Adéquation d'indices statistiques à l'interprétation de règles d'association

Hacène Cherfi, Yannick Toussaint

Équipe ORPAILLEUR (LORIA - INRIA Lorraine)
Campus scientifique - B.P. 239 - Vandœuvre-lès-Nancy F-54506 cedex

Abstract

This paper aims at defining a methodology of access and reading of association rules extracted from texts. The corpus which we used is a set of scientific abstracts in the field of molecular biology. Our research deals with: i) the extraction of association rules from frequent sets generated by the "*Close*" algorithm; ii) the computation of statistical indices for each rule which allows us to order them; iii) the interpretation of the rules by an expert of the domain; iv) the mapping of the items ii) and iii). This article will relate primarily to the last three items. We underline the importance to ease the task of the expert in this interpretation by classifying the rules according to the indices. We insist upon two of these: "*interest*" and "*dependence*" that are meaningful while dealing with the rules known as "*valid*". A discussion about our results identifies some points having an impact on the interpretability of the association rules.

Résumé

Nous proposons, dans cet article, la description d'une méthodologie d'accès et de lecture des règles d'association extraites à partir de textes. Le corpus qui a servi à notre expérience est une collection de textes sous forme de résumés d'articles scientifiques dans le domaine de la biologie moléculaire. Notre recherche porte sur : i) l'extraction des règles d'association à partir de la construction des ensembles fermés fréquents générés par l'algorithme "*Close*" ; ii) l'association d'indices statistiques à chaque règle, ce qui permet de les ordonner ; iii) l'interprétation des règles par un expert du domaine ; iv) la mise en correspondance des points ii) et iii). Cet article portera essentiellement sur les trois derniers points. Nous montrons l'importance d'aider l'expert, grâce aux indices, dans son interprétation des règles. Nous insistons, plus particulièrement, sur deux de ces indices : l'*intérêt* et la *dépendance* pour les règles dites "*totales*" que nous extrayons. Une discussion sur nos résultats identifie quelques points ayant un impact sur l'interprétabilité des règles d'association.

Mots-clés : Règles d'association, fouille de textes, indices statistiques, interprétation, biologie moléculaire.

1. Introduction

Notre travail porte sur la fouille de données dans les textes. La fouille de texte s'adresse à un utilisateur qui, dans notre cas, est expert d'un domaine particulier. Elle donne à cet expert une vue synthétique du contenu d'une collection d'un ou plusieurs milliers de textes, 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 et de tendance entre les textes. L'objectif de la fouille est donc de permettre à l'expert de retrouver, à travers la collection de textes, des relations connues dans le domaine, de pouvoir les localiser rapidement dans les documents, d'observer des familles de documents construites à partir d'une ou plusieurs de ces relations. Plus rarement, elle permet aussi de découvrir des relations non encore connues. Au delà du processus de fouille, nous insistons dans cet article sur l'étude de l'adéquation des résultats de ce processus aux besoins de l'expert.

Nous recherchons l'expression de ces relations par le biais des règles d'association extraites à partir des textes. Notre objectif est alors double :

(i) il s'agit d'identifier des règles d'association qui présentent un intérêt pour l'expert. Nous définissons à ce titre la notion de règle *interprétable*.

(ii) nous cherchons à savoir s'il existe des indices formels associés aux règles d'association qui refléteraient la classification de l'expert de la règle la plus interprétable à la moins interprétable.

Le premier impératif de la fouille de texte est donc de partir d'une collection de textes. Dans la section 2, nous décrivons les caractéristiques que doit avoir un texte en entrée de notre processus de fouille, les caractéristiques de cette collection et nous définissons la représentation des textes telle qu'elle est fournie au processus de fouille.

Dans un second temps, la section 3 concerne le processus de fouille. Nous optons pour le paradigme de la représentation symbolique pour extraire des règles d'association. De ce point de vue, nous nous situons dans la lignée des travaux de (Simon, 2000) et (Kodratoff, 1999) qui se sont intéressés à cette même problématique mais sur des données de nature différente. Nous définissons, au paragraphe 3.1, les règles d'association qui constituent des indicateurs précieux pour la veille technologique et l'analyse de l'information. Du fait de leur grand nombre, nous cherchons à les ordonner. Nous associons à ces règles des indices formels, que nous définissons au paragraphe 3.2, destinés à classer les règles entre elles.

Nous disposons alors de toutes les informations permettant d'aborder le point (i) de notre objectif en sections 4 et 5. Nous y définissons la notion de règle interprétable et demandons à l'expert d'évaluer chacune des règles par rapport à ce critère. Nous avons défini un environnement de navigation dans l'ensemble des règles. Nous évaluons (section 6) l'adéquation entre le processus de fouille et les besoins de l'expert par le biais du point (ii) de notre objectif. Nous étudions donc les valeurs des différents indices des règles pour chercher si certains reflètent l'intérêt que l'expert porte à certaines règles (Cherfi and Toussaint, 2001). La confrontation des résultats formels (calcul des règles d'association, calculs des indices) à la réalité du domaine (l'appréciation de l'expert) est, en ce sens, inédite.

