Vers une méthodologie de fouille de textes s'appuyant sur l'extraction de motifs fréquents et de règles d'association

Hacène Cherfi, Amedeo Napoli, Yannick Toussaint

Vers une méthodologie de fouille de textes s'appuyant sur l'extraction de motifs fréquents et de règles d'association

Hacène Cherfi, Amedeo Napoli et Yannick Toussaint

LORIA (UMR 7503 – CNRS, INRIA, Universités nancéennes)
Campus scientifique - BP 239
Vandœuvre-lès-Nancy - F-54506 - France
{cherfi,napoli,yannick}@loria.fr et
http://www.loria.fr/equipes/orpailleur/

Résumé : Nous proposons la description d'une m´ethodologie d'interpr´etation des règles d'association extraites à partir de textes. Le corpus qui a servi à notre exp´erience est une collection de textes sous forme de r´esum´es d'articles scientifiques dans le domaine de la biologie mol´eculaire. Notre recherche porte sur : i) l'extraction des règles d'association à partir de la construction des motifs ferm´es fr´equents g´en´er´es par l'algorithme « *Close* » ; ii) l'association de mesures *qualitatives* à chaque règle, ce qui permet de les ordonner ; iii) l'interpr´etation des règles par un analyste (expert du domaine) ; iv) la mise en correspondance des points ii) et iii). Nous montrons comment aider l'analyste, grâce à des mesures de qualit´e, dans l'interpr´etation des règles. Une discussion sur nos r´esultats met en valeur des points qui nous paraissent fondamentaux dans l'interpr´etation des règles d'association.

Mots-clés : Fouille de textes, règles d'association, mesures de qualit´e, interpr´etation, biologie mol´eculaire.

1 Introduction

La fouille de données dans les textes ou fouille de textes (FdT) fait l'objet d'une attention particulière de la part de la communauté d'extraction des connaissances à partir de données. La disponibilité d'une grande masse de données, principalement en provenance des bibliothèques électroniques et du Web, mais également dans les domaines industriels (documentations techniques aéronautique, automobile, etc.) ou médicaux (dossiers cliniques) rend nécessaire l'utilisation de techniques de fouille de données pour en extraire les éléments de connaissances les plus pertinents.

Notre travail porte sur la fouille de textes. Nous présentons dans cet article une méthodologie de fouille de textes qui s'appuie sur le processus de fouille de données. Nous mettons l'accent sur l'extraction et la classification de règles d'association ; puis sur l'interprétation des résultats par un analyste. Nous montrons que les étapes de

sélection, de prétraitement et d'indexation des textes orientent, globalement le processus de fouille. Ces étapes ont une grande influence sur la qualité des connaissances extraites à partir des textes. De plus, les textes étudiés ont été indexés automatiquement puis nettoyés manuellement et nous montrons comment le résultat du processus de fouille peut améliorer *in fine* cette indexation.

Ce travail est original à plus d'un titre puisqu'il présente une méthodologie complète et opérationnelle de fouille de textes s'appuyant sur des méthodes symboliques. De ce point de vue, un certain parallèle existe avec des travaux qui se sont intéressés à cette même problématique (Azé & Roche, 2003).

En section 2, nous situons le contexte de fouille de textes dans un processus plus général d'extraction des connaissances dans des bases de données. En section 3, nous donnons la définition de ce que nous appelons la « fouille de textes ». Les étapes de fouille sont décrites à travers : la modélisation des textes (section 4), le processus opérationnel d'extraction des motifs fréquents et des règles d'association (section 5), l'utilisation de mesures de qualité pour faire des tris sur les règles extraites (section 6) et l'étape d'interprétation et de validation des résultats (section 7). Enfin, des éléments de discussion sur nos résultats ainsi que des approches similaires sont présentées en section 8.

2 Fouille de textes : un paradigme de l'ECBD

L'extraction de connaissances dans des bases de données — abrégée en ECBD — est une activité qui consiste à analyser un ensemble de données brutes de façon à en extraire des connaissances exploitables. C'est un expert du domaine relatif aux données — l'« analyste » — qui est chargé de diriger l'extraction. En fonction de ses objectifs, l'analyste va sélectionner des données et utiliser des outils de *fouille de données* pour construire des modèles expliquant les données. L'analyste peut ensuite sélectionner et exploiter les modèles qui représentent un point de vue satisfaisant. Pour mener à bien son activité, l'analyste met à contribution ses connaissances mais aussi un ensemble d'outils regroupés au sein d'un système d'ECBD. Un système d'ECBD s'articule autour de quatre composantes principales :

1) les bases de données et leurs systèmes de gestion ; 2) un système à base de connaissances pour la gestion des connaissances et la résolution de problèmes sur le domaine relatif aux données ; 3) un système de fouille de données pouvant s'appuyer sur des techniques symboliques ou numériques comme les classifications par treillis et par arbres de décision, l'induction, l'analyse des données ou les statistiques ; 4) une interface se chargeant des interactions et de la visualisation des résultats.

Un système d'ECBD vise à traiter des bases de données volumineuses et évolutives, et il peut, pour ce faire, s'appuyer sur des connaissances du domaine lors du processus d'extraction des connaissances. L'ECBD peut être 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. Dans la suite, nous considérons les textes comme des données et nous montrons comment extraire, à partir de ces données, des éléments d'information, qui deviennent par la suite des connaissances. Nous faisons l'hypothèse qu'un texte contient

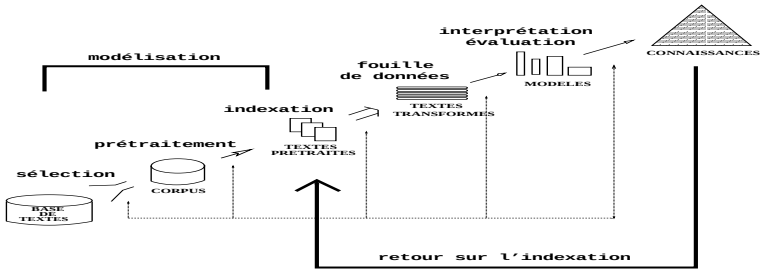

FIG. 1 – *Le processus de* F*d*T.

des connaissances explicites et implicites : il faut donc savoir représenter les connaissances explicites pour les exploiter afin de pouvoir représenter les connaissances implicites.

