

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

Hacène Cherfi, Dietmar Janetzko, Amedeo Napoli, Yannick Toussaint

► **To cite this version:**

Hacène Cherfi, Dietmar Janetzko, Amedeo Napoli, Yannick Toussaint. Sélection de règles d'association par un modèle de connaissances pour la fouille de textes. M. Liquière et M. Sebban. Conférence d'Apprentissage - CAp 2004, 2004, Montpellier, France, Presses Universitaires de Grenoble, pp.191-206, 2004. <inria-00107775>

HAL Id: inria-00107775

<https://hal.inria.fr/inria-00107775>

Submitted on 19 Oct 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

Hacène Cherfi¹, Dietmar Janetzko²,
Amedeo Napoli¹ et Yannick Toussaint¹

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 schéma général de référence d'extraction de connaissances à partir de bases de données introduit dans (Fayyad *et al.*, 1996), le processus de fouille de données est suivi d'une phase d'interprétation. Le processus de fouille que nous adoptons — ici appliqué 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'interprétation, ces règles sont évaluées par un analyste pour déterminer si une nouvelle connaissance peut être ajoutée au modèle de connaissances. La facilité de lecture d'une règle d'association est un point positif pour cette méthode de fouille de données ; en revanche, le

très grand nombre de règles extraites constitue un problème au sens où l'ensemble des règles devient trop difficile à appréhender pour un humain.

Dans cet article, nous souhaitons classer les règles d'association qui sont présentes à l'analyste en les ordonnant par qualité d'écroissante en fonction des connaissances disponibles sur le domaine des données. Nous considérons qu'une règle d'association est de bonne qualité si elle contient potentiellement des informations de nature à enrichir le modèle de connaissances. Les autres règles sont qualifiées de triviales puisqu'elles reflètent une connaissance déjà présente dans le modèle. Notre objectif est de faciliter la tâche de l'analyste en définissant une méthodologie de sélection des règles de qualité et de rejet des règles triviales qui exploite un modèle de connaissances. Notre modèle de connaissances est dans le cas présent un modèle terminologique qui, de façon analogue à un thésaurus, est caractérisé par un réseau de termes structurés hiérarchiquement par une relation de généralisation appelée EST-UN. Par exemple, une "pomme" EST-UN "fruit". Le sens du lien hiérarchique suivant la relation EST-UN dans le modèle de connaissances est important. Ainsi, nous considérons que la règle "fruit" \implies "pomme" n'est pas *triviale* par rapport au même modèle de connaissances et ne doit pas être rejetée.

Peu d'approches pour la sélection des règles d'association exploitent un modèle de connaissances. Or, la qualité d'une règle ne doit pas être définie uniquement à l'aide d'indices statistiques mais en évaluant l'apport de la règle par rapport aux connaissances du domaine déjà acquises : il existe souvent des sources de connaissances (*ex*, des ontologies) disponibles qui peuvent être exploitées. Le modèle de connaissances doit donc aider à interpréter les résultats du processus de fouille et réciproquement, les résultats de la fouille doivent contribuer à la mise à jour des connaissances. Nous considérons ainsi que la construction et l'enrichissement d'un modèle de connaissances sont des processus itératifs d'extraction de connaissances. À partir de données stables au cours des itérations, l'ensemble des règles d'associations extraites et leurs indices statistiques restent identiques alors que classement des règles suivant notre mesure de qualité évolue en fonction des mises à jour successives du modèle de connaissances.

La sélection des règles a été abordée par des mesures statistiques (Cherfi *et al.*, 2003). Diverses mesures de qualité ont été proposées et font l'objet d'une présentation synthétique 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'évaluer la qualité 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 données (Azé, 2003). Les limites de ces approches viennent du fait que l'évaluation se fait sans prendre en compte les connaissances du domaine et qu'il est impossible de développer une approche statistique indépendante pour l'évaluation des règles puisque ces mesures s'appuient sur le même ensemble de données que le processus d'extraction de règles d'association.

