Deux méthodologies de classification de règles d'association pour la fouille de textes

Hacène Cherfi, Amedeo Napoli et Yannick Toussaint
LORIA, BP 239, 54506 Vandœuvre-lès-Nancy cedex
{cherfi,napoli,yannick}@loria.fr,
http ://www.loria.fr/equipes/orpailleur/

Résumé. Parmi les inconvénients d'un processus de fouille de données textuelles fondé sur l'extraction de règles d'association figurent le grand nombre de règles extraites et la difficulté d'affecter à une règle un critère de qualité fiable par rapport aux connaissances de l'analyste (*i.e.*, l'expert du domaine). La plupart des approches pour la classification des règles d'association utilisent des méthodes statistiques pour juger de la qualité d'une règle et ne s'appuient pas sur les connaissances du domaine des données disponibles *a priori* pour classer les règles extraites. Dans cet article nous définissons la notion de qualité d'une règle d'association. Nous étudions en premier lieu les mesures statistiques permettant de classer les règles et nous proposons un algorithme combinant ces différentes mesures. Nous introduisons ensuite une nouvelle méthodologie de classification des règles exploitant un modèle de connaissances. Nous expérimentons cette mesure sur un exemple formel puis nous l'évaluons sur des données réelles.

1 Introduction

Les textes, du point de vue de la fouille de données, sont des données complexes qui posent de nouveaux défis. En premier lieu, les textes sont des données peu structurées contrastant avec les données des bases de données pour lesquelles un travail de modélisation est réalisé au préalable. Ils sont rédigés en langue naturelle avec tout ce que cela suppose en termes d'implicite, d'ambiguïté, d'imprécision, etc. Il n'existe pas de représentation standard décrivant l'intégralité du contenu d'un texte et sur laquelle il est possible d'appliquer des méthodes de fouille de données ; les représentations habituelles sont le plus souvent partielles et bruitées. Un second facteur de complexité vient du fait qu'expliciter l'implicite véhiculé par un texte nécessite le recours à un modèle de connaissances du domaine. Pour pouvoir déduire de nouvelles connaissances à partir de textes, il faut dépasser le principe de la co-occurrence des mots-clés dans les textes et pouvoir appliquer des opérations de généralisation ou de spécialisation sur les mots-clés en fonction de connaissances disponibles, et cela afin de pouvoir manipuler, regrouper ou dissocier les textes par exemple.

Dans cet article, nous présentons un processus de *fouille de textes*, qui s'aligne sur le schéma de référence de l'extraction de connaissances dans des bases de données introduit dans [Fayyad *et al.*, 1996]. Ce processus de fouille s'appuie sur une boucle itérative et interactive et place l'expert du domaine, appelé *analyste* par la suite, au centre du processus de fouille. Le processus de fouille de textes vise à construire, par des enrichissements successifs, un modèle du domaine. Réciproquement, à chaque itération,

le modèle du domaine est exploité pour guider le processus de fouille.

Le processus de fouille de textes présenté ici extrait des règles d'association portant sur les termes-clés contenus dans les textes. Un des intérêts des règles d'association vient de la facilité de lecture et d'interprétation de ces règles qui donnent une vision des régularités existant dans les textes, que l'analyste peut ensuite plus facilement prendre en compte, en vue par exemple de l'enrichissement d'un modèle de connaissances. Les termes-clés constituent une représentation (primaire) du contenu des textes. L'extraction des termes-clés nécessite une transformation des données textuelles reposant sur des techniques de traitement automatique du langage naturel décrites dans [Cherfi *et al.*, 2003a].

Le processus d'extraction de règles d'association produit un très grand nombre de règles qui est difficile à appréhender pour une personne. Afin de fournir à l'analyste les moyens d'interpréter convenablement les règles extraites, nous proposons dans cet article deux méthodologies de classification des règles. La première méthodologie repose sur un classement des règles d'association par des mesures de qualité statistiques [Cherfi *et al.*, 2003a]. De telles mesures de qualité ont été introduites et ont fait l'objet de présentations synthétiques dans [Lavrač *et al.*, 1999, Kuntz *et al.*, 2000, Tan *et al.*, 2002]. Une mesure permet de mettre en valeur certaines qualités des règles d'association : celles qui portent sur des signaux d'information faibles [Feldman et Hirsh, 1997], celles qui soulignent des dépendances fonctionnelles [Lehn *et al.*, 2004] ; ou bien celles qui ont le moins de contre-exemples [Azé et Kodratoff, 2004]. Une des limites de l'approche vient du fait que l'évaluation est effectuée sans prendre en compte de connaissances du domaine.

La seconde méthodologie proposée s'appuie justement sur l'exploitation de connaissances du domaine pour la fouille de textes [Cherfi *et al.*, 2004, Janetzko *et al.*, 2004]. Dans ce cas, les règles d'association présentées à l'analyste sont ordonnées en fonction de leur qualité, qui dépend de connaissances disponibles sur le domaine des textes. Une règle d'association est de bonne qualité si elle contient potentiellement des informations de nature à enrichir le modèle de connaissances. Une règle est qualifiée de « triviale » ou de conforme au modèle de connaissances du domaine si elle reflète un élément de connaissance déjà présent dans le modèle. Le modèle de connaissances est, à l'image d'un thésaurus, caractérisé par une hiérarchie de termes structurés par une relation de spécialisation notée « estUn ». Par exemple, une "pomme" estUn "fruit", qui se traduit encore par la règle "pomme" $\Longrightarrow$ "fruit". Le sens du lien hiérarchique de la relation estUn dans le modèle de connaissances est important, car il traduit la conformité ou non conformité d'un lien entre deux termes.

À partir de données stables au cours des itérations du processus de fouille de textes, l'ensemble des règles d'associations extraites et les mesures statistiques associées restent stables, alors que le classement des règles en fonction de la mesure de qualité des règles, que nous proposons, évolue suivant les mises à jour successives du modèle de connaissances, ce qui souligne le caractère dynamique du processus de fouille de textes et fait une des originalités de notre approche.

Dans cet article, nous décrivons le processus de fouille de textes et la classification d'un ensemble de règles d'association extraites en fonction d'un modèle probabiliste de connaissances du domaine. Nous définissons une mesure de qualité des règles par

rapport au modèle de connaissances appelée la *vraisemblance*. Enfin, nous évaluons le comportement de la vraisemblance sur un modèle formel puis sur une expérimentation sur des données textuelles issues d'un corpus de biologie moléculaire.

2 Processus de fouille de textes

2.1 Les règles d'association

Nous introduisons les règles d'association dans le contexte de la fouille de textes. Soit $\mathcal{T} = \{\mathtt{t}_1, \ldots, \mathtt{t}_n\}$ un ensemble de termes qui sert de vocabulaire de référence pour indexer un ensemble de textes $\mathcal{D} = \{\mathtt{d}_1, \ldots, \mathtt{d}_m\}$. Chaque texte est représenté par un ensemble de termes qui indexent son contenu. $\mathcal{I} \subseteq \mathcal{T}$ désigne le sous-ensemble des termes qui indexent au moins un texte (*cf.* FIG. 1).

Une règle d'association est une implication de la forme $\mathtt{B} \overset{\mathtt{P}}{\Longrightarrow} \mathtt{H}$ où B est la prémisse (*body*) et H est la conclusion (*head*) avec $\mathtt{B} \subseteq \mathcal{I}$, $\mathtt{H} \subseteq \mathcal{I}$ et $\mathtt{B} \cap \mathtt{H} = \emptyset$. Soit $\mathtt{B} = \{\mathtt{t}_1, \ldots, \mathtt{t}_p\}$ l'ensemble des termes de la prémisse d'une règle d'association r et $\mathtt{H} = \{\mathtt{t}_{p+1}, \ldots, \mathtt{t}_q\}$ l'ensemble des termes de la conclusion. $r : \mathtt{t}_1 \ldots \mathtt{t}_p \overset{\mathtt{P}}{\Longrightarrow} \mathtt{t}_{p+1} \ldots \mathtt{t}_q$ s'interprète comme le fait que tous les textes de $\mathcal{D}$ contenant les termes $\mathtt{t}_1$ et $\mathtt{t}_2$ ...et $\mathtt{t}_p$ ont tendance à contenir aussi les termes $\mathtt{t}_{p+1}$ et $\mathtt{t}_{p+2}$ ...et $\mathtt{t}_q$ avec une probabilité P.

L'extraction de règles d'association a été utilisée pour fouiller des données du type « panier de la ménagère », mais elle a également été étudiée pour la fouille de textes [Feldman et Hirsh, 1997, Delgado *et al.*, 2002]. Plusieurs algorithmes comme « Apriori » [Agrawal et Srikant, 1994] ou « Close » [Pasquier *et al.*, 1999] permettent de mettre en œuvre le processus d'extraction de règles. En revanche, il est nécessaire de trouver un moyen d'identifier dans l'ensemble des règles celles qui sont de bonne qualité ou, à l'inverse, un moyen d'éliminer les règles triviales, c'est-à-dire *conformes* à un modèle de connaissances.

Le support de r est le nombre de textes contenant les termes de $\mathtt{B} \sqcap \mathtt{H} = \{\mathtt{t}_1, \ldots, \mathtt{t}_p, \ldots, \mathtt{t}_q\}$ (ici non normalisé par le nombre total de textes). La confiance de r est le rapport entre le nombre de textes contenant l'ensemble des termes $\mathtt{B} \sqcap \mathtt{H}$ et le nombre de textes contenant B, ce qui reflète la probabilité conditionnelle $\mathtt{P}(\mathtt{H}|\mathtt{B})$. 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 *exacte*. Sinon la règle est dite *approximative* et se voit attribuer une confiance variant entre 0 et 1. Par exemple :« Dans 60% des cas (c'est-à-dire avec une confiance de 0.6), les textes qui parlent de `quinupristine` parlent aussi de `dalfopristine` ».

Le support et la confiance sont deux mesures associées aux règles d'association [Agrawal et Srikant, 1994] et exploitées par les algorithmes d'extraction de règles pour en réduire la complexité. Deux valeurs de seuil sont définies : σ_s pour le support minimal et σ_c pour la confiance minimale. Du fait de ces seuils, tous les termes de $\mathcal{I}$ ne sont pas présents dans les règles d'association. Nous désignons par $\mathcal{R}$ le sous-ensemble de $\mathcal{I}$ des termes présents au moins une fois en partie gauche ou droite d'une règle d'association. Donc $\mathcal{R} = \bigcup_{\mathtt{r}_i} (\mathtt{B}_i \sqcap \mathtt{H}_i)$ où $\mathtt{r}_i$ parcourt l'ensemble des règles d'association considérées.

