Sélection de règles d'association par un modèle de connaissances pour la fouille de textes

Hacène Cherfi, Dietmar Janetzko, Amedeo Napoli, Yannick Toussaint

Sélection de règles d'association par un modèle de connaissances pour la fouille de textes

Hacène Cherfi[1], Dietmar Janetzko[2],
Amedeo Napoli[1] et Yannick Toussaint[1]

[1] LORIA BP 239 54506 Vandœuvre-lès-Nancy (France)
{cherfi,napoli,yannick}@loria.fr

[2] Institute of Computer Science and Social Research
University of Freiburg
Friedrichstr. 50 D-79098 Freiburg (Germany)
dietmarja@aol.com

Résumé : Parmi les inconvénients d'un processus de fouille de données textuelles fondé sur l'extraction de règles 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 sélection des règles d'association utilisent des méthodes statistiques pour juger de la qualité d'une règle. L'approche standard d'extraction de règles n'utilise pas les connaissances du domaine des données disponibles *a priori*.
Dans cet article, nous évaluons la qualité d'une règle d'association par rapport au modèle de connaissances en définissant une mesure de vraisemblance. Cette vraisemblance mesure l'adéquation des règles extraites au modèle de connaissances du domaine. Nous pouvons classer les règles en deux catégories. D'une part, les règles qui sont strictement conformes au modèle sont dites *triviales* et sont ignorées. D'autre part, les règles qui ne dérivent pas du modèle sont potentiellement porteuses de nouvelles connaissances. Ces règles sont présentées à l'analyste pour être validées et, ensuite, pour enrichir le modèle de connaissances.
Mots-clés : Fouille de textes, Règles d'association, Apprentissage, Raisonnement statistique, Modèle de connaissances.

1 Introduction

Dans le sch´ema g´en´eral de r´ef´erence d'extraction de connaissances à partir de bases de donn´ees introduit dans (Fayyad *et al.*, 1996), le processus de fouille de donn´ees est suivi d'une phase d'interpr´etation. Le processus de fouille que nous adoptons — ici appliqu´e aux textes — cherche à extraire, d'un ensemble de textes, des règles d'association portant sur les termes contenus dans les textes. Au cours de la phase d'interpr´etation, ces règles sont ´evalu´ees par un analyste pour d´eterminer si une nouvelle connaissance peut être ajout´ee au modèle de connaissances. La facilit´e de lecture d'une règle d'association est un point positif pour cette m´ethode de fouille de donn´ees ; en revanche, le

très grand nombre de règles extraites constitue un problème au sens où l'ensemble des règles devient trop diffi cile à appr´ehender pour un humain.

Dans cet article, nous souhaitons classer les règles d'association qui sont pr´esent´ees à l'analyste en les ordonnant par qualit´e d´ecroissante en fonction des connaissances disponibles sur le domaine des donn´ees. Nous consid´erons qu'une règle d'association est de bonne qualit´e si elle contient potentiellement des informations de nature à enrichir le modèle de connaissances. Les autres règles sont qualifi ´ees de triviales puisqu'elles reflètent une connaissance d´ejà pr´esente dans le modèle. Notre objectif est de faciliter la tâche de l'analyste en d´efi nissant une m´ethodologie de s´election des règles de qualit´e et de rejet des règles triviales qui exploite un modèle de connaissances. Notre modèle de connaissances est dans le cas pr´esent un modèle terminologique qui, de façon analogue à un th´esaurus, est caract´eris´e par un r´eseau de termes structur´es hi´erarchiquement par une relation de g´en´eralisation appel´ee EST-UN. Par exemple, une "pomme" EST-UN "fruit". Le sens du lien hi´erarchique suivant la relation EST-UN dans le modèle de connaissances est important. Ainsi, nous consid´erons que la règle "fruit" $\Longrightarrow$ "pomme" n'est pas *triviale* par rapport au même modèle de connaissances et ne doit pas être rejet´ee.

Peu d'approches pour la s´election des règles d'association exploitent un modèle de connaissances. Or, la qualit´e d'une règle ne doit pas être d´efi nie uniquement à l'aide d'indices statistiques mais en ´evaluant l'apport de la règle par rapport aux connaissances du domaine d´ejà acquises : il existe souvent des sources de connaissances (*ex*, des ontologies) disponibles qui peuvent être exploit´ees. Le modèle de connaissances doit donc aider à interpr´eter les r´esultats du processus de fouille et r´eciproquement, les r´esultats de la fouille doivent contribuer à la mise à jour des connaissances. Nous consid´erons ainsi que la construction et l'enrichissement d'un modèle de connaissances sont des processus it´eratifs d'extraction de connaissances.À partir de donn´ees stables au cours des it´erations, l'ensemble des règles d'associations extraites et leurs indices statistiques restent identiques alors que classement des règles suivant notre mesure de qualit´e ´evolue en fonction des mises à jour successives du modèle de connaissances.

La s´election des règles a ´et´e abord´ee par des mesures statistiques (Cherfi *et al.*, 2003). Diverses mesures de qualit´e ont ´et´e propos´ees et font l'objet d'une pr´esentation synth´etique dans (Lavrač *et al.*, 1999; Kuntz *et al.*, 2000; Tan *et al.*, 2002). Cependant, l'accord entre ces mesures statistiques est faible lorsqu'il s'agit d'´evaluer la qualit´e d'une règle. Chaque mesure permet de mettre en valeur un certain type de règles d'association : celles qui portent sur des signaux d'information faibles ; ou bien celles qui ont le moins de contre-exemples et qui sont stables lorsque nous introduisons du bruit dans les donn´ees (Az´e, 2003). Les limites de ces approches viennent du fait que l'´evaluation se fait sans prendre en compte les connaissances du domaine et qu'il est impossible de d´evelopper une approche statistique ind´ependante pour l'´evaluation des règles puisque ces mesures s'appuient sur le même ensemble de donn´ees que le processus d'extraction de règles d'association.

Dans cet article, nous d´ecrivons, en premier lieu, notre processus de fouille de textes qui introduit un modèle de connaissances, sa mise en œuvre probabiliste ainsi que la d´efi nition des règles d'association. Nous d´efi nissons, ensuite, une mesure de qualit´e des règles par rapport au modèle de connaissances que nous appelons la *vraisemblance*.

