, 2001] C. Nédellec, M. Ould Abdel Vetah et P. Bessière. Sentence Filtering for Information Extraction in Genomics, a Classification Problem. Dans *Proc. of the 5^{th} Eur. Conf. on Principles of Data Mining and Knowledge Discovery (PKDD'01)*, rédacteurs L. De Raedt et A. Siebes, volume 2168 de *Lecture Notes in Computer Science – LNCS*, pages 326–338, Freiburg, Germany, 2001. Springer Verlag.

[Nebel, 1990] B. Nebel. Reasoning and Revision in Hybrid Representation Systems. Dans *Lecture Notes in Artificial Intelligence – LNAI*, 422. Springer-Verlag, Berlin, 1990.

[Norris, 1978] E.M. Norris. An algorithm for computing the maximal rectangles in a binary relation. *Revue Roumaine de Mathématiques Pures et Appliquées*, 23(2) :243–250, 1978.

[Padmanabhan et Tuzhilin, 2000] B. Padmanabhan et A. Tuzhilin. Small is Beautiful : Discovering the Minimal Set of Unexpected Patterns. Dans *Proc. of the 6^{th} ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD'00)*, rédacteurs R. Ramakrishnan, S. Stolfo, R. Bayardo et I. Parsa, pages 54–63, Boston, 2000. ACM Press.

[Pasquier *et al.*, 1999a] N. Pasquier, Y. Bastide, R. Taouil et L. Lakhal. Closed sets based discovery of small covers for association rules. Dans *Proc. of the 15^{th} French Conf. on Advanced Databases (BDA'99)*, pages 361–381, Bordeaux, 1999.

[Pasquier *et al.*, 1999b] N. Pasquier, Y. Bastide, R. Taouil et L. Lakhal. Efficient mining of association rules using closed itemset lattices. *Information Systems*, 24(1) :25–46, 1999.

[Pearl, 1988] J. Pearl. *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. Morgan Kauffman, San Mateo, USA, 1988.

[Pearson, 1998] J. Pearson. *Terms In Context*. John Benjamins Publisher, Amsterdam, 1998.

[Pei *et al.*, 2000] J. Pei, J. Han et R. Mao. CLOSET : An Efficient Algorithm for Mining Frequent Closed Itemsets. Dans *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 21–30. ACM Press, 2000.

[Piatetsky-Shapiro et Matheus, 1994] G. Piatetsky-Shapiro et C. Matheus. The interestingness of deviations. Dans *Proc. of the AAAI Workshop on Knowledge Discovery in Databases (KDD'94)*, pages 25–36, Seattle, USA, 1994. AAAI Press.

[Piatetsky-Shapiro, 1991] G. Piatetsky-Shapiro. *Knowledge Discovery in Databases*, chapitre Discovery, Analysis, and Presentation of Strong Rules (chapter 13), pages 229–248. AAAI/MIT Press, Menlo Park, G. Piatetsky-Shapiro and W.J. Frawley édition, 1991.

[Piwowarski *et al.*, 2002] B. Piwowarski, L. Denoyer et P. Gallinari. Un modèle pour la recherche d'information sur des documents structurés. Dans *Actes des $6^{èmes}$ Journées internationales d'Analyse statistique de Données Textuelles (JADT'02)*, rédacteurs A. Morin et P. Sébillot, volume 2, pages 605–616, Saint-Malo, 2002. INRIA.

[Quinlan, 1986] J. R. Quinlan. Induction of Decision Trees. *Machine Learning*, 1 :81–106, 1986.

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

[Rajman et Besançon, 1997] M. Rajman et R. Besançon. Text Mining : Natural Language Techniques and Text Mining Applications. Dans *Proc. of the 7th IFIP 2.6 Working Conference on Database Semantics (DS-7)*, rédacteur H. Prade, Leysin (Switzerland), 1997. Chapman & Hall. 15 pages.

[Rastier, 1995] F. Rastier. *Le terme : entre ontologie et linguistique*, volume 7 de *La banque des mots*, pages 35–65. CILF, Paris, 1995.

[Resnik, 1999] P. Resnik. Semantic Similarity in a Taxonomy : An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Artificial Intelligence Research*, 11 :95–130, 1999. Morgan Kaufmann Publishers.

[Riloff et Jones, 1999] E. Riloff et R. Jones. Learning Dictionaries for Information Extraction Using Multi-level Boot-strapping. Dans *Proc. of the 16th National Conf. on Artificial Intelligence AAAI-99*, pages 474–479, Orlando, USA, 1999. American Association for Artificial Intelligence, AAAI Press.

[Roche *et al.*, 2003] M. Roche, J. Azé, O. Matte-Tailliez et Y. Kodratoff. Mining texts by association rules discovery in a technical corpus. Dans *Proc. of the Int'l IIS Conf. on Intelligent Information Processing and Web Mining (IIPWM´03)*, rédacteurs M.A. Klopotek, S.T. Wierzchon et K. Trojanowski, Advances in Soft Computing, pages 89–98, Zakopane, Poland, 2003.

[Sabah, 2000] G. Sabah. *Sens et traitement automatique des langues*, chapitre 3, pages 77–108. Informatique et systèmes d'information. Traité IC2 : Information–Commande–Communication. Hermès Science Publications, J. M. Pierrel édition, 2000.

[Sahar, 1999] S. Sahar. Interestingness via What is Not Interesting. Dans *Proc. of the 5th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD'99)*, rédacteurs S. Chaudhuri et D. Madigan, pages 332–336, San Diego, USA, 1999. ACM Press.

[Salton, 1989] G. Salton. *Automatic Text Processing : The Transformation, Analysis ans Retrieval of Information by Computer*. Addison Wesley, MA, 1989.

[Sebastiani, 2003] F. Sebastiani. Text Categorization. Dans *Text Mining and its Applications*, rédacteur A. Zanasi, page 23 pages. WIT Press, Southampton, UK, 2003. Chapitre invité : À paraître.

[Séguéla et Aussenac-Gilles, 1999] P. Séguéla et N. Aussenac-Gilles. Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine. Dans *Actes de la Conférence Ingénierie des Connaissances (IC'99)*, pages 79–88, École Polytechnique, Paris, 1999.

[Shah *et al.*, 1999] D. Shah, L. V. S. Lakshmanan, K. Ramamritham et S. Sudarshan. Interestingness and Pruning of Mined Patterns. Dans *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD'99)*, rédacteurs K. Shim et R. Srikant, page 5 pages, Philadelphia, USA, 1999.

[Shieber, 1986] S. M. Shieber. *An Introduction to Unification-Based Approaches to Grammar*. Center for the Study of Language and Information, Stanford University, Stanford, CA, 1986.

[Silberschatz et Tuzhilin, 1996] A. Silberschatz et A. Tuzhilin. What Makes Patterns Interesting in Knowledge Discovery Systems. *IEEE Trans. on Knowledge and Data Engineering*, 8 :970–983, 1996.

[Simon et Napoli, 1999] A. Simon et A. Napoli. Building Viewpoints in an Object-Based Representation System for Knowledge Discovery in Databases. Dans *Proc. of the 1st Int'l Conf. on Information Reuse and Integration (IRI'99)*, rédacteur S. Rubin, pages 104–108. Int'l Society for Computers and their Applications (ISCA), 1999.

[Simon, 2000] A. Simon. *Outils classificatoires par objets pour l'extraction de connaissances dans les bases de données*. Thèse de doctorat, Université Henri Poincaré - Nancy 1, Nancy, 2000.