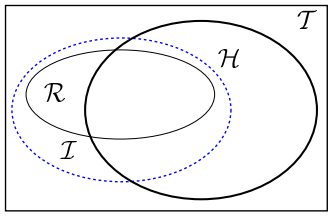


FIG. 1 – Les différents ensembles de termes : $\mathcal{T}$: ensemble des termes, $\mathcal{I}$: ensemble des termes d'indexation des textes, $\mathcal{R}$: sous-ensemble de $\mathcal{I}$ des termes d'indexation apparaissant dans les règles d'association, et $\mathcal{H}$: ensemble des termes du modèle M.

FIG. 1 montre les intersections et inclusions existant entre $\mathcal{I}, \mathcal{R}$ et $\mathcal{H}$. L'ensemble des termes $\mathcal{H}$ du modèle est à dissocier de l'ensemble des termes indexant les textes $\mathcal{I}$. En effet, nous ne pouvons pas garantir une parfaite adéquation entre le modèle de connaissances initial et le contenu des textes. Notamment, il nous semble intéressant de considérer qu'un modèle n'est pas complet au sens où il ne contient pas de façon exhaustive tous les termes du domaine. Ainsi, plus $\mathcal{I} \cap \mathcal{H}$ est grand, plus le modèle est complet par rapport à l'ensemble des textes ; plus $\mathcal{H} \cap \mathcal{R}$ est grand, meilleure est la couverture du modèle par rapport à l'ensemble des règles extraites.

2.2 Les mesures statistiques liées à une règle

Les mesures statistiques associées aux règles sont calculées à partir des mêmes données que celles utilisées pour l'extraction des règles d'associations. Certaines de ces mesures, comme le support, sont simplement relatives à la fréquence du motif, c'est-à-dire l'union des termes apparaissant en partie gauche et des termes en partie droite de la règle. D'autres mesures reposent sur le fait qu'une règle est composée de deux parties $\mathtt{B}$ et $\mathtt{H}$ mais ne sont pas sensibles au sens de la règle ou, au contraire, le prennent en compte. Ces différentes mesures reflètent chacune des propriétés différentes des règles en lien avec la distribution initiale des termes dans les textes.

Soit $\mathcal{D}(\mathtt{B})$, $\mathcal{D}(\mathtt{H})$ et $\mathcal{D}(\mathtt{B} \sqcap \mathtt{H}) = \mathcal{D}(\mathtt{B}) \cap \mathcal{D}(\mathtt{H})$ les sous-ensembles de textes de $\mathcal{D}$ possédant respectivement tous les termes de $\mathtt{B}$, $\mathtt{H}$ et $\mathtt{B} \sqcap \mathtt{H}$ (*cf.* FIG. 2). Trois valeurs de probabilités ont un impact sur la valeur des mesures que nous utilisons : $\mathtt{P}(\mathtt{B})$, $\mathtt{P}(\mathtt{H})$ et $\mathtt{P}(\mathtt{B} \cap \mathtt{H})$ qui se définissent par la formule générale suivante : $\left(\mathtt{P}(\mathtt{X}) = \frac{|\mathcal{D}(\mathtt{X})|}{|\mathcal{D}|}\right)$ compris entre 0 et 1. $\mathtt{P}(\mathtt{B} \cap \mathtt{H})$ est le support de la règle. La probabilité conditionnelle $\mathtt{P}(\mathtt{H}|\mathtt{B}) = \frac{\mathtt{P}(\mathtt{B} \cap \mathtt{H})}{\mathtt{P}(\mathtt{B})}$ en est la confiance.

Plus $\mathcal{D}(\mathtt{X})$ est grand et couvre l'ensemble $\mathcal{D}$, plus $\mathtt{X}$ est fréquent et plus $\mathtt{P}(\mathtt{X})$ est fort et donc proche de 1. Si la règle est constituée de motifs $\mathtt{B}$ et $\mathtt{H}$ très fréquents, alors ces motifs sont partagés par presque tous les textes. Par conséquent, le volume de connaissances convoyé par ces motifs, du point de vue de la découverte de connaissances, est faible ou nul.

– Pour le cas (a) de FIG. 2, $\mathtt{P}(\mathtt{B})$ et $\mathtt{P}(\mathtt{H})$ sont toutes deux proches de 1, les règles as-

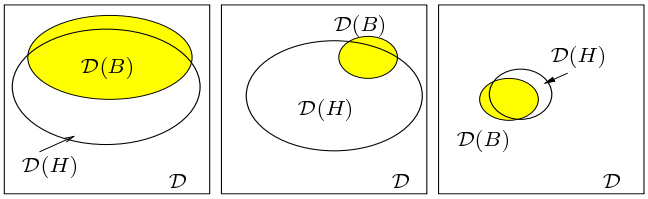


FIG. 2 – Principaux cas illustrant les variations de $\mathcal{D}(\mathtt{B})$ et $\mathcal{D}(\mathtt{H})$.

sociées sont considérées comme peu informatives. Un ensemble de termes présent dans presque tous les textes implique, très probablement, un autre ensemble présent dans tous les textes. Il y a de grandes chances que les termes de B et H soient des termes 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" ;

- Le cas (b), lorsque $\mathtt{P}(\mathtt{B})$ est plus faible, 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 de $\mathtt{P}(\mathtt{B})$ et $\mathtt{P}(\mathtt{H})$ y sont rares et apparaissent presque à chaque fois ensemble ($\mathtt{P}(\mathtt{B} \cap \mathtt{H}) \simeq \mathtt{P}(\mathtt{B}) \simeq \mathtt{P}(\mathtt{H})$). Ces termes sont donc vraisemblablement reliés dans un contexte du domaine ;
- Le quatrième cas possible (B fréquent, H rare) n'est pas considéré ici. Ce cas correspond à un seuil de *confiance* faible $\left(\frac{\mathtt{P}(\mathtt{B} \cap \mathtt{H})}{\mathtt{P}(\mathtt{B})} \ll 1\right)$.

2.2.1 Le support et la confiance

Les mesures de support et de confiance ne différencient pas complètement les cas (a), (b) et (c) de FIG. 2. Le support représente l'intersection $\mathcal{D}(\mathtt{B}) \cap \mathcal{D}(\mathtt{H})$ et peut 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 mesures de support et de confiance ne sont pas suffisantes pour identifier les cas : du plus significatif (c) vers le moins significatif (a). La suite du paragraphe montre que d'autres mesures statistiques de qualité sont capables de différencier les trois cas possibles de la FIG. 2.

2.2.2 L'intérêt

L'**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 $\mathtt{P}(\mathtt{B} \cap \mathtt{H}) = \mathtt{P}(\mathtt{B}) \times \mathtt{P}(\mathtt{H})$. La valeur de l'intérêt est donnée par :

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

L'intérêt varie dans l'intervalle $[0, +\infty[$. Si B et H sont indépendants alors $\texttt{int}\ [\texttt{B} \implies \texttt{H}] = 1$. Plus B et H sont incompatibles, plus $\texttt{P}(\texttt{B} \cap \texttt{H})$ tend vers 0 et donc l'intérêt est proche de 0. Plus B et H sont dépendants, plus l'intérêt est supérieur à 1.

Puisque $\mathcal{D}(\texttt{B}) \cap \mathcal{D}(\texttt{H}) \subseteq \mathcal{D}(\texttt{B})$ et $\mathcal{D}(\texttt{B}) \cap \mathcal{D}(\texttt{H}) \subseteq \mathcal{D}(\texttt{H})$, plus $\mathcal{D}(\texttt{B})$ et $\mathcal{D}(\texttt{H})$ sont petits dans $\mathcal{D}$ et sont proches de leur intersection, plus la valeur de l'intérêt augmente. Si $\texttt{P}(\texttt{B} \cap \texttt{H}) \simeq \texttt{P}(\texttt{B})$ alors $\texttt{int}\ [\texttt{B} \Longrightarrow \texttt{H}] \simeq \frac{\texttt{P}(\texttt{B})}{\texttt{P}(\texttt{B}) \times \texttt{P}(\texttt{H})} = \frac{1}{\texttt{P}(\texttt{H})}$, de la même manière $\texttt{int}\ [\texttt{B} \Longrightarrow \texttt{H}] = \frac{1}{\texttt{P}(\texttt{B})}$. Ainsi, quand $\texttt{P}(\texttt{B})$ ou $\texttt{P}(\texttt{H})$ tendent vers 0, l'intérêt tend vers $+\infty$. Par conséquent, les règles qui se trouvent dans le cas (c) sont classées en début. L'intérêt est symétrique : $\texttt{int}\ [\texttt{B} \Longrightarrow \texttt{H}] = \texttt{int}\ [\texttt{H} \Longrightarrow \texttt{B}]$.

2.2.3 La conviction

La **conviction** permet de privilégier parmi les deux règles $\texttt{B} \Longrightarrow \texttt{H}$ et $\texttt{H} \Longrightarrow \texttt{B}$ celle qui a le moins de contre-exemples. Dans notre contexte, un contre-exemple correspond au motif $\texttt{B} \sqcap \neg\texttt{H}$ tel que $\neg\texttt{H}$ signifie l'absence d'au moins un terme du motif dans au moins un texte de $\mathcal{D}(H)$. $|\mathcal{D}(\neg\texttt{H})| = |\mathcal{D}| - |\mathcal{D}(\texttt{H})|$ et $\texttt{P}(\neg\texttt{H}) = 1 - \texttt{P}(\texttt{H})$.

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

La conviction vaut $\left(\frac{1}{\texttt{int}[\texttt{B} \Longrightarrow \neg\texttt{H}]}\right)$, n'est pas symétrique et mesure la validité de la direction de l'implication de B vers H. La valeur de conviction augmente lorsque $\texttt{P}(\neg\texttt{H})$ est élevé ($\texttt{P}(\texttt{H})$ faible), $\texttt{P}(\texttt{B})$ est élevé et lorsque $\texttt{P}(\texttt{B} \cap \texttt{H}) \simeq \texttt{P}(\texttt{B})$ car $\texttt{P}(\texttt{B}) = \texttt{P}(\texttt{B} \cap \texttt{H}) + \texttt{P}(\texttt{B} \cap \neg\texttt{H})$. Ce qui classe les règles du cas (c) en premier.