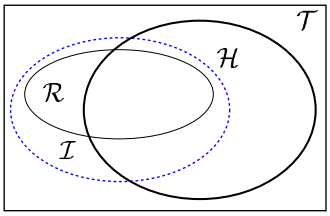


FIG. 1 – *Les différents ensembles de termes :* $\mathcal{T}$ *: ensemble des termes,* $\mathcal{I}$ *: ensemble des termes d'indexation des textes,* $\mathcal{R}$ *: sous-ensemble des termes d'indexation* $\mathcal{I}$ *apparaissant dans les r`egles d'association, et* $\mathcal{H}$ *: ensemble des termes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

Enfi n, nous ´evaluons le comportement de la vraisemblance sur un modèle formel puis sur une exp´erimentation avec des donn´ees textuelles r´eelles.

2 Processus de fouille de textes

2.1 Un modèle terminologique hiérarchique

Soit $\mathcal{T} = \{\mathtt{t}_1, \ldots, \mathtt{t}_n\}$ un ensemble de termes (*cf.* FIG.1) qui sert de vocabulaire de r´ef´erence pour indexer un ensemble de textes — not´e $\mathtt{d}$ pour *document* — $\mathcal{D} = \{\mathtt{d}_1, \ldots, \mathtt{d}_m\}$.

Nous d´efi nissons notre modèle de connaissances comme ´etant un modèle terminologique qui, de façon analogue à un th´esaurus, est caract´eris´e par un r´eseau de termes $\mathcal{H} \subseteq \mathcal{T}$ structur´es hi´erarchiquement par une relation de g´en´eralisation EST-UN. Un terme appartenant à ce modèle peut être seul ou être reli´e par la relation de g´en´eralisation EST-UN à un ou plusieurs autres termes.

La relation EST-UN est d´efi nie sur $\mathcal{H} \times \mathcal{H}$. Cette relation est r´eflexive, antisym´etrique et transitive. La relation EST-UN constitue un ordre partiel. Le modèle ainsi d´efi ni n'est pas n´ecessairement connexe.

2.2 Les règles d'association

Nous introduisons les règles d'association dans le contexte de la fouille de textes. Chaque texte est repr´esent´e par un ensemble de termes qui indexent son contenu. $\mathcal{I}$ d´esigne le sous-ensemble de termes qui indexent au moins un texte.

Une règle d'association est une implication de la forme $\mathtt{B} \overset{\mathtt{P}}{\Longrightarrow} \mathtt{H}$ où $\mathtt{B}$ est la pr´emisse (*body*) et $\mathtt{H}$ 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 pr´emisse d'une règle d'association r et $\mathtt{H} = \{\mathtt{t}_{p+1}, \ldots, \mathtt{t}_q\}$ l'ensemble des termes de la conclusion. $r : \mathtt{t}_1 \ldots \mathtt{t}_p \overset{\mathtt{P}}{\Longrightarrow} \mathtt{t}_{p+1} \ldots \mathtt{t}_q$ signifi e que tous les textes de $\mathcal{D}$ contenant les termes $\mathtt{t}_1$ et $\mathtt{t}_2 \ldots$ et $\mathtt{t}_p$ ont tendance à contenir aussi les termes $\mathtt{t}_{p+1}$ et $\mathtt{t}_{p+2} \ldots$ et $\mathtt{t}_q$ avec une probabilit´e $\mathtt{P}$.

L'extraction de règles d'association a ´et´e utilis´ee pour fouiller des donn´ees du type « panier de la m´enagère », mais elle a ´egalement ´et´e ´etudi´ee pour la fouille de textes (Feldman & Hirsh, 1997; Delgado *et al.*, 2002). Plusieurs algorithmes comme « Apriori » (Agrawal & Srikant, 1994) ou « Close » (Pasquier *et al.*, 1999) permettent de mettre en œuvre le processus d'extraction de règles. En revanche, il est n´ecessaire de trouver un moyen d'identifier dans l'ensemble des règles celles qui sont de bonne qualit´e ou, à l'inverse, un moyen d'´eliminer les règles triviales.

Le support de r est le nombre de textes contenant les termes de $\mathtt{B} \cup \mathtt{H} = \{\mathtt{t_1}, \ldots, \mathtt{t_p}, \ldots, \mathtt{t_q}\}$. La confiance de r est le rapport entre le nombre de textes contenant l'ensemble des termes $\mathtt{B} \cup \mathtt{H}$ et le nombre de textes contenant $\mathtt{B}$ $(\mathtt{t_1}, \ldots, \mathtt{t_p})$. Ce rapport d´efinit la probabilit´e conditionnelle $\mathtt{P}(\mathtt{H}|\mathtt{B})$. Le support et la confiance sont deux mesures associ´ees aux règles d'association (Agrawal & Srikant, 1994) et exploit´ees par les algorithmes d'extraction de règles pour en r´eduire la complexit´e. Deux valeurs de seuil sont d´efinies : g 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 pr´esents dans les règles d'association. Nous d´esignons par $\mathcal{R}$, le sous-ensemble des termes pr´esents, au moins une fois, en partie gauche ou droite, de l'ensemble des règles d'association, *i.e.*, $\mathcal{R} = \bigcup_{\mathtt{r_i}} (\mathtt{B_i} \cup \mathtt{H_i})$.

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

2.3 Les règles triviales

Nous cherchons à supprimer les règles d'association fond´ees sur la relations EST-UN, c'est-à-dire pour lesquelles les termes de B sont li´es par une relation de g´en´eralisation avec les termes de H dans le modèle de connaissances. Par exemple, si le terme "fruit" est plus g´en´eral que le terme "pomme", une règle du type "pomme" $\Longrightarrow$ "fruit" est fortement candidate pour être rejet´ee puisqu'elle exprime une g´en´eralisation connue. En revanche, une règle du type "tarte à la cerise" $\Longrightarrow$ "chocolat", "beurre" doit être conserv´ee. Cette règle exprime potentiellement une relation int´eressante entre "tarte à la cerise", "chocolat" et "beurre" entre lesquels il n'existe pas de lien hi´erarchique EST-UN.

Comment d´efinir ce qu'est une règle triviale par rapport au modèle de connaissances introduit? La construction ou la mise à jour d'un modèle passe par l'ajout de nouveaux termes et l'ajout de relations de g´en´eralisation entre les termes. Une règle d'association $a \Longrightarrow b$ est triviale si la relation de g´en´eralisation que l'analyste peut ajouter au modèle à partir de cette règle, i.e. (a EST-UN b) est d´ejà pr´esente dans le modèle. Il est remarquer que dans le cadre de la fouille de texte, une règle $a \Longrightarrow b$ signifie que chaque fois qu'il y a une occurrence de a dans un texte, il y a (avec une certaine confiance) aussi

une occurrence de b. Cela ne signifie pas formellement que, dans le modèle terminologique, on peut systématiquement ajouter que (a EST-UN b). L'ajout d'une telle relation de généralisation est à l'initiative et sous le contrôle de l'analyste.

