Fouille de textes par combinaison de règles d'association et d'indices statistiques

Hacène Cherfi, Yannick Toussaint

Fouille de textes par combinaison de règles d'association et d'indices statistiques

Hacène Cherfi — Yannick Toussaint

Équipe ORPAILLEUR (LORIA - INRIA Lorraine)
Campus scientifique - B.P. 239 - Vandœuvre-lès-Nancy F-54506 cedex
{cherfi, yannick}@loria.fr

RÉSUMÉ. *Nous proposons la description d'une méthodologie d'accès et de lecture des règles d'association extraites à partir de textes. Le corpus ayant servi à notre expérimentation est constitué de résumés d'articles scientifiques dans le domaine de la biologie moléculaire. Ce processus génère un trop grand nombre de règles et nous amène à chercher à les trier de la plus informative à la moins informative. Le classement est établi suivant des indices statistiques. Une discussion sur nos résultats identifie quelques points ayant un impact sur l'interprétabilité des règles d'association.*

ABSTRACT. *This paper aims at defining a methodology of access and reading of association rules extracted from texts. The corpus used is a set of scientific abstracts in the field of molecular biology. The mining process often generates a huge number of rules. This issue leads us to raise the question of how they could be ranked from the most to the least significant one with the help of statistical indices. A discussion about our results identifies some points having an impact on the interpretability of the association rules.*

MOTS-CLÉS : *Règles d'association, fouille de textes, indices statistiques, interprétation, terminologie.*

KEYWORDS: *Association rules, text mining, statistical indices, interpretation, terminology.*

1. Introduction

Notre travail porte sur la fouille de données dans les textes. Nous présentons dans cet article une chaîne complète de traitement, sélection et indexation des textes ; puis nous décrivons des processus de fouille et d'interprétation par un expert. Nous montrons également que la fouille de textes (FdT) est sensible au traitement des textes et à la qualité de l'information qui en a été extraite. La FdT s'adresse à un utilisateur qui, dans notre cas, est expert d'un domaine particulier. Elle donne à celui-ci une vue synthétique du contenu d'un corpus, 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 travers le corpus, des relations connues dans son 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 de nouvelles relations. Ces relations sont fondées sur le principe de la cooccurrence de *termes* dans un même texte. La cooccurrence peut refléter des liens sémantiques du domaine comme cela est souvent considéré en travaux de *sémantique lexicale*. Nous observons ces relations à travers des *règles d'association* et nous optons pour le paradigme de la représentation symbolique pour extraire ces règles. De ce point de vue, nous nous situons dans la lignée des travaux de [Kod99] qui se sont intéressés à cette même problématique mais sur des données de nature différente. Cependant, le nombre de règles extrait croît de manière exponentielle par rapport au nombre de termes du corpus. Leur lecture est alors une tâche difficile. Nous nous intéressons à trouver le moyen de sélectionner, parmi ces règles, celles qui présentent un intérêt particulier pour l'expert. Afin de réaliser cet objectif, nous procédons en deux étapes :

1) L'expert identifie, dans l'ensemble des règles, un sous-ensemble de règles qui présentent, pour ses besoins, un intérêt particulier ;

2) nous cherchons les indices formels associés à chacune des règles et qui refléteraient l'ordre de préférence établi par l'expert.

Différentes approches ont été proposées pour gérer ce grand nombre de règles. La première approche cherche à réduire leur nombre en calculant une base minimale de règles à partir de laquelle il est possible de retrouver la totalité des règles. Cette réduction s'opère soit durant le processus de fouille [GD86, Lux91] ou après avoir organisé les données dans une structure hiérarchique, par exemple un treillis de Galois [TS00, STB+01]. Une deuxième approche consiste à filtrer ces règles et ne chercher que celles dont la prémisse et/ou la conclusion sont des termes d'un « type » particulier [KMR+94]. Les techniques incrémentales [CHNW96] permettent de générer les règles par étapes. Cela consiste à ajouter, un à un, les documents dans le corpus. Un critère de maintenance permet alors d'observer les nouvelles règles construites par rapport à celles générées à l'étape précédente. [BA99] définissent, quant à eux, deux ordres partiels sur les règles en combinant le support et la confiance. Les règles qui satisfont ces deux ordres sont censées être les plus pertinentes parmi l'ensemble des règles.