Dans cet article, nous écrivons, en premier lieu, notre processus de fouille de textes qui introduit un modèle de connaissances, sa mise en œuvre probabiliste ainsi que la définition des règles d'association. Nous définissons, ensuite, une mesure de qualité des règles par rapport au modèle de connaissances que nous appelons la *vraisemblance*.

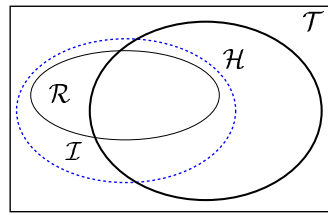


FIG. 1 – Les différents ensembles de termes : \mathcal{T} : ensemble des termes, \mathcal{I} : ensemble des termes d'indexation des textes, \mathcal{R} : sous-ensemble des termes d'indexation \mathcal{I} apparaissant dans les règles d'association, et \mathcal{H} : ensemble des termes du modèle M .

Enfin, nous évaluons le comportement de la vraisemblance sur un modèle formel puis sur une expérimentation avec des données textuelles réelles.

2 Processus de fouille de textes

2.1 Un modèle terminologique hiérarchique

Soit $\mathcal{T} = \{t_1, \dots, t_n\}$ un ensemble de termes (cf. FIG.1) qui sert de vocabulaire de référence pour indexer un ensemble de textes — noté \mathcal{D} pour *document* — $\mathcal{D} = \{d_1, \dots, d_m\}$.

Nous définissons notre modèle de connaissances comme étant un modèle terminologique qui, de façon analogue à un thésaurus, est caractérisé par un réseau de termes $\mathcal{H} \subseteq \mathcal{T}$ structurés hiérarchiquement par une relation de généralisation EST-UN. Un terme appartenant à ce modèle peut être seul ou être relié par la relation de généralisation EST-UN à un ou plusieurs autres termes.

La relation EST-UN est définie sur $\mathcal{H} \times \mathcal{H}$. Cette relation est réflexive, antisymétrique et transitive. La relation EST-UN constitue un ordre partiel. Le modèle ainsi défini n'est pas nécessairement connexe.

2.2 Les règles d'association

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

Une règle d'association est une implication de la forme $B \xrightarrow{P} H$ où B est la prémisse (*body*) et H est la conclusion (*head*) avec $B \subseteq \mathcal{I}$, $H \subseteq \mathcal{I}$ et $B \cap H = \emptyset$. Soit $B = \{t_1, \dots, t_p\}$ l'ensemble des termes de la prémisse d'une règle d'association r et $H = \{t_{p+1}, \dots, t_q\}$ l'ensemble des termes de la conclusion. $r : t_1 \dots t_p \xrightarrow{P} t_{p+1} \dots t_q$ signifie que tous les textes de \mathcal{D} contenant les termes t_1 et $t_2 \dots$ et t_p ont tendance à contenir aussi les termes t_{p+1} et $t_{p+2} \dots$ et 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 & 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 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.

Le support de r est le nombre de textes contenant les termes de $B \cup H = \{t_1, \dots, t_p, \dots, t_q\}$. La confiance de r est le rapport entre le nombre de textes contenant l'ensemble des termes $B \cup H$ et le nombre de textes contenant B (t_1, \dots, t_p). Ce rapport définit la probabilité conditionnelle $P(H|B)$. Le support et la confiance sont deux mesures associées aux règles d'association (Agrawal & Srikant, 1994) et exploitées par les algorithmes d'extraction de règles pour en réduire la complexité. Deux valeurs de seuil sont définies : σ 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 des termes présents, au moins une fois, en partie gauche ou droite, de l'ensemble des règles d'association, *i.e.*, $\mathcal{R} = \bigcup_{r_i} (B_i \cup H_i)$.