[Simpson, 1951] E.H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society*, 13(B) :238–241, 1951.

[Soderland *et al.*, 1995] S. Soderland, D. Fisher, J. Aseltine et W. Lehnert. CRYSTAL : Inducing a Conceptual Dictionary. Dans *Proc. of the 14th Int'l Joint Conf. on Artificial Intelligence (IJCAI'95)*, rédacteur C. Mellish, pages 1314–1319, San Francisco, 1995. Morgan Kaufmann.

[Soderland, 1999] S. Soderland. Learning Information Extraction Rules for Semi-Structured and Free Text. *Machine Learning*, 34(1-3) :233–272, 1999. Kluwer Academic Publishers.

[Sowa, 1992] J.F. Sowa. *Semantic networks*. Encyclopedia of Artificial Intelligence. John Wiley & Sons, New York, 1992. 2^{nd} edition by S. C. Shapiro.

[Sowa, 2002] J. F. Sowa. Architectures for intelligent systems. *IBM Systems Journal : Artificial Intelligence*, 41(3) :331–349, 2002.

[Srikant et Agrawal, 1995] R. Srikant et R. Agrawal. Mining Generalized Association Rules. Dans *Proc. of the 21^{st} Int'l Conf. on Very Large Databases (VLDB'95)*, pages 407–419, Zurich, 1995. Morgan Kaufmann Press. Expanded version available as IBM Research Report RJ 9963.

[Stumme *et al.*, 2001] G. Stumme, R. Taouil, Y. Bastide, N. Pasquier et L. Lakhal. Intelligent structuring and reducing of association rules with formal concept analysis. Dans *Proc. of Advances in Artificial Intelligence (KI'01) : Joint German/Austrian Conference on AI*, rédacteurs F. Baader, G.Brewker et T. Eiter, volume 2174 de *Lecture Notes in Artificial Intelligence – LNAI*, pages 335–350, Vienna, 2001. Springer-Verlag.

[Stumme *et al.*, 2002] G. Stumme, R. Taouil, Y. Bastide, N. Pasquier et L. Lakhal. Computing Iceberg Concept Lattices with Titanic. *Journal of Data and Knowledge Engineering*, 42(2) :189–222, 2002.

[Stumme et Maedche, 2001] G. Stumme et A. Maedche. FCA-MERGE : Bottom-Up Merging of Ontologies. Dans *Proc. of the 17^{th} Int'l Joint Conf. on Artificial Intelligence (IJCAI'01)*, rédacteur B. Nebel, pages 225–234, Seattle, USA, 2001. Morgan Kaufmann.

[Subramonian, 1998] R. Subramonian. Defining diff as a Data Mining Primitive. Dans *Proc. of the 4^{th} Int'l Conf. on Knowledge Discovery and Data Mining (KDD'98)*, rédacteurs R. Agrawal, P. E. Stolorz et G. Piatetsky-Shapiro, pages 334–338, New York, USA, 1998. AAAI Press.

[Suzuki et Kodratoff, 1998] E. Suzuki et Y. Kodratoff. Discovery of Surprising Exception Rules based on Intensity of Implication. Dans *Proc. of the 2^{nd} Eur. Symp. on Principles of Data Mining and Knowledge Discovery (PKDD'98)*, rédacteurs J. M. Zytkow et M. Quafafou, volume 1510 de *Lecture Notes in Artificial Intelligence – LNAI*, pages 10–18, Nantes, 1998. Springer-Verlag.

[Tan *et al.*, 2002] P.N. Tan, V. Kumar et J. Srivastava. Selecting the right interestingness measure for associaton patterns. Dans *Proc. of the 8^{th} ACM Int'l Conf. on Knowledge Discovery and Data Mining (KDD'02)*, pages 183–193, Edmonton, Canada, 2002. ACM Press.

[Tan, 1999] A.-H. Tan. Text mining : The state of the art and the challenges. Dans *Proc. of the Workshop on Knowledge Discovery from Advanced Databases*, pages 65–70, Beijing, China, 1999. In conjunction the 3^{rd} Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD'99).

[Taouil, 2000] R. Taouil. *Algorithmique du treillis des fermés : application à l'analyse formelle de concepts et aux bases de données*. Thèse de doctorat, Université Blaise Pascal - Clermont-Ferrand II, 2000.

[Tazi et Virbel, 1985] S. Tazi et J. Virbel. Formal Representation of Textual Structures for an Intelligent Text-Editing System. Dans *Proc. of Natural Language Understanding and Logic Programming Workshop*, rédacteurs V. Dahl et P. Saint-Dizier, pages 191–205, Rennes, 1985. Elsevier Science (North-Holland).

[Toussaint *et al.*, 1998] Y. Toussaint, F. Namer, B. Daille, C. Jacquemin, J. Royauté et N. Hathout. Une approche linguistique et statistique pour l'analyse de l'information en corpus. Dans *Actes de la $5^{ème}$ Conf. nationale sur le Traitement Automatique des Langues Naturelles (TALN'98)*, pages 182–191, Paris, 1998.

[Toussaint *et al.*, 2000] Y. Toussaint, A. Simon et H. Cherfi. Apport de la fouille de données textuelles pour l'analyse de l'information. Dans *Actes de la conférence IC'2000, Ingénierie des Connaissances*, pages 335–344, Toulouse, France, 2000.

[UMLS, 2000] UMLS. The Unified Medical Language System. National Library of Medicine, 11^{th} edition, 2000.

[Van-Rijsbergen, 1979] C. J. Van-Rijsbergen. *Information Retrieval*. Butterworths and Co, 1979. 2^{nd} edition.

[Vapnik, 1995] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

[Vijay-Shankar, 1992] K. Vijay-Shankar. Using descriptions of trees in a tree-adjoining grammar. *Computational Linguistics*, 18 :481–518, 1992.

[Wilkinson, 1994] R. Wilkinson. Effective retrieval of structured documents. Dans *Proc. of the 17^{th} Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval*, rédacteurs W. Croft et C. van Rijsbergen, volume 9, pages 311–317, Dublin, 1994. ACM Press.

[Wilks, 1997] Y. Wilks. Senses and Texts. *Computers and the Humanities*, 31(2), 1997.

[Zaccai et Garrec, 1998] J. Zaccai et C. Garrec. *Les macromolécules du vivant - Structure, dynamique et fonctions*. Éditions CNRS, Paris, 1998.

[Zaki et Hsiao, 1999] M. J. Zaki et C.-J. Hsiao. CHARM : An efficient algorithm for association rule mining. Rapport Technique 99-10, Rensselaer Polytechnic Institute – Computer Science Dept., Troy, USA, 1999.

[Zaki, 2000] M. J. Zaki. Generating non-redundant association rules. Dans *Proc. of the 6^{th} Int'l Conf. on Knowledge Discovery and Data Mining (KDD'00)*, pages 34–43, Boston, 2000. ACM Press.

[Zhang et Zhang, 2002] C. Zhang et S. Zhang. Association Rule Mining : Models and Algorithms. Dans *Tutorial*, volume 2307 de *Lecture Notes in Artificial Intelligence – LNAI*. Springer, 2002. 238 pages.

[Zheng *et al.*, 2001] Z. Zheng, R. Kohavi et L. Mason. Real World Performance of Association Rule Algorithms. Dans *Proc. of the 7^{th} ACM Int'l Conf. on Knowledge Discovery and Data Mining (SIGKDD'01)*, rédacteurs F. Provost et R. Srikant, pages 401–406. ACM Press, 2001. Poster version.