Ces approches ont toutes un point commun : l'élimination de règles dites moins informatives ou d'autres qui sont redondantes. À l'opposé, notre approche consiste à garder toutes les règles possibles car nous ne pouvons préjuger de celles qui seront, au final, retenues par l'expert à l'étape (1). Nous cherchons à aider l'expert dans la lecture de ces règles en les triant suivant les valeurs de leurs indices statistiques (2).

La section 2 décrit les caractéristiques que doivent avoir les textes du corpus. Nous précisons le format de représentation des textes. La section 3 concerne le processus de FdT ainsi que la définition des règles d'association. Nous associons à ces règles des

indices formels, que nous définissons en section 4. Ces indices sont destinés à classer les règles entre elles ; puis nous introduisons le critère d'interprétabilité d'une règle et nous demandons à l'expert de toutes les évaluer par rapport à ce critère. Cette analyse est présentée en section 5. En section 6, nous évaluons la variation des différents indices et la confrontons à l'analyse de l'expert. Enfin, nous donnons, en section 7, quelques éléments de discussion qui ont suivi l'évaluation de ces résultats. La confrontation des résultats formels (calcul des règles d'association, calculs des indices) à la réalité du domaine (l'appréciation de l'expert) est un travail inédit en FdT.

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 donc 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). Notre représentation repose donc sur un réseau terminologique et sur la liste des *termes* extraite 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. L'ensemble des termes constitue une terminologie du domaine et chaque terme indexant un texte participe à la construction du contenu du texte.*

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. Comme le soulignent [FGM+96] : *« 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 et homogène dans un domaine de spécialité. Le cadrage du sujet permet d'avoir une terminologie délimitée. [Har68] a montré qu'un corpus spécialisé est toujours caractérisé par un vocabulaire restreint ;

– 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 mieux repérer les termes dans les textes, nous collectons non seulement le *terme préférentiel* (*i.e.* celui qui est décrit dans la terminologie de référence) mais également toutes ses formes variantes. Par exemple, on veut 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 [Jac94]. C'est un analyseur syntaxique fondé sur les grammaires d'unification et, plus précisément, sur la forme logique des Grammaires d'Arbres Adjoints [VS92]. FASTR recherche, dans des séquences textuelles acceptables, le maximum de termes qui s'y trouvent par identification des termes à partir d'une *liste contrôlée* (appelée nomenclature terminologique). Les formes variantes de termes reconnues dans les textes sont, par la suite, ramenées à leur terme préférentiel.

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

Figure 1: *Exemple d'un document du corpus (version du texte raccourcie).*

Notre corpus est constitué de 1 361 documents d'environ 200 000 mots, soit 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. Figure1 donne l'exemple du document n° 391 de notre corpus.

3. Processus de fouille de textes

Définition 2 (Fouille de Textes) *Notre processus de fouille est fondé sur l'utilisation :*

1) d'une méthode formelle d'extraction des règles d'association ;

2) d'un classement des règles suivant des indices statistiques ;

3) d'un mécanisme interactif d'accès aux règles et au contenu des documents.

L'extraction des règles d'association (1) se fait en deux étapes. Premièrement, nous calculons les ensembles fermés fréquents en utilisant l'algorithme *Close* [PBTL99]. Ces ensembles correspondent aux intensions tels qu'elles sont définis en analyse formelle de concepts (AFC) [GW00]. Puis nous en déduisons les règles d'association. Les indices statistiques calculés en (2) sont des mesures de pondération affectés aux règles. Ces indices donnent un poids à chaque règle et permettent alors de les « classer ». Un environnement de navigation hypertextuelle (3) aide l'expert du domaine à

interpréter les règles d'association obtenues en (1). Il lui permet d'accéder au contenu des documents (*cf.* Figure1) liés à une règle.

3.1. *Règles d'association*

Les règles d'association ont été, initialement, utilisées en analyse de données [GD86] ; puis en fouille de données afin de trouver des régularités, des corrélations dans des bases de données relationnelles de grandes tailles [AS94]. Elles ont été, par la suite, appliquées à la fouille de textes [FD95].