3 Fouille de textes : vers une définition

Nous définissons la fouille de textes à travers les trois étapes de la figure 1. La modélisation des textes, l'activation du processus de fouille de données, et l'interprétation des informations extraites.

La FdT s'adresse à un analyste qui est expert dans un domaine particulier. Elle donne à celui-ci une vue synthétique du contenu d'un corpus (*i.e.* une base de textes sélectionnée), exhibe des relations entre les différentes notions impliquées 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 aussi de découvrir de nouvelles relations.

Dans ce qui suit, 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 règles d'association extraites peuvent être interprétées comme des cooccurrences de termes dans les textes ; et par conséquent, refléter des liens sémantiques entre termes d'un texte, comme cela est le cas en sémantique lexicale (Anick & Pustejovsky, 1990).

Cependant, le nombre de règles extrait croît de manière exponentielle par rapport au nombre de termes du corpus. L'interprétation des règles est alors une tâche difficile. Nous suggérons des moyens de sélectionner, parmi ces règles, celles qui présentent un intérêt particulier pour l'expert. Pour cela, nous procédons en deux étapes :

1. L'expert identifie, dans l'ensemble des règles, un sous-ensemble de règles qui présente, pour ses besoins et pour ses connaissances, un intérêt particulier ;
2. nous calculons des mesures de qualité associées à chacune des règles qui suggèrent à l'expert une classification des règles extraites et une sélection de celles qui lui semblent les plus pertinentes de son point de vue.

Les trois prochaines sections détaillent les étapes de fouille de textes.

4 Du texte à sa modélisation

Si la fouille de textes est proche de la fouille de données, elle s'en différencie sur quelques points majeurs. Même s'il n'y a pas un consensus sur la dénomination des différents niveaux d'analyse d'un texte, les travaux en Traitement Automatique de la Langue (TAL) décomposent généralement un texte en plusieurs niveaux de structure : la structure logique du texte (introduction, hypothèses, développement, conclusion, etc.) ; la structure du discours pour parler généralement de ce qui gère l'articulation entre les paragraphes ou l'enchaînement des phrases dans les paragraphes ; la sémantique de la phrase ; sa syntaxe ; et finalement le lexique qui peut contenir des informations sémantiques plus ou moins fines pour établir des liens entre mots et permettre de construire la sémantique de la phrase.

Tous ces niveaux font des textes des données beaucoup plus complexes à traiter que des bases de données formelles 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. C'est d'ailleurs en partie cette complexité d'analyse des textes qui fait l'intérêt et la complémentarité de la FdT comparé à la fouille dans des bases de données puisque le contenu qui est extrait d'un texte peut être différent en fonction de la représentation choisie pour chacun de ses niveaux.

Nous décomposons le passage du texte à sa modélisation en deux étapes. La première, le prétraitement, est chargé d'extraire dans les textes les parties textuelles intéressantes et de les annoter pour que des outils robustes des TAL puissent être mis en œuvre dans la seconde étape. Cette seconde étape doit représenter le texte dans un système formel sur lequel les outils de fouille de données pourront être appliqués.

4.1 Le prétraitement des textes

4.1.1 La sélection des champs textuels dans les sources

Les textes que nous traitons sont extraits de la base de données documentaires PASCAL-BIOMED de l'INIST [1] constituée de notices bibliographiques d'articles scientifiques. De même qu'aux données peuvent être associées des métadonnées, les textes sont également caractérisés par un ensemble de données contextuelles qui se trouvent dans des champs codés en XML : titre, auteur(s), date, statut (publié ou non), mots-clés, etc. La figure 2 en donne une vue partielle.La première étape du prétraitement porte donc sur l'extraction pour chaque document des deux champs constitués de textes moins formellement structurés : le titre et le résumé. Nous utilisons la librairie DILIB (Ducloy, 1999) qui manipule des structures XML.

4.1.2 L'étiquetage morpho-syntaxique

L'étiquetage morpho-syntaxique correspond à la préparation des textes pour l'application d'outils de TAL dans la phase de modélisation du contenu . Cet étiquetage associe à chaque mot d'une phrase, sa catégorie morpho-syntaxique (nom, adjectif, verbe,

1. INstitut de l'Information Scientifi que et Technique, ´etablissement partenaire qui nous a fourni le corpus.

Document 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 ofbxacin (0.5 microg/ml) and sparfbxacin (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 ofbxacin and sparfbxacin.

FIG. 2 – *Vue partielle d'une notice bibliographique.*

etc.). Plusieurs étiqueteurs existent à l'heure actuelle sur l'anglais et atteignent des performances autour de 99,5% de correction, ce qui en fait des outils fiables. Ils utilisent généralement, à la base, un modèle statistique de langage appris qui peut prédire la catégorie d'un mot en fonction de la catégorie du mot précédemment rencontré. Brill (Brill, 1999) intègre également un lexique, des règles lexicales et contextuelles qui le rend plus facilement adaptable à un nouveau domaine scientifique pour lequel le vocabulaire ou les tournures langagières sont plus spécifiques. Par exemple, la phrase (1) étiquetée donne la phrase (2):

1. Les fractions pectiques contiennent des proportions hautement estérifiées
2. Les/DTN:pl fractions/SBC:pl pectiques/ADJ:pl contiennent/VCJ:pl des/PREP:pl proportions/SBC:pl hautement/ADV estérifiées/ADJ2PAR:pl

Outre l'adaptation au domaine de ces outils, la forme même des textes peut influencer la qualité de l'étiquetage : ils sont initialement prévus pour fonctionner sur des phrases complètes, syntaxiquement correctes mais isolées. Ainsi, nous avons dû constituer une nouvelle configuration de l'étiqueteur de Brill pour qu'il soit adapté au traitement de séquences nominales isolées comme c'est le cas lorsque l'on souhaite étiqueter des listes de termes issues d'un thésaurus.