Annexe A

Étiquetage, variations terminologiques et codage XML du corpus

Sommaire

Cette annexe apporte des précisions sur la façon de prétraiter les textes de notre corpus de biologie moléculaire donné en entrée de l'outil TAMIS. Le prétraitement se fait à l'aide d'outils de traitement automatique des langues (TAL). Il s'agit de l'étiqueteur de Brill qui associe la catégorie morpho-syntaxique à chaque mot des textes et de l'analyseur syntaxique FASTER qui indexe les textes par des termes-clés du domaine. les étiquettes pour l'anglais et le français sont données en annexe A.1 ; puis quelques exemples de variations terminologiques, repérées par FASTER, entre les termes présents dans les textes et les termes-index sont présentées en annexe A.2 ; enfin, le format des textes en entrée du processus de FdT est donné en annexe A.3.

A.1 Étiquettes de Brill

Nous donnons les étiquettes de l'outil développé par E. Brill [Brill et Pop, 1999] et que nous avons utilisé pour affecter une catégorie morpho-syntaxique (nom, verbe, etc.) à tous les mots des `1 361` résumés de notre corpus de biologie moléculaire[28].

L'étiqueteur de Brill intègre des lexiques issus de différents corpus (*Wall Street Journal*, *The Brown Corpus*, *Switchboard*, etc.), des règles lexicales et contextuelles – appelées patrons lexicaux –, ce qui rend cet étiqueteur adaptable à un nouveau domaine scientifique pour lequel le vocabulaire est plus spécifique car il suffit d'adapter les règles lexicales.

[28] Ce travail a été fait en collaboration avec l'URI : Unité Recherche et Innovation de l'INIST : INstitut de l'Information Scientifique et Technique.

TAB. A.1 – Étiquettes pour corpus spécialisés (à gauche) et pour les textes d'anglais général (à droite)

N°	Étiquette	Signification
1	CC	Coordinating conjunction
2	CD	Cardinal number
3	DT	Determiner
4	EX	Existential "there"
5	FW	Foreign word
6	IN	Preposition or subordinating conj.
7	JJ	Adjective
8	JJR	Adjective, comparative
9	JJS	Adjective, superlative
10	LS	List item marker
11	MD	Modal
12	NN	Noun, singular or mass
13	NNS	Noun, plural
14	NP	Proper noun, singular
15	NPS	Proper noun, plural
16	PDT	Predeterminer
17	POS	Possessive ending
18	PP	Personal pronoun
19	PP$	Possessive pronoun
20	RB	Adverb
21	RBR	Adverb, comparative
22	RBS	Adverb, superlative
23	RP	Particle
24	SYM	Symbol
25	TO	"to"
26	UH	Interjection
27	VB	Verb, base form
28	VBD	Verb, past tense
29	VBG	Verb, gerund or present participle
30	VBN	Verb, past participle
31	VBP	Verb, non-3rd person singular present
32	VBZ	Verb, 3rd person singular present
33	WDT	Wh-determiner
34	WP	Wh-pronoun
35	WP$	Possessive wh-pronoun
36	WRB	Wh-adverb

N°	Étiquette	Signification
1	AN	Auxiliary Verb (untensed)
2	AT	Auxiliary Verb (tensed)
3	CC	Coordinating conjunction
4	CO	Complementizer
5	DT	Determiner
6	JJ	Adjective
7	NE	Negation
8	NN	Noun
9	PDT	Predeterminer
10	PN	Pronoun
11	PR	Preposition and subordinat. conjunc.
12	RB	Adverb
13	RP	Particle
14	TO	To
15	UH	Exclamation
16	VBG	Present Participle
17	VBN	Past Participle
18	VN	Main Verb (untensed)
19	VT	Main Verb (tensed)

Les étiquettes sont empruntées, et donc compatibles, avec celles utilisées par l'étiqueteur plus général du projet *Penn Treebank* [Marcus *et al.*, 1994] et d'autres étiquettes de l'anglais classique *Old English.*

A.1.1 Étiquettes de Brill pour l'anglais

Nous reprenons l'exemple en anglais de (§ L'étiquetage morpho-syntaxique 2.4.1.2). La phrase (1) est étiquetée en (2).

(1) Two resistant strains were isolated after four rounds of selection.

(2) Two/CD resistant/JJ strains/NNS :pl were/VBD isolated/VBN after/IN four/CD rounds/NNS :pl of/IN selection/NN ./.

La liste non exhaustive des étiquettes pour l'anglais apprises à partir de corpus spécialisés et celles issues de l'anglais classique est donnée TAB. A.1.

A.1.2 Étiquettes de Brill spécifiques au français

Nous reprenons l'exemple en français de (§ 2.4.1.2). La phrase (a) est étiquetée en (b).

(a) Les fractions pectiques contiennent des proportions hautement estérifiées

(b) Les/DTN :pl fractions/SBC :pl pectiques/ADJ :pl contiennent/VCJ :pl des/PREP :pl proportions/SBC :pl hautement/ADV estérifiées/ADJ2PAR :pl ./.

La liste non exhaustive des étiquettes du français pour l'outil WINBRILL développé à l'INALF/CNRS est donnée TAB. A.2.

TAB. A.2 – Étiquettes de Brill pour le français

Étiq.	Signification	Étiq.	Signification
ABR	abréviation	SBCsg	substantif, nom commun singulier
ADJsg	adjectif (sauf Participe passé) au singulier	SBCpl	substantif, nom commun pluriel
ADJpl	adjectif (sauf Participe Passé) au pluriel	SBPsg	substantif, nom propre ou à majuscule, singulier
ADV	adverbe	SBPpl	substantif, nom propre ou à majuscule, pluriel
CAR	cardinal (en chiffres ou en lettres)	SYM	symbole ou signe mathématique
COO	coordonnant	ACJsg	verbe « avoir », conjugué, singulier
DTNsg	dét. de groupe nominal, sing., non contracté	ACJpl	verbe « avoir », conjugué, pluriel
DTNpl	dét. de groupe nominal, plur., non contracté	ANCFF	verbe « avoir », non conjugué, infinitif
DTCsg	dét. de groupe nominal, sing., contracté	ANCNT	verbe « avoir », non conj., gérond. / part. présent
DTCpl	dét. de groupe nominal, plur., contracté	APARsg	verbe « avoir »,non conj., part. passé, singulier
FGW	mot étranger	APARpl	verbe « avoir »,non conj., part. passé, pluriel
INJ	interjection, onomatopée, etc.	ECJsg	verbe « être », conjugué, singulier
PFX	préfixe détaché	ECJpl	verbe « être », conjugué, pluriel
PREP	préposition	ENCFF	verbe « être », non conj., infinitif
PRVsg	pronom supporté par verbe (conjoint, clitique)	ENCNT	verbe « être », non conj., gérond. ou part. présent
PRVpl	pronom supporté par verbe (conjoint, clitique)	EPARsg	verbe « être », non conj., part. passé, sing.
PRV++	pronom supporté par verbe (clitique, réfléchi)	VCJsg	autre Verbe, conjugué, singulier
PROsg	autre pronom, singulier	VCJ	plautre Verbe, conjugué, pluriel
PROpl	autre pronom, pluriel	VNCFF	autre Verbe, non conj., infinitif
PRO++	autre pronom, genre indéterminé	VNCNT	autre Verbe, non conj., gérond. ou part. présent
PUL	particule non indépendante	VPARsg	autre Verbe, non conj., part. passé apr. « avoir »
REL	relatif (pronom, adjectif ou adverbe)	VPARpl	autre Verbe, non conj., part. passé apr. « avoir »
SUB	subordonnant	ADJ1PAR	part. passé apr « être », adjectival ou verbal,
SUB$	subordonnant possible = Code défaut de « que »	ADJ2PAR	Part. passé adjectival, sing./plu. (non apr. auxil.)