Définition 3 (Règle d'association) *Une règle d'association est du type :*

$$\mathtt{R} : \mathtt{t}_1 \wedge \ldots \wedge \mathtt{t}_i \Longrightarrow \mathtt{t}_{i+1} \wedge \ldots \wedge \mathtt{t}_n$$

(où $\mathtt{t}_1, \ldots, \mathtt{t}_n$ *sont des* termes*)*

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 R est que, si des documents possèdent les termes $\{\mathtt{t}_1, \ldots, \mathtt{t}_i\}$ alors ils ont tendance à posséder également $\{\mathtt{t}_{i+1}, \ldots, \mathtt{t}_n\}$.

4. Indices statistiques liés aux règles

Une correspondance entre les « statistiques exploratoires » et la « théorie des ensembles » est définie comme une fonction d'interprétation de la règle : $\mathtt{R} : \mathtt{B} \Longrightarrow \mathtt{H}$. En effet, nous pouvons représenter les résultats d'une expérience en utilisant des ensembles dans un espace de possibilité S. Lorsque S est fini, nous pouvons attribuer, à chaque élément de cet espace, une quantité positive appelée « probabilité » [Spi88]. La valeur informative de $\mathtt{R} : \mathtt{B} \Longrightarrow \mathtt{H}$ dépend de la distribution des termes en B et H sur les documents. Soient $\mathtt{S}_\mathtt{B}$, $\mathtt{S}_\mathtt{H}$ et $\mathtt{S}_{\mathtt{B}\wedge\mathtt{H}}$ les ensembles de documents possédant respectivement les termes B, H et $\mathtt{B} \wedge \mathtt{H}$. Trois valeurs de probabilités [1] ont un impact sur la valeur des indices $\mathtt{P}(\mathtt{B})$, $\mathtt{P}(\mathtt{H})$ et $\mathtt{P}(\mathtt{B} \wedge \mathtt{H})$:

$$\mathtt{P}(\mathtt{X}) = \frac{\mathtt{nombre\ de\ documents\ ayant\ X}}{\mathtt{nombre\ total\ de\ documents}}$$

Figure 2 représente trois principales distributions de termes qui nous intéressent particulièrement. Le quatrième cas (H rare et B fréquent) n'arrive pas dans ce contexte puisque nous manipulons des règles à confiance élevée[2].

Du point de vue de l'extraction des connaissances, plus $\mathtt{P}(\mathtt{H})$ est grand, plus la règle sera triviale et, par conséquent, moins informative.

1. Plus précisément, il s'agit là de fréquences relatives correspondant à des estimations de probabilités.
2. *cf.* sections 4.1 pour la défi nition de `minconf` et 5.1 pour les valeurs utilisées dans nos expériences.

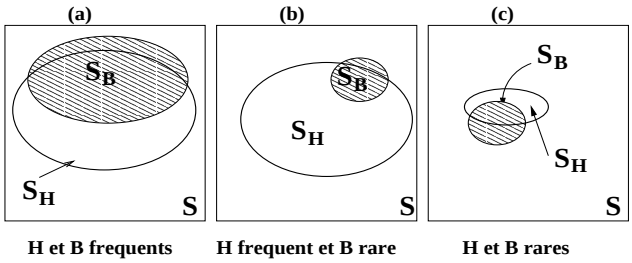

Figure 2: Trois principaux cas illustrant la variabilité de S_B et S_H –S est l'espace des possibilités, ici les documents–.

– En (a), $P(B)$ et $P(H)$ sont toutes deux élevées, ceci indique que les termes de la règle sont répandus dans le corpus. Les règles de ce cas seront considérées comme les moins intéressantes ;

– comme $P(B)$ est plus faible, le cas (b) paraît, en ce sens, plus intéressant. L'inconvénient est que tout document ayant B aura tendance à avoir H également ;

– le cas (c) est, finalement, le plus intéressant. Les termes y sont rares et apparaissent souvent ensemble (*i.e.* $P(B \wedge H)$ est grand). Ces termes seront, plus probablement, reliés dans un contexte.

Les deux paragraphes suivants montrent que les valeurs d'indices sont capables de différencier les trois cas de la Figure 2.