4.2 Modélisation du contenu des textes : l'indexation terminologique

Dans la mesure où l'analyste n'avait pas formulé d'objectifs de fouille particuliers, nous avons adopté une modélisation du contenu des résumés de textes dont la sémantique est simple : chaque résumé est modélisé par l'ensemble des termes qu'il possède. C'est une indexation contrôlée à partir d'une liste de termes attestée. Cette indexation terminologique contrôlée permet d'associer à un groupe de mots, un concept – *i.e.* une notion participant à une base de connaissance – et permet ainsi de passer d'un élément de nature linguistique à un élément de nature connaissance. De plus, l'indexation terminologique constitue la première étape de représentation de la sémantique d'un énoncé sans nécessiter de choix *a priori* sur la nature de la représentation. Elle présuppose simplement que la cooccurrence de termes dans un même texte reflète une certaine proximité sémantique entre ces termes. Enfin, les résumés essaient de situer les travaux et les avancées scientifiques en peu de mots. On y retrouve donc une forte densité de termes, un maximum de contenu informationnel et un minimum d'information inutile.

4.2.1 L'identification des termes et de leurs variantes

FASTR (Jacquemin, 1994) est un outil d'identification dans les textes des termes à partir d'une nomenclature attestée. Dans notre cas, cette nomenclature résulte de la fusion de plusieurs thésaurus du domaine. Dans le but de réduire le silence (ne pas réussir à reconnaître un terme dans un texte), FASTR permet de reconnaître un terme sous des formes variantes. Par exemple, le terme "*transfer of capsular biosynthesis genes*" doit être considéré comme une forme variante du terme attesté "*gene transfer*". Cependant, certaines variantes ne sont pas acceptables aux yeux de l'expert et l'intérêt de FASTR est de ne garder dans les variantes que celles qui sont issues d'une transformation linguistique préservant le sens.

Chaque terme de la nomenclature attestée est caractérisé par sa structure syntaxique (étiquetage morpho-syntaxique). Étant donné une forme variante rencontrée dans un texte, FASTR 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). Ainsi, la forme attestée "*gene transfer*" peut être reconnue (opération d'inversion en anglais) 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 FASTR.

4.2.2 Description des données

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

Deux indexations ont été menées avec ce corpus sur la biologie moléculaire. La première expérience a eu lieu avec une indexation entièrement automatique utilisant FASTR. 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 %) étaient 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 texte en langage naturel.

Une seconde expérience 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, près de la moitié, des termes considérée comme du bruit. 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).

5 Le processus de fouille de textes

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.*, 1999). Ces motifs fréquents permettent de construire des règles d'association. Les mesures de qualité des règles calculées en (2) sont des mesures affectées aux règles. Ces mesures pondèrent chaque règle et permettent donc de les « classer ». Un environnement de navigation (3) aide l'expert du domaine à interpréter les règles d'association obtenues en (1). Il lui permet d'accéder au contenu des textes liés à une règle (*cf.* figure 2, complétée par la liste des termes issus de l'indexation du titre et du résumé).

5.1 Motifs fréquents pour la FdT

Les algorithmes de fouille que nous utilisons sont fondés sur l'extraction des motifs fréquents afin de générer un ensemble de règles d'association.

Un motif fréquent est défini comme un ensemble d'items présents dans un nombre « suffisamment grand » d'objets d'une base de données formelle. Pour qu'un motif soit *fréquent*, il suffit que le nombre de fois où il apparaît soit $\geq$ à un seuil σ fixé en paramètre. Cette fréquence d'apparition du motif est appelée *support* du motif.

Dans notre cas, un objet est un texte, un item 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 en relation par l'intermédiaire d'une relation d'indexation. Cette relation peut être représentée sous la forme d'une matrice ($\mathcal{D} \times \mathcal{T}$) de booléens ($\mathtt{1}$ = présence et $\mathtt{0}$ = absence du terme $\mathtt{t}$ dans un texte particulier $\mathtt{d}$). Cette matrice constitue la structure en entrée du processus de fouille de données.

L'approche naïve pour chercher les motifs fréquents consiste à compter le nombre de fois que chaque ensemble des parties de $\mathcal{T}$ apparaît. Ce qui donne $2^{\mathcal{T}}$ sous-ensembles à tester. 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 *Close* qui minimise cet espace de recherche et qui est décrit en détails dans (Bastide *et al.*, 2002).

5.2 Règles d'association pour la FdT

Les règles d'association ont été initialement étudiées en analyse de données ; puis en fouille de données afin de trouver des régularités, des corrélations dans des bases de données relationnelles de grandes tailles. Les règles d'association ont été, par la suite, appliquées à la fouille de textes (Feldman *et al.*, 1998; Kodratoff, 1999).

Définition 1 (Règle d'association)
Une r`egle d'association est du type :

$\mathtt{R}$: $\mathtt{t_1}$,..., $\mathtt{t_i} \Longrightarrow \mathtt{t_{i+1}}$,..., $\mathtt{t_n}$ *(o`u* $\{\mathtt{t_1}, ..., \mathtt{t_n}\}$ *est un ensemble de* termes*)*

Une règle $\mathtt{R}$: $\mathtt{B} \Longrightarrow \mathtt{H}$ est constituée d'un ensemble de termes ($\mathtt{B}$) impliquant un ensemble de termes ($\mathtt{H}$). Dans notre contexte, l'explication intuitive de la règle $\mathtt{R}$ est : si des textes possèdent tous les termes de l'ensemble $\{\mathtt{t_1}, ... , \mathtt{t_i}\}$ alors ils possèdent tous les termes de $\{\mathtt{t_{i+1}}, ... , \mathtt{t_n}\}$ avec une probabilité $\mathtt{p}$.