A.2 Variations repérées par l'outil FASTER

Nous illustrons, dans TAB. A.3, certaines variations repérées entre le terme dans le texte et le terme d'indexation ainsi qu'un commentaire sur la variation opérée par le système d'indexation automatique FASTER. Nous utilisons ce système durant notre processus de FdT pour représenter le contenu des textes de notre corpus (*cf.* § 2.4.2.3).

TAB. A.3 – Variations dans l'indexation entre la graphie du terme trouvée dans le texte et la graphie d'indexation

Terme d'origine	Terme d'indexation	Variation
amoxicillin	Amoxicillin	changement de graphie
after treatment	After-Treatment Aftercare	composition du terme synonymie
model	Models	terme attesté est au pluriel
mice numbers colonies sensitivities	Mouse Number Colony Sensitivity	terme simple mis au singulier
antibiotic resistant strains	antibiotic resistant strain	terme composé mis au singulier
sensitive helicobacter pylori strain	sensitive strain	terme composé mis au singulier et suppression de mots (XX,16,Ins)
isolated started	Isolate Starting	verbe mis à l'infinitf conjugué Isolate ou la forme progressif verbe mis à la forme progressive "ing"
higher dosage	high dosage	remplacement de l'adjectif
inoculation	Vaccination	synonymie
dosage	dosage	terme simple inchangé
treatment group	treatment group	terme composé inchangé
treatment failure	failure of treatment	terme de trois mots permutation et insertion de "of" (XXX,10,Perm)
strains were tested mouse were treated	tested strain treat mouse	nominalisation verbe conjugué + sujet nominalisation verbe infinitif + sujet (XX,31,Perm)
not	not gene	introduction de l'ambiguïté : transformation négation en nom de gène

A.3 Codage XML d'un texte du corpus

Nous donnons ici le format de codage d'une notice bibliographique (correspondant à un texte de notre corpus d'expérimentation). Par exemple en FIG. A.1, la notice `n°000867` est composée : du titre $<$ `TI` $>$, du texte qui est un résumé d'article scientifique $<$ `AB` $>$, de la liste des termes d'indexation produits par l'outil FASTR $<$ `MC` $>$, de la liste des auteurs $<$ `PA` $>$, de leur affiliation $<$ `AF` $>$, de la base de textes dont est issu le texte $<$ `SO` $>$, de l'année de la publication $<$ `PY` $>$, du pays de la publication $<$ `CP` $>$ et de la langue utilisée $<$ `LA` $>$.

<record>
<TI>Appearance of a metronidazole-resistant Helicobacter pylori strain in an infected-ICR-mouse model and difference in eradication of metronidazole-resistant and -sensitive strains</TI>
<AB>The numbers of colonies isolated from 56 ICR mice 2 weeks after 4 days of treatment with metronidazole (3.2, 10, or 32 mglkg of body weight) or amoxicillin (1, 3.2 or 10 mg/kg), with treatment started 4 days after H. pylori CPY2052 inoculation, were counted, and the isolated strains were tested for their sensitivities to two antibiotics to rule out the presence of antibiotic-resistant strains. We tested whether antibiotic-resistant strains appeared in vivo after the **failure of treatment** using the Helicobacter pylori-infected euthymic mouse model. The numbers of colonies isolated from 56 ICR mice 2 weeks after 4 days of treatment with metronidazole (3.2, 10, or 32 mglkg of body weight) or amoxicillin (1, 3.2 or 10 mg/kg), with treatment started 4 days after H. pylori CPY2052 inoculation, were counted, and the isolated strains were tested for their sensitivities to two antibiotics to rule out the presence of antibiotic-resistant strains. One metronidazole-resistant strain was detected in a mouse treated with 10 mg of metronidazole per kg, and the MIC of metronidazole for this strain was 25 μg/ml, compared to a MIC of 1.56 μg/ml for the original strain. However, no resistant **strain was detected** in the amoxicillin treatment group. After the examination described above, mice challenged with a metronidazole-resistant or -sensitive strain isolated from the stomach of a mouse were treated with metronidazole or amoxicillin. The metronidazole-resistant strain was more difficult to eradicate in vivo than the sensitive strain **after treatment** with metronidazole but **not** after treatment with amoxicillin. Thus, a metronidazole-resistant H. pylori strain was selected by insufficient treatment, but no resistant strain was selected with amoxicillin. Eradication of a metronidazole-resistant H. pylori strain in vivo required a higher dosage than eradication of a metronidazole-sensitive H. pylori strain. These results may explain one of the reasons for H. pylori **treatment failure**. </AB>
<MC>
<terme><mot>failure of treatment</mot>
<var><st>failure of treatment</st><tr>**0**</tr></var>
<var><st>treatment failure</st><tr>**XXX,10,Perm**</tr></var></terme>

<terme><mot>detect strain</mot>
<var><st>strain was detected</st><tr>**XX,31,Perm**</tr></var></terme>

<terme><mot>aftercare</mot><sy>after-treatment</sy>
<var><st>after treatment</st><tr>**0**</tr></var></terme>

<terme><mot>NOT gene</mot><sy>NOT</sy>
<var><st>not</st><tr>**0**</tr></var></terme>

<terme><mot>treatment failure</mot>
<var><st>failure of treatment</st><tr>**XX,37,Perm**</tr></var>
<var><st>treatment failure</st><tr>**0**</tr></var></terme>
</MC>
<PA>Matsumoto-S ; Washizuka-Y ; Matsumoto-Y ; Tawara-S ; Ikeda-F ; Yokota-Y ; Karita-M</PA>
<AF>Division of Chemotherapy, New Drug Research Laboratories, Fujisawa Pharmaceutical Co., Ltd., 2-1-6, Kashima, Yodogawa-ku, Osaka, Japan ; Hofuonsen Hospital, 1640, Daidou, Houfu, Yamaguchi, Japan</AF>
<SO>Antimicrobial-agents-and-chemotherapy. 1997 ;41 (12) :2602-2605</SO>
<IS>0066-4804</IS>
<PY>1997</PY>
<CP>United-States</CP>
<LA>English</LA>
</record>

FIG. A.1 – Exemple au format XML de la notice n°000867 de notre corpus.

Annexe B

Justifications mathématiques et démonstrations

Sommaire

B.1 Nombre maximal de règles générable

Nous montrons que le nombre maximal $\mathtt{N}$ de règles que nous pouvons engendrer à partir des données est égal à : $\mathtt{N} = 3^{\mathtt{n}} - 2^{\mathtt{n}+1} + 1$.

Soit $|\mathcal{P}| = \mathtt{n}$, $\mathtt{N}$ est le nombre total de règles sur $\mathcal{P}$.