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

2.2.4 La dépendance

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

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

Cette mesure varie dans l'intervalle $[0, 1[$. Plus cette mesure est proche de 0 (resp. 1) plus B et H sont indépendants (resp. dépendants). Ce qui augmente le plus sa valeur est la taille de $\mathcal{D}(\texttt{H})$. Les valeurs sont sensiblement égales pour les cas (a) et (b). C'est particulièrement notable pour les règles exactes où la confiance $\texttt{P}(\texttt{H}|\texttt{B})$ vaut 1 et donc $\texttt{dep}\ [\texttt{B} \implies \texttt{H}] = 1 - \texttt{P}(\texttt{H})$ ne dépend pas de $\texttt{P}(\texttt{B})$. 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 différencier les cas (a) et (b), les deux mesures suivantes sont définies.

2.2.5 La nouveauté et la satisfaction

La **nouveauté** est définie par :

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

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

La nouveauté varie entre $]-1, 1[$ et prend une valeur négative quand $\texttt{P}(\texttt{B} \cap \texttt{H}) < \texttt{P}(\texttt{B}) \times \texttt{P}(\texttt{H})$. La nouveauté s'approche de -1 pour des règles de faibles supports $\texttt{P}(\texttt{B} \cap \texttt{H}) \simeq 0$. Nous sommes intéressés par les petites valeurs absolues de cette mesure, autour de la valeur d'indépendance 0. La nouveauté est symétrique alors que la règle $\texttt{B} \Longrightarrow \texttt{H}$ peut avoir plus de contre-exemples que la règle $\texttt{H} \Longrightarrow \texttt{B}$. Pour cette raison, nous introduisons la **satisfaction** :

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

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

Cette mesure varie dans l'intervalle $]-\infty, 1]$ et vaut 0 en cas d'indépendance de B et H. La satisfaction n'est pas utile pour classer les règles exactes car sa valeur est 1 (puisque les règles exactes ont une confiance $\texttt{P}(\texttt{H}|\texttt{B}) = 1$). Pour cette mesure, $\texttt{P}(\texttt{H})$ apparaît au numérateur et au dénominateur, donc la variation de cette mesure dépend de $\texttt{P}(\texttt{B})$. Plus $\texttt{P}(\texttt{B})$ est faible, plus cette mesure est élevée. Par l'intermédiaire de cette mesure, les règles du cas (a) sont en fin de classement et sont différenciées du cas (b). Nous sommes intéressés par les valeurs élevées de cette mesure (autour de la valeur 1).

En résumé, ces deux mesures peuvent être consultées simultanément lorsqu'on se trouve dans les cas (a) ou (b) (pour des règles à faible dépendance). Plus la nouveauté est faible et la satisfaction forte, plus la règle est considérée comme significative.

TAB. 1 – (a) La base de données textuelles – (b) Ensemble des règles extraites.

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

n°	règle	n°	règle
r_1	b⇒e	r_{11}	a⇒c
r_2	b⇒c, e	r_{12}	b, c⇒a, e
r_3	a, b⇒c, e	r_{13}	d⇒a, c
r_4	a⇒b, c, e	r_{14}	c⇒b, e
r_5	b, c⇒e	r_{15}	c⇒a, d
r_6	b⇒a, c, e	r_{16}	c⇒a, b, e
r_7	e⇒b, c	r_{17}	c, e⇒b
r_8	a, e⇒b, c	r_{18}	c, e⇒a, b
r_9	a⇒c, d	r_{19}	e⇒b
r_{10}	e⇒a, b, c	r_{20}	c⇒a

2.3 Calcul des mesures sur un exemple formel et synthèse

Nous introduisons un exemple formel, repris de [Pasquier *et al.*, 1999], afin d'étudier le comportement des mesures statistiques sur un petit ensemble de données. Soit un ensemble de six textes $\{d_1, \ldots, d_6\}$ décrits par un ensemble de termes d'indexation $\{a, \ldots, e\}$. TAB. 1(a) donne la répartition des termes dans les textes.

Vingt règles d'association, numérotées $r_1, \ldots, r_{20}$, sont extraites avec un support minimal de $\sigma_s = 1$ et une confiance minimale de $\sigma_c = 0.1$. L'ensemble des règles d'association est donné en TAB. 1(b). L'extraction des règles d'association a été réalisée avec l'algorithme *Close* [Pasquier *et al.*, 1999], qui fait une recherche par niveau dans le tableau booléen décrivant le produit cartésien $\mathcal{T} \times \mathcal{I}$. L'algorithme commence par le plus petit motif fréquent *fermé* et calcule, par niveau, les motifs fermés plus longs dans $\mathcal{I}$. Un motif est fermé s'il correspond à un ensemble maximal de termes partagés par un ensemble de textes. Le motif est fréquent s'il apparaît dans au moins σ_s textes. Une fois les motifs fermés fréquents calculés, les règles d'association en sont dérivées. Les règles extraites sont celles qui ont la partie `B` minimale et la partie `H` maximale, et ce, uniquement à partir des motifs *fermés*. Par exemple, nous obtenons les règles "`b`" $\Longrightarrow$ "`e`" et "`b`" $\Longrightarrow$ "`c, e`" car les motifs {`b, e`} et {`b, c, e`} sont des fermés ; en revanche, nous n'avons pas la règle "`b`" $\Longrightarrow$ "`c`" car le motif {`b, c`} n'est pas un fermé dans notre corpus. Le classement de ces règles par valeurs décroissantes selon ces différentes mesures est donné en TAB. 2.

TAB. 2 – Valeurs des mesures statistiques pour chaque règle, classées par valeurs décroissantes

n°	sup.	n°	conf.	n°	int.	n°	conv.	n°	dep.	n°	nouv.	n°	sat.
r_1	5	r_1	1.000	r_9	2.000	r_7	1.667	r_{13}	0.500	r_1	0.139	r_1	1.000
r_2	5	r_3	1.000	r_{13}	2.000	r_2	1.667	r_3	0.333	r_{19}	0.139	r_3	1.000
r_6	5	r_5	1.000	r_3	1.500	r_{12}	1.333	r_8	0.333	r_2	0.111	r_5	1.000
r_7	5	r_8	1.000	r_8	1.500	r_{18}	1.333	r_1	0.167	r_3	0.111	r_8	1.000
r_{10}	5	r_{11}	1.000	r_{12}	1.500	r_9	1.250	r_5	0.167	r_5	0.111	r_{11}	1.000
r_{14}	5	r_{13}	1.000	r_{18}	1.500	r_{20}	1.250	r_9	0.167	r_7	0.111	r_{13}	1.000
r_{15}	5	r_{17}	1.000	r_1	1.200	r_6	1.111	r_{11}	0.167	r_8	0.111	r_{17}	1.000
r_{16}	5	r_{19}	1.000	r_2	1.200	r_{10}	1.111	r_{12}	0.167	r_9	0.111	r_{19}	1.000
r_{19}	5	r_2	0.800	r_5	1.200	r_{16}	1.111	r_{17}	0.167	r_{11}	0.111	r_2	0.400
r_{20}	5	r_7	0.800	r_6	1.200	r_{15}	1.042	r_{18}	0.167	r_{12}	0.111	r_7	0.400
r_5	4	r_{14}	0.800	r_7	1.200	r_4	1.000	r_{19}	0.167	r_{13}	0.111	r_{12}	0.250
r_{12}	4	r_4	0.667	r_{10}	1.200	r_{14}	0.833	r_2	0.133	r_{17}	0.111	r_{18}	0.250
r_{17}	4	r_{20}	0.600	r_{11}	1.200	r_1	0.000	r_7	0.133	r_{18}	0.111	r_9	0.200
r_{18}	4	r_{12}	0.500	r_{15}	1.200	r_3	0.000	r_{20}	0.100	r_{20}	0.111	r_{20}	0.200
r_4	3	r_{18}	0.500	r_{16}	1.200	r_5	0.000	r_6	0.067	r_6	0.056	r_6	0.100
r_9	3	r_6	0.400	r_{17}	1.200	r_8	0.000	r_{10}	0.067	r_{10}	0.056	r_{10}	0.100
r_{11}	3	r_{10}	0.400	r_{19}	1.200	r_{11}	0.000	r_{16}	0.067	r_{16}	0.056	r_{16}	0.100
r_3	2	r_{16}	0.400	r_{20}	1.200	r_{13}	0.000	r_{14}	0.033	r_{15}	0.028	r_{15}	0.040
r_8	2	r_9	0.333	r_4	1.000	r_{17}	0.000	r_{15}	0.033	r_4	0.000	r_4	0.000
r_{13}	1	r_{15}	0.200	r_{14}	0.960	r_{19}	0.000	r_4	0.000	r_{14}	-0.028	r_{14}	-0.200

Nous synthétisons les propriétés de ces différentes mesures en TAB. 3. Nous donnons également un algorithme définissant l'usage des mesures statistiques que nous avons présentées en vue d'identifier les règles illustrant le cas (c) de FIG. 2. `valeur(X)` renvoie la valeur de la mesure « `X` » associée à la règle `r`. On dit que `valeur(X)` pour `r` est élevée si `r` apparaît en début de colonne en TAB. 2 pour la mesure `X`.