2. Description des données

La fouille de textes commence par la sélection des textes et la représentation de leur « contenu ». La représentation d'un texte doit être indépendante de sa syntaxe et refléter majoritairement sa sémantique. Il est indispensable de pouvoir relier, entre elles, les notions citées dans le(s) texte(s). Cette représentation repose donc sur un réseau terminologique et sur la liste des ***termes*** extraits à partir des textes.

Définition 1 (Terme) *Un terme 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. Les termes font partie d'une terminologie qui est un réseau sémantique.*

Lorsqu'on veut caractériser un texte, rendre compte de son contenu, l'indexation par les termes est plus appropriée que l'indexation par des mots simples uniquement. Comme le soulignent (Faraj et al., 1996) : « *Les termes composés permettent généralement de limiter l'ambiguïté et*

d'augmenter la précision » grâce au repérage, dans les textes, de notions mieux dénommées ainsi qu'au réseau sémantique constitué par la terminologie. Quelles sont les caractéristiques pour choisir une collection de textes candidats à la fouille ?

- L'ensemble des textes doit refléter un contenu cohérent ou homogène dans un domaine de spécialité. Les textes peuvent décrire des expériences faites sur un sujet particulier, dans un but précis ou montrer des méthodologies pour résoudre un problème cible. Le cadrage du sujet permet d'avoir une terminologie délimitée.
- Chaque texte doit être caractérisé par une forte *densité* de termes. Plus il y a de termes dans un texte, plus le réseau sémantique reflétant le contenu du texte sera complet. Ainsi préférera-t-on le résumé d'un article scientifique à une thèse.

Ce sont ces deux principaux critères qui font de notre collection de textes un ***corpus***. Pour augmenter le nombre de termes extraits des textes, nous collectons non seulement le terme *préférentiel* (*i.e.* celui qui est décrit dans la terminologie) mais également toutes ses formes variantes. Par exemple, on voudrait que le terme "*transfer of capsular biosynthesis genes*" indexe le texte par son terme préférentiel "*gene transfer*". Pour faire cette indexation, nous avons opté pour l'outil FASTR (Jacquemin, 1994). C'est un analyseur syntaxique fondé sur les grammaires d'unification (Shieber, 1986) et, plus précisément, sur la forme logique des Grammaires d'Arbres Adjoints (Vijay-Shankar, 1992). FASTR recherche, dans des séquences textuelles acceptables, le maximum de termes qui s'y trouvent par identification de ces termes à partir d'une *liste contrôlée* (appelée nomenclature terminologique). Les formes variantes de termes reconnues dans les textes sont ramenées à leur terme préférentiel.

Notre corpus est constitué de 1 407 documents d'environ 200 000 mots, soit environ 6 Mø.

Document 000391
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
Texte : 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.
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. 1 – *Exemple d'un document du corpus (version du texte raccourcie)*

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 de la mutation de gènes provoquant une résistance aux antibiotiques. La figure (FIG.1) donne l'exemple du document n° 000391 de notre corpus.

3. Processus de fouille de textes

Définition 2 (Fouille de Textes) *Le processus de fouille est fondé sur l'utilisation de méthodes symboliques. C'est la combinaison :*

(a) d'une méthode formelle d'extraction des règles d'association ;
(b) d'un classement des règles suivant des indices statistiques ;
(c) d'un mécanisme interactif d'accès aux règles et au contenu des documents.

L'extraction des règles d'association (a) se fait, grâce à l'analyse formelle de concepts (AFC) (Ganter and Wille, 2000), par la construction des « *ensembles fermés fréquents* » générés par l'algorithme *Close* (Pasquier et al., 1999). Les concepts formels obtenus permettent le calcul de règles d'association. Les indices statistiques calculés en (b) sont des mesures de pondération affectés aux règles. Ces indices donnent un poids à chaque règle et permettent alors de "classer" ces règles. Tous les indices introduits (voir section 3.2) ne s'avéreront pas utilisables pour notre processus de fouille de textes. L'environnement de navigation hypertextuelle (c) aide l'expert du domaine à interpréter les règles d'association obtenues en (a). Il lui permet d'accéder au contenu des documents (*cf.* FIG.1) liés à une règle.

3.1. Règles d'association

Les règles d'association ont été, initialement, utilisées 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 (Simon and Napoli, 1999). Elles ont été, par la suite, appliquées à la fouille de textes [(Feldman and Dagan, 1995),(Toussaint et al., 2000)].

Définition 3 (Règle d'association) *Une règle d'association [(Guigues and Duquenne, 1986), (Luxenburger, 1991), (Agrawal and Srikant, 1994)] est du type :*

$$R \ : \ t_1 \wedge \ t_2 \Longrightarrow t_3 \wedge \ t_4 \wedge \ t_5 \quad (t_1, \ ..., \ t_5 \ sont \ des \ termes) \tag{1}$$

Une règle est constituée d'une conjonction de termes en partie gauche (que nous appelons B*) impliquant une conjonction de termes en partie droite (appelée* H*). La règle sera donc notée* $\mathtt{R} \ : \ \mathtt{B} \Longrightarrow \mathtt{H}$.