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

2.4 Le modèle de connaissances probabiliste

Le modèle de connaissances pour la sélection des règles d'association exploitée dans un modèle probabiliste et construit sur $(\mathcal{H} \times \mathcal{H}$, EST-UN$)$ est notée M. L'objectif est de définir la vraisemblance d'une règle $r : a \Longrightarrow b$ comme la probabilité de trouver un chemin de "a" vers "b" dans le modèle de connaissances (*i.e.*, la probabilité de transition de a vers b).

Pour définir $P_M(a \Longrightarrow b)$, nous définissons une distribution de probabilités qui est conforme à la théorie de la propagation de l'activation, « spreading activation theory » (Collins & Loftus, 1975), selon laquelle un marqueur d'information part d'un nœud du réseau et se propage à travers ce réseau. La force de ce marqueur est fonction du nombre de relations existant entre le terme de départ et les termes d'arrivée du marqueur. De plus cette force s'affaiblit de façon proportionnelle à la distance parcourue par le marqueur. La force du marqueur partant de a pour aller à b dans le modèle est définie par la probabilité de transition d'un terme "a" vers un terme "b" La distribution de probabilités attribue à chaque couple de termes (t_i,t_j) du modèle M une probabilité de transition de t_i vers t_j.

Conformément aux principes de théorie de la propagation de l'activation, nous définissons une distribution de probabilités sur le modèle M par :

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

$$(\mathtt{t_i},\mathtt{t_j}) \longmapsto \left[\ell(\mathtt{t_i},\mathtt{t_j}) \times \left(\sum_{\{\mathtt{x} \in \mathcal{H} \mid \mathtt{t_i Est - un^+ x}\}} \frac{1}{\ell(\mathtt{t_i},\mathtt{x})} \right) \right]^{-1} \qquad (1)$$

où $\ell(\mathtt{t_i},\mathtt{t_j})$ est la longueur minimale des chemins entre $\mathtt{t_i}$ et $\mathtt{t_j}$. Soit $\mathcal{B}_\mathtt{i} = \{\mathtt{x} \in \mathcal{H} \mid \mathtt{t_i Est - un^+ x}\}$ l'ensemble des termes de M reliés à $\mathtt{t_i}$ par un chemin, de longueur $\geqslant 1$. La cardinalité $|\mathcal{B}_\mathtt{i}|$ est appelée par la suite le *facteur de branchement* de $\mathtt{t_i}$.

Le premier facteur de (1) assure que plus le chemin entre $\mathtt{t_i}$ et $\mathtt{t_j}$ est court, plus la probabilité est forte. Le second facteur assure que la somme de toutes les probabilités de transitions de $\mathtt{t_i}$ vers les autres termes auxquels il est relié dans le modèle est égale à 1. Si $\mathtt{t_i}$ est un terme relié par EST-UN à beaucoup d'autres termes du modèle, alors les probabilités de transitions à partir de $\mathtt{t_i}$ seront plus faibles que s'il était relié à peu de termes.

3 Définition de la vraisemblance d'une règle

Nous d´efi nissons la vraisemblance d'une règle d'association *simple* du type $a \Longrightarrow b$ par la probabilit´e $P_M(a,b)$. Plus le lien hi´erarchique entre a et b est fort dans le modèle, plus cette vraisemblance est forte, et donc plus la règle peut être consid´er´ee comme triviale pour un analyste. Nous observons cependant que cette d´efi nition ne permet de calculer que la vraisemblance de règles simples pour lesquelles on suppose que les termes indexant les textes et pr´esents dans les règles sont ´egalement d´ecrits dans le modèle, c'est-à-dire que $a,b \in \mathcal{R} \cap \mathcal{H}$.

Prenons l'exemple du modèle M introduit en FIG.2(a), la distribution de probabilit´es est repr´esent´ee par la matrice donn´ee FIG. 2(b). Cette table nous donne pour tout couple (t_i,t_j) la valeur de $P_M(t_i,t_j)$. Pour un chemin court entre deux termes – par exemple (b,c) – la distribution donne une forte probabilit´e (0,30) alors que pour un chemin long – par exemple (b,d) – elle donne une valeur faible (0,05). Le calcul de vraisemblance pour une règle simple se fait par un simple accès à cette matrice de probabilit´es. Par exemple, la règle $(b \Longrightarrow e)$ a pour vraisemblance dans le modèle M : $P_M(b \Longrightarrow e) = P_M(b,e) = 0{,}30$.

Nous g´en´eralisons dans la section suivante le calcul de vraisemblance des règles d'association par rapport à un modèle M. En effet, pour que le processus de s´election des règles puisse être utilis´e sur des donn´ees r´eelles, nous devons le rendre robuste et r´esoudre deux problèmes :

- les règles simples où $|\mathtt{B}| \times |\mathtt{H}| = 1$ ne repr´esentent qu'un sous-ensemble r´eduit par rapport à l'ensemble des règles extraites. Il est donc imp´eratif de g´en´eraliser la d´efi nition de la vraisemblance pour traiter les règles *complexes* (où $|\mathtt{B}| \times |\mathtt{H}| > 1$) ;
- nous consid´erons qu'un modèle de connaissances peut toujours être enrichi. Nous devons pouvoir calculer la vraisemblance d'une règle même si certains termes de la règle n'appartiennent pas au modèle de connaissances. Nous devons donc ´etendre notre distribution de probabilit´es pour prendre en compte le modèle M (dont les termes sont dans $\mathcal{H}$) et l'ensemble des termes pr´esents dans les règles (que nous avons not´e $\mathcal{R}$). La distribution de probabilit´es doit donc être ´etendue à l'ensemble $\mathcal{R} \cup \mathcal{H}$.

Pour en faciliter la compr´ehension, nous dissocions la pr´esentation de ces deux points. La section 3.1 suppose que la distribution de probabilit´es est ´etendue à $\mathcal{R} \cup \mathcal{H}$ et d´efi nit la vraisemblance pour les règles complexes. La section 3.2 propose trois possibilit´es pour ´etendre la distribution de probabilit´es et discute de l'impact de ces choix sur la vraisemblance des règles complexes.