Une règle $\mathtt{r}$= $\mathtt{p}_1 \Longrightarrow \mathtt{p}_2$ où $\mathtt{p}_1 \subseteq \mathcal{P}$, $\mathtt{p}_2 \subseteq \mathcal{P}$ et $\mathtt{p}_1 \cap \mathtt{p}_2 = \emptyset$ et $\mathtt{p}_1, \mathtt{p}_2 \neq \emptyset$.

Nous cherchons à trouver $\mathtt{N}$?

Soit $\mathtt{m} \in 2^{\mathcal{P}}$ avec $|\mathtt{m}| \geqslant 2$.

Le nombre de règles générables pas $\mathtt{m}$: Il y en a autant que de sous-ensembles $\mathtt{s}$ de $\mathtt{m}$ tels que $\mathtt{s} \neq \emptyset$ et $\mathtt{s} \neq \mathtt{m}$, c'est-à-dire : $2^{|\mathtt{m}|} - 2$.

Or, pour une règle $\mathtt{p}_1 \Longrightarrow \mathtt{p}_2$ donnée, il existe un motif $\mathtt{m}$ unique tel que $\mathtt{m} = \mathtt{p}_1 \cup \mathtt{p}_2$.

Donc :

$$\mathtt{N} = \sum_{\substack{\mathtt{m} \in 2^{\mathcal{P}} \\ |\mathtt{m}| \geqslant 2}} (2^{|\mathtt{m}|} - 2)$$

$$\mathtt{N} = \sum_{\alpha=2}^{\alpha=\mathtt{n}} (2^{\alpha} - 2) \ \times \ \mathtt{nombre\ de\ motifs\ de\ taille}\ \alpha$$

$$\mathtt{N} = \sum_{\alpha=2}^{\alpha=\mathtt{n}} (2^{\alpha} - 2) \ \times \ C_n^{\alpha}$$

$$\mathtt{N} = \sum_{\alpha=2}^{\alpha=\mathtt{n}} C_n^{\alpha} 2^{\alpha} - 2 \sum_{\alpha=2}^{\alpha=\mathtt{n}} C_n^{\alpha}$$

Or

$$\sum_{\alpha=2}^{\alpha=\mathtt{n}} C_n^\alpha = \sum_{\alpha=0}^{\alpha=\mathtt{n}} C_n^\alpha - C_n^1 - C_n^0 = 2^{\mathtt{n}} - \mathtt{n} - 1$$

Donc

$$\sum_{\alpha=2}^{\alpha=\mathtt{n}} C_n^\alpha 2^\alpha = \sum_{\alpha=0}^{\alpha=\mathtt{n}} C_n^\alpha 2^\alpha - 2C_n^1 - C_n^0 = \sum_{\alpha=0}^{\alpha=\mathtt{n}} C_n^\alpha 2^\alpha - 2\mathtt{n} - 1$$

De plus

$$\sum_{\alpha=0}^{\alpha=\mathtt{n}} C_n^\alpha 2^\alpha = (2+1)^{\mathtt{n}} = 3^{\mathtt{n}}$$

Car par le principe du triangle de Pascal :

$$(\mathtt{a}+\mathtt{b})^{\mathtt{n}} = \sum_{\mathtt{p}=0}^{\mathtt{p}=\mathtt{n}} \mathtt{C}_{\mathtt{n}}^{\mathtt{p}} \times \mathtt{a}^{\mathtt{p}} \times \mathtt{b}^{\mathtt{n}-\mathtt{p}}$$

D'où

$$\mathtt{N} = 3^{\mathtt{n}} - 2\mathtt{n} - 1 - 2\left(2^{\mathtt{n}} - \mathtt{n} - 1\right)$$

Après simplification, nous obtenons :

$$\mathtt{N} = 3^{\mathtt{n}} - 2^{\mathtt{n}+1} + 1$$

■

B.2 Probabilité conditionnelle

Quelle est la probabilité qu'un individu possède l'ensemble des propriétés `H` sachant qu'il possède l'ensemble des propriétés `B` ?

Soit `x` une variable aléatoire appartenant à $\mathcal{O}$.
Soit m un motif tel que $m \subseteq \mathcal{P}$.

$$\mathtt{P}(\mathtt{x\ possède\ H} \mid \mathtt{x\ possède\ B}) = \left(\frac{\mathtt{P(x\ possède\ H\ «\ ET\ »\ x\ possède\ B)}}{\mathtt{P(x\ possède\ B)}}\right)$$

Soit `f` la fonction qui relis m et `x` :

$$\mathtt{f} : 2^{\mathcal{P}} \longrightarrow 2^{\mathcal{O}}$$

$$m \longmapsto \mathtt{f}(m) = \{x \in \mathcal{O} \mid \forall p \in m \ \ p\mathcal{R}x \ (x\ possède\ m)\}$$

$$\mathtt{P(x\ possède\ m)} = \frac{|\mathtt{f(m)}|}{|\mathcal{O}|} = \mathtt{support(m)}$$

d'où $\left(\frac{\mathtt{P(x\ possède\ B} \cup \mathtt{H)}}{\mathtt{P(x\ possède\ B)}}\right) = \left(\frac{\mathtt{support(B} \cup \mathtt{H)}}{\mathtt{support(B)}}\right)$ /* Ce qui montre que la confiance $\in [0,\ 1]$ */.

Dans le cadre de la fouille de textes, par exemple :

Ensemble de termes :
$\mathcal{T} = \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$,
$\mathtt{B} = \{\mathtt{a}, \mathtt{b}\}$, $\mathtt{H} = \{\mathtt{c}\}$
Ensemble de textes :
$\mathcal{D} = \{\mathtt{d}_1, \mathtt{d}_2, \mathtt{d}_3\}$

$\not\Vdash R$	d_1	d_2	d_3
a	×		×
b	×	×	×
c		×	×

Soit $\mathtt{B} \subseteq \mathcal{P}$, $\mathtt{H} \subseteq \mathcal{P}$
$\mathtt{B}(\mathtt{x}) = \{\mathtt{x} \in \mathcal{O} \mid \forall \mathtt{p} \in \mathtt{B},\ \mathtt{p}\mathcal{R}\mathtt{x}\ (\mathtt{x}\ \mathtt{possède}\ \mathtt{B})\}$
$\mathtt{H}(\mathtt{x}) = \{\mathtt{x} \in \mathcal{O} \mid \forall \mathtt{p} \in \mathtt{H},\ \mathtt{p}\mathcal{R}\mathtt{x}\ (\mathtt{x}\ \mathtt{possède}\ \mathtt{H})\}$

$$\mathtt{confiance}(\mathtt{B} \Longrightarrow \mathtt{H}) = \left(\frac{|\mathtt{B}(\mathtt{X}) \cap \mathtt{H}(\mathtt{X})|}{|(\mathtt{X})|}\right)$$

$(\mathtt{x}\ \mathtt{possède}\ \mathtt{B}\ \mathtt{et}\ \mathtt{x}\ \mathtt{possède}\ \mathtt{H})$
$\Longleftrightarrow (\forall \mathtt{p} \in \mathtt{B},\ \mathtt{p}\mathcal{R}\mathtt{x}) \wedge (\forall \mathtt{p} \in \mathtt{H},\ \mathtt{p}\mathcal{R}\mathtt{x})$
$\Longleftrightarrow (\forall \mathtt{p} \in (\mathtt{B} \cup \mathtt{H}),\ \mathtt{p}\mathcal{R}\mathtt{x}) \Longleftrightarrow \mathtt{x}\ \mathtt{possède}\ (\mathtt{B} \cup \mathtt{H})$ ■