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 adéquation entre le modèle de connaissances initial et le contenu des textes. Notamment, il nous semble intéressant de considérer 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 fondées sur la relations EST-UN, c'est-à-dire pour lesquelles les termes de B sont liés par une relation de généralisation avec les termes de H dans le modèle de connaissances. Par exemple, si le terme « fruit » est plus général que le terme « pomme », une règle du type « pomme » \implies « fruit » est fortement candidate pour être rejetée puisqu'elle exprime une généralisation connue. En revanche, une règle du type « tarte à la cerise » \implies « chocolat », « beurre » doit être conservée. Cette règle exprime potentiellement une relation intéressante entre « tarte à la cerise », « chocolat » et « beurre » entre lesquels il n'existe pas de lien hiérarchique EST-UN.

Comment définir 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 généralisation entre les termes. Une règle d'association $a \implies b$ est triviale si la relation de généralisation que l'analyste peut ajouter au modèle à partir de cette règle, *i.e.* (a EST-UN b) est déjà présente dans le modèle. Il est remarquable que dans le cadre de la fouille de texte, une règle $a \implies b$ 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 $(a \text{ EST-UN } b)$. 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 au modèle de connaissances donné. 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 $(\mathcal{H} \times \mathcal{H}, \text{EST-UN})$ est notée M . L'objectif est de définir la vraisemblance d'une règle $r : a \implies b$ 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 $P_M(a \implies b)$, 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 attribuée 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 :

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

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

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

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

3 Définition de la vraisemblance d'une règle

Nous définissons la vraisemblance d'une règle d'association *simple* du type $a \implies b$ par la probabilité $P_M(a,b)$. Plus le lien hiérarchique entre a et b est fort dans le modèle, plus cette vraisemblance est forte, et donc plus la règle peut être considérée comme triviale pour un analyste. Nous observons cependant que cette définition ne permet de calculer que la vraisemblance de règles simples pour lesquelles on suppose que les termes indexant les textes et présents dans les règles sont également écrits 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 probabilités est représentée par la matrice donnée 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 probabilité (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 probabilités. Par exemple, la règle $(b \implies e)$ a pour vraisemblance dans le modèle M : $P_M(b \implies e) = P_M(b,e) = 0,30$.

Nous généralisons 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 sélection des règles puisse être utilisé sur des données réelles, nous devons le rendre robuste et résoudre deux problèmes :

- les règles simples où $|\mathcal{B}| \times |\mathcal{H}| = 1$ 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 $|\mathcal{B}| \times |\mathcal{H}| > 1$) ;
- nous considérons 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 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 3.1 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 propose trois possibilités pour étendre la distribution de probabilités et discute de l'impact de ces choix sur la vraisemblance des règles complexes.

3.1 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 est fondé sur le calcul de la probabilité dans la théorie de la propagation de l'activation pour chaque couple de termes issu du produit cartésien de la partie droite

avec la partie gauche de la règle. Étant donnée une règle complexe $r : t_1 \dots t_i \rightarrow t_{i+1} \dots t_p$ (où $B = \{t_1, \dots, t_i\}$ et $H = \{t_{i+1}, \dots, t_p\}$), la probabilité du produit cartésien est :

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

qui s'écrit en extension :

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

Nous observons cependant que plus le nombre de termes présents dans une règle est important, plus la probabilité $P_M(B \times H)$ est faible. Or le nombre de termes présents dans une règle ne doit pas affecter la vraisemblance d'une règle. L'équation 3 généralise donc l'équation 2 en prenant la moyenne géométrique de la probabilité du produit cartésien. Nous définissons ainsi la vraisemblance d'une règle :

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

Nous soulignons que le calcul de vraisemblance de l'équation 3 pour les règles complexes est également applicable aux règles simples puisque $|B| \times |H| = 1$.

3.2 L'extension de la distribution de probabilités

La distribution de probabilités qui a été introduite en paragraphe 2.4 doit être étendue pour traiter deux cas de figure :