3.1 La vraisemblance des règles complexes

Supposons à pr´esent que la distribution de probabilit´es est d´efi nie pour tous les couples de termes $(t_k,t_l) \in (\mathcal{R} \cup \mathcal{H}) \times (\mathcal{R} \cup \mathcal{H})$. Le calcul de la vraisemblance pour une règle complexe est fond´e sur le calcul de la probabilit´e dans la th´eorie de la propagation de l'activation pour chaque couple de termes issu du produit cart´esien de la partie droite

avec la partie gauche de la règle. Étant donn´ee une règle complexe $r : t_1...t_i \rightarrow t_{i+1}...t_p$ (où $\mathtt{B} = \{t_1,...,t_i\}$ et $\mathtt{H} = \{t_{i+1},...,t_p\}$), la probabilit´e du produit cart´esien est :

$$P_M(\mathtt{B} \times \mathtt{H}) = \prod_{(t_k,t_l)\in \mathtt{B}\times\mathtt{H}} P_M(t_k,t_l) \qquad (2)$$

qui s'´ecrit en extension :

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

Nous observons cependant que plus le nombre de termes pr´esents dans une règle est important, plus la probabilit´e $P_M(\mathtt{B} \times \mathtt{H})$ est faible. Or le nombre de termes pr´esents dans une règle ne doit pas affecter la vraisemblance d'une règle. L'´equation 3 g´en´eralise donc l'´equation 2 en prenant la moyenne g´eom´etrique de la probabilit´e du produit cart´esien. Nous d´efinissons ainsi la vraisemblance d'une règle :

$$P_M(r_i) = \sqrt[|\mathtt{B}_i|\times|\mathtt{H}_i|]{P_M(\mathtt{B} \times \mathtt{H})} = \sqrt[|\mathtt{B}_i|\times|\mathtt{H}_i|]{\prod_{(t_k,t_l)\in \mathtt{B}\times\mathtt{H}} P_M(t_k,t_l)} \qquad (3)$$

Nous soulignons que le calcul de vraisemblance de l'´equation 3 pour les règles complexes est ´egalement applicable aux règles simples puisque $|\mathtt{B}| \times |\mathtt{H}| = 1$.

3.2 L'extension de la distribution de probabilités

La distribution de probabilit´es qui a ´et´e introduite en paragraphe 2.4 doit être ´etendue pour traiter deux cas de figure :

1 – L'´equation 1 (section 2.4) qui d´efinit la distribution de probabilit´es permet de calculer la probabilit´e d'une transition entre deux termes du modèle de connaissances M qui sont reli´es par au moins un chemin. Lorsqu'il n'existe pas de chemin entre deux termes t_k et t_l, la probabilit´e n'est pas calculable.

2 – Il existe des termes pr´esents dans les règles d'association qui ne font pas (encore) partie du modèle de connaissances M. Ce sont les termes $t \in \mathcal{R} \setminus \mathcal{H}$. Pour les prendre en compte, il est possible d'´etendre la distribution de probabilit´es 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 isol´e : aucune relation n'a ´et´e d´efinie dans le modèle pour ce terme. Ce point nous ramène donc au problème ´evoqu´e au point 1 ci-dessus.

Nous consid´erons à pr´esent que le modèle de connaissances est ´etendu de $\mathcal{H}$ à $\mathcal{R} \cup \mathcal{H}$. Trois strat´egies 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 probabilit´e de transition nulle pour tout couple de termes entre lesquels il n'existe pas de chemin dans le modèle de connaissances. Cette approche est int´eressante lorsqu'il s'agit de règles simples. En effet, la

valeur 0 permet d'identifier facilement les règles simples taxinomiques à partir de la matrice des probabilités de transitions. L'analyste peut alors chercher à interpréter une règle taxinomique simple et, éventuellement, en déduire qu'il faut mettre à jour le modèle, c'est-à-dire, introduire un lien taxinomique entre les deux termes. En revanche, cette méthode défavorise les règles complexes. Il suffit qu'il existe un couple de termes sans transition pour que la vraisemblance de la règle soit nulle. Il n'y a donc pas de continuité dans la vraisemblance. Dès qu'il y a un couple de termes non taxinomique, la vraisemblance est nulle ; inversement, lorsque la vraisemblance est non nulle, tous les couples de termes sont taxinomiques.

Valeur de pénalité :

Afin de réduire le nombre de règles dont la vraisemblance serait nulle, une seconde stratégie consiste à attribuer une valeur de pénalité aux couples de termes pour lesquels il n'existe pas de transition. Cette valeur est la probabilité minimale pour M : $\mathtt{P_M(t_k,t_l)} = \frac{1}{n+1}$ avec n le nombre de termes de $\mathcal{H}$. Ainsi, dans FIG. 2 (b), $P_M(\mathtt{b,c}) = \frac{1}{\mathtt{n}+1} = \frac{1}{6}$ car $\mathtt{n} = 5$.
Par exemple pour (b,c), nous avons : $P_M(\mathtt{b,c}) = 1 \times \left(\left(3 \times \frac{1}{1}\right) + \left(2 \times \frac{1}{6}\right)\right)^{-1} = 0{,}3$
et pour (b,d) : $P_M(\mathtt{b,d}) = \frac{1}{6} \times \left(\left(3 \times \frac{1}{1}\right) + \left(2 \times \frac{1}{6}\right)\right)^{-1} = 0{,}05$.

Stratégie mixte :

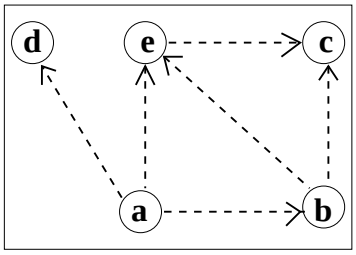
Une troisième stratégie — celle que nous adoptons dans la suite de l'article — consiste à associer une probabilité nulle dans le cas de règles simples et à appliquer une valeur de pénalité dans le cas de règles complexes.

4 Exemple formel

Prenons un exemple formel repris de (Pasquier *et al.*, 1999) afin d'étudier le comportement de l'équation (3) pour identifier les règles d'association triviales. Soit un modèle de connaissances décrit par la FIG. 2(a). Ce modèle doit être interprété de la façon suivante : "a" EST-UN "b", "e" EST-UN "c", etc. Chaque nœud est relié à lui-même par une relation réflexive. Sur ce modèle, nous construisons la distribution de probabilités dont la matrice est donnée par FIG. 2(b). Pour ce modèle, nous présentons un ensemble de textes $\{d_1, \ldots, d_6\}$ décrits par un ensemble de termes d'indexation $\{a, \ldots, e\}$ (*cf.* FIG. 3(1)).