Une règle se construit à partir du motif $\mathtt{B} \bigcup \mathtt{H}$. Le **support** de $\mathtt{B} \Longrightarrow \mathtt{H}$ est défini comme le support de $\mathtt{B} \bigcup \mathtt{H}$. Il est d'usage de définir le support d'une règle comme

$\texttt{support}(\texttt{B} \bigcup \texttt{H})$, mais il dénote le $\texttt{support}(\texttt{B et H})$ 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 de B "et" tous les termes de H. Nous choisissons de prendre la notation $\texttt{support}(\texttt{B} \bigcap \texttt{H})$.

L'indice de **confiance** de R est défini par $\frac{\texttt{support}(\texttt{B} \bigcap \texttt{H})}{\texttt{support}(\texttt{B})}$. La confiance donne une mesure du pourcentage d'exemples (et de contre-exemples) de la règle. Un contre-exemple signifie qu'il y a des textes qui possèdent les termes B mais pas nécessairement tous les termes H. Lorsque la confiance vaut 1, la règle est dite *totale* (ou exacte). Elle s'exprime sous la forme d'une condition : « S'il pleut dehors, alors le sol sera mouillé ». Sinon la règle est dite *partielle* et se voit attribuer une confiance variant entre 0 et 1. Par exemple :« Dans 60% des cas (*i.e.* avec une confiance de 0,6), les textes qui parlent de quinupristine parlent aussi de dalfopristine ».

Une règle est valide si son support est supérieur à un seuil minsup (au-dessus duquel le motif est considéré comme fréquent), et un seuil minconf pour ne générer que des règles dont la valeur de confiance est comprise entre minconf et 1. Les algorithmes de construction de règles d'association utilisent par définition ces deux seuils.

Le nombre de règles extrait croît de manière exponentielle par rapport au nombre de termes du corpus.

Notre approche consiste à conserver toutes les règles car nous ne pouvons préjuger de celles qui seront, au final, retenues par l'expert. Nous cherchons à aider l'expert 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 » différents sur l'ensemble des règles extrait.

6 Mesures de qualité des règles d'association

Soient $\mathcal{D}(\texttt{B})$, $\mathcal{D}(\texttt{H})$ et $(\mathcal{D}(\texttt{B} \bigcup \texttt{H}) = \mathcal{D}(\texttt{B}) \bigcap \mathcal{D}(\texttt{H}))$ les ensembles de textes de $\mathcal{D}$ possédant respectivement tous les termes de B, H et $\texttt{B} \bigcup \texttt{H}$ (*cf.* figure 3). Trois valeurs de probabilités ont un impact sur la valeur des mesures que nous utilisons. Il s'agit de : $\texttt{P}(\texttt{B})$, $\texttt{P}(\texttt{H})$ et $\texttt{P}(\texttt{B},\texttt{H})$ qui se définissent par la formule générale suivante : $\left(\texttt{P}(\texttt{X}) = \frac{|\mathcal{D}(\texttt{X})|}{|\mathcal{D}|}\right)$ compris entre 0 et 1. $\texttt{P}(\texttt{B},\texttt{H})$ est le support de la règle. La probabilité conditionnelle $\texttt{P}(\texttt{H}|\texttt{B}) = \frac{\texttt{P}(\texttt{B},\texttt{H})}{\texttt{P}(\texttt{B})}$ en est la confiance.

Plus $\mathcal{D}(\texttt{X})$ est grand (*i.e.* couvre l'espace $\mathcal{D}$), plus $\texttt{P}(\texttt{X})$ est fort (*i.e.* proche de 1). La règle met donc en présence des motifs, en parties B et H, très fréquents qui décrivent presque tous les textes. Par conséquent, les connaissances potentiellement apportées par ces motifs, du point de vue de l'extraction et de la découverte de connaissances par l'expert, sont considérées comme des connaissances non informatives.

– Pour le cas (a), $\texttt{P}(\texttt{B})$ et $\texttt{P}(\texttt{H})$ sont toutes deux proches de 1, 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") ;

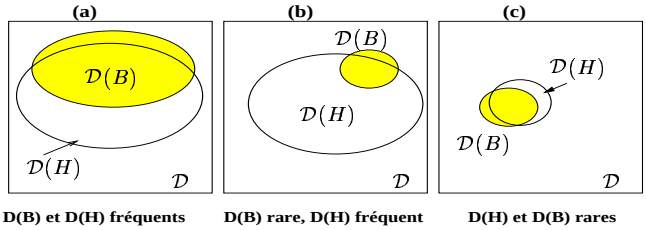

FIG. 3 – *Principaux cas illustrant les variations de* $\mathcal{D}(\mathtt{B})$ *et* $\mathcal{D}(\mathtt{H})$

– 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 B aura tendance à posséder H ;
– le cas (c) est le plus intéressant. Les termes y sont rares et apparaissent presque à chaque fois ensemble (*i.e.* $\mathtt{P}(\mathtt{B},\mathtt{H}) \simeq \mathtt{P}(\mathtt{B}) \simeq \mathtt{P}(\mathtt{H})$). Ces termes sont donc vraisemblablement reliés dans un contexte du domaine ; le quatrième cas possible ($\mathcal{D}(\mathtt{B})$ fréquent, $\mathcal{D}(\mathtt{H})$ rare) n'est pas dans notre contexte. Pour avoir ce cas, il faut prendre un seuil de *confiance* faible $\left(\frac{\mathtt{P}(\mathtt{B},\mathtt{H})}{\mathtt{P}(\mathtt{B})} \ll 1\right)$.