- 1 – L'équation 1 (section 2.4) 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 n'est pas 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 $t \in \mathcal{R} \setminus \mathcal{H}$. 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 ramène 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 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 :

$$P_M(t_k, t_l) = \frac{1}{n+1} \text{ avec } n \text{ le nombre de termes de } \mathcal{H}. \text{ Ainsi, dans FIG. 2 (b),}$$

$$P_M(b, c) = \frac{1}{n+1} = \frac{1}{6} \text{ car } n = 5.$$

$$\text{Par exemple pour (b,c), nous avons : } P_M(b, c) = 1 \times \left(\left(3 \times \frac{1}{1} \right) + \left(2 \times \frac{1}{6} \right) \right)^{-1} = 0,3$$

$$\text{et pour (b,d) : } P_M(b, d) = \frac{1}{6} \times \left(\left(3 \times \frac{1}{1} \right) + \left(2 \times \frac{1}{6} \right) \right)^{-1} = 0,05.$$

Stratégie mixte :

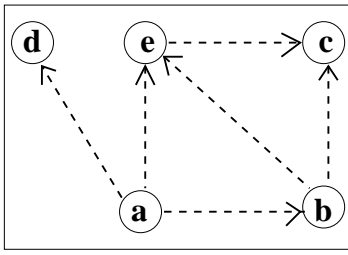
Une troisième stratégie — celle que nous adoptons dans la suite de l'article — consiste à associer une probabilité nulle dans le cas de règles simples et à appliquer une valeur de pénalité dans le cas de règles complexes.

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 $\{d_1, \dots, d_6\}$ décrits par un ensemble de termes d'indexation $\{a, \dots, e\}$ (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, \dots, r_{20} , sont extraites avec un support minimal $\sigma_s = 1$ et une confiance minimale $\alpha_c = 0,1$ et leurs valeurs de vraisemblance sont calculées à la FIG. 3 (2). Par exemple, pour la règle r_6 , nous avons :



r	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 modèle de connaissances M – (b) Probabilités de transition pour M.

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

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 \neg 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 présents dans B et H: en ligne 1 se trouvent les règles simples T-règles (1,1) taxinomiques et \neg T-règles (1,1) non taxinomiques, en ligne 2 les T-règles (1,n) et \neg T-règles (1,n), puis (n,1) et (n,m) avec $n, m > 1$.

Texte	Termes	T			\neg T				
		n°	n/d	$P_M(r)$	n°	n/d	$P_M(r)$		
		r_1	$b \Rightarrow e$	0/1	0,300	r_{19}	$e \Rightarrow b$	1/0	0,000
		r_{11}	$a \Rightarrow c$	0/0	0,111	r_{20}	$c \Rightarrow a$	1/0	0,000
d_1	{acd}	r_2	$b \Rightarrow c, e$	0/2	0,300	r_{13}	$d \Rightarrow a, c$	2/0	0,100
d_2	{bce}	r_4	$a \Rightarrow b, c, e$	0/2	0,176	r_{14}	$c \Rightarrow b, e$	2/0	0,100
d_3	{abce}	r_6	$b \Rightarrow a, c, e$	1/2	0,165	r_{15}	$c \Rightarrow a, d$	2/0	0,100
d_4	{be}	r_7	$e \Rightarrow b, c$	1/1	0,163	r_{16}	$c \Rightarrow a, b, e$	3/0	0,100
d_5	{abce}	r_9	$a \Rightarrow c, d$	0/1	0,157				
d_6	{bce}	r_{10}	$e \Rightarrow a, b, c$	2/1	0,121	r_{17}	$c, e \Rightarrow b$	2/0	0,081
		r_5	$b, c \Rightarrow e$	1/1	0,173	r_{12}	$b, c \Rightarrow a, e$	3/1	0,110
		r_3	$a, b \Rightarrow c, e$	0/3	0,217	r_{18}	$c, e \Rightarrow a, b$	4/0	0,081
		r_8	$a, e \Rightarrow b, c$	1/2	0,160				

FIG. 3 – (1) La base de données textuelles – (2) Mesure de vraisemblance pour les règles de l'exemple FIG.2(a) et le modèle M.

Nous observons, de façon empirique, un seuil $s = 0,111$ qui sépare les T-règles ($P_M(r_i) \geq 0,111$) des \neg T-règles ($P_M(r_i) < 0,111$). Ce point sera discuté en section 4.2.