L'utilisation de mesures statistiques sur un ensemble qui ne contient que 20 règles

TAB. 3 – Synthèse des mesures de qualité que nous utilisons pour une règle d'association

Mesure	Définition	Domaine	Indépendance	Situations de référence	Symétrie
int [B $\Longrightarrow$ H]	$\frac{P(B \cap H)}{P(B) \times P(H)}$	$[0, +\infty[$	1	incompatible = 0	×
conv [B $\Longrightarrow$ H]	$\frac{P(B) \times P(\neg H)}{P(B \cap \neg H)}$	$[0, +\infty[$	1	> 1 dépendant, $[0, 1[$ non dépendant	
dep [B $\Longrightarrow$ H]	$\lvert P(H\lvert B) - P(H) \rvert$	$[0, 1[$	0	$\simeq 1$ dépendant	
nov [B $\Longrightarrow$ H]	$P(B \cap H) - P(B) \times P(H)$	$]-1, 1[$	0	$\simeq -1$, faible support	×
sat [B $\Longrightarrow$ H]	$\frac{(P(\neg H) - P(\neg H\lvert B))}{P(\neg H)}$	$]-\infty, 1]$	0	$= 1$ règle exacte	

Algorithm 1 Utilisation des mesures statistiques

E := $\emptyset$ /* L'ensemble initial des règles identifiées */
pour chaque règle extraite r **faire**
  si valeur(int) **ou** valeur(conv) élevée
    alors E := E $\cup$ {r}
      sinon si valeur(dep) élevée
      alors E := E $\cup$ {r}
        sinon si valeur(nov) faible **ou** valeur(sat) élevée
        alors E := E $\cup$ {r}
fin_pour
renvoyer E /* L'ensemble final des règles identifiées */

est bien sûr peu significatif. [Cherfi *et al.*, 2003b] présente une expérimentation de cet algorithme sur des données réelles et une évaluation par un analyste et nous reprenons ces mêmes données pour l'évaluation de la vraisemblance en section 5. Il n'en demeure pas moins que pour un analyste en charge de l'étude d'un grand nombre de règles, il est difficile de définir une méthodologie précise de l'utilisation de ces mesures. Nous introduisons dans la section suivante la mesure de vraisemblance d'une règle par rapport à un modèle de connaissances du domaine des données. Le comportement de cette mesure pour une règle, contrairement aux mesures statistiques, n'est sensible qu'au modèle de connaissances.

3 Définition de la vraisemblance d'une règle

3.1 Définition du modèle de connaissances

Le modèle de connaissances est caractérisé par un réseau de termes $\mathcal{H} \subseteq \mathcal{T}$ structurés hiérarchiquement par une relation de spécialisation estUn (de façon analogue à un thésaurus). Un terme dans le modèle peut être isolé ou relié par la relation de spécialisation estUn à un ou plusieurs autres termes (le modèle n'est pas nécessairement

connexe). La relation estUn, définie sur $\mathcal{H} \times \mathcal{H}$, est réflexive, antisymétrique et transitive (ordre partiel). Les modèles de connaissances que nous exploitons sont construits à partir de thésaurus existants du domaine. Entre deux termes liés par la relation `estUn`, il peut exister plusieurs chemins. L'existence d'arcs de transitivité pour la relation `estUn` est donc autorisée (voir l'exemple de FIG. 3(a)).

Nous cherchons à identifier les règles d'association traduisant une relation estUn, c'est-à-dire celles pour lesquelles les termes de `B` sont liés par une relation de spécialisation avec les termes de `H` dans le modèle de connaissances. Par exemple, si le terme "fruit" est plus général que le terme "pomme" dans le modèle, une règle comme "pomme" $\Longrightarrow$ "fruit" exprime une relation de spécialisation connue et peut être identifiée comme une règle conforme au modèle ou triviale. En revanche, une règle comme "tarte à la cerise" $\Longrightarrow$ "chocolat", "beurre" exprime potentiellement une relation intéressante s'il n'existe pas de lien estUn entre "tarte à la cerise", "chocolat" et "beurre", et donc doit être conservée. Identifier les règles d'association conformes au modèle permet donc d'éviter de les présenter pour avis à l'expert.

Si l'évolution du modèle de connaissances passe par l'adjonction de nouveaux termes et de nouvelles relations de spécialisation entre les termes, alors une règle d'association $a \Longrightarrow b$ est conforme ou triviale si la relation de spécialisation `a` estUn `b` que l'analyste est en passe d'ajouter au modèle est déjà présente dans le modèle. Par ailleurs, le fait qu'une règle $a \Longrightarrow b$ soit extraite par le processus de fouille de textes ne signifie pas que la relation `a` estUn `b` doive nécessairement être ajoutée au modèle de connaissances, car tout enrichissement du modèle est fait sous le contrôle de l'analyste.

Le modèle de connaissances pour la classification des règles d'association est un modèle probabiliste construit sur $(\mathcal{H} \times \mathcal{H}, \texttt{estUn})$ et noté M. La vraisemblance d'une règle $r : a \Longrightarrow b$ notée $P_M(a \Longrightarrow b)$ se définit comme la probabilité de trouver un chemin de "`a`" vers "`b`" dans le modèle de connaissances (la probabilité de transition de a vers b). Pour définir $P_M(a \Longrightarrow b)$, nous nous sommes inspirés de la théorie de la « propagation de l'activation » (*spreading activation theory* [Collins et Loftus, 1975]), selon laquelle un marqueur d'information part d'un nœud du réseau (de transitions) et se propage à travers ce réseau avec une certaine force. La force d'un marqueur est fonction du nombre de relations existant entre le terme de départ et tous les termes d'arrivée du marqueur. Cette force s'affaiblit de façon proportionnelle à la distance parcourue par le marqueur. Pour notre part, la force du marqueur partant de `a` pour aller à `b` dans le modèle est définie par la probabilité de transition d'un terme "`a`" vers un terme "`b`".

Étant donné un modèle de connaissances M du domaine, nous définissons une distribution de probabilités qui va servir de base à la mesure de vraisemblance entre deux termes. La distribution de probabilités est calculée en utilisant le chemin de longueur minimale qui relie, dans le modèle M, un terme $\texttt{t}_\texttt{i}$ à un terme $\texttt{t}_\texttt{j}$. La longueur du chemin minimal entre $\texttt{t}_\texttt{i}$ et $\texttt{t}_\texttt{j}$ dans M est notée $d(\texttt{t}_\texttt{i}, \texttt{t}_\texttt{j})$, et la probabilité de transition entre $\texttt{t}_\texttt{i}$ et $\texttt{t}_\texttt{j}$ est calculée en fonction de $d(\texttt{t}_\texttt{i}, \texttt{t}_\texttt{j})$. Il existe deux cas particuliers :

(1) par convention, pour tout terme $\texttt{t}_\texttt{i}$, $\texttt{d}(\texttt{t}_\texttt{i}, \texttt{t}_\texttt{i}) = 1$; ceci pour prendre en compte la réflexivité de la relation de spécialisation estUn et éviter des probabilités anormalement élevées en cas d'absence d'arc sortant (ce qui est illustré par le cas du sommet `c` de l'exemple FIG. 3 (a)) ;

(2) s'il n'existe pas de chemin entre un terme $\mathtt{t_i}$ et un terme $\mathtt{t_j}$, alors $\mathtt{d}(\mathtt{t_i}, \mathtt{t_j}) = \mid \mathtt{M} \mid +1$ où $\mid \mathtt{M} \mid$ est le cardinal de l'ensemble des termes-clés M (M est un ensemble *fini)*.

La vraisemblance entre $\mathtt{t_i}$ et $\mathtt{t_j}$, notée $P_M(\mathtt{t_i}, \mathtt{t_j})$ pour la règle $\mathtt{t_i} \Longrightarrow \mathtt{t_j}$ repose sur le produit de deux facteurs : la distance de $\mathtt{t_i}$ à $\mathtt{t_j}$ et le *poids* de $\mathtt{t_i}$ dans le modèle noté $\delta(\mathtt{t_i})$. En outre, il faut encore tenir compte de deux éléments : (1) plus la distance entre deux termes est grande, plus la valeur de la vraisemblance doit être faible, (2) le poids d'un terme $\mathtt{t_i}$ dépend de l'ensemble des termes du modèle, qu'ils soient atteignables ou non depuis $\mathtt{t_i}$ Ainsi, la formule qui calcule la vraisemblance entre $\mathtt{t_i}$ à $\mathtt{t_j}$ est la suivante :

$$P_M(\mathtt{t_i}, \mathtt{t_j}) = [\mathtt{d}(\mathtt{t_i}, \mathtt{t_j}) \times \delta(\mathtt{t_i})]^{-1} \tag{6}$$

où le poids de $\mathtt{t_i}$ $\delta(\mathtt{t_i}) = \sum_{\mathtt{x} \in \mathtt{M}} 1/\mathtt{d}(\mathtt{t_i}, \mathtt{x})$. Le poids $\delta(\mathtt{t_i})$ d'un terme $\mathtt{t_i}$ est dépendant du nombre d'arcs sortants associés à $\mathtt{t_i}$ dans le modèle M : plus ce nombre est élevé plus le poids est élevé (ce nombre est aussi appelé *facteur de branchement*). À l'inverse, lorsqu'il n'existe aucun arc sortant, l'élément lui-même devient prépondérant car $\mathtt{d}(\mathtt{t_i}, \mathtt{t_i}) = 1$. Il faut encore remarquer que le poids $\delta(\mathtt{t_i})$ est calculé une seule fois pour tout $\mathtt{t_i}$ et que l'équation suivante est vérifiée : $\sum_{\mathtt{x} \in \mathtt{M}} \mathtt{P_M}(\mathtt{t_i}, \mathtt{x}) = 1$.

3.2 Définition de la vraisemblance

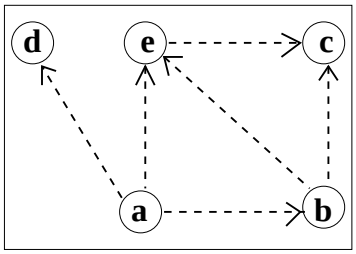
Nous définissons la vraisemblance d'une règle d'association *simple* du type $a \Longrightarrow b$ par la probabilité $P_M(a, b)$. Plus le lien hiérarchique entre a et b est direct dans le modèle, plus la vraisemblance est forte, et donc plus la règle peut être considérée comme triviale pour un analyste. Cela permet de calculer la vraisemblance des règles simples pour lesquelles on suppose que les termes indexant les textes et présents dans les règles sont également décrits dans le modèle, c'est-à-dire que $a, b \in \mathcal{R} \cap \mathcal{H}$.