L'explication intuitive de la règle (1) est que : si les documents possèdent les termes $\{\mathtt{t}_1, \mathtt{t}_2\}$ alors ils possèdent également $\{\mathtt{t}_3, \mathtt{t}_4, \mathtt{t}_5\}$. À l'origine, deux indices ont été associés aux règles : le *support* et la *confiance*.

Définition 4 (Support) *Le support d'une règle d'association est l'ensemble des documents qui participent à sa génération. Le support représente le nombre de documents qui sont décrits par les termes présents en partie gauche et droite de la règle (on dira par la suite : le nombre de documents qui vérifient* B *et* H*). Pour la règle (1) le support est :*

$$sup \ [B \Longrightarrow H] = nombre \ de \ documents \ vérifiant \ \{t_1, t_2, t_3, t_4, t_5\} \tag{2}$$

Le support peut également être exprimé relativement au nombre total de documents du corpus. C'est la probabilité d'apparition de l'ensemble des documents correspondant à $\mathtt{B} \wedge \mathtt{H}$ et que nous noterons par la suite $\mathtt{P(B \ , \ H)}$:

$$P(B, \ H) = \frac{sup \ [B \Longrightarrow H]}{nombre \ total \ de \ documents \ du \ corpus} \in [0, \ 1] \tag{3}$$

Définition 5 (Confiance) *La confiance mesure le degré de validité d'une règle, c'est-à-dire lorsqu'il existe des contre-exemples de documents qui vérifient* B *mais pas nécessairement tous les termes de* H*. Pour la règle (1), la confiance vaut :*

$$conf \ [B \Longrightarrow H] = \frac{nombre \ de \ documents \ vérifiant \{t_1, t_2, t_3, t_4, t_5\}}{nombre \ de \ documents \ vérifiant \{t_1, t_2\}} \in [0, \ 1] \tag{4}$$

En termes probabilistes, la confiance mesure la probabilité conditionnelle de H sachant B :

$$conf \ [B \Longrightarrow H] = P(H \mid B) = \frac{sup \ [B \Longrightarrow H]}{nombre \ de \ documents \ vérifiant \{t_1, t_2\}} \in [0, \ 1] \tag{5}$$

Lorsque la confiance vaut 1, la règle est dite ***totale***. Dans le cas contraire, la règle est dite ***partielle*** à *x* %.

3.2. Indices statistiques associés aux règles d'association

Le support et la confiance ne permettent pas, à eux seuls, d'indiquer la « qualité » d'une règle. Nous introduisons d'autres indices statistiques qui apportent des informations supplémentaires et permettent différents classements des règles.

Lorsque B et H sont indépendants (*i.e.* $\mathtt{P}(\mathtt{B},\mathtt{H}) = \mathtt{P}(\mathtt{B}) \times \mathtt{P}(\mathtt{H})$), la confiance (5) vaut $\mathtt{P}(\mathtt{H})$. Une règle d'association générée entre deux ensembles de termes indépendants n'apporte pas d'information autre que le fait que ces termes sont mis ensemble dans beaucoup de textes. L'indice de dépendance est classiquement utilisé en probabilités, il permet de calculer l'apport de B dans la règle.

Définition 6 (Dépendance) *L'indice de dépendance renforce une règle en mesurant le fait que* B *et* H *soient dépendants ou pas :*

$$dep\,[B \implies H] = \mid P(H \mid B) - P(H) \mid \tag{6}$$

Les termes très fréquents dans le corpus n'apportent pas d'information « particulière » puisque tout terme du corpus impliquera un terme fréquent. Alors que les termes rares, possiblement porteurs d'information, apparaissent dans des règles à faible *support* auxquelles on ne s'intéresse pas en premier lieu. Cette différence d'apparition des termes a conduit à la définition de l'indice suivant :

Définition 7 (Intérêt) *L'intérêt mesure la dépendance entre* B *et* H*. Cet indice privilégie les termes rares aux dépens de termes trop répandus dans le corpus.*

$$int\,[B \implies H] = \frac{P(B,H)}{P(B) \times P(H)} \tag{7}$$

L'*intérêt* a un comportement *symétrique* pour B et pour H, c'est-à-dire que : $\mathtt{int}\,[\mathtt{B} \Longrightarrow \mathtt{H}]$ est égal à $\mathtt{int}\,[\mathtt{H} \Longrightarrow \mathtt{B}]$. Il ne reflète pas, par définition, l'implication $\mathtt{B} \Longrightarrow \mathtt{H}$. (Brin et al., 1997) proposent l'indice suivant :

Définition 8 (Conviction) *La conviction mesure également la dépendance mais pour les contre-exemples* $\mathtt{B} \wedge \neg\mathtt{H}$[1].

$$conv\,[B \implies H] = \frac{P(B) \times P(\neg H)}{P(B,\neg H)} \tag{8}$$

Cet indice n'est applicable qu'aux règles *partielles* car $\mathtt{P}(\mathtt{B},\ \neg\mathtt{H})$ vaut 0 et $\mathtt{conv}\,[\mathtt{B} \Longrightarrow \mathtt{H}]$ n'est pas une valeur calculable.