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. Vingt règles d'association, numérotées $r_1, \ldots, r_{20}$, sont extraites avec un support minimal $\sigma_\mathtt{s} = 1$ et une confiance minimale $\sigma_\mathtt{c} = 0{,}1$ et leurs valeurs de vraisemblance sont calculées à la FIG. 3 (2). Par exemple, pour la règle r_6, nous avons :



↱	a	b	c	d	e
a	0.22	0.22	0.11	0.22	0.22
b	0.05	0.30	0.30	0.05	0.30
c	0.10	0.10	0.60	0.10	0.10
d	0.10	0.10	0.10	0.60	0.10
e	0.06	0.06	0.40	0.06	0.40

FIG. 2 – *(a) Le mod`ele de connaissances M– (b) Probabilit´es de transition pour M.*

$P_M(\mathtt{b} \Longrightarrow \mathtt{a,c,e}) = (P_M(b,a) \times P_M(b,c) \times P_M(b,e))^{1/3} = (0{,}05 \times 0{,}3 \times 0{,}3)^{1/3} = 0{,}165$.

Nous classons ces règles en 8 classes. La colonne de gauche de FIG. 3 (2) contient des règles taxinomiques T-règles et la colonne de droite des règles non taxinomiques ¬T-règles, c'est-à-dire des règles non triviales qui relient des termes entre lesquels il n'existe pas de lien taxinomique. Les lignes de cette table regroupent les règles en fonction de leur structure, i.e., le nombre de termes pr´esents dans B et H: en ligne 1 se trouvent les règles simples T-règles (1,1) taxinomiques et ¬T-règles (1,1) non taxinomiques, en ligne 2 les T-règles (1,n) et ¬T-règles (1,n), puis (n,1) et (n,m) avec $\mathtt{n}, \mathtt{m} > 1$.

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

n°	T	n/d	$P_M(r)$	n°	¬T	n/d	$P_M(r)$
r_1	b⇒e	0/1	0,300	r_{19}	e⇒b	1/0	0,000
r_{11}	a⇒c	0/0	0,111	r_{20}	c⇒a	1/0	0,000
r_2	b⇒c, e	0/2	0,300				
r_4	a⇒b, c, e	0/2	0,176	r_{13}	d⇒a, c	2/0	0,100
r_6	b⇒a, c, e	1/2	0,165	r_{14}	c⇒b, e	2/0	0,100
r_7	e⇒b, c	1/1	0,163	r_{15}	c⇒a, d	2/0	0,100
r_9	a⇒c, d	0/1	0,157	r_{16}	c⇒a, b, e	3/0	0,100
r_{10}	e⇒a, b, c	2/1	0,121				
r_5	b, c⇒e	1/1	0,173	r_{17}	c, e⇒b	2/0	0,081
r_3	a, b⇒c, e	0/3	0,217	r_{12}	b, c⇒a, e	3/1	0,110
r_8	a, e⇒b, c	1/2	0,160	r_{18}	c, e⇒a, b	4/0	0,081

FIG. 3 – *(1) La base de donn´ees textuelles – (2) Mesure de vraisemblance pour les r`egles de l'exemple* FIG.*2(a) et le mod`ele M.*

Nous observons, de façon empirique, un seuil $s = 0{,}111$ qui s´epare les T-règles ($P_M(r_i) \geqslant 0{,}111$) des ¬T-règles ($P_M(r_i) < 0{,}111$). Ce point sera discut´e en section 4.2.

Les T-règles (1,1) sont purement taxinomiques. En accord avec la d´efi nition 3 de la vraisemblance, plus la longueur du lien taxinomique est importante (la longueur est 1 pour r_1 et 2 pour r_{11}), plus la valeur de vraisemblance est faible ($P_M(r_1) > P_M(r_{11})$). Ainsi, r_{11} est moins triviale que r_1 (selon la propri´et´e en § 2.4). À l'inverse, pour les ¬T-règles (1,1), il n'y a pas de chemin de "e" vers "b" (règle r_{19}) ni de "c" vers "e" (règle r_{20}). Nous avons donc $P_M(r_{19}) = P_M(r_{20}) = 0$. Notons que le sens des relations

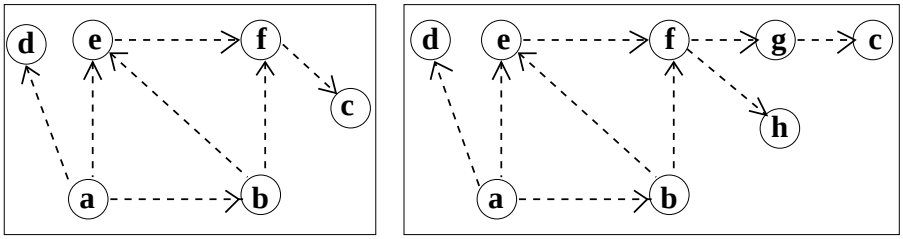


FIG. 4 – *Les variantes M_1 et M_2 du modèle de connaissances M de* FIG. *2 (a).*

taxinomiques est respect´e.

Les T-règles (1,n), (n,1) et (n,m) de FIG. 3 (2), nous pouvons v´erifier deux principes d´ecoulant des propri´et´es attendues de la vraisemblance que nous avons d´efinie : (i) moins il y a de liens non taxinomiques entre les termes de B et de H, plus la valeur de la vraisemblance est ´elev´ee. (ii) plus les liens taxinomiques sont directs, plus la valeur de vraisemblance est ´elev´ee. La colonne n/d de FIG. 3 (2) donne le nombre de couples de termes de $\mathtt{B} \times \mathtt{H}$ qui ne sont pas des relations taxinomiques (not´e n) et le nombre de relations taxinomiques directes avec un chemin de longueur 1 (not´e d). Par exemple, pour la règle r_8, 1/2 pour n/d signifie que, parmi les $|\mathtt{B}| \times |\mathtt{H}| = 2 \times 2 = 4$ couples de termes, un couple est non taxinomique, *i.e.* (e,b) et que deux couples sont taxinomiques directs, *i.e.* (a,b) et (e,c). Il y a donc un couple taxinomique indirect, *i.e.* (a,c). L'analyse de cet exemple formel montre que la vraisemblance permet d'attribuer une valeur forte aux règles triviales par rapport au modèle M et une valeur faible aux règles qui sont faiblement taxinomiques.

4.2 Discussion

Les deux colonnes de FIG. 3 (2) s´eparent les règles taxinomiques des règles non taxinomiques. Cependant, la question de l'existence d'un seuil s pour la valeur de vraisemblance se pose. Nous montrons, dans cette section, que ce seuil ne peut être d´efini formellement et d´epend du modèle choisi. Nous souhaitons ´egalement caract´eriser le comportement de notre m´ethodologie lorsque le modèle ´evolue. Pour cela, nous prenons le même ensemble de règles $\{r_1, \ldots, r_{20}\}$.