4.1. *Indices de support et de confiance*

L'indice $P(B \wedge H)$ est appelé **support** de la règle R. Plus il est grand, plus $S_{B \wedge H}$ l'est aussi (*i.e.* plus nombreux sont les documents de S qui ont contribué à l'extraction de la règle).

La probabilité conditionnelle $P(H|B) = \frac{P(B \wedge H)}{P(B)}$ est appelée la **confiance** de R. La confiance mesure la validité de la règle : plus cet indice est faible, plus il y a de contre-exemples (*i.e.* des documents qui possèdent les termes B mais pas tous les termes H). Lorsque la confiance vaut 1, la règle est dite *totale*, autrement elle est dite *partielle*.

Les algorithmes de construction de règles d'association utilisent des seuils de support et de confiance (resp. `minsup` et `minconf`) pour gagner en efficacité. Cependant, ces deux indices ne différencient pas, à eux seuls, les cas (a), (b) et (c) de Figure 2. Le support représente l'intersection $S_B \cap S_H$, il peut alors distinguer (a) de (b) et (c). La confiance représente l'inclusion de S_B dans S_H et n'est pas un facteur discriminant des trois cas.

4.2. *Autres indices*

Nous présentons d'autres indices, proposés dans la littérature, qui permettent différents classements des règles.

4.2.1. *L'intérêt*

L'indice d'**intérêt** mesure la déviation par rapport à l'indépendance de $\mathtt{B}$ et $\mathtt{H}$. Sa valeur est donnée par :

$$\mathtt{int}\,[\mathtt{B} \implies \mathtt{H}] = \frac{\mathtt{P}(\mathtt{B} \wedge \mathtt{H})}{\mathtt{P}(\mathtt{B}) \times \mathtt{P}(\mathtt{H})} \tag{1}$$

L'intérêt a un comportement symétrique pour $\mathtt{B}$ et pour $\mathtt{H}$, c'est-à-dire que : $\mathtt{int}\,[\mathtt{B} \implies \mathtt{H}]$ est égal à $\mathtt{int}\,[\mathtt{H} \implies \mathtt{B}]$. Cet indice varie entre $[0, +\infty[$. D'après notre définition en section 4, on a $\mathtt{P}(\mathtt{B} \wedge \mathtt{H}) \leq \mathtt{P}(\mathtt{B})$ et $\mathtt{P}(\mathtt{B} \wedge \mathtt{H}) \leq \mathtt{P}(\mathtt{H})$ qui sont toujours vérifiées. Par conséquent, l'intérêt est fort lorsque $\mathtt{S}_\mathtt{B}$ et $\mathtt{S}_\mathtt{H}$ sont petits. De plus, quand $\mathtt{S}_\mathtt{B} \cap \mathtt{S}_\mathtt{H} \approx \mathtt{S}_\mathtt{B}$ et $\mathtt{S}_\mathtt{B} \cap \mathtt{S}_\mathtt{H} \approx \mathtt{S}_\mathtt{H}$, cet indice a de fortes valeurs également.

4.2.2. *La conviction*

La **conviction**, proposée par [BMUT97], mesure l'indépendance de $\mathtt{B}$ et $\neg\mathtt{H}$.

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

Lorsque $\mathtt{B} \wedge \neg\mathtt{H}$ est vrai, cela constitue un contre-exemple à la règle puisque c'est le cas qui rend l'implication logique fausse. La conviction mesure alors la validité de l'implication de $\mathtt{B}$ vers $\mathtt{H}$. Elle favorise les règles qui ont un faible $\mathtt{P}(\mathtt{H})$ ou un fort $\mathtt{P}(\mathtt{B})$ en combinaison avec une confiance élevée (*i.e.* $\mathtt{P}(\mathtt{B} \wedge \mathtt{H}) \approx \mathtt{P}(\mathtt{B})$). Cet indice n'est pas calculable pour les règles partielles puisque $\mathtt{P}(\mathtt{B} \wedge \mathtt{H})$ vaut 0. Il varie entre $[0, +\infty[$ et a l'avantage de ne pas être symétrique.

Lorsque $\mathtt{B}$ et $\mathtt{H}$ sont indépendants, les indices d'intérêt et de conviction valent 1.