6.1 Indices de support et de confiance

Les indices de support et de confiance ne différencient pas, complètement, les cas (a), (b) et (c) de la figure 3. Le support représente l'intersection $\mathcal{D}(\mathtt{B}) \cap \mathcal{D}(\mathtt{H})$, il peut alors distinguer (a) de ((b) et (c)). La confiance représente l'inclusion de $\mathcal{D}(\mathtt{B})$ dans $\mathcal{D}(\mathtt{H})$ et n'est pas un facteur discriminant de ces trois cas.
Pour ces raisons, les indices de support et de confiance ne sont pas suffisants pour identifier les cas du plus significatif (c) vers le moins significatif (a). Leurs caractéristiques statistiques ne reflètent pas totalement une significativité de la règle. Le paragraphe suivant montre que d'autres valeurs d'*indices statistiques* (ou mesures de qualité) sont capables de différencier les trois cas possibles, dans cette étude, de la figure 3.

6.2 Autres indices de qualité des règles

Nous présentons d'autres indices, dont une synthèse se trouve dans (Lavrač *et al.*, 1999), qui permettent différents classements des règles.

6.2.1 L'intérêt

L'indice d'**intérêt** (ou *lift*) mesure la déviation du support de la règle par rapport au cas d'indépendance. Rappelons que pour deux événements indépendants B et H, $\mathtt{P}(\mathtt{H}|\mathtt{B}) = \mathtt{P}(\mathtt{H})$ et donc $\mathtt{P}(\mathtt{B},\mathtt{H}) = \mathtt{P}(\mathtt{B}) \times \mathtt{P}(\mathtt{H})$.
La valeur de l'intérêt est donnée par :

$$\mathtt{int}\left[\mathtt{B} \implies \mathtt{H}\right] = \frac{\mathtt{P}(\mathtt{B},\mathtt{H})}{\mathtt{P}(\mathtt{B}) \times \mathtt{P}(\mathtt{H})} \qquad (1)$$

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

Par définition, on a $(\mathcal{D}(\mathtt{B}) \bigcap \mathcal{D}(\mathtt{H})) \subseteq \mathcal{D}(\mathtt{B})$ et $(\mathcal{D}(\mathtt{B}) \bigcap \mathcal{D}(\mathtt{H})) \subseteq \mathcal{D}(\mathtt{H})$. 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 $\mathtt{P}(\mathtt{B},\mathtt{H}) \simeq \mathtt{P}(\mathtt{B})$ alors $\mathtt{int}\,[\mathtt{B} \Longrightarrow \mathtt{H}] \simeq \frac{\mathtt{P}(\mathtt{B})}{\mathtt{P}(\mathtt{B}) \times \mathtt{P}(\mathtt{H})} = \frac{1}{\mathtt{P}(\mathtt{H})}$, de la même manière $\mathtt{int}\,[\mathtt{B} \Longrightarrow \mathtt{H}] = \frac{1}{\mathtt{P}(\mathtt{B})}$. 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) seront classées en premier. L'intérêt est une mesure symétrique $\mathtt{int}\,[\mathtt{B} \Longrightarrow \mathtt{H}] = \mathtt{int}\,[\mathtt{H} \Longrightarrow \mathtt{B}]$.

6.2.2 La conviction

La **conviction** mesure la déviation du support du contre-exemple à la règle dû au motif $\mathtt{B} \bigcap \neg\mathtt{H}$ 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 $\mathtt{P}(\neg\mathtt{H}) = 1 - \mathtt{P}(\mathtt{H})$.

$$\mathtt{conv}\,[\mathtt{B} \implies \mathtt{H}] = \frac{\mathtt{P}(\mathtt{B}) \times \mathtt{P}(\neg\mathtt{H})}{\mathtt{P}(\mathtt{B},\neg\mathtt{H})} \tag{2}$$

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 $\mathtt{P}(\neg\mathtt{H})$ est fort (*i.e.* $\mathtt{P}(\mathtt{H})$ faible), $\mathtt{P}(\mathtt{B})$ est fort et lorsque $\mathtt{P}(\mathtt{B},\mathtt{H}) \simeq \mathtt{P}(\mathtt{B})$ car $\mathtt{P}(\mathtt{B}) = \mathtt{P}(\mathtt{B},\mathtt{H}) + \mathtt{P}(\mathtt{B},\neg\mathtt{H})$. Ce qui classe les règles du cas (c) en premier.

Comme l'intérêt, cet indice varie dans l'intervalle $[0, +\infty[$. Cet indice dénote une dépendance entre $\mathtt{B}$ et $\mathtt{H}$ s'il est > 1, une indépendance s'il est $= 1$ et pas de dépendance si cet indice est compris dans $[0,1[$. Il n'est pas calculable pour les règles totales puisque nous ne pouvons diviser par $\mathtt{P}(\mathtt{B},\neg\mathtt{H})$ qui vaut 0, car il n'y a aucun contre-exemples à la règle.

6.2.3 La dépendance

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

$$\mathtt{dep}\,[\mathtt{B} \implies \mathtt{H}] = |\mathtt{P}(\mathtt{H}|\mathtt{B}) - \mathtt{P}(\mathtt{H})| \tag{3}$$

Cet indice varie dans l'intervalle $[0,1[$ car c'est une valeur absolue, toujours positive. Plus cet indice 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 totales 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) d'une part et des cas ((a) et (b)) d'autre part. Pour cette raison, les deux indices suivants qui mesurent également des dépendances sont définis.