Si nous op´erons sur le modèle des modifications majeures, par exemple, en cr´eant un lien taxinomique entre deux termes (t_u, t_v) intervenant dans le calcul de la vraisemblance d'une règle r, alors l'analyse faite en section 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 d´efinissons une modification mineure comme suit : prenons les couples de termes (t_u, t_v) du modèle M qui interviennent dans le calcul de la vraisemblance des diff´erentes règles. S'il existe un chemin entre c_u et c_v, alors le nouveau modèle que nous d´efinissons pr´eserve l'existence d'un chemin (´eventuellement diff´erent du chemin dans M). S'il n'existe pas de chemin entre

TAB. 1 – *Mesures P_{M_1} (à gauche) et P_{M_2} (à droite) pour les* 20 *règles de* TAB. *3*

n°	T	P_{M_1}	n°	¬T	P_{M_1}	n°	T	P_{M_2}	n°	¬T	P_{M_2}
r_1	b⇒e	0,286	**r_{12}**	**b,c⇒a,e**	**0,091**	r_1	b⇒e	0,231	**r_{12}**	**b,c⇒a,e**	**0,105**
r_2	b⇒c,e	0,187	r_{13}	d⇒a,c	0,083	r_2	b⇒c,e	0,127	r_{13}	d⇒a,c	0,062
r_3	a,b⇒c,e	0,149	r_{14}	c⇒b,e	0,083	r_5	b,c⇒e	0,117	r_{14}	c⇒b,e	0,062
r_5	b,c⇒e	0,148	r_{15}	c⇒a,d	0,083	r_4	a⇒b,c,e	0,116	r_{15}	c⇒a,d	0,062
r_4	a⇒b,c,e	0,143	r_{16}	c⇒a,b,e	0,083	r_3	a,b⇒c,e	0,108	r_{16}	c⇒a,b,e	0,062
r_9	a⇒c,d	0,119	**r_{10}**	**e⇒a,b,c**	0,074	r_9	a⇒c,d	0,092	**r_7**	**e⇒b,c**	0,052
r_6	b⇒a,c,e	0,109	**r_{11}**	**a⇒c**	0,069	r_6	b⇒a,c,e	0,073	**r_{11}**	**a⇒c**	0,046
r_8	a,e⇒b,c	0,104	r_{17}	c,e⇒b	0,063	r_8	a,e⇒b,c	0,069	**r_{10}**	**e⇒a,b,c**	0,044
r_7	e⇒b,c	0,092	r_{18}	c,e⇒a,b	0,063				r_{17}	c,e⇒b	0,043
			r_{19}	e⇒b	0,000				r_{18}	c,e⇒a,b	0,043
			r_{20}	c⇒a	0,000				r_{19}	e⇒b	0,000
									r_{20}	c⇒a	0,000

c_u et c_v, alors le nouveau modèle pr´eserve ´egalement le fait que ce chemin n'existe pas.

Nous souhaitons montrer avec l'ensemble $\{r_1, \ldots, r_{20}\}$ que :

1. ces modifications mineures ont une incidence sur la valeur du seuil ;
2. une règle class´ee dans M comme taxinomique peut se trouver class´ee parmi les règles non taxinomiques.

Nous introduisons deux modèles M_1 et M_2 (cf. FIG. 4) l´egèrement diff´erents de M. Pour assurer que les modifications sur le modèle M sont mineures, ces modifications portent sur les termes *puits* (en th´eorie des graphes (Gross & Yellen, 2003)), *i. e.*, des termes qui ne sont à l'origine d'aucune relation avec un autre terme. "c" et "d" v´erifient cette propri´et´e. Dans M_1, l'introduction du terme "f" rallonge tous les chemins entre un terme quelconque (diff´erent 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 augment´es par rapport à M_1.

Nous observons l'´evolution des valeurs de vraisemblance affect´ees aux règles d'association $\{r_1, \ldots, r_{20}\}$. Dans M, les règles $\{r_1, \ldots, r_{11}\}$ ´etaient class´ees comme T-règles et les règles $\{r_{12}, \ldots, r_{20}\}$ comme ¬T-règles. Dans la mesure où la nature des liens entre termes dans M_1 et M_2 reste inchang´ee, nous consid´erons que la règle r_{12} reste la règle seuil s´eparant les T-règles et les ¬T-règles. Dans ces conditions, on observe un abaissement du seuil de $s = 0{,}111$ pour M, à $s_1 = 0{,}091$ pour M_1 et $s_2 = 0{,}105$ pour M_2.

Nous observons, particulièrement les règles où le terme "c" est pr´esent et nous remarquons que :

– la règle r_{10} est consid´er´ee comme taxinomique dans M. r_{10} a deux liens non taxinomiques ((e,a),(e,b)) contre un lien taxinomique direct (e,c). De ce fait, cette règle devrait être non taxinomique. L'affaiblissement du lien taxinomique (e,c) dans M_1 suffit à faire passer la règle r_{10} parmi les règles non taxinomiques. *A fortiori*, dans M_2 ;

– la règle r_7 a une valeur de vraisemblance l´egèrement sup´erieure, *i.e.*, plus taxinomique, dans M que r_{10} puisqu'elle implique un lien non taxinomique (e,b) pour

un lien taxinomique direct (e,c). Elle reste class´ee taxinomique dans M_1 mais devient non taxinomique dans M_2 ;

– pour M_1 et M_2 , la règle r_{11} purement taxinomique indirecte passe ´egalement parmi les règles non taxinomiques ;

– seules les règles d'association ayant le terme "c" en partie droite H changent de statut, ce qui est conforme à nos attentes compte tenu des modifi cations choisies pour d´efi nir M_1 et M_2.

Ces r´esultats s'analysent comme suit :

1 – la mesure de vraisemblance que nous proposons se comporte de façon coh´erente par rapport à sa d´efi nition lorsque nous l'appliquons sur les donn´ees et que la hi´erarchie du modèle subit des modifi cations « mineures ». Le score de vraisemblance ne modifi e pas le classement des règles d'association purement taxinomiques ou purement non taxinomiques ;

2 – les valeurs pour le seuil des règles taxinomiques et non taxinomiques ne sont pas ind´ependantes du modèle de connaissances choisi. Par cons´equent, les valeurs de seuils ne peuvent être fi x´ees *a priori* ;

3 – si les règles pr´esentent des connaissances nouvelles, alors le modèle de connaissances peut être enrichi de façon incr´ementale. Et nous avons le moyen de compl´eter ce modèle avec de nouveaux termes identifi ´es grâce aux règles d'association.

