Méthodologie de sélection et de lecture de règles d'association pour la fouille de textes

Hacène Cherfi, Yannick Toussaint

Méthodologie de sélection et de lecture de règles d'association pour la fouille de textes

Hacène Cherfi — Yannick Toussaint

Équipe Orpailleur — LORIA (UMR 7503 - CNRS - INRIA - Universités nancéennes)

{cherfi,yannick}@loria.fr

Tél. : 03 83 59 20 86

Mots-clés : Fouille de textes, règles d'association, indices statistiques, biologie moléculaire.

Notre travail porte sur la conception et la réalisation d'un outil de fouille dans les textes (FdT) pour aider un utilisateur, expert d'un domaine donné, dans sa tâche de veille technologique et scientifique. L'objectif de la FdT est de permettre à l'expert de retrouver, à travers un corpus, des relations connues entre les concepts du domaine, de pouvoir les localiser rapidement dans les documents et d'observer des familles de documents construites à partir d'une ou plusieurs de ces relations. Elle permet aussi de découvrir de nouvelles relations. Nous recherchons ces relations à travers des règles d'association sur les documents caractérisés par des ensembles de termes.

La finalité de notre travail est de trouver le moyen de sélectionner, parmi ces règles, celles qui présentent un intérêt particulier pour l'expert. En effet, le nombre de règles extrait croît de manière exponentielle par rapport au nombre de termes du corpus. Leur lecture est, par conséquent, une tâche difficile. Nous procédons donc en deux étapes :

(1) l'expert identifie manuellement un sous-ensemble de règles qu'il arrive à relier à ses connaissances ;

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

Différentes approches ont été proposées pour gérer ce grand nombre de règles [4, 7, 5, 3]. Ces approches ont un point commun : l'élimination de règles non-informatives ou redondantes. À l'opposé, notre approche privilégie le bruit au silence et consiste donc à garder toutes les règles. En revanche, elles sont triées selon des indices statistiques.

Notre corpus $\mathcal{D}$ est constitué de 1 361 documents ($\approx$ 200 000 mots) indexés par 632 termes différents. Les textes sont des résumés d'articles scientifiques, en anglais, traitant de la biologie moléculaire ; plus particulièrement des mutations génétiques en résistance à des antibiotiques. Nous avons obtenu 347 règles.

Une règle d'association est du type $\mathtt{R} : \mathtt{B} \Longrightarrow \mathtt{H}$ où B et H sont des conjonctions de termes. Les règles sont obtenues par le calcul des motifs fermés fréquents en utilisant l'algorithme *Close* [8]. Une règle d'association est la réalisation de l'événement « si j'ai B, alors j'ai tendance à avoir H ».

La valeur informative de R dépend de la distribution des termes de B et de H sur les documents (voir FIG. 1). Le cas (d) est rendu impossible en raison du seuil de *confiance* élevé pour les règles. Des indices statistiques ont été définis pour mesurer la qualité d'une règle [1]. Tous ces indices utilisent trois estimations de probabilités : $\mathtt{P}(\mathtt{B})$, $\mathtt{P}(\mathtt{H})$ et $\mathtt{P}(\mathtt{B} \wedge \mathtt{H})$. La confrontation avec l'expert dans le domaine de la biologie moléculaire a montré que les règles qu'il retenait se trouvaient majoritairement dans les cas Fig. 1(c) mais également dans (b). Plus les termes sont présents dans le corpus (Fig. 1(a)), plus la règle est triviale et par conséquent moins intéressante. Nous illustrons les résultats de cette analyse par quelques exemples.

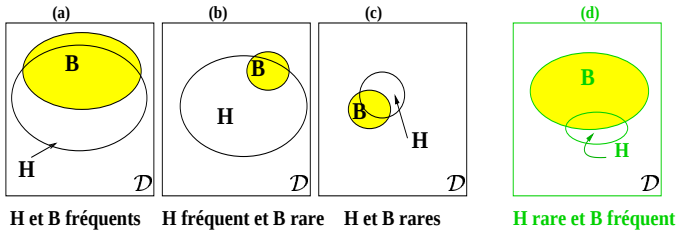

FIG. 1 – *Les différents cas de distribution de* `B` *et* `H` *sur l'ensemble* $\mathcal{D}$ *des documents —le cas (d) est impossible dans notre contexte—.*

En confrontant deux règles [1] à forte valeur d'*intérêt* [2], nous avons mis en valeur des similitudes de comportement de populations de bactéries en résistance à deux antibiotiques (`quinupristine` et `dalfopristine`). Le tri selon l'*intérêt* privilégie les règles informatives (cas (c)). De plus, la symétrie de cet indice a associé la même valeur à ces deux règles. Elles ont donc été présentées simultanément à l'expert, ce qui en a facilité l'interprétation.