Définition 9 (Étonnement)

$$spr\,[B \implies H] = \frac{(P(B,\ H) - P(B,\ \neg H))}{P(H)} \tag{9}$$

Cet indice est défini pour mesurer l'*affirmation* : différence entre la *confirmation* $\mathtt{P}(\mathtt{B},\ \mathtt{H})$ et l'*infirmation* $\mathtt{P}(\mathtt{B},\ \neg\mathtt{H})$ d'une règle. Cet indice est présenté dans (Kodratoff and Azé, 2002) et permet de rechercher les règles dites « étonnantes ». Moins H est répandu, plus il est étonnant de trouver une bonne affirmation de la règle. Les auteurs remarquent que les règles dont l'indice est supérieur à un certain seuil sont insensibles au bruit, c'est-à-dire aux données non désirées liées à des biais dans l'indexation dans notre cas. Comme (8), cet indice n'est valable que pour les règles *partielles*, toujours à cause de la présence de $\mathtt{P}(\mathtt{B},\ \neg\mathtt{H})$.

1. $(\mathtt{B} \Longrightarrow \mathtt{H})$ est logiquement équivalent à $\neg(\mathtt{B} \wedge \neg\mathtt{H})$

4. Expérimentations

4.1. Description des expériences

Deux expériences furent menées avec le corpus sur la biologie moléculaire :

- une première expérience avec une indexation automatique par FASTR. L'ensemble des documents a été indexé par un total de $22\,885$ termes qui correspondent à $3\,337$ termes différents, avec une moyenne de $16,26$ termes par document. Parmi ces termes, $1\,762$ (soit $52,8\,\%$) étaient des termes n'apparaissant qu'une seule fois en index (*i.e.* des termes *hapax*) et beaucoup de bruit lié au découpage des termes en sous-termes par FASTR. Cette distribution des termes dans le corpus, que nous considérons comme un « éparpillement », est un biais bien connu en analyse de l'information textuelle. Il est dû, notamment, au bruit lié aux termes, périphériques du domaine, utilisés par les auteurs du texte ;
- une seconde expérience eut lieu avec des termes filtrés à la main par les documentalistes de l'INIST[2]. Ce filtrage manuel permet d'éliminer une grande partie du bruit. Il résulte que l'ensemble des documents a été indexé par un total de $14\,374$ termes pour $1\,361$ documents (le reste des documents n'ayant plus de termes d'indexation). Ces termes correspondent à 632 termes différents (soit $18,94\,\%$ du nombre de termes différents de la $1^{\text{ère}}$ expérience), avec une moyenne de $10,56$ termes par document. À noter qu'il n'y a pas de termes apparaissant moins de 5 fois dans l'indexation et $49\,\%$ des termes apparaissent entre 5 et 15 fois.

4.2. Résultats obtenus

Pour notre première expérience, les règles *totales* générées de support minimal 10 étaient au nombre de $1\,202$ dont 713 étaient des règles avec un support $\in [10, 15]$ (soit $59,31\,\%$ des règles obtenues). Lorsqu'on fixe le support à 1, nous obtenons plus de $460\,000$ règles. Cette première expérience nous a permis de tester la robustesse de nos calculs sur un corpus de taille moyenne. Les règles obtenues n'étaient pas intelligibles et beaucoup trop nombreuses. 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 $2^{\text{ème}}$ expérience, les règles *totales* générées de support minimal 10 étaient au nombre de 128. Lorsqu'on ramène le support à 1, nous obtenons $163\,175$ règles.

5. Interprétation par l'expert

Nous avons soumis les 128 règles obtenues lors de la seconde expérience à un expert documentaliste de l'INIST. Les règles n'ont pas été classées pour laisser à l'expert une libre appréciation. Il est important, pour nous, de repérer quelles règles lui paraissent "interprétables". Puis nous avons confronté ces règles aux indices calculés pour chacune d'elles. Les règles ont été, dans l'ensemble, bien interprétées. Ceci nous amène à une première définition de l'interprétabilité.