Prenons l'exemple du modèle de connaissances décrit par la FIG. 3(a) qui doit être interprété de la façon suivante : "a" estUn "b", "e" estUn "c", etc. Sur ce modèle, nous calculons la distribution de probabilités en appliquant la fonction (6) et dont la matrice est donnée dans FIG. 3(b). Cette table nous donne pour tout couple (t_i, t_j) la valeur de $P_M(t_i, t_j)$. Pour un chemin court entre deux termes – par exemple (a,e) – le calcul de la probabilité de transition est le suivant :

$$\begin{aligned}
\mathtt{P_M(a,e)} &= \left[\mathtt{d(a,e)} \times \left(\sum_{\mathtt{x} \in \{\mathtt{a,b,c,d,e}\}} \frac{1}{\mathtt{d(a,x)}}\right)\right]^{-1} \\
&= \left[1 \times \left(\left(\frac{1}{\mathtt{d(a,a)}} + \frac{1}{\mathtt{d(a,b)}} + \frac{1}{\mathtt{d(a,d)}} + \frac{1}{\mathtt{d(a,e)}}\right) + \left(\frac{1}{\mathtt{d(a,c)}}\right)\right)\right]^{-1} \\
&= \left[1 \times \left(\left(4 \times \frac{1}{1}\right) + \left(1 \times \frac{1}{2}\right)\right)\right]^{-1} \\
&= \left[\frac{9}{2}\right]^{-1} = 0.22.
\end{aligned}$$

alors que pour un chemin plus long – par exemple (a,c) – elle donne une valeur plus faible :

$$\mathtt{P_M(a,c)} = \left[\mathtt{d(a,c)} \times \left(\sum_{\mathtt{x} \in \{\mathtt{a,b,c,d,e}\}} \frac{1}{\mathtt{d(a,x)}}\right)\right]^{-1}$$



$\vec{r}$	a	b	c	d	e	Σ
a	0.22	0.22	0.11	0.22	0.22	1
b	0.05	0.30	0.30	0.05	0.30	1
c	0.10	0.10	0.60	0.10	0.10	1
d	0.10	0.10	0.10	0.60	0.10	1
e	0.06	0.06	0.40	0.06	0.40	1

FIG. 3 – (a) Le modèle de connaissances M– (b) Probabilités de transition pour M.

$$= \left[2 \times \left(\left(\frac{1}{\mathtt{d(a,\,a)}} + \frac{1}{\mathtt{d(a,\,b)}} + \frac{1}{\mathtt{d(a,\,d)}} + \frac{1}{\mathtt{d(a,\,e)}}\right) + \left(\frac{1}{\mathtt{d(a,\,c)}}\right)\right)\right]^{-1}$$
$$= \left[2 \times \left(\left(4 \times \frac{1}{1}\right) + \left(1 \times \frac{1}{2}\right)\right)\right]^{-1} = [9]^{-1} = 0.11.$$

Le calcul de vraisemblance pour une règle simple se fait par une lecture de la table des probabilités de transitions dans M de FIG. 3(b). Par exemple, la règle $a \Longrightarrow c$ possède la valeur de vraisemblance dans le modèle M de : $P_M(a \Longrightarrow c) = P_M(a, c) = 0.11$.

Nous étendons dans la section suivante le calcul de vraisemblance des règles d'association par rapport à un modèle M aux règles non simples. En effet, pour que le processus de classification des règles puisse être utilisé sur des données réelles, nous devons résoudre deux problèmes :

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

Pour en faciliter la compréhension, nous dissocions la présentation de ces deux points. La section 3.2.1 suppose que la distribution de probabilités est étendue à $\mathcal{R} \cup \mathcal{H}$ et définit la vraisemblance pour les règles complexes. La section 3.2.2 propose trois possibilités pour étendre la distribution de probabilités et discute de l'impact de ces choix sur la vraisemblance des règles complexes.

3.2.1 La vraisemblance des règles complexes

Supposons à présent que la distribution de probabilités est définie pour tous les couples de termes $(t_k, t_l) \in (\mathcal{R} \cup \mathcal{H}) \times (\mathcal{R} \cup \mathcal{H})$. Le calcul de la vraisemblance pour une règle complexe s'appuie sur le produit cartésien des termes de la partie droite avec ceux de la partie gauche de la règle. Étant donnée une règle complexe $r : t_1 ... t_i \rightarrow t_{i+1} ... t_p$, la probabilité du produit cartésien est :

$$P_M(\mathtt{B} \times \mathtt{H}) = \prod_{(t_k,t_l)\in \mathtt{B}\times\mathtt{H}} P_M(t_k, t_l) \quad (7)$$

qui s'écrit en extension :

$$P_M(\mathtt{B} \times \mathtt{H}) = P_M(t_1, t_{i+1})...P_M(t_1, t_p)...P_M(t_i, t_{i+1})...P_M(t_i, t_p)$$

Plus le nombre de termes présents dans une règle est important, plus la probabilité $P_M(\mathtt{B} \times \mathtt{H})$ est faible. Or le nombre de termes présents dans une règle ne doit pas affecter la vraisemblance d'une règle. L'équation 8 généralise donc l'équation 7 en prenant la moyenne géométrique de la probabilité du produit cartésien. Nous définissons ainsi la vraisemblance d'une règle :

$$P_M(r_i) = \sqrt[|\mathtt{B_i}|\times|\mathtt{H_i}|]{P_M(\mathtt{B_i} \times \mathtt{H_i})} = \sqrt[|\mathtt{B_i}|\times|\mathtt{H_i}|]{\prod_{(t_k,t_l)\in \mathtt{B_i}\times\mathtt{H_i}} P_M(t_k, t_l)} \quad (8)$$

Remarquons que le calcul de vraisemblance donné par la formule (8) pour les règles complexes généralise de façon naturelle la vraisemblance des règles simples puisque $|\mathtt{B}| = |\mathtt{H}| = 1$.

3.2.2 L'extension de la distribution de probabilités

La distribution de probabilités doit être étendue pour traiter deux cas :

1 – La formule (6) qui définit la distribution de probabilités permet de calculer la probabilité d'une transition entre deux termes du modèle de connaissances M qui sont reliés par au moins un chemin. Lorsqu'il n'existe pas de chemin entre deux termes t_k et t_l, la probabilité doit être calculable.

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

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

Probabilité nulle. La première solution consiste à associer une probabilité de transition nulle pour tout couple de termes entre lesquels il n'existe pas de chemin dans le modèle de connaissances. Cette approche est intéressante lorsqu'il s'agit de règles simples. En effet, la valeur `0` permet d'identifier facilement les règles simples non triviales à partir de la matrice des probabilités de transitions. L'analyste peut alors chercher à interpréter une règle non triviale simple et, éventuellement, en déduire qu'il faut

mettre à jour le modèle, c'est-à-dire, introduire un lien trivial entre les deux termes. En revanche, cette méthode défavorise les règles complexes. Il suffit qu'il existe un couple de termes sans transition pour que la vraisemblance de la règle soit nulle. Il n'y a donc pas de continuité dans la vraisemblance.

Valeur fixe arbitraire. Afin de ne pas avoir de vraisemblance nulle, une seconde stratégie consiste à attribuer une valeur fixée arbitrairement petite aux couples de termes pour lesquels il n'existe pas de transition dans M. Cette valeur est la probabilité minimale pour M (maximale en terme de longueur de chemin si nous considérons que tous les termes de M forment une ligne et que les termes t_k et t_l sont respectivement sommet et feuille de $\mathcal{H}$) : $d(t_k, t_l) = \mathtt{N} + 1$ avec $\mathtt{N}$ le nombre de termes de $\mathcal{H}$. Ainsi, pour le terme $\mathtt{b}$: $\mathtt{d(b,a)} = \mathtt{d(b,d)} = 6$ (car $\mathtt{N} = 5$). L'inexistence d'un chemin entre deux termes d'une règle est discutée sur les exemples du couple (b,e) pour la règle $b \Longrightarrow e$ et (b,d) pour une règle éventuelle $b \Longrightarrow d$. Par exemple pour $(\mathtt{b}, \mathtt{c})$, nous avons :

$$\begin{aligned}
\mathtt{P_M(b,e)} &= \left[\mathtt{d(b,e)} \times \left(\sum_{\mathtt{x} \in \{\mathtt{b,c,e,a,d}\}} \frac{1}{\mathtt{d(b,x)}}\right)\right]^{-1} \\
&= \left[1 \times \left(\left(\frac{1}{\mathtt{d(b,b)}} + \frac{1}{\mathtt{d(b,c)}} + \frac{1}{\mathtt{d(b,e)}}\right) + \left(\frac{1}{\mathtt{d(b,a)}} + \frac{1}{\mathtt{d(b,d)}}\right)\right)\right]^{-1} \\
&= \left[1 \times \left(\left(3 \times \frac{1}{1}\right) + \left(2 \times \frac{1}{6}\right)\right)\right]^{-1} = \left[\frac{10}{3}\right]^{-1} = 0.30.
\end{aligned}$$

et pour $(\mathtt{b}, \mathtt{d})$: $P_M(b,d) = \left[\frac{1}{6} \times \left(\left(3 \times \frac{1}{1}\right) + \left(2 \times \frac{1}{6}\right)\right)\right]^{-1} = 0.05$. Cette faible valeur signifie donc bien qu'il n'y a pas de relation hiérarchique entre les termes de la règle. De la même façon, $P_M(a,a)$ est différent de $P_M(b,b)$, ...selon qu'un terme est relié ou pas à un ou plusieurs autres termes du modèle. Par ailleurs, une règle de type "a" $\Longrightarrow$ "a" ne peut pas être extraite par notre processus.

Stratégie mixte. Une troisième stratégie consiste à associer une probabilité nulle dans le cas de règles simples et à appliquer une valeur fixe arbitraire dans le cas de règles complexes.

Critère de choix d'une stratégie. La stratégie d'extension de la distribution de probabilités pour les règles complexes — choix de la valeur nulle ou de la stratégie mixte — dépend de l'objectif de fouille. Si nous choisissons la valeur nulle, nous insistons sur une séparation stricte entre les règles purement triviales (où tous les liens sont triviaux) et les règles ayant au moins un lien non trivial. Dès qu'il y a un couple de termes entre lesquels il n'existe pas de relation triviale dans le modèle, la vraisemblance de la règle est nulle ; inversement, lorsque la vraisemblance est non nulle, tous les couples de termes sont reliés par des relations hiérarchiques. La stratégie mixte permet une continuité et un classement graduel dans laquelle un lien trivial direct peut éventuellement être plus important que l'absence d'un lien dans le modèle. Sur un grand nombre de règles, le but est d'évaluer les règles non triviales pour enrichir le modèle. La valeur nulle peut donc être choisie — celle que nous adoptons pour l'évaluation sur un exemple réel —. Sur un plus petit exemple, la stratégie mixte se révèle intéressante — celle que nous adoptons pour l'évaluation sur un exemple formel —.