Notons que pour des valeurs d'*intérêt* faibles, cet indice ne permet pas de différencier les règles moins informatives (cas (b)), des règles non-informatives (cas (a)). En revanche la *nouveauté*, qui est une variante de la *dépendance*, permet cette distinction. Elle différencie, par exemple, deux règles [3], la première, non retenue par l'expert, illustre le cas(a) alors que l'autre, retenue, illustre le cas (b).

En confrontant plusieurs règles ayant une forte valeur de *conviction*, nous avons retracé dans les textes une antériorité dans la découverte du gène `GyrA` par rapport à `ParC`. La *conviction* est l'inverse de l'*intérêt* d'une règle qui porterait sur les contre-exemples ($\texttt{B} \Longrightarrow \neg\texttt{H}$). Moins les contre-exemples sont fondés, plus la *conviction* est grande. Nous avons vérifié sur nos donnés que cet indice renforce le côté implicatif de `B` vers `H`. Dans notre exemple, l'expert avait souligné que `ParC` et `GyrA` sont deux gènes régulièrement présents ensemble dans les règles et il le justifiait par leurs comportements comparables du point de vue de la mutation. Pourtant, le sens de l'implication $\ldots\texttt{ParC}\ldots \Longrightarrow \ldots\texttt{GyrA}\ldots$ dans des règles de forte *conviction* contribuait à les différencier. Finalement, l'explication réside dans le fait que les textes les plus anciens de notre corpus ne traitent que de `GyrA` alors que les textes plus récents traitent de `GyrA` et de `ParC`.

Les règles, les indices associés et, par conséquent, l'interprétation qu'en fait l'expert sont sensibles à la qualité de la représentation du contenu des textes. Par exemple, dans la formulation en langue naturelle, l'auteur peut introduire des termes généraux, pas indispensables, pour décrire un phénomène de résistance. À l'opposé, un terme important, mais sous-entendu, ne fait pas partie de la description du texte. Cette variabilité de présence/absence des termes dans les textes se répercute dans les règles puisque celles-ci sont construites à partir des occurrences des termes dans les textes.

Grâce à la réalisation d'un environnement interactif de consultation des règles d'association et des différents indices qui leurs sont associés, nous avons montré que certains indices statistiques aident les experts à interpréter un ensemble de règles en proposant différents tris, des plus informatives vers les moins informatives. Cependant, nous n'avons pas observé de seuils pour ces indices en dessous desquels les règles sont toutes rejetées par l'expert. Trier les règles semble donc mieux adapté que de les éliminer.

1. "`quinupristin`" $\Longrightarrow$ "`dalfopristin`" et "`dalfopristin`" $\Longrightarrow$ "`quinupristin`"

2. Les indices d'intérêt, de conviction et de nouveauté sont défi nis dans [2, 6] par :
$\texttt{int}\,[\texttt{B} \Longrightarrow \texttt{H}] = \frac{\texttt{P(B}\wedge\texttt{H)}}{\texttt{P(B)}\times\texttt{P(H)}}$ $\texttt{conv}\,[\texttt{B} \Longrightarrow \texttt{H}] = \frac{\texttt{P(B)}\times\texttt{P(}\neg\texttt{H)}}{\texttt{P(B}\wedge\neg\texttt{H)}} = \frac{1}{\texttt{int}\,[\texttt{B} \Longrightarrow \neg\texttt{H}]}$ $\texttt{nov}\,[\texttt{B} \Longrightarrow \texttt{H}] = \texttt{P(H}\wedge\texttt{B)} - \texttt{P(B)} \times \texttt{P(H)}$

3. "`meticillin`" $\Longrightarrow$ "`staphyloccocus aureus`" et "`mecA gene`" "`meticillin`" $\Longrightarrow$ "`staphyloccocus aureus`"

Références

[1] Jérôme Azé and Yves Kodratoff. A Study of the Effect of Noisy Data in Rule Extraction Systems. In *Proc. of EMCSR 2002: 16th European Meeting on Cybernetics and Systems Research*, Vienna, 2002. 6 pages.

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

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

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

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

[6] N. Lavrač, P. Flach, and B. Zupan. Rule Evaluation Measures: A Unifying View. In S. Džeroski and P. Flech, editors, *Proc. of ILP'99: 9th International Workshop on Inductive Logic Programming*, volume 1634 of *LNAI*, pages 174–185, Bled, Slovenia, 1999. Springer-Verlag, Heidelberg. Co-located with ICML'99.

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

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