4.2.3. *La dépendance*

L'indice de **dépendance** est couramment utilisé en statistiques pour mesurer que le fait de connaître $\mathtt{B}$ influe sur la probabilité d'avoir $\mathtt{H}$.

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

Plus $\mathtt{H}$ dépend de $\mathtt{B}$, plus cet indice est grand. Ce qui augmente le plus sa valeur est la taille de $\mathtt{S}_\mathtt{H}$. Nous obtenons alors des valeurs sensiblement égales pour les (a) et (b). Ceci est particulièrement visible pour les règles totales où $\mathtt{dep}\,[\mathtt{B} \implies \mathtt{H}] = 1 - \mathtt{P}(\mathtt{H})$ ne dépend pas de $\mathtt{B}$. Pour cela, nous introduisons les deux indices suivants qui sont également des dépendances.

4.2.4. *La nouveauté et la satisfaction*

L'indice de **nouveauté** [PS91] est défini par :

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

La valeur absolue de cet indice vaut $\mathtt{dep}\,[\mathtt{B} \Longrightarrow \mathtt{H}] \times \mathtt{P}(\mathtt{B})$. Plus $\mathtt{P}(\mathtt{B})$ est faible, plus cet indice est petit. Ainsi les règles des cas (b) et (c) sont moins bien classées. Nous sommes intéressés par les petites valeurs de cet indice. Il varie entre $]-1, 1[$ et prend une valeur négative quand $\mathtt{P}(\mathtt{B} \wedge \mathtt{H}) \approx 0$.

L'indice de nouveauté est symétrique. Pour cette raison, nous introduisons l'indice suivant appelé **satisfaction** :

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

qui s'écrit également : $|\mathtt{sat}\,[\mathtt{B} \Longrightarrow \mathtt{H}]| = \frac{\mathtt{dep}\,[\mathtt{B} \Longrightarrow \neg\mathtt{H}]}{\mathtt{P}(\neg\mathtt{H})}$. Plus $\mathtt{P}(\mathtt{B})$ est faible, plus cet indice est élevé. Il n'est pas utile pour classer les règles totales car il vaut toujours 1. Lorsque B et H sont indépendants, les indices de dépendance, de nouveauté et de satisfaction valent 0.

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.

5. Expérimentations

Deux expériences ont été menées avec le corpus sur la biologie moléculaire.

Une première expérience eut lieu avec une indexation automatique utilisant FASTR. L'ensemble des documents 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, utilisés par les auteurs du texte.

Une seconde expérience eut lieu avec les 22 885 termes filtrés à la main par les documentalistes de l'INIST[3]. 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 dont 632 différents (soit $18,94\,\%$ du nombre de termes différents par rapport à la 1ère expérience). À noter que $49\,\%$ des termes apparaissent entre 5 et 15 fois.

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

Tableau 1: Pourcentage de règles obtenues par cas de distribution des termes

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

5.1. *Description des résultats*

Pour notre première expérience, nous avons fixé `minsup` à $0,7\%$ (correspondant à un minimum de 10 documents) et `minconf` à 100% (*i.e.* règles totales), nous avons obtenu 1 202 règles. Les règles étaient trop nombreuses pour être finement toutes analysées. Comme le soulignent [GKCG01] : *« ... 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 gardé `minsup` équivalent à 10 documents. Nous avons obtenu 347 règles, dont 128 totales. C'est un nombre raisonnablement interprétable par l'expert.

Le Tableau 1 montre que plus de 60% des règles représentent le Figure 2 case (c), le plus interessant à notre sens.

5.2. *Interprétation par l'expert*

Nous avons soumis les 347 règles obtenues lors de la seconde expérience à notre expert. Les règles n'ont pas été classées afin de lui laisser une libre appréciation. Il est important, pour nous, de repérer quelles règles lui paraissent « interprétables ». Ceci nous amène à une définition de l'interprétabilité.