6.2.4 La nouveauté et la satisfaction

L'indice de **nouveauté** est défini par :

$$\mathtt{nov}\,[\mathtt{B} \Longrightarrow \mathtt{H}] = \mathtt{P}(\mathtt{B},\mathtt{H}) - \mathtt{P}(\mathtt{B}) \times \mathtt{P}(\mathtt{H}) \tag{4}$$

La valeur absolue de cet indice vaut $\mathtt{dep}\,[\mathtt{B} \implies \mathtt{H}] \times \mathtt{P}(\mathtt{B})$. Plus $\mathtt{P}(\mathtt{B})$ est faible, plus cet indice est petit. Ainsi, les règles des cas (b) sont rejetées en fin de classement et sont

différenciées du cas (a), alors que la dépendance ne le fait pas. Nous sommes intéressés par les petites valeurs absolues de cet indice (*i.e.* autour de la valeur d'indépendance 0). La nouveauté varie entre $]-1,1[$ et prend une valeur négative quand $P(B,H) < P(B) \times P(H)$. La nouveauté s'approche de -1 pour des règles de faibles supports $P(B,H) \simeq 0$.
La nouveauté est symétrique. Nous avons une même valeur de cet indice pour les règles $B \Longrightarrow H$ et $H \Longrightarrow B$, alors que l'une peut avoir plus de contre-exemples que l'autre. Pour cette raison, nous introduisons l'indice suivant appelé **satisfaction** :

$$\texttt{sat}\,[B \implies H] = \frac{(P(\neg H) - P(\neg H|B))}{P(\neg H)} \tag{5}$$

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

Cette mesure varie dans l'intervalle $[0,1]$ et vaut 0 en cas d'indépendance de B et H. La satisfaction n'est pas utile pour classer les règles totales car sa valeur est 1 (puisque les règles totales ont une confiance $P(H|B) = 1$).

Pour cet indice, $P(H)$ apparaît au numérateur et au dénominateur, donc la variation de cet indice dépend de $P(B)$. Plus $P(B)$ est faible, plus cet indice est élevé. Par l'intermédiaire de cet indice, les règles du cas (a) sont rejetées en fin de classement et sont différenciées du cas (b). Nous sommes intéressés par de fortes valeurs de cet indice (*i.e.* autour de la valeur 1).

En somme, ces deux indices peuvent être consultés 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 7.3.

7 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 expert. Il faut noter que chaque indice ne couvre qu'un sous-intervalle de ses valeurs possibles. Par exemple, nous n'observons pas de cas d'indépendance pour les indices. De même, nous n'avons pas de valeurs négatives pour la *nouveauté*.

7.1 Description des résultats

Dans le paragraphe 4.2.2, nous avons mentionné deux indexations (automatique et filtrée manuellement) faites sur un corpus de résumés de textes. Nous avons appliqué le processus de fouille sur ce corpus avec ces deux indexations.

Pour la première expérience, nous avons fixé `minsup` à $0{,}7\%$ (correspondant à un seuil minimum de support de 10 textes pour les règles extraites) car $49\ \%$ des termes apparaissent entre 5 et 15 fois dans les textes. Nous avons donc choisi de prendre la valeur moyenne de 10. `minconf` correspond à 100% (*i.e.* règles totales). Nous avons obtenu 1 202 règles. Les règles étaient trop nombreuses pour être toutes analysées finement. 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, sur des termes filtrés, nous avons fixé `minconf` à 80% et nous avons conservé `minsup` équivalent à 10 textes. Nous avons obtenu 347 règles, dont 128 règles totales. Nous diminuons le nombre de règles extrait de plus d'un tiers. C'est un nombre de règles raisonnablement interprétable, en quelques heures, par l'expert. Plus de 60% des règles représentent la figure 3 cas (c), le plus intéressant comme expliqué en début de section 6.

7.2 Interprétation par l'expert

Nous avons proposé les 347 règles obtenues lors de la seconde expérience à un expert du domaine. Les règles n'ont pas été classées afin de lui laisser une libre appréciation. La confrontation à l'avis de l'expert a montré que les règles qu'il retenait se trouvaient dans les cas de la figure 3 (c) et de la figure 3 (b). Il est important de repérer quelles sont les règles qui lui paraissent « interprétables ». Une règle est *interprétable* si l'expert 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'expert consiste à expliquer pourquoi il est normal, de son point de vue, que tel terme apparaisse avec tel autre.

Analyse par l'expert

Les textes décrivent le phénomène de la mutation des gènes dans les bactéries provoquant une résistance aux antibiotiques. Voici quelques commentaires sur les règles :

Num´ero: **120**
R`egle: "determine region" "gyrA gene" "gyrase" "mutation" $\Longrightarrow$ "Quinolone"
pB: "0.008" *pH*: "0.059" *pBH*: "0.008" *Support*: "11" *Confi ance*: "1.000" *Int´er^et*: "17.012"
Conviction: "ind´efi nie" *D´ependance*: "0.941" *Nouveaut´e*: "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 le nom de la bactérie ($\neq$ pour les 11 textes).

Num´ero: **279**
R`egle: "mutation" "parC gene" "Quinolone" $\Longrightarrow$ "gyrA gene"
pB: "0.015" *pH*: "0.046" *pBH*: "0.014" *Support*: "21" *Confi ance*: "0.952" *Int´er^et*: "20.574"
Conviction: "20.028" *D´ependance*: "0.906" *Nouveaut´e*: "0.014" *Satisfaction*: "0.950"

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".

Num´ero: **202**
R`egle: "grlA gene" $\Longrightarrow$ "mutation" "Staphylococcus Aureus"
pB: "0.009" *pH*: "0.023" *pBH*: "0.008" *Support*: "12" *Confi ance*: "0.917" *Int´er^et*: "40.245"
Conviction: "11.727" *D´ependance*: "0.894" *Nouveaut´e*: "0.008" *Satisfaction*: "0.915"

Num´ero: **270**
R`egle: "mecA" "meticillin" $\Longrightarrow$ "mecA gene" "Staphylococcus Aureus"
pB: "0.009" *pH*: "0.012" *pBH*: "0.009" *Support*: "12" *Confi ance*: "1.000" *Int´er^et*: "80.059"
Conviction: "ind´efi nie" *D´ependance*: "0.988" *Nouveaut´e*: "0.009" *Satisfaction*: "1.000"

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”.