Définition 10 (Règle interprétable) *Une règle est* ***interprétable*** *si l'expert peut relier tous les termes apparaissant dans* `B` *et* `H`*. Le travail de l'expert consiste à expliquer pourquoi il est normal, de son point de vue, que tel terme apparaisse avec tel autre. Cette explication va souvent refléter ce qui est exprimé par des liens sémantiques dans le réseau terminologique. La règle*

2. INstitut de l'Information Scientifique et Technique, établissement qui nous a également fourni le corpus

*est **moins interprétable** lorsque l'expert ne voit pas de relation sémantique entre le termes dans son domaine.*

Avant de décrire les règles, nous rappelons quelques notions utiles à la compréhension des règles à interpréter. 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. Certains antibiotiques permettent d'inhiber la synthèse protéique en empêchant, par exemple, la division cellulaire de la bactérie ou alors sa production d'enzymes. Cette bactérie sera donc neutralisée. Au sein du génome de la bactérie, certaines mutations peuvent provoquer une résistance aux antibiotiques qui ne pourront, entre autres, plus se fixer sur la bactérie. C'est un des schémas du phénomène de résistance des bactéries aux antibiotiques. Ce qui suit est une sélection de certaines règles interprétées par l'expert.

5.1. Les meilleures règles interprétées

Nous commençons par présenter la règle qui reflète, le plus, une description du domaine d'activité (*i.e.* du phénomène de résistance aux antibiotiques).

Numéro :000045
Règle : "determine region" "gyra gene" "gyrase" "mutation" $\Longrightarrow$ "quinolone"
Support : "11" *Intérêt* : "17.012" *Dépendance* : "0.941"

Cette règle indique que les 11 documents cités décrivent la mutation du gène "gyrA" qui correspond à l'enzyme "gyrase" dans un fragment ou 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, qui n'est pas le même pour les 11 documents (pour le document 000391 (*cf.* FIG.1), il s'agit de "Chlamydia trachomatis" alors que pour le document 000491 c'est "Pseudomonas aeruginosa", etc.).

Numéro :000114
Règle : "parc gene" "sequence" $\Longrightarrow$ "gyra gene"
Support : "11" *Intérêt* : "21.603" *Dépendance* : "0.954"

Cette règle fait ressortir le fait que le gène "parC" a été découvert plus récemment que le gène "gyrA". Ces deux gènes sont liés dans leurs mutations (par mutation combinée). Chaque fois qu'on parle de "parC", les auteurs font référence aussi à "gyrA".

Numéro :000011
Règle : "bla gene" "escherichia coli" $\Longrightarrow$ "β-Lactamase"
Support : "12" *Intérêt* : "10.007" *Dépendance* : "0.900"

La bactérie "Escherichia Coli" possède le gène "bla" donc fabrique l'enzyme β-Lactamase et sera résistante à la famille des β-lactams. Enfin, la règle suivante nous informe que grâce à l'utilisation de la "Meticillin", on inhibe le gène "mecA" et on combat la "Staphylococcus Aureus".

Numéro :000095
Règle : "meca" "meticillin" $\Longrightarrow$ "meca gene" "staphylococcus aureus"
Support : "12" *Intérêt* : "80.059" *Dépendance* : "0.988"

Cette bactérie, chez l'homme, est à l'origine d'un problème de santé publique grave puisqu'elle est responsable de milliers de morts dans le monde.

5.2. Les règles indésirables

5.2.1. Bruit lié à l'indexation

L'analyseur FASTR, dans son processus d'extraction de termes, procède par reconnaissance de termes les plus longs puis par découpage en sous-termes. Deux règles illustrent ce biais :

Numéro : 000108
Règle : "mycobacterium tuberculosis" ⟹ "tuberculosis"
Support : "72" *Intérêt* : "14.956" *Dépendance* : "0.933"

Dans cette règle à fort support de 72 documents, on voit que le re-découpage de "Mycobacterium Tuberculosis" en "Tuberculosis" est généré par FASTR car le terme "Tuberculosis", dans cette règle, n'empêche pas son interprétation. La "Tuberculosis" (*tuberculose*) reste cohérente avec la bactérie "Mycobacterium Tuberculosis" qui la provoque. Les choses sont différentes pour la règle suivante :

Numéro : 000087
Règle : "infection" "urinary infection" ⟹ "urinary tract"
Support : "10" *Intérêt* : "104.692" *Dépendance* : "0.990"

L'unique terme d'index exact serait : "urinary tract infection".

5.2.2. Bruit lié à la synonymie

Malgré l'utilisation d'une nomenclature terminologique par FASTR (*cf.* section 2), tous les synonymes n'ont pas été ramenés au terme préférentiel de la nomenclature du fait du manque d'exhaustivité de cette nomenclature. Les deux règles suivantes en sont une illustration :

Numéro : 000073
Règle : "epidemic strain" ⟹ "outbreak"
Support : "16" *Intérêt* : "17.449" *Dépendance* : "0.943"

Numéro : 000127
Règle : "topoisomerase" ⟹ "gyrase"
Support : "15" *Intérêt* : "30.932" *Dépendance* : "0.968"

Les termes en partie gauche et droite sont des synonymes. Le fort support de ces règles par rapport aux précédentes souligne le fait que les biologistes manipulent dans leurs textes indifféremment un terme ou son synonyme. C'est particulièrement gênant pour des analyses automatiques de texte ou pour la fouille.