4 Étude de la vraisemblance sur l'exemple formel

Reprenons l'exemple précédent des textes $d_1, \ldots, d_6$ décrits par les termes $a, \ldots, e$ en TAB.1 afin d'étudier le comportement de la fonction (8) pour identifier les règles d'association triviales par rapport au modèle formel introduit en FIG. 3(a).

4.1 Comportement de la vraisemblance par rapport au modèle

Le but de cet exemple est de pouvoir évaluer le comportement de la vraisemblance sur un ensemble réduit de règles et un modèle de connaissances de petite taille. Reprenons les vingt règles d'association de TAB. 1 (b). Leur valeur de vraisemblance est calculée en TAB. 4. Par exemple, pour la règle r_6, nous avons :

$$P_M(\mathtt{b} \Longrightarrow \mathtt{a},\mathtt{c},\mathtt{e}) = (P_M(b,a) \times P_M(b,c) \times\ P_M(b,e))^{1/3} = (0.05 \times 0.3 \times 0.3)^{1/3} = 0.165.$$

Nous classons les règles en 8 classes. La colonne de gauche de TAB. 4 contient des règles triviales dites T-règles et la colonne de droite des règles non triviales dites ¬T-règles, c'est-à-dire des règles non triviales qui relient des termes entre lesquels il n'existe pas de lien trivial. Les lignes de cette table regroupent les règles en fonction de leur structure, c'est-à-dire du nombre de termes présents dans B et H : en ligne 1 se trouvent les règles simples triviales, T-règles $(1,1)$, ou simples non triviales ¬T-règles $(1,1)$, en ligne 2 les règles complexes triviales (respectivement non triviales) dont la prémisse est constituée d'un seul terme T-règles $(1,\mathtt{m})$ (respectivement ¬T-règles $(1,\mathtt{m})$), en ligne 3 les règles complexes triviales (respectivement non triviales) dont la conclusion est constituée d'un seul terme T-règles $(\mathtt{n},1)$ (respectivement ¬T-règles $(\mathtt{n},1)$), et enfin, en ligne 4 les règles complexes quelconques triviales T-règles $(\mathtt{n},\mathtt{m})$ et non triviales ¬T-règles $(\mathtt{n},\mathtt{m})$.

TAB. 4 – Mesure de vraisemblance pour les règles de l'exemple FIG.3(a) et le modèle M

n°	T	n/d	$P_M(r)$	n°	¬T	n/d	$P_M(r)$
r_1	b⇒e	0/1	0.300	r_{19}	e⇒b	1/0	0.000
r_{11}	a⇒c	0/0	0.111	r_{20}	c⇒a	1/0	0.000
r_2	b⇒c, e	0/2	0.300				
r_4	a⇒b, c, e	0/2	0.176	r_{13}	d⇒a, c	2/0	0.100
r_6	b⇒a, c, e	1/2	0.165	r_{14}	c⇒b, e	2/0	0.100
r_7	e⇒b, c	1/1	0.163	r_{15}	c⇒a, d	2/0	0.100
r_9	a⇒c, d	0/1	0.157	r_{16}	c⇒a, b, e	3/0	0.100
r_{10}	e⇒a, b, c	2/1	0.121				
r_5	b, c⇒e	1/1	0.173	r_{17}	c, e⇒b	2/0	0.081
r_3	a, b⇒c, e	0/3	0.217	r_{12}	b, c⇒a, e	3/1	0.110
r_8	a, e⇒b, c	1/2	0.160	r_{18}	c, e⇒a, b	4/0	0.081

Nous observons, de façon empirique, un seuil $s = 0.111$ qui sépare les T-règles $(P_M(r_i) \geqslant 0.111)$ des ¬T-règles $(P_M(r_i) < 0.111)$. Ce point sera discuté en section 4.2.

Les T-règles $(1,1)$ sont purement triviales. En accord avec la définition 8 de la vraisemblance, plus la longueur du lien est importante (la longueur est 1 pour r_1 et 2 pour r_{11}), plus la valeur de vraisemblance est faible ($P_M(r_1) > P_M(r_{11})$). Ainsi, r_{11} est moins triviale que r_1. À l'inverse, pour les $\neg$T-règles $(1,1)$, il n'y a pas de chemin de "e" vers "b" (règle r_{19}) ni de "c" vers "e" (règle r_{20}). Nous avons donc $P_M(r_{19}) = P_M(r_{20}) = 0$. Notons que le sens des relations triviales est respecté.

Sur les T-règles $(1,m)$, $(n,1)$ et (n,m) en TAB. 4, nous pouvons vérifier les deux principes duaux découlant des propriétés attendues de la vraisemblance que nous avons définie : (i) moins il y a de liens non triviaux entre les termes de B et de H, plus la valeur de la vraisemblance est élevée, (ii) plus les liens triviaux sont directs, plus la valeur de vraisemblance est élevée. La colonne n/d de TAB. 4 donne le nombre de couples de termes de $B \times H$ qui ne sont pas des relations triviales (noté n) et le nombre de relations triviales directes avec un chemin de longueur 1 (noté d). Par exemple, pour la règle r_8, 1/2 pour n/d signifie que, parmi les $|B| \times |H| = 2 \times 2 = 4$ couples de termes, un couple est relié par une relation non triviale, ici (e, b), et que deux couples sont reliés par une relation triviale directe, ici (a, b) et (e, c). Il y a donc un couple ayant une relation triviale indirecte, ici (a, c). L'analyse de cet exemple formel montre que la vraisemblance permet d'attribuer une valeur forte aux règles triviales ou fortement triviales par rapport au modèle M et une valeur faible aux règles qui sont faiblement triviales.

4.2 Discussion

Les deux colonnes de TAB. 4 séparent les règles triviales des règles non triviales. Ainsi, la question de l'existence d'un seuil s pour la valeur de vraisemblance se pose. Nous montrons, dans cette section, que ce seuil dépend du modèle choisi. Nous allons également caractériser le comportement de notre méthodologie lorsque le modèle évolue, avec le même ensemble de règles $\{r_1, \ldots, r_{20}\}$.

Si nous opérons sur le modèle des modifications majeures, par exemple, en créant un lien trivial entre deux termes (t_u, t_v) intervenant dans le calcul de la vraisemblance d'une règle r, alors l'analyse faite en section 4.1 montre que le calcul de vraisemblance sur le nouveau modèle donne une valeur plus forte pour r.

L'impact de modifications mineures du modèle engendre des changements pour la vraisemblance d'une règle qui sont plus subtils. Nous définissons une modification mineure comme suit. Soient les couples de termes (t_u, t_v) du modèle M qui interviennent dans le calcul de la vraisemblance des différentes règles : s'il existe un chemin entre t_u et t_v, alors le nouveau modèle que nous définissons préserve l'existence d'un chemin (éventuellement différent du chemin dans M). S'il n'existe pas de chemin entre t_u et t_v, alors le nouveau modèle préserve également le fait que ce chemin n'existe pas.

Nous allons voir que :

1. les modifications mineures ont une incidence sur la valeur du seuil ;
2. une règle classée triviale dans M peut se trouver classée parmi les règles non triviales dans un autre modèle.

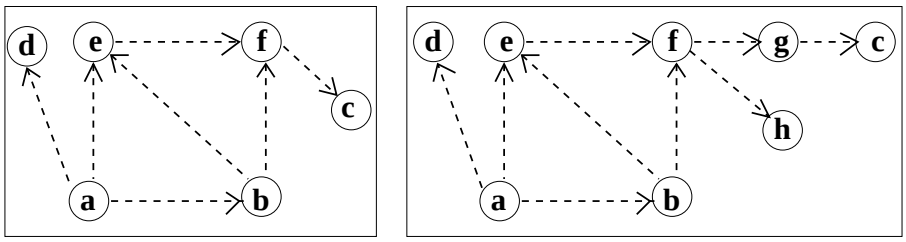


FIG. 4 – Les variantes M_1 et M_2 du modèle de connaissances M de FIG. 3 (a).

Nous introduisons deux modèles M_1 et M_2 (cf. FIG. 4) légèrement différents de M. Pour assurer que les modifications sur le modèle M sont mineures, ces modifications portent sur les termes *puits* (en théorie des graphes [Gondran et Minoux, 1995]), c'est-à-dire des termes qui ne sont à l'origine d'aucune relation avec un autre terme. "c" et "d" vérifient cette propriété. Dans M_1, l'introduction du terme "f" rallonge tous les chemins entre un terme quelconque (différent de "c") et le terme "c". Le fait de n'introduire qu'un seul nouveau terme, augmente faiblement le facteur de branchement. Dans M_2, la longueur des chemins et le facteur de branchement sont augmentés par rapport à M_1.

Dans M, les règles $\{r_1, \ldots, r_{11}\}$ sont classées comme T-règles et les règles $\{r_{12}, \ldots, r_{20}\}$ comme $\neg$T-règles. Dans la mesure où la nature des liens entre les termes dans M_1 et M_2 reste inchangée, nous considérons que la règle r_{12} reste la règle seuil séparant les T-règles et les $\neg$T-règles (*cf.*, TAB. 5). Dans ces conditions, le seuil passe de $s = 0.111$ pour M, à $s_1 = 0.091$ pour M_1 et $s_2 = 0.105$ pour M_2.

Pour les règles où le terme "c" est présent, nous remarquons que :

- la règle r_{10} est considérée comme triviale dans M. r_{10} a deux liens non triviaux (e,a),(e,b) contre un lien trivial direct (e,c). De ce fait, cette règle devrait être non triviale. L'affaiblissement du lien trivial (e,c) dans M_1 suffit à faire passer la règle r_{10} parmi les règles non triviales. *A fortiori*, dans M_2 ;
- la règle r_7 a une valeur de vraisemblance légèrement supérieure, donc plus triviale, dans M que r_{10} puisqu'elle implique un lien non trivial (e, b) pour un lien de relation triviale directe (e,c). Elle reste classée triviale dans M_1 mais devient non triviale dans M_2 ;
- pour M_1 et M_2 , la règle r_{11} purement triviale indirecte passe également parmi les règles non triviales ;
- seules les règles d'association ayant le terme "c" en partie droite changent de statut, ce qui est conforme à nos attentes compte tenu des modifications choisies pour définir M_1 et M_2.