Num´ero: **293**
R`egle: ”mycobacterium tuberculosis” $\Longrightarrow$ ”tuberculosis”
pB: ”0.053” *pH*: ”0.067” *pBH*: ”0.053” *Support*: ”72” *Confi ance*: ”1.000” *Int´er ˆet*: ”14.956”
Conviction: ”ind´efi nie” *D´ependance*: ”0.933” *Nouveaut´e*: ”0.049” *Satisfaction*: ”1.000”

Num´ero: **335**
R`egle: ”restriction enzyme” $\Longrightarrow$ ”enzyme”
pB: ”0.008” *pH*: ”0.112” *pBH*: ”0.008” *Support*: ”11” *Confi ance*: ”1.000” *Int´er ˆet*: ”8.954”
Conviction: ”ind´efi nie” *D´ependance*: ”0.888” *Nouveaut´e*: ”0.007” *Satisfaction*: ”1.000”

Certaines règles ont été jugées inintéressantes. La plupart sont dues à un artefact de 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).

Num´ero: **183**
R`egle: ”epidemic strain” $\Longrightarrow$ ”outbreak”
pB: ”0.012” *pH*: ”0.057” *pBH*: ”0.012” *Support*: ”16” *Confi ance*: ”1.000” *Int´er ˆet*: ”17.449”
Conviction: ”ind´efi nie” *D´ependance*: ”0.943” *Nouveaut´e*: ”0.011” *Satisfaction*: ”1.000”

Num´ero: **2**
R`egle: ”agar dilution” $\Longrightarrow$ ”dilution method”
pB: ”0.019” *pH*: ”0.025” *pBH*: ”0.019” *Support*: ”26” *Confi ance*: ”1.000” *Int´er ˆet*: ”40.029”
Conviction: ”ind´efi nie” *D´ependance*: ”0.975” *Nouveaut´e*: ”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 courante.

7.3 Confrontation des indices 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'expert a globalement réussi à interpréter, par rapport au domaine, les règles que nous lui avons présentées.

L'indice d'*intérêt*, par définition, classe en premier les règles ayant des termes rares en B et en H (de figure 3 cas (c)). On s'attend à ce que l'expert préfère ce genre de règle. L'expérience vérifie 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 cet indice. Ces deux règles ont respectivement comme valeur d'intérêt 80,059 et 40,245. Ces règles sont porteuses d'information du point de vue de l'expert puisqu'il a réussi à les commenter très facilement (voir en début du paragraphe 7.2). Par ailleurs, la règle 159 (”dna” ”gyrA gene” $\Longrightarrow$ ”mutation”) qui illustre la figure 3 cas (b) ainsi que la règle 228 (”Gyrase” ”protein” $\Longrightarrow$ ”mutation”) pour la figure 3 cas (a) sont moins informatives. Leur intérêt et leur conviction sont plus proches de 1 (respectivement 4,929 et 5,086). Le comportement symétrique de l'indice d'intérêt peut se révéler intéressant. Par exemple, la règle 108 ”dalfopristin” $\Longrightarrow$ ”quinupristin” et la règle 332 ”quinupristin” $\Longrightarrow$ ”dalfopristin” ont la même forte valeur d'intérêt (de 75,611). Cet indice a permis de les rapprocher dans le classement. Nous avons mis en valeur des simili-

tudes de comportement de populations de bactéries en résistance à deux antibiotiques (”quinupristine” et dalfopristine). Ce qui est confirmé par l'expert.

En confrontant plusieurs règles à fortes valeurs de *conviction*, nous avons retracé dans les textes une antériorité dans la découverte du gène GyrA par rapport à ParC. Nous avons vérifié sur nos données que cet indice renforce la direction de l'implication de B vers H. Dans notre exemple, l'expert 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 $\ldots\texttt{ParC}\ldots \Longrightarrow \ldots\texttt{GyrA}\ldots$[2] dans des règles de fortes valeurs de *conviction* contribuait 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 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 classera de façon plus proche car toutes les deux illustrent la figure 3 cas (c). 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 $\texttt{P(H)}$, ce qui nous place également dans la figure 3 cas (c). Les règles totales 270, 120 ou à valeur de confiance proche de 1 (*ex.* règle 279) sont celles qui sont les plus dépendantes (car $\texttt{dep[B} \Longrightarrow \texttt{H]} = 1 - \texttt{P(H)}$).

Enfin, les deux règles suivantes illustrent le comportement de la *nouveauté* et de la *satisfaction*. La règle non informative 273 (”meticillin” $\Longrightarrow$ ”staphylococcus Aureus”) correspond à la figure 3 cas (a), alors que celle qui est mieux interprétée 265 (”mecA gene” ”meticillin” $\Longrightarrow$ ”Staphylococcus Aureus”) — à cause de la présence du gène — correspond à la figure 3 cas (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 indices peuvent donc distinguer le cas moins informatif (b) du cas non informatif (a), là où la dépendance ne peut aider à les différencier.

8 Éléments de discussion

En extraction des 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'expert juge, par exemple, que la règle : (”aztreonam” ”clavulanic acid” ”enzyme” $\Longrightarrow$ ”β-lactamase”) est plus interprétable que : (”aztreonam” ”enzyme” $\Longrightarrow$ ”β-lactamase”), qui se trouve être plus générique et couvre plus d'exemples (16 *vs.* 11).

Les deux règles totales 219 (”gyrA gene” ”resistance mechanism” $\Longrightarrow$ ”quinolone”) et 326 (”quinolone” ”resistance mechanism” $\Longrightarrow$ ”gyrA gene”) portent sur exactement les mêmes textes. Comme le mécanisme de résistance porte sur les quinolones, l'expert préfère la seconde règle. Tous les indices discriminants (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 la contradiction à l'interprétation des deux règles : « No changes in the *quinolone*-resistant *determing region*s of parC, parE, *gyrA*, or gyrB were found in this mutant. ». Ce qui montre que

2. Plus, ´eventuellement d'autres termes not´es par (. . .)

la *négation*, si elle n'est pas prise en compte dans le processus de fouille, reste un problème entier et n'est pas encore abordée dans nos travaux.