B.3 Règles redondantes

Nous reprenons ici TAB. 3.3 de § 3.2.4.4 et nous montrons que lorsque les règles d'association sont considérées avec le formalisme de la logique propositionnelle, nous voyons que les règles redondantes sont vraies lorsque la règle de départ est vraie :

TAB. B.1 – Ensemble de règles d'association redondantes engendrées par une règle valide

Règle de départ	Règle redondante
$\mathtt{ab} \Longrightarrow \mathtt{cd}$	$\mathtt{ab} \Longrightarrow \mathtt{c}$
	$\mathtt{ab} \Longrightarrow \mathtt{d}$
	$\mathtt{abc} \Longrightarrow \mathtt{d}$
	$\mathtt{abd} \Longrightarrow \mathtt{c}$

Soit $\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}$ quatre propositions et soit la règle d'association $\mathtt{a} \wedge \mathtt{b} \rightarrow \mathtt{c} \wedge \mathtt{d}$ notée sous sa forme logique. Nous avons :

$\mathtt{a} \wedge \mathtt{b} \rightarrow \mathtt{c} \wedge \mathtt{d}$
$\Longleftrightarrow \neg\mathtt{a} \vee \neg\mathtt{b} \vee (\mathtt{c} \wedge \mathtt{d})$
$\Longleftrightarrow (\neg\mathtt{a} \vee \neg\mathtt{b} \vee \neg\mathtt{c}) \wedge (\neg\mathtt{a} \vee \neg\mathtt{b} \vee \neg\mathtt{d})$
Si $(\neg\mathtt{a} \vee \neg\mathtt{b} \vee \neg\mathtt{c})$ alors $(\neg\mathtt{a} \vee \neg\mathtt{b} \vee \neg\mathtt{c} \vee \neg\mathtt{d})$
Si $(\neg\mathtt{a} \vee \neg\mathtt{b} \vee \neg\mathtt{d})$ alors $(\neg\mathtt{a} \vee \neg\mathtt{b} \vee \neg\mathtt{c} \vee \neg\mathtt{d})$
Pour que $\mathtt{a} \wedge \mathtt{b} \rightarrow \mathtt{c} \wedge \mathtt{d}$ soit vraie, il faut que les quatre sous-règles de TAB. B.1 soient vraies. ■

Annexe C

Description de l'outil TAMIS

« Un joli exemple vaut mieux qu'un long discours. »
Anonyme

Comme nous le signalons en § 4.3.2.2, nous présentons ici l'outil TAMIS. Nous commençons par la version Java puis la version disponible sur le Web à l'adresse :
`http : //www.loria.fr/~cherfi/GoldMine /MicroBio/Resultats/pagenavigation.html`

La FIG. C.1 montre une règle d'association. Deux listes de termes correspondant respectivement aux motifs `B` et `H` sont suivies d'un tableau qui donne la liste des mesures de qualité attachées à la règle ; puis la liste des textes qui ont permis d'extraire la règle (à gauche) et celle des textes qui sont des contre-exemples à la règle (à droite). Nous pouvons choisir de proposer un classement des règles suivant chacune des mesures (par ordre croissant et décroissant). Afin de faciliter la visualisation, nous pouvons choisir le nombre de règles à afficher par page (curseur en haut à gauche) et nous pouvons commencer la visualisation en milieu ou fin de classement (curseur en haut à droite).

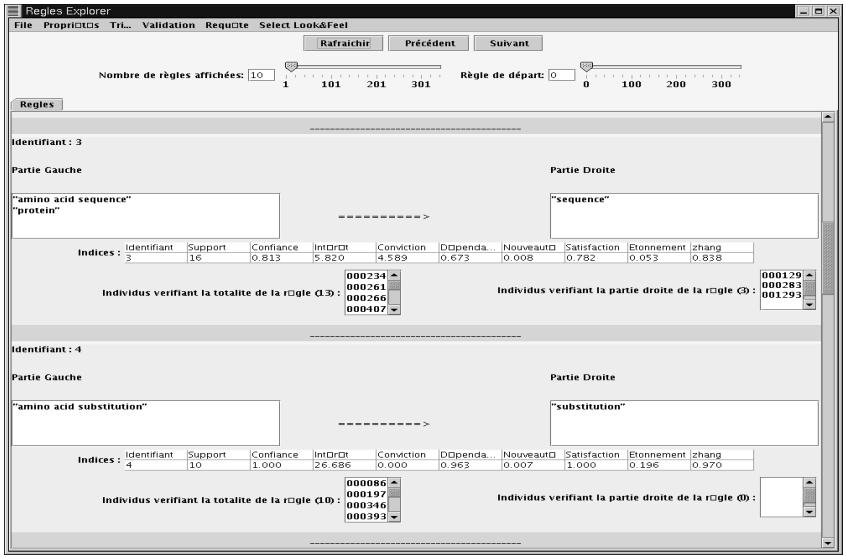


FIG. C.1 – Aperçu 1 de l'interface de navigation Java de l'outil TAMIS.

Pour se focaliser sur certaines règles d'association, le classement par mesures de qualité proposé ci-dessus ne suffit pas, la FIG. C.2 montre l'utilisation d'un langage de requête que nous proposons à l'analyste. Ce langage comprend l'opérateur de présence de termes `IN` (les termes sont sélectionnés par l'analyste et choisis parmi les termes en partie `B`, `H` ou les deux), de seuils pour les valeurs des mesures de qualité (les règles ayant une mesure de conviction supérieure/inférieure à une valeur donnée) ; ce sont les opérateurs disponibles pour les requêtes atomiques. De plus, nous proposons l'opérateur unaire `NOT` (pour exclure les règle qui possèdent un terme en partie `B`, `H` ou les deux), enfin nous proposons des opérateurs binaires `AND`, `N − AND`, `OR` et `N − OR` pour combiner des résultats de requêtes atomiques.

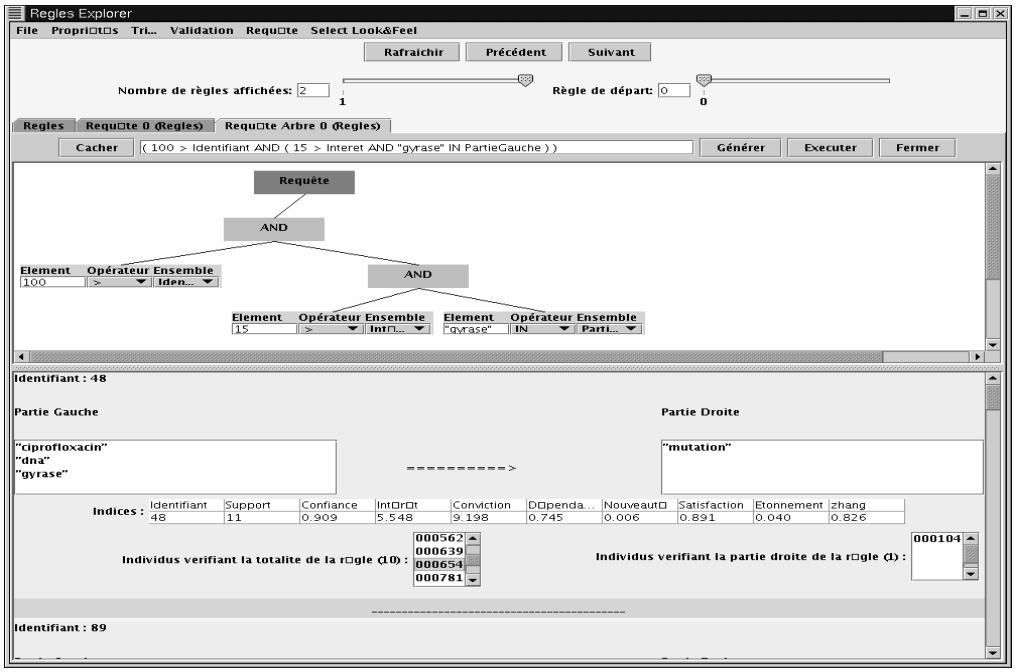


FIG. C.2 – Aperçu 2 de l'interface de navigation Java de l'outil TAMIS.