Ces résultats s'analysent comme suit :

1 – la mesure de vraisemblance que nous proposons se comporte de façon cohérente par rapport à sa définition lorsque nous l'appliquons sur les données et que la hiérarchie du modèle subit des modifications « mineures ». Le score de vraisemblance ne

TAB. 5 – Mesures de vraisemblance P_{M_1} et P_{M_2} pour les 20 règles de TAB. 1

n°	T	P_{M_1}	n°	¬T	P_{M_1}
r_1	b⇒e	0.286	**r_{12}**	**b,c⇒a,e**	**0.091**
r_2	b⇒c,e	0.187	r_{13}	d⇒a,c	0.083
r_3	a,b⇒c,e	0.149	r_{14}	c⇒b,e	0.083
r_5	b,c⇒e	0.148	r_{15}	c⇒a,d	0.083
r_4	a⇒b,c,e	0.143	r_{16}	c⇒a,b,e	0.083
r_9	a⇒c,d	0.119	**r_{10}**	**e⇒a,b,c**	0.074
r_6	b⇒a,c,e	0.109	**r_{11}**	**a⇒c**	0.069
r_8	a,e⇒b,c	0.104	r_{17}	c,e⇒b	0.063
r_7	e⇒b,c	0.092	r_{18}	c,e⇒a,b	0.063
			r_{19}	e⇒b	0.000
			r_{20}	c⇒a	0.000

n°	T	P_{M_2}	n°	¬T	P_{M_2}
r_1	b⇒e	0.231	**r_{12}**	**b,c⇒a,e**	**0.105**
r_2	b⇒c,e	0.127	r_{13}	d⇒a,c	0.062
r_5	b,c⇒e	0.117	r_{14}	c⇒b,e	0.062
r_4	a⇒b,c,e	0.116	r_{15}	c⇒a,d	0.062
r_3	a,b⇒c,e	0.108	r_{16}	c⇒a,b,e	0.062
r_9	a⇒c,d	0.092	**r_7**	**e⇒b,c**	0.052
r_6	b⇒a,c,e	0.073	**r_{11}**	**a⇒c**	0.046
r_8	a,e⇒b,c	0.069	**r_{10}**	**e⇒a,b,c**	0.044
			r_{17}	c,e⇒b	0.043
			r_{18}	c,e⇒a,b	0.043
			r_{19}	e⇒b	0.000
			r_{20}	c⇒a	0.000

modifie pas le classement des règles d'association purement triviales ou purement non triviales ;

2 – les valeurs pour le seuil des règles triviales et non triviales dépendent du modèle de connaissances choisi. Par conséquent, les valeurs de seuils ne peuvent être fixées *a priori* ;

3 – si les règles reflètent des connaissances nouvelles, alors le modèle de connaissances peut être enrichi de façon incrémentale. Et l'analyste a le moyen de compléter ce modèle avec les nouveaux termes identifiés grâce aux règles d'association.

5 Expérimentation sur des données textuelles

Nous présentons les résultats de la classification des règles d'association en fonction d'un modèle de connaissances lors d'une expérimentation sur des données textuelles. Ces règles sont extraites à partir d'un ensemble de notices bibliographiques décrivant des articles scientifiques en biologie moléculaire. On y trouve des données et des méta-données codées en XML comme le titre, les auteurs, la date, le statut (publié/non

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 ofloxacin and sparfloxacin 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 of both resistant strains, leading to a Ser83->Ile substitution (Escherichia coli numbering) in the corresponding protein. The gyrB, parC, and parE 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.
Concept(s) : "characterization" "chlamydia trachomatis" "determine region" "dna" "escherichia coli" "gyra gene" "gyrase" "gyrb gene" "mutation" "ofloxacin" "parc gene" "pare gene" "point mutation" "protein" "quinolone" "sparfloxacin" "substitution" "topoisomerase"

FIG. 5 – Un exemple de la notice bibliographique n°391 (texte raccourci).

publié), les termes d'indexation et le résumé (*cf.* FIG. 5). Notre corpus a été constitué à partir de 1 361 notices soit environ 240 000 mots (1,6 M-octets).

Deux champs textuels ont été extraits des notices : le titre et le résumé. Nous traitons ces textes par un processus automatique d'indexation terminologique à partir d'un thésaurus de référence (FASTR [Jacquemin, 1994]) qui extrait les termes et leurs variantes linguistiquement acceptables. Chaque texte peut être représenté par un ensemble de termes et il devient alors possible d'appliquer des algorithmes de fouille de données comme *Apriori* ou *Close*. L'ensemble des textes a été indexé par $|\mathcal{I}| = 632$ termes. Nous avons obtenu 347 règles d'association avec les seuils $\sigma_s = 10$ et $\sigma_c = 0.8$.

Le modèle de connaissances utilisé pour la classification des règles est issu du métathésaurus UMLS [UMLS, 2000]. Il contient quelques 125 000 termes venant d'environ 100 thésaurus médicaux et biologiques. Alors que le métathésaurus contient 11 relations différentes, nous nous sommes limités à la relation estUn ("PAR" : parent). Ce modèle ne couvre $\mathcal{I}$ que partiellement. Au total, 438 termes de $\mathcal{H}$ sont identiques à ceux de $\mathcal{I}$ ($\approx 70\%$ des termes). Le modèle est donc incomplet. Parmi les 347 règles, 136 d'entre elles (soit $\approx 40\%$) ont une vraisemblance non nulle et ce sont toutes des règles complexes. Nous nous sommes focalisés sur les règles de vraisemblance nulle — stratégie de la valeur nulle choisie pour les règles complexes —, soit 211 règles dont 46 simples et 165 complexes. Le seuil s de vraisemblance des règles est donc fixé à 0.

Nous avons réalisé une classification des règles d'association selon le calcul de vraisemblance et nous avons demandé à un analyste d'en évaluer le résultat. Certaines règles classées non triviales par le calcul sont triviales du point de vue de l'analyste en raison de l'incomplétude du modèle. De même, certaines règles classées comme triviales par calcul de la vraisemblance ne sont pas triviales pour l'analyste. Cela provient des probabilités de transition très élevées de certains liens triviaux par rapport à d'autres liens non triviaux.

Pour évaluer la qualité de cette classification par rapport aux attentes de l'analyste, nous devons déterminer 4 classes de règles : les vraies-positives (non triviales $\neg$T par calcul de la vraisemblance et évaluées comme non triviales par l'analyste), les fausses-positives ($\neg$T par calcul, mais qui sont triviales pour l'analyste), les vraies-négatives (triviales par calcul et triviales pour l'analyste) et les fausses-négatives (triviales par

calcul, mais non triviales pour l'analyste).

L'évaluation des règles afin de leur assigner une de ces 4 classes a été réalisée par l'analyste. Parmi les 136 règles qui ont été identifiées comme triviales par notre processus de classification des règles, 122 (soit 90%) sont vraies-négatives et 14 (soit 10%) sont fausses-négatives. Le faible pourcentage de règles fausses-négatives montre la capacité de la mesure de vraisemblance à identifier les règles triviales.

Parmi les 211 règles d'association qui sont identifiées comme étant non triviales ($\neg$T), il y a 115 (soit 55%) qui sont vraies-positives et 96 (soit 44%) qui sont fausses-positives (déjà connues de l'analyste). Le fort pourcentage de règles vraies-positives s'explique par l'incomplétude du modèle disponible dans le domaine traité par les textes. En revanche, le fort pourcentage de fausses-positives nous permet de souligner les termes absents du modèle et de pouvoir l'enrichir de ces nouveaux termes. TAB. 6 résume les résultats obtenus sur nos textes de biologie moléculaire.

TAB. 6 – Confrontation : Connaissances de l'analyste / Calcul de la vraisemblance selon $\mathcal{M}$

	Évaluation par rapport aux connaissances de l'analyste	
Vraisemblance calculée sur $\mathcal{M}$	Triviale	Non triviale
Triviale	90% (vraies négatives)	10% (fausses négatives)
Non triviale	45% (fausses positives)	55% (vraies positives)

6 Approches comparables

De nombreux travaux de recherche en fouille de textes se sont concentrés sur la façon de gérer le très grand nombre de règles d'association extraites à partir de corpus de textes. Cependant, la plupart de ces travaux ont abordé le problème du point de vue statistique, sans chercher à y introduire des connaissances. Les travaux de [Basu *et al.*, 2001] constituent sur ce point une exception puisqu'ils proposent une approche exploitant une base de connaissances pour réduire l'ensemble des règles. Au lieu de s'appuyer sur une approche probabiliste comme la nôtre, ils introduisent une mesure de similarité sémantique entre mots.

Un parallèle peut être fait avec des approches faisant appel à des connaissances explicites de l'analyste. Les connaissances de l'analyste permettent d'élaguer des familles de règles spécifiques qu'il ne souhaite pas extraire ou au contraire de mettre en valeur des règles qui dévient de ses propres connaissances [Sahar, 1999, Liu *et al.*, 2003]. Ces approches sont difficilement reproductibles lorsqu'on change de domaine de fouille. En effet, il faut définir à nouveau, avec l'analyste, les connaissances qu'il ne souhaite pas extraire, et par la suite, solliciter son avis pour l'interprétation des résultats. Le processus de fouille de textes que nous définissons est fondé sur une gestion *a posteriori* des règles extraites par le classement des règles par rapport à un modèle de connaissances du domaine.

Les règles d'association généralisées [Srikant et Agrawal, 1995], [Han, 1995], [Hipp *et al.*, 2002] suivent une approche différente puisque l'extraction des règles exploite le fait que les termes appartiennent à différents niveaux d'une ontologie. Si l'on connaît les ancêtres d'un terme, alors un critère est appliqué afin de contraindre le processus d'extraction (bloquer les règles qui introduisent à la fois un terme et son ancêtre, par exemple). Ce processus reste d'une grande complexité calculatoire et le nombre de règles générées est au final encore plus élevé. Un travail similaire exploitant un modèle de connaissances pour la classification de termes est proposé par [Resnik, 1999]. La similarité est fondée sur l'information mutuelle. Elle sert à désambiguïser (affecter un seul sens) des termes selon la proximité sémantique qu'ils ont avec leurs voisins dans un thésaurus (WORDNET). Les travaux de [Savasere *et al.*, 1998] permettent également d'exploiter les connaissances du domaine (une taxinomie) pour générer les règles dites négatives : $X \not\Longrightarrow Y$. Le nombre de règles négatives est potentiellement plus exponentiel que le nombre de règles d'association que nous manipulons (*i.e.* $X \Longrightarrow Y$). Les auteurs proposent d'utiliser la taxonomie pour confronter les familles de propriétés (*i.e.*, les catégories) qui sont en association pour prédire des fréquences attendues des propriétés élémentaires. Les propriétés qui dévient des fréquences attendues serviront à générer les règles négatives correspondantes.