5.2.3. Bruit lié aux énumérations

Certains noms de gènes sont systématiquement associés à d'autres noms. La raison n'est pas liée au fait que l'article s'intéresse à tous les gènes cités dans le texte. Il peut s'agir seulement de situer un gène par rapport à des gènes analogues ou des sous-unités de gènes qui ont un comportement similaire.

Numéro : 000048
Règle : "determine region" "gyra gene" "pare gene" ⟹ "parc gene" "quinolone"
Support : "10" *Intérêt* : "45.367" *Dépendance* : "0.978"

Certains textes parmi les 10 ont étudié la sous-unité "E" du gène "par" et citent les autres noms (*parC gene*, *gyrA gene*) pour situer ou définir le gène "parE".

5.2.4. Choix entre deux règles proches

Alors qu'en analyse formelle de concepts, nous recherchons des règles décrivant des concepts « génériques » ayant un fort *support*, nous avons été étonnés de voir qu'entre les deux règles suivantes :

Numéro : 000006
Règle : "aztreonam" "enzyme" $\Longrightarrow$ "β-lactamase"
Support : "16" *Intérêt* : "10.007" *Dépendance* : "0.900"

Numéro : 000005
Règle : "aztreonam" "clavulanic acid" "enzyme"$\Longrightarrow$ "β-lactamase"
Support : "11" *Intérêt* : "10.007" *Dépendance* : "0.900"

l'expert a préféré la seconde règle car le nom de l'acide aminé "Clavunalic", qui inhibe l'enzyme β-Lactamase, y est cité. Pourtant, la première règle possède un *support* plus fort.

6. Bilan de l'expérimentation

Dans cette section, nous allons mettre en évidence des liens entre les règles extraites et les différents indices des paragraphes 3.1 et 3.2. Ensuite, nous ferons quelques remarques sur cette confrontation entre les indices et les règles interprétées par l'expert.

6.1. Adéquation entre les règles présentées et les indices

Nous nous sommes intéressés aux règles listées en section 5, que l'expert a interprétées, pour essayer de les confronter aux valeurs des indices. Tout d'abord, il faut noter que les règles présentées à l'expert étaient des règles *totales*. Elles ont donc toutes une *confiance* de 100 %. Les règles présentées, ont majoritairement des *supports* assez proches compris entre 10 et 12. C'est d'ailleurs une caractéristique globale de toutes les règles extraites sauf celles dues au bruit lié à la synonymie (*cf.* paragraphe 5.2.2). En général, nous constatons qu'un fort support ne signifie pas nécessairement un intérêt de l'expert pour la règle (*cf.* paragraphe 4.1, 1ère expérience). En ce qui concerne l'*intérêt*, on vérifie bien qu'il caractérise une différence de représentativité des termes en B et H. La règle 000087 : ("infection" "urinary infection" $\Longrightarrow$ "urinary tract"), d'*intérêt* : 104.692, possède le terme "infection" qui domine par sa fréquence dans cette règle. Il apparaît 180 fois dans le corpus, contre seulement 13 fois pour "urinary tract" et 12 fois pour "urinary infection". Pour la règle 000095 : ("meca" "meticillin" $\Longrightarrow$ "meca gene" "staphylococcus aureus"), d'*intérêt* : 80.059, le déséquilibre est plutôt constaté du côté de H : le terme "Staphylococcus Aureus" est présent 180 fois dans le corpus alors que les termes "mecA gene" et "meticillin" apparaissent beaucoup moins: 18 et 52 fois respectivement. Les indices d'*intérêt* de ces règles sont les valeurs les plus grandes pour l'ensemble des règles. En ce qui concerne la *dépendance*, nous avons constaté que les règles les plus dépendantes étaient celles relatives au biais dû à l'indexation dans les règles : 000087 citée plus haut et 0000120 : ("polymyxin b" $\Longrightarrow$ "polymyxin") avec toutes deux 99 %. Les règles découlant du biais introduit par la synonymie ont également un fort indice de *dépendance*, la règle 000127 : ("topoisomerase" $\Longrightarrow$ "gyrase") a un indice de 96 % et la règle 000073 : ("epidemic strain" $\Longrightarrow$ "outbreak") a un indice de 94 %.