Définition 4 (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.*

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 extraites (voir Figure 3) :

La règle 120 reflète le plus le phénomène de résistance. Elle indique que les documents cités décrivent la mutation du gène "gyrA" qui contrôle le comportement de l'enzyme "gyrase" dans une zone précise de l'ADN. Cet enzyme est responsable de la résistance aux antibiotiques de la famille des "Quinolones". Pour avoir le schéma complet du mécanisme de résistance, il manque dans la règle le nom de la bactérie, puisqu'il n'est pas le même pour les 11 documents. La règle 279 fait ressortir le fait

Numéro : 120
Règle : "determine region" "gyrA gene" "gyrase" "mutation" $\Longrightarrow$ "Quinolone"
pB : "0.008" *pH* : "0.059" *pBH* : "0.008"
Support : "11" *Confiance* : "1.000" *Intérêt* : "17.012" *Conviction* : "indéfi nie"
Dépendance : "0.941" *Nouveauté* : "0.008" *Satisfaction* : "1.000"

Numéro : 183
Règle : "epidemic strain" $\Longrightarrow$ "outbreak"
pB : "0.012" *pH* : "0.057" *pBH* : "0.012"
Support : "16" *Confiance* : "1.000" *Intérêt* : "17.449" *Conviction* : "indéfi nie"
Dépendance : "0.943" *Nouveauté* : "0.011" *Satisfaction* : "1.000"

Numéro : 202
Règle : "grlA gene" $\Longrightarrow$ "mutation" "Staphylococcus Aureus"
pB : "0.009" *pH* : "0.023" *pBH* : "0.008"
Support : "12" *Confiance* : "0.917" *Intérêt* : "40.245" *Conviction* : "11.727"
Dépendance : "0.894" *Nouveauté* : "0.008" *Satisfaction* : "0.915"

Numéro : 270
Règle : "mecA" "meticillin" $\Longrightarrow$ "mecA gene" "Staphylococcus Aureus"
pB : "0.009" *pH* : "0.012" *pBH* : "0.009"
Support : "12" *Confiance* : "1.000" *Intérêt* : "80.059" *Conviction* : "indéfi nie"
Dépendance : "0.988" *Nouveauté* : "0.009" *Satisfaction* : "1.000"

Numéro : 279
Règle : "mutation" "parC gene" "Quinolone" $\Longrightarrow$ "gyrA gene"
pB : "0.015" *pH* : "0.046" *pBH* : "0.014"
Support : "21" *Confiance* : "0.952" *Intérêt* : "20.574" *Conviction* : "20.028"
Dépendance : "0.906" *Nouveauté* : "0.014" *Satisfaction* : "0.950"

Numéro : 335
Règle : "restriction enzyme" $\Longrightarrow$ "enzyme"
pB : "0.008" *pH* : "0.112" *pBH* : "0.008"
Support : "11" *Confiance* : "1.000" *Intérêt* : "8.954" *Conviction* : "indéfi nie"
Dépendance : "0.888" *Nouveauté* : "0.007" *Satisfaction* : "1.000"

Figure 3: *Quelques règles commentées.*

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'on parle de "parC", les auteurs font référence aussi à "gyrA". Les deux règles informatives 270 et 202 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".

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.* règle 335). D'autres règles relient des termes et leurs synonymes. Cela vient du fait que les auteurs emploient indifféremment des termes et leurs synonymes pour décrire un même concept (*ex.* règle 183).

6. Confrontation des indices à l'analyse de l'expert

Nous supposons que le corpus reflète les connaissances du domaine, et que l'indexation reflète le contenu des textes. L'expert a globalement approuvé les règles que nous lui avons présenté par rapport au domaine.

L'indice d'*intérêt*, par définition, classe en premier les règles ayant des termes rares en B et en H (cas (c) Figure 2). On s'attend à ce que l'expert préfère ce genre de règle. Or l'expérience vérifie que les règles 270 et 202, qui illustrent ce cas avec une forte valeur pour cet indice (respectivement $80,059$ et $40,245$), sont porteuses d'information du point de vue de l'expert. Par ailleurs, la règle 159 ("dna" "gyrA gene" $\Longrightarrow$ "mutation") qui illustre le cas (b) ainsi que la règle 228 ("Gyrase" "protein" $\Longrightarrow$ "mutation") pour le cas (a) sont moins informatives. Leur intérêt et leur conviction sont faibles (respectivement $4,929$ et $5,086$).