5 Expérimentation sur des données textuelles

Nous pr´esentons les r´esultats de la s´election des règles d'association par un modèle de connaissances pour une exp´erimentation sur des donn´ees textuelles du monde r´eel. Ces règles sont extraites à partir d'un ensemble de notices bibliographiques d´ecrivant des articles scientifi ques en biologie mol´eculaire. On y trouve des donn´ees et des m´etadonn´ees cod´ees en XML comme le titre, les auteurs, la date, le statut (publi´e/non publi´e), les termes d'indexation et le r´esum´e (*cf.* FIG. 5). Notre corpus a ´et´e constitu´e à partir de 1 361 notices soit environ 240 000 mots (1,6 M-octets).

Deux champs textuels ont ´et´e extraits des notices : le titre et le r´esum´e. Nous traitons ces textes par un processus automatique d'indexation terminologique à partir d'un th´esaurus de r´ef´erence (FASTR (Jacquemin, 1994)) qui extrait les termes et leurs variantes linguistiquement acceptables. Chaque texte peut être repr´esent´e par un ensemble de termes et il est possible d'appliquer des algorithmes de fouille de donn´ees classiques comme *Apriori* ou *Close*. L'extraction des règles d'association a ´et´e r´ealis´ee en s'appuyant sur l'algorithme *Close* (Pasquier *et al.*, 1999). *Close* fait une recherche par niveau dans un tableau bool´een d´ecrivant le produit cart´esien $\mathcal{T} \times \mathcal{I}$ des motifs fr´equents. L'algorithme commence par le plus petit motif fr´equent ferm´e et calcule de façon incr´ementale les motifs ferm´es plus long dans $\mathcal{I}$. Un motif est ferm´e s'il correspond à un ensemble maximal de termes partag´e par un ensemble de textes. Le motif est fr´equent s'il apparaît dans au moins σ_s textes. Une fois les motifs ferm´es fr´equents calcul´es, les règles d'association en sont d´eriv´ees.

L'ensemble des textes a ´et´e index´e 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

Document: **#391**
Titre: Sequencing of gyrase and topoisomerase IV quinolone-resistance-determining regions of Chlamydia trachomatis and characterization of quinolone-resistant mutants obtained In vitro.
Auteur(s): Dessus-Babus-S; Bebear-CM; Charron-A; Bebear-C; de-Barbeyrac-B
Résumé: The L2 reference strain of Chlamydia trachomatis was exposed to subinhibitory concentrations of ofbxacin and sparfbxacin 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 ofbxacin and sparfbxacin.
Concept(s): "characterization" "chlamydia trachomatis" "determine region" "dna" "escherichia coli" "gyra gene" "gyrase" "gyrb gene" "mutation" "ofbxacin" "parc gene" "pare gene" "point mutation" "protein" "quinolone" "sparfbxacin" "substitution" "topoisomerase"

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

utilis´e pour la s´election des règles est issu du m´etath´esaurus UMLS (UMLS, 2000). Il contient quelques 125 000 termes venant d'environ 100 th´esaurus m´edicaux et biologiques. Alors que le m´etath´esaurus contient 11 relations diff´erentes, nous nous sommes limit´es aux relations de type EST-UN ("PAR": parent). Ce modèle ne couvre $\mathcal{I}$ que partiellement. Au total, 438 termes de $\mathcal{H}$ sont identiques à ceux de $\mathcal{I}$ ($\approx$ 70% des termes). Le modèle est donc incomplet. Parmi les 347 règles, 136 d'entre elles (soit $\approx$ 40%) ont une vraisemblance $\neq 0$ et ce sont toutes des règles complexes. Nous nous sommes focalis´es sur des règles de vraisemblance nulle soit 211 règles dont 46 simples et 165 complexes.

Nous avons r´ealis´e une classifi cation des règles d'association en règles triviales / non triviales. Certaines règles non rejet´ees sont triviales mais l'incompl´etude du modèle n'a pas permis de les identifi er. De même, certaines règles rejet´ees ne sont pas triviales en raison des probabilit´es de transition très ´elev´ees de certains liens taxinomiques par rapport à d'autres liens non taxinomiques.

Pour ´evaluer la qualit´e de cette classifi cation, nous devons d´eterminer 4 classes de règles : les vraies-positives (non taxinomiques $\neg$T et ´evalu´ees comme non triviales), les fausses positives ($\neg$T, mais qui sont triviales), les vraies-n´egatives (taxinomiques et triviales) et les fausses-n´egatives (taxinomiques, mais non triviales).

L'´evaluation des règles afi n de leur assigner une de ces 4 classes a ´et´e r´ealis´ee par l'analyste. Parmi les 136 règles qui ont ´et´e rejet´ees par notre processus de s´election des règles, 122 (soit 90%) sont vraies-n´egatives et 14 (soit 10%) sont fausses-n´egatives. Le faible pourcentage de règles fausses-n´egatives montre la robustesse de la mesure de vraisemblance pour identifi er les règles taxinomiques.

Parmi les 211 règles d'association qui sont non rejet´ees ($\neg$T), il y a 115 (soit 55%) qui sont vraies-positives et 96 (soit 44%) qui sont fausses-positives, *i.e.*, d´ejà connues de l'analyste). Le fort pourcentage de règles vraies-positives s'explique par l'incompl´etude du modèle disponible dans le domaine trait´e 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.

6 Approches comparables

De nombreux travaux de recherche en fouille de textes se sont concentr´es sur la façon de g´erer le très grand nombre de règles d'association extraites à partir de corpus de textes. Cependant, la plupart de ces travaux ont abord´e 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 r´eduire l'ensemble des règles. Au lieu de s'ancrer dans une approche probabiliste comme la nôtre, ils introduisent une mesure de similarit´e s´emantique entre mots.

Les règles d'association g´en´eralis´ees (Srikant & Agrawal, 1995; Han, 1995; Hipp *et al.*, 2002) constituent une approche diff´erente puisque l'extraction des règles exploite le fait que les termes appartiennent à diff´erents niveaux d'une ontologie. Si l'on connaît les ancêtres d'un terme, alors un critère est appliqu´e afi n 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 complexit´e calculatoire et le nombre de règles g´en´er´e est au fi nal encore plus ´elev´e. Enfi n, un travail similaire exploitant un modèle de connaissances pour la classifi cation de termes est propos´e par (Resnik, 1999). La similarit´e est fond´ee sur l'information mutuelle. Elle sert à d´esambiguïser (affecter un seul sens) des termes ambigus selon la proximit´e s´emantique qu'ils ont avec leurs voisins dans un th´esaurus (*i.e.*, WORDNET).