7 Conclusion et perspectives

Nous proposons une méthodologie de classification des règles d'association en fonction d'un modèle de connaissances qui s'appuie successivement sur un processus symbolique d'extraction de règles d'association et sur un processus probabiliste de calcul d'une mesure de vraisemblance pour classer les règles. Nous avons montré que le comportement de la vraisemblance est de nature différente du comportement des mesures statistiques déjà existantes. Nous avons appliqué cette mesure de vraisemblance pour l'identification de règles non triviales de celles qui sont triviales par rapport à un domaine de spécialité. Nous avons étudié et montré que les propriétés de la mesure de vraisemblance sont cohérentes avec les attentes d'un analyste en fouille de textes. Cette mesure est robuste à des variations légères du modèle. Enfin, la méthodologie que nous avons présentée permet une démarche incrémentale en fouille de textes car le modèle est progressivement enrichi et la valeur de la mesure de vraisemblance d'une règle est modifiée par cet enrichissement.

Le travail actuel peut être étendu dans plusieurs directions. Tout d'abord, nous souhaitons prendre en compte d'autres relations que la relation hiérarchique estUn. Les liens de causalité entre termes sont également transitifs et peuvent, à ce titre, être exploités de la même manière que la relation estUn. En revanche, composer simultanément plusieurs relations de natures différentes peut se révéler impossible car la transitivité n'est pas nécessairement conservée et le calcul des probabilités ne peut plus se faire de façon adéquate. La méthodologie que nous avons présentée ne considère pas les liens pouvant exister entre les termes composant la prémisse ou la conclusion d'une règle extraite. Il nous semble intéressant de construire une variante de la mesure qui prenne en compte les liens entre termes apparaissant du même côté d'une règle. Le choix d'un seuil empirique demeure fixé par jugement de l'analyste. Nous envisageons

de définir un moyen pour apprendre à déterminer ce seuil à partir de la topologie du modèle choisi. Par exemple, la probabilité qu'un terme du modèle apparaisse dans une règle, le nombre de termes dans le modèle et le nombre de termes présents dans les règles peuvent constituer des paramètres pour l'apprentissage du seuil de vraisemblance.

Références

[Agrawal et Srikant, 1994] R. Agrawal et R. Srikant. Fast algorithms for mining association rules in large databases. In *Proc. of the 20th Int'l Conf. on Very Large Databases (VLDB'94)*, pages 478–499, Santiago, Chile, 1994. Extended version : IBM Research Report RJ 9839.

[Azé et Kodratoff, 2004] J. Azé et Y. Kodratoff. Extraction de "pépites" de connaissance dans les données : une nouvelle approche et une étude de la sensibilité au bruit. In H. Briand, M. Sebag, R. Gras, et F. Guillet, editors, *Numéro spécial RNTI-E-1, Mesures de qualité pour la fouille de données*, volume 1 of *Revue des Nouvelles Technologies de l'Information*, pages 247–270. Cépaduès Éditions, Toulouse, 2004.

[Basu *et al.*, 2001] S. Basu, R. J. Mooney, K. V. Pasupuleti, et J. Ghosh. Evaluating the Novelty of Text-Mined Rules using Lexical Knowledge. In *Proc. of the 7th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD'01)*, pages 233–238, San Francisco, 2001. ACM Press.

[Cherfi *et al.*, 2003a] H. Cherfi, A. Napoli, et Y. Toussaint. Towards a Text Mining Methodology Using Frequent Itemsets and Association Rule Extraction. In *Journées d'informatique Messine (JIM'03)*, pages 285–294, Metz, France, 2003. E. SanJuan, INRIA Lorraine.

[Cherfi *et al.*, 2003b] H. Cherfi, A. Napoli, et Y. Toussaint. Vers une méthodologie de fouille de textes s'appuyant sur l'extraction de motifs fréquents et de règles d'association. In R. Gilleron, editor, *Actes de la Conférence d'Apprentissage (CAp'03)*, pages 61–76, Laval, 2003. dans le cadre de la plate-forme (AFIA'03), Presses universitaires de Grenoble.

[Cherfi *et al.*, 2004] H. Cherfi, D. Janetzko, A. Napoli, et Y. Toussaint. Sélection de règles d'association par un modèle de connaissances pour la fouille de textes. In M. Liquière et M. Sebban, editors, *Actes de CAp'04 : Conférence d'Apprentissage*, pages 191–206, Montpellier, 2004. Presses Universitaires de Grenoble.

[Collins et Loftus, 1975] A. Collins et E. Loftus. A Spreading-Activation Theory of Semantic Processing. *Psychological Review*, 82(6) :407–428, 1975.

[Delgado *et al.*, 2002] M. Delgado, M. J. Martin-Bautista, D. Sanchez, et M.A. Vila. Mining Text Data : Special Features and Patterns. In D.J. Hand, N.M. Adams, et R.J. Bolton, editors, *Pattern Detection and Discovery : Proc. of ESF Exploratory Workshop*, volume 2447 of *Lecture Notes in Artificial Intelligence – LNAI*, pages 140–153, London, 2002. Springer-Verlag.

[Fayyad *et al.*, 1996] U.M. Fayyad, G. Piatetsky-Shapiro, et P. Smyth. From data mining to knowledge discovery. *AI Magazine*, 17(3) :37–54, 1996.

[Feldman et Hirsh, 1997] R. Feldman et H. Hirsh. Exploiting Background Information in Knowledge Discovery from Text. *Journal of Intelligent Information Systems*, 9(1) :83–97, 1997.

[Gondran et Minoux, 1995] M. Gondran et M. Minoux. *Graphes et algorithmes (3^{me} édition revue et augmentée)*. Éditions Eyrolles, Paris, 1995.

[Han, 1995] J. Han. Mining Knowledge at Multiple Concept Levels. In *Proc. of 4th the Int'l Conf. on Information and Knowledge Management (CIKM'95)*, pages 19–24, Baltimore, USA, 1995. ACM Press. Invited talk.

[Hipp *et al.*, 2002] J. Hipp, U. Güntzer, et G. Nakhaeizadeh. Data mining of association rules and the process of knowledge discovery in databases. In *Data Mining in E-Commerce, Medicine, and Knowledge Management*, pages 15–36. Springer, Heidelberg, 2002.

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

[Janetzko *et al.*, 2004] D. Janetzko, H. Cherfi, R. Kennke, A. Napoli, et Y. Toussaint. Knowledge-based selection of association rules for text mining. In R. López de Mántaras et L. Saitta, editors, *Proc. of the 16th European Conference on Artificial Intelligence (ECAI'04)*, pages 485–489, Valencia, Spain, 2004. IOS Press.

[Kuntz *et al.*, 2000] P. Kuntz, F. Guillet, R. Lehn, et H. Briand. A User-Driven Process for Mining Association Rules. In D.A. Zighed, H.J Komorowski, et J.M. Zytkow, editors, *Proc. of the 4th Eur. Conf. on Principles of Data Mining and Knowledge Discovery (PKDD'00)*, volume 1910 of *Lecture Notes in Artificial Intelligence – LNAI*, pages 483–489, Lyon, 2000. Springer-Verlag.

[Lavrač *et al.*, 1999] N. Lavrač, P. Flach, et B. Zupan. Rule Evaluation Measures : A Unifying View. In *Proc. of the 9th Int'l Workshop on Inductive Logic Programming (ILP'99)*, volume 1634 of *Lecture Notes in Artificial Intelligence – LNAI*, pages 174–185, Bled, Slovenia, 1999. Springer-Verlag, Heidelberg. Co-located with ICML'99.

[Lehn *et al.*, 2004] R. Lehn, F. Guillet, et H. Briand. Qualité d'un ensemble de règles : élimination des règles redondantes. In H. Briand, M. Sebag, et F. Guillet, editors, *Mesures de Qualité pour la Fouille de Données*, Revue des Nouvelles Technologies de l'Information (RNTI), pages 141–167. Cépaduès Éditions, Toulouse, 2004.

[Liu *et al.*, 2003] B. Liu, Y. Ma, C.K. Wong, et P.S. Yu. Scoring the Data Using Association Rules. *Applied Intelligence*, 18(2) :119–135, 2003.

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

[Resnik, 1999] P. Resnik. Semantic Similarity in a Taxonomy : An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Artificial Intelligence Research*, 11 :95–130, 1999. Morgan Kaufmann Publishers.

[Sahar, 1999] S. Sahar. Interestingness via What is Not Interesting. In S. Chaudhuri et D. Madigan, editors, *Proc. of the 5th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD'99)*, pages 332–336, San Diego, USA, 1999. ACM Press.

[Savasere *et al.*, 1998] A. Savasere, E. Omiecinski, et S. B. Navathe. Mining for Strong Negative Associations in a Large Database of Customer Transactions. In *Proc. of the 14th IEEE Int'l Conf. on Data Engineering (ICDE'98)*, pages 494–502, Orlando, USA, 1998. IEEE Computer Society.

[Srikant et Agrawal, 1995] R. Srikant et R. Agrawal. Mining Generalized Association Rules. In *Proc. of the 21st Int'l Conf. on Very Large Databases (VLDB'95)*, pages 407–419, Zurich, 1995. Morgan Kaufmann Press.

[Tan *et al.*, 2002] P.-N Tan, V. Kumar, et J. Srivastava. Selecting the right interestingness measure for associaton patterns. In *Proc. of the 8th ACM Int'l Conf. on Knowledge Discovery and Data Mining (KDD'02)*, pages 183–193, Edmonton, Canada, 2002. ACM Press.

[UMLS, 2000] UMLS. The Unified Medical Language System. National Library of Medicine, 11^{th} edition, 2000.

Summary

A reoccuring problem in mining association rules is the identification of interesting association rules within the overall, and possibly huge set of extracted rules. The majority of previous work in this area uses statistical methods for quality estimation and classification of association rules. However, strictly bottom-up approaches are oblivious of knowledge, though, rule extraction may profit from the usage of knowledge. In this paper, we conceive of this problem as classification of association rules that represent knowledge which is not available in a formal way. Our methodology uses domain knowledge models to carry out the rule-mining task.