Les T-règles (1,1) sont purement taxinomiques. En accord avec la définition 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 propriété en § 2.4) 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

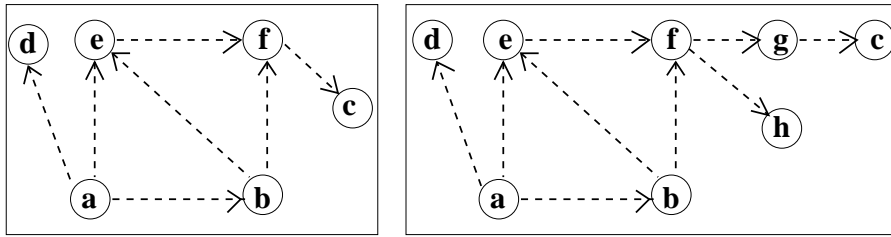


FIG. 4 – Les variantes M_1 et M_2 du modèle de connaissances M de FIG. 2 (a).

taxinomiques est respecté.

Les T-règles $(1,n)$, $(n,1)$ et (n,m) de FIG. 3 (2), nous pouvons vérifier deux principes d’écoulant des propriétés attendues de la vraisemblance que nous avons définie : (i) moins il y a de liens non taxinomiques entre les termes de B et de H , plus la valeur de la vraisemblance est élevée. (ii) plus les liens taxinomiques sont directs, plus la valeur de vraisemblance est élevée. La colonne n/d de FIG. 3 (2) donne le nombre de couples de termes de $B \times H$ qui ne sont pas des relations taxinomiques (noté n) et le nombre de relations taxinomiques 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 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) séparent 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 défini formellement et dépend du modèle choisi. Nous souhaitons également caractériser le comportement de notre méthodologie lorsque le modèle évolue. Pour cela, nous prenons le même ensemble de règles $\{r_1, \dots, r_{20}\}$.

Si nous opérons sur le modèle des modifications majeures, par exemple, en créant 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 définissons 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 différentes règles. S’il existe un chemin entre c_u et c_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

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

c_u et c_v , alors le nouveau modèle préserve également le fait que ce chemin n'existe pas.

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

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

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 (Gross & Yellen, 2003)), *i. e.*, 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 .

Nous observons l'évolution des valeurs de vraisemblance affectées aux règles d'association $\{r_1, \dots, r_{20}\}$. Dans M , les règles $\{r_1, \dots, r_{11}\}$ étaient classées comme T-règles et les règles $\{r_{12}, \dots, r_{20}\}$ comme $\neg T$ -règles. Dans la mesure où la nature des liens entre termes dans M_1 et M_2 reste inchangée, nous considérons que la règle r_{10} reste la règle seuil séparant les T-règles et les $\neg T$ -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 présent et nous remarquons que :

- la règle r_{10} est considérée comme taxinomique dans M . Elle 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 légèrement supérieure, *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 classée 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 également 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 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 modifie 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 indépendantes du modèle de connaissances choisi. Par conséquent, les valeurs de seuils ne peuvent être fixées *a priori* ;
- 3 – si les règles présentent des connaissances nouvelles, alors le modèle de connaissances peut être enrichi de façon incrémentale. Et nous avons le moyen de compléter ce modèle avec de nouveaux termes identifiés grâce aux règles d’association.

5 Expérimentation sur des données textuelles

Nous présentons les résultats de la sélection des règles d’association par un modèle de connaissances pour une expérimentation sur des données textuelles du monde réel. Ces règles sont extraites à partir d’un ensemble de notices bibliographiques d’écrivant 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 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 est possible d’appliquer des algorithmes de fouille de données classiques comme *Apriori* ou *Close*. L’extraction des règles d’association a été réalisée en s’appuyant sur l’algorithme *Close* (Pasquier *et al.*, 1999). *Close* fait une recherche par niveau dans un tableau booléen d’écrivant le produit cartésien $\mathcal{T} \times \mathcal{I}$ des motifs fréquents. L’algorithme commence par le plus petit motif fréquent fermé et calcule de façon incrémentale les motifs fermés plus long dans \mathcal{I} . Un motif est fermé s’il correspond à un ensemble maximal de termes partagé par un ensemble de textes. Le motif est fréquent s’il apparaît dans au moins q textes. Une fois les motifs fermés fréquents calculés, les règles d’association en sont dérivées.