6.2. Éléments de discussion

Nous donnons ici quelques éléments de réflexion qui ont suivi notre expérimentation. Notre approche s'appuie sur une description booléenne (présence *vs.* absence) des termes dans les documents. Cette représentation ne prend pas en compte la fréquence d'apparition des termes

à l'intérieur du document et/ou plus globalement dans l'ensemble des documents. L'approche est différente en Recherche d'Information où l'on associe à un terme une *pondération*. Cette pondération est fonction de la fréquence par rapport aux autres termes dans le document et de la fréquence dans la globalité du corpus. Cela permet de faire un classement entre documents contenant les mêmes termes en réponse à une requête de l'utilisateur. Notre méthode paraît en ce sens plus sensible à la phase d'indexation. Si un terme est absent de l'indexation (*i.e.* silence), cela peut entraîner la disparition d'une règle du fait des seuils de *support* et de *confiance* choisis. 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'avions souligné en paragraphe 4.1, un texte intégral, du fait de sa structure informelle en langage naturel, fait souvent référence à divers termes périphériques au domaine considéré qui introduisent du bruit. Enfin, nous remarquons que l'*implication* dans une règle ne porte pas d'information particulière pour le jugement de sa qualité par l'expert. Peut-être que le réflexe d'indexation lié au statut de documentaliste de notre expert fait qu'il voit les règles comme une liste de termes. Si nous lui présentions des règles ayant un minimum de termes en `B` et un maximum en `H`, peut-être que l'expert verrait un sens à l'*implication*. Nous n'avons pas trouvé ce genre de règle, mais en prenant un *support* plus faible de `5`, nous obtenons la règle suivante :

Règle : "Sequence" "parE gene" $\Longrightarrow$ "Quinolone" "Resistance" "determine region" "gyrA gene" "parC gene" "parE gene"

Cela est à rapprocher du phénomène de disparité des termes souligné précédemment. Un autre point de discussion concerne l'utilité, voire la nécessité, d'une navigation hypertextuelle dans les règles. Cette navigation correspond au besoin de l'expert de pouvoir accéder rapidement au contenu des textes auxquels font référence les documents associés à une règle à interpréter. Cela dénote, parfois, l'ambiguïté des termes en `B` et en `H`. Les règles, les valeurs d'indices (*cf.* section 5) et les documents associés sont visualisables grâce à un navigateur Web.

7. Approches comparables

Certains travaux se sont intéressés à l'extraction des règles d'association par une structuration préalable des données dans un espace de généralisation (Bournaud and Courtine, 2001), d'autres comme (Simon, 2000) retrouve les règles à partir de la construction explicite du système classificatoire - ici, il s'agit d'un treillis de concepts - et de l'organisation des relations d'héritage entre les concepts du treillis. Tout concept formel (ou nœud) du treillis permet de générer une règle de la forme :

$$t_i \Longrightarrow TH \tag{10}$$

où $\mathtt{t_i} \in \mathtt{TP} = \{\mathtt{t_1}, ..., \mathtt{t_p}\}$ constitue l'ensemble des termes *propres* d'un concept et `TH` constitue l'ensemble des termes *hérités* (*i.e.* les termes appartenant aux concepts « pères » du nœud courant). Mais ces méthodes demeurent liées à la gestion, très coûteuse en espace mémoire et en temps de calcul, d'une structure de données en amont (espace de généralisation ou treillis de concepts) que nous n'exploitons pas. 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 :

$$[Secher]\ COD\ < aliment > \tag{11}$$

$$[Secher]\ CC\ < air > \tag{12}$$

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. (Suzuki and 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 $\neg \mathtt{H} \Longrightarrow \neg \mathtt{B}$. 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 tous les établissements industriels qui ont fait alliance ou qui ont fusionné : $\mathtt{"intuit\ corp""novell\ corp"} \Longrightarrow \mathtt{"merger"}$.

8. Conclusion et perspectives

L'extraction de règles d'association est souvent exploitée dans le processus de fouille de données et de textes. Cependant, l'interprétation de ces règles et l'évaluation de leur qualité aussi bien par des indices statistiques que par des experts du domaine restent difficiles à maîtriser. Le nombre de règles extrait ne permet pas une vue globale, synthétique et une exploitation « efficace » des régularités et d'éventuelles "connaissances" qui émergent d'un grand corpus de textes. Nous avons combiné l'utilisation d'indices statistiques avec une approche symbolique. L'objectif étant de sélectionner les règles à analyser en vue de les présenter à un expert du domaine pour leur validation. Cette sélection permet de réduire le nombre de règles à analyser. Nous nous sommes intéressés à l'extraction des règles dites *totales*. Nous avons trouvé que deux de ces indices : l'*intérêt* et la *dépendance* correspondent à des règles que l'expert a jugé *interprétables*. Les autres indices cités dans l'article ne sont applicables et/ou significatifs que pour les règles dites *partielles*. La réalisation d'une interface Web pour la navigation entre règles, indices et documents correspondants constitua une autre aide pour l'expert.

Il est nécessaire de poursuivre ce travail, par la génération des règles *partielles* pour ce même corpus afin de tester l'apport des deux autres indices : *conviction* et *étonnement*. La réduction du bruit lié à la synonymie peut être fait en complétant la nomenclature utilisée durant l'indexation automatique grâce avec un thésaurus plus riche, comme par exemple : l'UMLS (UMLS, 2000).

Références

Agrawal R. and Srikant R. (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94)*, pages 478–499, Santiago, Chile. Extended version: IBM Research Report RJ 9839.