On vérifie que la *conviction* renforce le côté implicatif de B vers H : la règle 279, déjà présentée, souligne une antériorité dans la découverte des gènes cités et possède une forte valeur de conviction ($20,028$). En revanche, la règle 215 dans le sens "gyrA gene" vers "parC" est moins bien classée ($11,735$). Cet indice 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 le cas (c).

La *dépendance* est forte pour de faibles valeurs de $\mathrm{P(H)}$, ce qui nous place également dans le cas (c) Figure 2 (*ex.* règle 279). Les règles 270 et 120, comme toutes les règles totales, sont celles qui sont les plus dépendantes (car $\mathrm{dep}\,[\mathrm{B} \Longrightarrow \mathrm{H}]=1-\mathrm{P(H)}$).

Enfin, deux règles illustrent le comportement de la *nouveauté* et de la *satisfaction*. La règle non informative 273 ("meticillin" $\Longrightarrow$ "staphylococcus Aureus") correspond au cas (a), alors que celle qui est mieux interprétée 265 ("mecA gene" "meticillin" $\Longrightarrow$ "Staphylococcus Aureus") – grâce à la présence du gène – correspond à Figure 2 (b). Ces deux règles ont une dépendance très faible. Néanmoins, le classement par nouveauté place 273 devant 265, alors que la satisfaction les classe inversement. Ces deux indices peuvent donc distinguer le cas (a) du cas (b), là où la dépendance ne peut aider à les discriminer.

Le Tableau 2 montre que plus de la moitié des règles extraites (53.3%) ont une valeur de *dépendance* au-dessus de la valeur moyenne. Près du tiers seulement ont des valeurs d'*intérêt* et de *conviction* supérieures à la moyenne. Ces pourcentages, confirmés par l'étape d'interprétation par l'expert, soulignent la nécessité du filtrage par classement des règles.

Tableau 2: Pourcentage de règles ayant de fortes valeurs pour les trois indices

Indice	% de règles fortes
Intérêt	29.4
Conviction	28.8
Dépendance	**53.3**

7. Éléments de discussion

La distribution des termes dans notre corpus est telle que peu de termes apparaissent souvent et beaucoup d'autres très rarement. Nous savons, depuis l'établissement de la loi de Zipf (*voir* [GS01] pour sa définition) que c'est, en général, le cas pour les textes écrits en langage naturel. Par conséquent, chaque indice statistique ne couvre qu'un intervalle de valeurs possibles très limité (*ex.*, pas de valeurs négatives pour la nouveauté ou la satisfaction). Des expérimentations sur des distributions de termes différentes devraient produire une plus grande variation des indices.

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 documents *vs.* 11).

Notre approche s'appuie sur une description booléenne (présence *vs.* absence) des termes dans le document. Notre méthode est, en cela, assez sensible à la qualité de l'indexation. L'expert a repéré des règles non fondées, mais pour la plupart liées au bruit de l'indexation. Par conséquent, les règles peuvent révéler, de façon ad hoc, la qualité de l'indexation. Ainsi, la combinaison d'une indexation automatique et d'un filtrage manuel aident à améliorer la qualité des règles obtenues.

8. Conclusion

Cet article présente une expérience complète mettant en présence une méthode de traitement automatique de corpus, un processus de fouille de textes et qui prend en compte une évaluation des résultats par l'utilisateur final (*i.e.* l'expert). Nous soulignons l'exigence d'avoir une bonne qualité d'indexation pour extraire des règles informatives. Bien que cela induise une subjectivité liée à toute expertise humaine, nous avons confronté de façon satisfaisante 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 illustrant le cas (c) Figure 2. C'est ce cas que nous avons identifié comme étant le plus informatif du

point de vue de l'expert. Enfin, les indices de *nouveauté* et de *satisfaction* permettent de distinguer le cas (a) du cas (b) pour des règles à faible *dépendance*.

Remerciements

Les auteurs tiennent à remercier toute l'équipe URI de l'INIST, notamment Jean Royauté et Alain Zasadzinski pour le corpus de textes fourni et l'expertise sur l'interprétation des règles. Nous tenons également à remercier les différents relecteurs de cet article.

9. Bibliographie

[AS94] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proc. of the 20th Int'l Conf. on Very Large Data Bases (VLDB'94)*, pages 478–499, Santiago, Chili, 1994.