La FIG. C.3 présente la version Web de l'outil TAMIS. La version Web propose à l'analyste les mêmes informations et fonctionnalités que la version Java de FIG. C.1. Les classements par mesure de qualité sont disponibles dans le cadre gauche de l'interface Web. Nous pouvons également choisir le nombre de règle affichées par page ainsi que la règle de départ. Un clic sur un des numéros de textes nous amène à visualiser l'interface de FIG. C.3 qui présente la notice bibliographique du corpus (titre, auteurs, résumé, termes d'indexation). Un code de couleur différent, et dégradés de couleurs, sur-ligne chaque terme d'indexation ainsi que ses occurrences dans le texte. Ce qui permet à l'analyste de repérer plus facilement le terme qu'il cherche dans le texte.

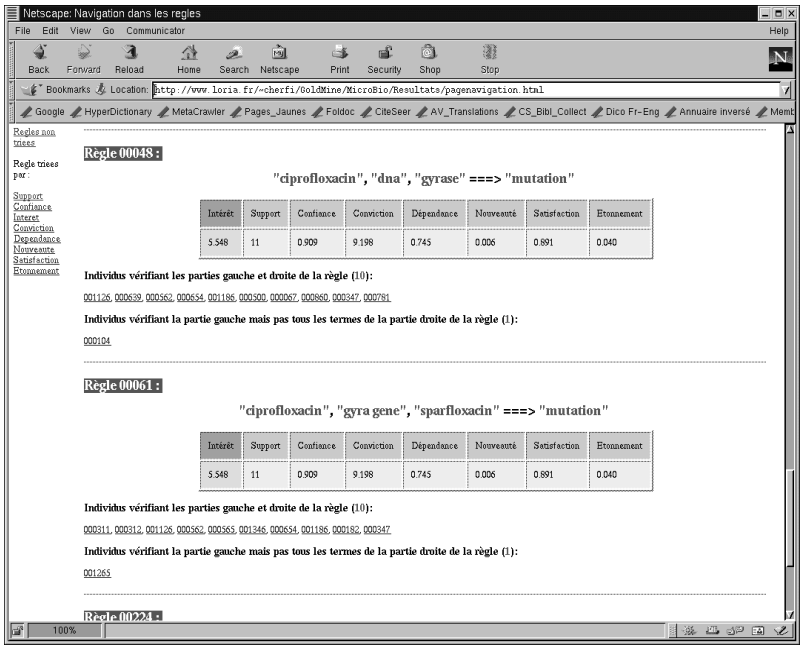


FIG. C.3 – Aperçu de l'interface de navigation Web de l'outil TAMIS.

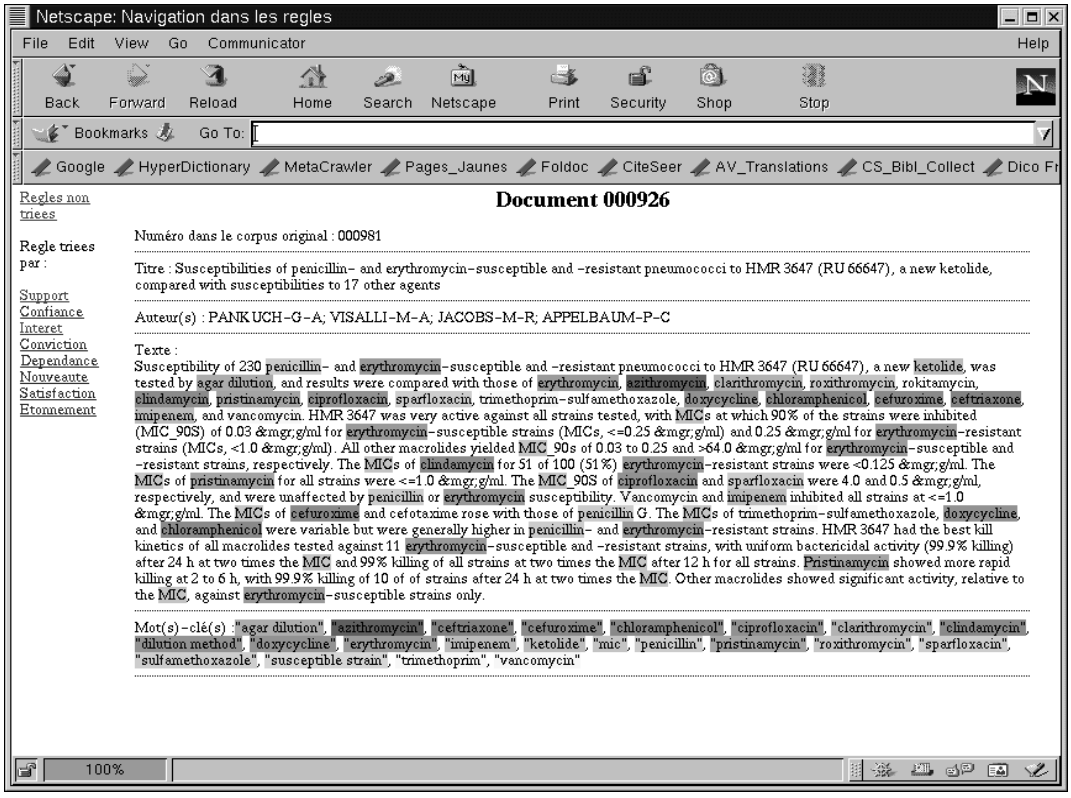


FIG. C.4 – Aperçu de l'interface présentant un texte et ses termes-index.

Annexe D

Détail des règles d'association extraites du corpus de biologie moléculaire

Nous donnons, dans cette annexe, le détail des mesures correspondant aux règles que nous avons cité au chapitre 4, ainsi que nombre de textes qui possèdent le motif $\mathtt{B} \cup \mathtt{H}$ – constituant les textes qui ont permis d'extraire la règle – et le nombre de textes qui possèdent le motif $\mathtt{B}$ uniquement (noté $\mathtt{B} \cap \neg\mathtt{H}$) – ce sont des textes qui constituent de contre-exemples à une règle *approximative* –. Les règles et leurs valeurs de mesures sont données dans l'ordre de référence à partir du § 4.3.2.3.