Bournaud I. and Courtine M. (2001). Un Espace de Généralisation pour l'Extraction de Règles d'Association. In Briand H. and Guillet F. editors, *Actes EGC'01 : Journées Extraction et Gestion des Connaissances*, volume 1 of *1-2*, pages 129–140, Nantes, France. Éditions Hermès.

Brin S., Motwani R., Ullman J., and Tsur S. (1997). Dynamic Itemset Counting and Implication Rules for Market Basket Data. In *Proceedings of the ACM SIGMOD'97 Conference on Management of Data*, volume 36, Tucson, USA.

Cherfi H. and Toussaint Y. (2001). Extraction et Interprétation des Règles d'association pour la Fouille de Textes. In *Actes de l'Atelier A3CTE-01 : Applications, Apprentissage, Acquisition des connaissances à partir de textes électroniques*, pages 15–16, Grenoble. Plate-forme AFIA. Résumé (Version courte).

Faraj N., Godin R., Missaoui R., David S., and Plante P. (1996). Analyse d'une méthode d'indexation automatique basée sur une analyse syntaxique de texte. *Canadian Journal of Information and Library Science / Revue l'information et la bibliothéconomie*, 21(1):1–21.

Faure D., Nédellec C., and Rouveirol C. (1998). Acquisition of Semantic Knowledge using Machine learning methods: The System ASIUM. Technical Report ICS-TR-88-16, LRI Université Paris-Sud.

Feldman R. and Dagan I. (1995). Knowledge Discovery in Textual Databases (KDT). In Fayyad U. M. and Uthurusamy R. editors, *Proceedings of the 1st International Conference on Data Mining and Knowledge Discovery*, Montreal, CA. AAI Press.

Feldman R., Fresko M., Kinar Y., Lindell Y., Liphstat O., Rajman M., Schler Y., and Zamir O. (1998). Text mining at the term level. *Lecture Notes in Artificial Intelligence: Principles of Data Mining and Knowledge Discovery*, 1510(1):65–73.

Ganter B. and Wille R. (2000). *Formal Concept Analysis: Mathematical Foundations*. Springer-Verlag, New York.

Gras R., Kuntz P., Couturier R., and Guillet F. (2001). Une version entropique de l'intensité d'implication pour les corpus volumineux. In Briand H. and Guillet F. editors, *Actes EGC'01 : Journées Extraction et Gestion des Connaissances*, volume 1 of *1-2*, pages 69–80, Nantes, France. Éditions Hermès.

Guigues J. and Duquenne V. (1986). Familles minimales d'implication informatives résultant d'un tableau de données binaires. *Mathématiques, Informatique et Sciences Humaines*, 95:5–18.

Jacquemin C. (1994). FASTR : A Unification-Based Front-End to Automatic Indexing. In *Proceedings of Information Multimedia Information Retrieval Systems and Management*, pages 34–47, New-York. Rockfeller University.

Kodratoff Y. (1999). Knowledge Discovery in Texts : A definition, and Applications. In Ras Z. W. and Skowron A. editors, *Foundations of Intelligent Systems*, volume 1609 of *Lecture Notes in Artificial Intelligence*, pages 16–29, Warsaw, Poland. 11th International Symposium, ISMS'99, Springer.

Kodratoff Y. and Azé J. (2002). Rating the Interest of Rules Induced from Data and from Texts. À paraître.

Luxenburger M. (1991). Implications partielles dans un contexte. *Mathématiques, Informatique et Sciences Humaines*, 29(113):35–55.

Pasquier N., Bastide Y., Taouil R., and Lakhal L. (1999). Efficient mining of association rules using closed itemset lattices. *Information Systems*, 24(1):25–46.

Shieber S. M. (1986). *An Introduction to Unification-Based Approaches to Grammar*. Center for the Study of Language and Information, Stanford University, Stanford, CA.

Simon A. (2000). *Outils classificatoires par objets pour l'extraction de connaissances dans les bases de donnés*. PhD thesis, Université Henri Poincaré - Nancy 1, Nancy, France.

Simon A. and Napoli A. (1999). Building Viewpoints in an Object-Based Representation System for Knowledge Discovery in Databases. In Rubin S. editor, *1st International Conference on Information Reuse and Integration, IRI-99*, pages 104–108. International Society for Computers and their Applications, ISCA.

Suzuki E. and Kodratoff Y. (1998). Discovery of Surprising Exception Rules based on Intensity of Implication. In *Proc. of the 2nd Eur. Symp. on Principles of Data Mining and Knowledge Discovery PKDD'98*, pages 10–18, Nantes, France.

Toussaint Y., Simon A., and Cherfi H. (2000). Apport de la fouille de données textuelles pour l'analyse de l'information. In *Actes de la conférence IC'2000, Ingénierie des Connaissances*, pages 335–344, Toulouse, France.

UMLS (2000). The Unified Medical Language System. 11th edition, National Library of Medicine.

Vijay-Shankar K. (1992). Using descriptions of trees in a tree-adjoining grammar. *Computational Linguistics*, 18:481–518.