[BA99] R. J. Bayardo and R. Agrawal. Mining the most interesting rules. In *Proc. of the 5th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pages 145–154, 1999.

[BMUT97] S. Brin, R. Motwani, J. Ullman, and S. Tsur. Dynamic Itemset Counting and Implication Rules for Market Basket Data. In *Proc. of the ACM SIGMOD'97 Conference on Management of Data*, volume 36, pages 255–264, Tucson, USA, 1997.

[CHNW96] D.W. Cheung, J. Han, V. Ng, and C.Y. Wong. Maintenance of discovered association rules in large databases : An incremental updating techniques. In *Proc. 12th IEEE Int'l Conf. on Data Engineering (ICDE-96)*, Nouvelle-Orléans, USA, 1996.

[FD95] R. Feldman and I. Dagan. Knowledge Discovery in Textual Databases (KDT). In *Proceedings of the 1st Int'l Conf. on Data Mining and Knowledge Discovery*, Montréal, CA, 1995. AAI Press.

[FGM+96] N. Faraj, R. Godin, R. Missaoui, S. David, and P. Plante. 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, 1996.

[GD86] J.L. Guigues and V. Duquenne. Familles minimales d'implication informatives résultant d'un tableau de données binaires. *Mathématiques, Informatique et Sciences Humaines*, 95 :5–18, 1986.

[GKCG01] R. Gras, P. Kuntz, R. Couturier, and F. Guillet. Une version entropique de l'intensité d'implication pour les corpus volumineux. In *Actes EGC'01 : Extraction et Gestion des Connaissances*, volume 1, pages 69–80, Nantes, 2001. Éditions Hermès.

[GS01] A. Gelbukh and G. Sidorov. Zipf and Heaps laws' coeffi cients depend on language. In *LNCS : Proc. of Conf. on Intelligent Text Process. and Comp. Linguist. (CICLing'01)*, volume 2004, pages 332–335. Springer-Verlag, 2001.

[GW00] B. Ganter and R. Wille. *Formal Concept Analysis : Mathematical Foundations*. Springer-Verlag, 2000.

[Har68] Z. Harris. *Mathematical Structures of Languages*. Wiley-Interscience, New-York, 1968.

[Jac94] C. Jacquemin. FASTR : A Unification-Based Front-End to Automatic Indexing. In *Proc. of Computer-Assisted Information Retrieval (RIAO'94)*, pages 34–47, New-York, 1994. Rockfeller University.

[KMR+94] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A. I. Verkamo. Finding interesting rules from large sets of discovered association rules. In *Proc. of the 3rd Int'l Conf. on Knowledge Management*, pages 401–407, Gaithersburg, USA, 1994. ACM Press.

[Kod99] Y. Kodratoff. Knowledge Discovery in Texts : A definition, and applications. In *LNAI : Proc. of the 11th Int'l Symp. ISMS'99*, volume 1609, pages 16–29, Warsaw, 1999. Springer.

[Lux91] M. Luxenburger. Implications partielles dans un contexte. *Mathématiques, Informatique et Sciences Humaines*, 29(113) :35–55, 1991.

[PBTL99] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Efficient mining of association rules using closed itemset lattices. *Information Systems*, 24(1) :25–46, 1999.

[PS91] G. Piatetsky-Shapiro. *Discovery, Analysis, and Presentation of Strong Rules*. AAAI/MIT Press, 1991. Chapitre 13.

[Spi88] M. R. Spiegel. *Theory and Problems of Statistics*. Mc-Graw Hill, 1988. 2nd Edition.

[STB+01] G. Stumme, R. Taouil, Y. Bastide, N. Pasquier, and L. Lakhal. Intelligent structuring and reducing of association rules with formal concept analysis. In *LNAI : Proc. of KI 2001 Advances in Artificial Intelligence*, volume 2174, pages 335–350. Springer, 2001.

[TS00] Y. Toussaint and A. Simon. Building and interpreting term dependencies using association rules extracted from Galois lattices. In *Proc. of Content-Based Multimedia Information Access RIAO'00*, volume 2, pages 1686–1693, Paris, 2000.

[VS92] K. Vijay-Shankar. Using descriptions of trees in a tree-adjoining grammar. *Computational Linguistics*, 18 :481–518, 1992.