8.1 Retour sur l'indexation

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 (*i.e.* silence), cela peut entraîner la disparition d'une règle du fait du seuil *minsup*.

Bien que le corpus soit spécialisé (résistance des bactéries aux antibiotiques), nous constatons une assez grande disparité des termes retenus à l'indexation. On retrouve ce phénomène régulièrement en analyse automatique de corpus. Comme nous l'avons souligné en paragraphe (4.2.2, Description des données), un texte fait souvent référence à des termes périphériques au domaine considéré qui introduisent du bruit.

Comme le montre la figure 1, le processus de fouille de texte prévoit une boucle de réutilisation des connaissances extraites par retour à l'étape d'indexation. Certaines règles révèlent des termes qui sont périphériques au domaine et qui n'ont pas été repérés 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. Par conséquent, il est possible, par un processus itératif, de (1) filtrer un terme repéré dans certaines règles en éliminant toutes ses occurrences dans l'indexation des textes ; (2) extraire les règles, et retourner à (1).

8.2 Approches comparables

Les travaux de (Azé & Roche, 2003) se différencient par l'extraction de règles vérifiant une contrainte (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 » 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), on part 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` $>$) et ([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 et un terme comme "air" en complément circonstanciel. Enfin, dans (Feldman *et al.*, 1998), l'exploration des règles se fait par la sélection de celles pour lesquelles les termes dans `B` et `H` sont d'un certain type. Cela permet d'arriver jusqu'à des indices de *confiance* très faibles (de l'ordre de `0.1`). Par exemple, chercher tous les établissements industriels qui ont fait alliance ou qui ont fusionné : "`intuit corp`" "`novell corp`" $\Longrightarrow$ "`merger`".

9 Conclusion

Cet article présente une expérience complète mettant en présence une méthode de traitement automatique de corpus, un processus de fouille de données et qui prend en compte une évaluation des résultats par l'analyste (*i.e.* l'expert). L'ensemble de ces processus constitue 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 induise une subjectivité liée à toute expertise humaine, nous avons confronté de façon opérationnelle la valeur des indices présentés aux besoins de l'expert. Nous avons trouvé qu'une combinaison de *l'intérêt* et de la *conviction* permettent de bien classer les règles qui sont les plus significatives et qui illustrent la figure 3 cas (c). C'est ce cas que nous avons identifié comme étant le plus informatif du point de vue de l'expert. Les indices de *nouveauté* et de *satisfaction* permettent de distinguer le cas (a) du cas (b) pour des règles à faible *dépendance*. Le but de cette méthodologie est de s'assurer de l'apport de mesures qualitatives pour présenter en premier, celles qui sont les plus informatives. La reproduction de l'expérimentation avec un expert différent ou sur un corpus différent nous permettra d'avoir une validation systématique de cette approche.

Références

ANICK P. & PUSTEJOVSKY J. (1990). An Application of lexical Semantics to Knowledge Acquisition from Corpora. In *Proc. of COLING'90*, volume 3, p. 7–12, Helsinki.

AZÉ J. & ROCHE M. (2003). Une application de la fouille de textes : l'extraction des règles d'association à partir d'un corpus sp´ecialis´e. In *In Proc. of Extraction et Gestion des Connaissances (EGC'03)*, volume 17 of *RSTI/RIA-ECA*, p. 283–294, Lyon: Éditions Hermès.

BASTIDE Y., TAOUIL R., PASQUIER N., STUMME G. & LAKHAL L. (2002). Pascal : un algorithme d'extraction des motifs fr´equents. *Tech. et Sci. Informatiques*, **21**(1), 65–95.

BRILL E. (1999). Unsupervised learning of disambiguation rules for part of speech tagging. In *Proc. of Joint SIGDAT ACL'99 Conference on Empirical Methods in NLP and Very Large Corpora (EMNLP/VLC-99)*, College Park, University of Maryland.

DUCLOY J. (1999). DILIB, une plate-forme XML pour la g´en´eration de serveurs WWW et la veille scientifique et technique. In *Micro-Bulletin*. CNRS.

FAURE D., NÉDELLEC C. & ROUVEIROL C. (1998). *Acquisition of Semantic Knowledge using Machine learning methods: The System ASIUM*. Rapport interne ICS-TR-88-16, LRI.

FELDMAN R., FRESKO M., KINAR Y., LINDELL Y., LIPHSTAT O., RAJMAN M., SCHLER Y. & ZAMIR O. (1998). Text mining at the term level. *LNAI: Principles of Data Mining and Knowledge Discovery*, **1510**(1), 65–73.

GRAS R., KUNTZ P., COUTURIER R. & GUILLET F. (2001). Une version entropique de l'intensit´e d'implication pour les corpus volumineux. In H. BRIAND & F. GUILLET, Eds., *Actes EGC'01 : Journ´ees Extraction et Gestion des Connaissances*, volume 1 of *1-2*, p. 69–80, Nantes: Éditions Hermès.

JACQUEMIN C. (1994). FASTR : A Unification-Based Front-End to Automatic Indexing. In *Proc. of Computer-Assisted Information Retrieval (RIAO'94)*, p. 34–47, New-York.

KODRATOFF Y. (1999). Knowledge Discovery in Texts : A definition, and Applications. In *LNAI: Proc. of the 11th Int'l Symp. ISMS'99*, volume 1609, p. 16–29, Warsaw: Springer.

LAVRAČ N., FLACH P. & ZUPAN B. (1999). Rule Evaluation Measures: A Unifying View. In *Proc. of ILP'99: 9th International Workshop on Inductive Logic Programming*, volume 1634 of *LNAI*, p. 174–185, Bled, Slovenia: Springer-Verlag.

PASQUIER N., BASTIDE Y., TAOUIL R. & LAKHAL L. (1999). Efficient mining of association rules using closed itemset lattices. *Information Systems*, **24**(1), 25–46.