Numéro : **270**
Règle : **"meca"** $\sqcap$ **"meticillin"** $\Longrightarrow$ **"meca gene"** $\sqcap$ **"staphylococcus aureus"**

Support	Confiance	Intérêt	Conviction	Dépendance	Nouveauté	Satisfaction	Nb textes en $\mathtt{B} \cup \mathtt{H}$	Nb textes en $\mathtt{B} \cap \neg\mathtt{H}$
12	1,000	80,059	indéfinie	0,988	0,009	1,000	12	0

Numéro : **202**
Règle : **"grla gene"** $\Longrightarrow$ **"mutation"** $\sqcap$ **"staphylococcus aureus"**

Support	Confiance	Intérêt	Conviction	Dépendance	Nouveauté	Satisfaction	Nb textes en $\mathtt{B} \cup \mathtt{H}$	Nb textes en $\mathtt{B} \cap \neg\mathtt{H}$
12	0,917	40,245	11,727	0,894	0,008	0,915	11	1

Numéro : **159**
Règle : **"dna"** $\sqcap$ **"gyra gene"** $\Longrightarrow$ **"mutation"**

Support	Confiance	Intérêt	Conviction	Dépendance	Nouveauté	Satisfaction	Nb textes en $\mathtt{B} \cup \mathtt{H}$	Nb textes en $\mathtt{B} \cap \neg\mathtt{H}$
26	0,808	4,929	4,348	0,644	0,012	0,770	21	5

Numéro : **228**
Règle : **"gyrase"** $\sqcap$ **"protein"** $\Longrightarrow$ **"mutation"**

Support	Confiance	Intérêt	Conviction	Dépendance	Nouveauté	Satisfaction	Nb textes en $B \cup H$	Nb textes en $B \cap \neg H$
12	0,833	5,086	5,017	0,669	0,006	0,801	10	2

Numéro : **108**
Règle : **"dalfopristin"** $\Longrightarrow$ **"quinupristin"**

Support	Confiance	Intérêt	Conviction	Dépendance	Nouveauté	Satisfaction	Nb textes en $B \cup H$	Nb textes en $B \cap \neg H$
18	0,944	75,611	17,775	0,932	0,012	0,944	17	1

Numéro : **332**
Règle : **"quinupristin"** $\Longrightarrow$ **"dalfopristin"**

Support	Confiance	Intérêt	Conviction	Dépendance	Nouveauté	Satisfaction	Nb textes en $B \cup H$	Nb textes en $B \cap \neg H$
17	1,000	75,611	indéfinie	0,987	0,012	1,000	17	0

Numéro : **279**
Règle : **"mutation"** $\sqcap$ **"parc gene"** $\sqcap$ **"quinolone"** $\Longrightarrow$ **"gyra gene"**

Support	Confiance	Intérêt	Conviction	Dépendance	Nouveauté	Satisfaction	Nb textes en $B \cup H$	Nb textes en $B \cap \neg H$
21	0,952	20,574	20,028	0,906	0,014	0,950	20	1

Numéro : **215**
Règle : **"gyra gene"** $\sqcap$ **"pare gene"** $\Longrightarrow$ **"parc gene"** $\sqcap$ **"quinolone"**

Support	Confiance	Intérêt	Conviction	Dépendance	Nouveauté	Satisfaction	Nb textes en $B \cup H$	Nb textes en $B \cap \neg H$
12	0,917	41,586	11,735	0,895	0,008	0,915	11	1

Numéro : **120**
Règle : **"determine region"** $\sqcap$ **"gyra gene"** $\sqcap$ **"gyrase"** $\sqcap$ **"mutation"** $\Longrightarrow$ **"quinolone"**

Support	Confiance	Intérêt	Conviction	Dépendance	Nouveauté	Satisfaction	Nb textes en $B \cup H$	Nb textes en $B \cap \neg H$
11	1,000	17,012	indéfinie	0,941	0,008	1,000	11	0

Numéro : **273**
Règle : **"meticillin"** $\Longrightarrow$ **"staphylococcus aureus"**

Support	Confiance	Intérêt	Conviction	Dépendance	Nouveauté	Satisfaction	Nb textes en $B \cup H$	Nb textes en $B \cap \neg H$
52	0,865	6,543	6,446	0,733	0,028	0,845	45	7

Numéro : **265**
Règle : **"meca gene"** $\sqcap$ **"meticillin"** $\Longrightarrow$ **"staphylococcus aureus"**

Support	Confiance	Intérêt	Conviction	Dépendance	Nouveauté	Satisfaction	Nb textes en $B \cup H$	Nb textes en $B \cap \neg H$
15	0,933	7,057	13,016	0,801	0,009	0,923	14	1

Numéro : **329**
Règle : **"quinolone"** $\sqcap$ **"substitution"** $\Longrightarrow$ **"gyra gene"**

Support	Confiance	Intérêt	Conviction	Dépendance	Nouveauté	Satisfaction	Nb textes en $B \cup H$	Nb textes en $B \cap \neg H$
13	0,923	19,941	12,398	0,877	0,008	0,919	12	1

Glossaire

Nous donnons dans ce glossaire les définitions correspondant à certains sigles, acronymes et termes importants utilisés dans notre manuscrit.

FDT : Fouille de textes
ACF : Analyse de concepts formels, du terme anglais FCA : Formal Concept Analysis
EDCD : Extraction des connaissances à partir de bases de données
IDF : Inverted Document Frequency : mesure le nombre de classes ou de textes dans lequel un terme apparaît
IA : Intelligence artificielle
TF : Term Frequency : mesure le nombre d'occurrences d'un terme dans un texte
TAL : Traitement automatique de la langue

Corpus : Ensemble de textes choisi pour notre expérimentation

Itemset : motif constitué d'un ensemble de propriétés

Syntagme : Ensemble de mots constituant une unité syntaxique élémentaire

Thésaurus : Ressource terminologique constituée d'une nomenclature de termes

Index

Voici un index de termes-clés du domaine

Résumé

Ce travail de thèse porte sur la problématique d'extraction de connaissances à partir de textes, plus communément appelée la fouille de textes (FdT). Il s'articule autour des problèmes liés à l'analyse des textes, la fouille de textes proprement dite, et l'interprétation des éléments de connaissances extraits. Dans ce cadre, un système d'extraction des connaissances nécessaires pour analyser les textes en fonction de leur contenu est étudié et implanté. Les méthodes de fouille de données appliquées sont la recherche de motifs fréquents (avec l'algorithme « Close ») et l'extraction de règles d'association. Le mémoire s'attache à définir précisément le processus de fouille de textes et ses principales caractéristiques et propriétés en s'appuyant sur l'extraction de motifs fréquents et de règles d'association. En outre, une étude minutieuse d'un ensemble donné de mesures de qualité qu'il est possible d'attacher aux règles d'association est menée, toujours dans le cadre de la fouille de textes. Il est montré quel rôle ces mesures peuvent avoir sur la qualité et sur l'interprétation des règles extraites ; comment peuvent-elles influer sur la qualité globale du processus de fouille de textes. L'utilisation d'un modèle de connaissances vient appuyer et surtout compléter cette première approche. Il est montré, par la définition d'une mesure de vraisemblance, l'intérêt de découvrir de nouvelles connaissances en écartant les connaissances déjà répertoriées et décrites par un modèle de connaissances du domaine. Les règles d'association peuvent donc être utilisées pour alimenter un modèle de connaissances terminologiques du domaine des textes choisi. La thèse inclut la réalisation d'un système appelé TAMIS : « Text Analysis by Mining Interesting ruleS » ainsi qu'une expérimentation et une validation sur des données réelles de résumés de textes en biologie moléculaire.

Mots-clés: Fouille de textes, règles d'association, mesures de qualité, interprétation, apprentissage, raisonnement statistique, modèle de connaissances, biologie moléculaire.