7 Conclusion et perspectives

Nous proposons une m´ethodologie de s´election des règles d'association par l'utilisation d'un modèle de connaissances. Cette m´ethodologie applique successivement un processus symbolique d'extraction de règles d'association et un processus probabiliste de calcul d'une mesure de vraisemblance entre une règle et un modèle de connaissances. Nous avons appliqu´e cette mesure pour la s´election des règles non taxinomiques en rejetant celles qui sont taxinomiques car elles sont triviales dans un domaine de sp´ecialit´e. Nous avons ´etudi´e et montr´e que les propri´et´es de la mesure de vraisemblance est coh´erente avec les attentes d'un analyste en fouille de textes. Cette mesure est robuste à des variations l´egères du modèle. Enfi n, la m´ethodologie que nous avons pr´esent´ee permet une d´emarche incr´ementale en fouille de textes car le modèle est progressivement enrichi et la valeur de la mesure de vraisemblance d'une règle est modifi ´ee par cet enrichissement.

Le travail actuel peur être ´etendu dans plusieurs directions. Tout d'abord, nous souhaitons prendre en compte d'autres relations que EST-UN. Les liens de causalit´e reposent sur une transitivit´e des relations entre termes et peuvent, à ce titre, être exploit´es de la même manière que la relation taxinomique EST-UN. En revanche, composer simultan´ement plusieurs relations de natures diff´erentes peut se r´ev´eler d´elicat car il faut consid´erer et manipuler la transitivit´e dans les probabilit´es de transition de façon à garder une coh´erence de la s´emantique de ces relations. La m´ethodologie que nous avons pr´esent´ee ne considère pas les liens existant à l'int´erieur de `B` ou à l'int´erieur de `H`. Il nous semble int´eressant de proposer une variante de cette mesure qui prend en compte

les liens entre termes apparaissant du même côt´e d'une règle. Le choix d'un seuil empirique demeure d´elicat. Il est fi x´e par jugement de l'analyste. Nous envisageons de d´efi nir un moyen pour apprendre ce seuil à partir de la topologie du modèle choisi. Par exemple, la probabilit´e qu'un terme du modèle apparaisse dans une règle, le nombre de termes dans le modèle et le nombre de termes pr´esents dans les règles peuvent constituer des paramètres pour l'apprentissage du seuil de vraisemblance.

Références

AGRAWAL R. & SRIKANT R. (1994). Fast algorithms for mining association rules in large databases. In *Proc. of the 20^{th} Int'l Conf. on Very Large Databases (VLDB'94)*, p. 478–499, Santiago, Chile. Extended version: IBM Research Report RJ 9839.

AZÉ J. (2003). Une nouvelle mesure de qualité pour l'extraction de pépites de connaissances. *Extraction des connaissances et apprentissage (ECA)*, **17**(1), 171–182.

BASU S., MOONEY R. J., PASUPULETI K. V. & GHOSH J. (2001). Evaluating the Novelty of Text-Mined Rules using Lexical Knowledge. In *Proc. of the 7^{th} ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD'01)*, p. 233–238, San Francisco: ACM Press.

CHERFI H., NAPOLI A. & TOUSSAINT Y. (2003). Vers une méthodologie de fouille de textes s'appuyant sur l'extraction de motifs fréquents et de règles d'association. In R. GILLERON, Ed., *Actes de la Conf´erence d'Apprentissage (CAp'03)*, p. 61–76, Laval: dans le cadre de la plate-forme (AFIA'03) Presses universitaires de Grenoble.

COLLINS A. & LOFTUS E. (1975). A Spreading-Activation Theory of Semantic Processing. *Psychological Review*, **82**(6), 407–428.

DELGADO M., MARTIN-BAUTISTA M. J., SANCHEZ D. & VILA M. (2002). Mining Text Data: Special Features and Patterns. In D. HAND, N. ADAMS & R. BOLTON, Eds., *Pattern Detection and Discovery: Proc. of ESF Exploratory Workshop*, volume 2447 of *Lecture Notes in Artificial Intelligence – LNAI*, p. 140–153, London: Springer-Verlag.

FAYYAD U., PIATETSKY-SHAPIRO G. & SMYTH P. (1996). From data mining to knowledge discovery. *AI Magazine*, **17**(3), 37–54.

FELDMAN R. & HIRSH H. (1997). Exploiting Background Information in Knowledge Discovery from Text. *Journal of Intelligent Information Systems*, **9**(1), 83–97.

GROSS J. & YELLEN J. (2003). *Handbook of Graph Theory*, volume 25 of *Discrete Mathematics and Its Applications*. New York: CRC Press. 1192 pages.

HAN J. (1995). Mining Knowledge at Multiple Concept Levels. In *Proc. of 4^{th} the Int'l Conf. on Information and Knowledge Management (CIKM'95)*, p. 19–24, Baltimore, USA: ACM Press. Invited talk.

HIPP J., GÜNTZER U. & NAKHAEIZADEH G. (2002). Data mining of association rules and the process of knowledge discovery in databases. In *Data Mining in E-Commerce, Medicine, and Knowledge Management*, p. 15–36. Heidelberg: Springer.

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

KUNTZ P., GUILLET F., LEHN R. & BRIAND H. (2000). A User-Driven Process for Mining Association Rules. In D. ZIGHED, H. KOMOROWSKI & J. ZYTKOW, Eds., *Proc. of the 4^{th} Eur. Conf. on Principles of Data Mining and Knowledge Discovery (PKDD'00)*, volume 1910 of *Lecture Notes in Artificial Intelligence – LNAI*, p. 483–489, Lyon: Springer-Verlag.

LAVRAČ N., FLACH P. & ZUPAN B. (1999). Rule Evaluation Measures: A Unifying View. In *Proc. of the 9^{th} Int'l Workshop on Inductive Logic Programming (ILP'99)*, volume 1634 of *Lecture Notes in Artificial Intelligence – LNAI*, p. 174–185, Bled, Slovenia: Springer-Verlag, Heidelberg. Co-located with ICML'99.

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

RESNIK P. (1999). 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. Morgan Kaufmann Publishers.

SRIKANT R. & AGRAWAL R. (1995). Mining Generalized Association Rules. In *Proc. of the 21^{st} Int'l Conf. on Very Large Databases (VLDB'95)*, p. 407–419, Zurich: Morgan Kaufmann Press.

TAN P.-N., KUMAR V. & SRIVASTAVA J. (2002). Selecting the right interestingness measure for associaton patterns. In *Proc. of the 8^{th} ACM Int'l Conf. on Knowledge Discovery and Data Mining (KDD'02)*, p. 183–193, Edmonton, Canada: ACM Press.

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