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

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).

utilisé pour la sélection 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 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 focalisés sur des règles de vraisemblance nulle soit 211 règles dont 46 simples et 165 complexes.

Nous avons réalisé une classification des règles d'association en règles triviales / non triviales. Certaines règles non rejetées sont triviales mais l'incomplétude du modèle n'a pas permis de les identifier. De même, certaines règles rejetées ne sont pas triviales en raison des probabilités de transition très élevées de certains liens taxinomiques par rapport à d'autres liens non taxinomiques.

Pour évaluer la qualité de cette classification, nous devons déterminer 4 classes de règles: les vraies-positives (non taxinomiques $\neg T$ et évaluées comme non triviales), les fausses positives ($\neg T$, mais qui sont triviales), les vraies-négatives (taxinomiques et triviales) et les fausses-négatives (taxinomiques, mais non triviales).

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 rejetées par notre processus de sélection des règles, 122 (soit 90%) sont vraies-négatives et 14 (soit 10%) sont fausses-négatives. Le faible pourcentage de règles fausses-négatives montre la robustesse de la mesure de vraisemblance pour identifier les règles taxinomiques.

Parmi les 211 règles d'association qui sont non rejetées ($\neg T$), il y a 115 (soit 55%) qui sont vraies-positives et 96 (soit 44%) qui sont fausses-positives, *i.e.*, 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.

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'ancrer dans une approche probabiliste comme la nôtre, ils introduisent une mesure de similarité sémantique entre mots.

Les règles d'association généralisées (Srikant & Agrawal, 1995; Han, 1995; Hipp *et al.*, 2002) constituent 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 élevée. Enfin, un travail similaire exploitant un modèle de connaissances pour la classification de termes est proposée par (Resnik, 1999). La similarité est fondée sur l'information mutuelle. Elle sert à désambiguïser (affecter un seul sens) des termes ambigus selon la proximité sémantique qu'ils ont avec leurs voisins dans un thésaurus (*i.e.*, WORDNET).

7 Conclusion et perspectives

Nous proposons une méthodologie de sélection des règles d'association par l'utilisation d'un modèle de connaissances. Cette méthodologie 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 appliqué cette mesure pour la sélection des règles non taxinomiques en rejetant celles qui sont taxinomiques car elles sont triviales dans un domaine de spécialité. Nous avons étudié et montré que les propriétés de la mesure de vraisemblance est cohérente 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 EST-UN. Les liens de causalité reposent sur une transitivité des relations entre termes et peuvent, à ce titre, être exploités de la même manière que la relation taxinomique EST-UN. En revanche, composer simultanément plusieurs relations de natures différentes peut se révéler délicat car il faut considérer et manipuler la transitivité dans les probabilités de transition de façon à garder une cohérence de la sémantique de ces relations. La méthodologie que nous avons présentée ne considère pas les liens existant à l'intérieur de B ou à l'intérieur de H. Il nous semble intéressant de proposer une variante de cette mesure qui prend en compte

les liens entre termes apparaissant du même côté d'une règle. Le choix d'un seuil empirique demeure délicat. Il est fixé par jugement de l'analyste. Nous envisageons de définir un moyen pour apprendre 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 R. & SRIKANT R. (1994). Fast algorithms for mining association rules in large databases. In *Proc. of the 20th 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 7th 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 Conférence 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-SHAPIO 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 4th 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: Rockefeller University Press.
- KUNTZ P., GUILLET F., LEHN R. & BRIAND H. (2000). A User-Driven Process for Mining Association Rules. In D. ZIGHEH, H. KOMOROWSKI & J. ZYTKOW, Eds., *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*, p. 483–489, Lyon: Springer-Verlag.

LAVRAČ N., FLACH P. & ZUPAN B. (1999). 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*, 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 21st 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 association patterns. In *Proc. of the 8th 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, 11th edition.