Deux méthodes de classification de règles d'association en fouille de textes

Hacène Cherfi, Amedeo Napoli, Yannick Toussaint

Deux méthodes de classification de règles d'association en fouille de textes

Hacène Cherfi Amedeo Napoli
Yannick Toussaint
LORIA, BP 239, 54506 Vandœuvre-lès-Nancy cedex
{cherfi,napoli,yannick}@loria.fr,

Résumé

Un processus de fouille de données textuelles s'appuyant sur l'extraction de règles d'association engendre un très grand nombre de règles extraites. Il est alors nécessaire pour classifier les règles extraites de pouvoir disposer de critères fiables en rapport avec des connaissances du domaine. Dans cet article, nous présentons deux méthodes de classification : la première met en jeu des mesures statistiques tandis que la seconde, plus originale, repose sur un modèle de connaissances du domaine. Une discussion sur le bien-fondé de cette dernière approche termine cet article.

1 Introduction

Dans cet article, nous présentons le processus de *fouille de textes* (FDT) comme un processus d'extraction de connaissances à partir de données contrôlé et orienté d'une part par un *analyste*, expert du domaine des textes, et d'autre part par un modèle de connaissances du domaine des textes. L'objectif est d'enrichir progressivement les connaissances de l'analyste et celles du modèle, et réciproquement, de guider le processus d'extraction grâce au modèle de connaissances.

Dans ce qui suit, le processus de FDT cherche à extraire des règles d'association à partir de tableaux booléens $\texttt{Textes} \times \texttt{Termes}$, où `Textes` désigne un ensemble de textes et `Termes` un ensemble de termes-clés associés. Deux méthodes de FDT sont évoquées, qui reposent toutes deux sur l'extraction de règles d'association, par l'intermédiaire d'un algorithme d'extraction de motifs fréquents. La première méthode propose une classification des règles d'association extraites sur la base de mesures de qualité statistiques [4], où chaque mesure met en valeur des éléments de nature différente, comme des informations rares, stables au bruit, des dépendances fonctionnelles, etc. Cette approche relève d'un processus purement statistique, ne tenant aucun compte des connaissances du domaine. La seconde méthode propose au contraire une classification qualitative des règles extraites, où une règle d'association peut prétendre être de bonne qualité si elle contient des éléments d'information potentiellement aptes à enrichir le modèle des connaissances du domaine [3]. Le modèle de connaissances, noté $(\mathcal{K}, \sqsubseteq)$, se ramène à un ensemble *fini* de termes-clés muni d'un ordre partiel, qui doit servir à l'interprétation des résultats de la fouille ; réciproquement, les résultats de la fouille doivent contribuer à enrichir le modèle.

Dans cet article, nous présentons successivement les deux processus de fouille de textes et montrons en quoi ils diffèrent, le premier dépendant de critères numériques et le second dépendant de critères qualitatifs en rapport avec les connaissances du domaine des textes. Une brève discussion sur la problématique de la fouille de textes termine l'article.

2 Les règles d'association en FDT et les mesures statistiques

Soit $\mathcal{T} = \{\mathtt{t_1}, \mathtt{t_2}, \ldots, \mathtt{t_m}\}$ un ensemble de textes et $\mathcal{K} = \{\mathtt{k_1}, \mathtt{k_2}, \ldots, \mathtt{k_n}\}$ un ensemble de termes-clés associés à ces textes. Une règle d'association est une implication pondérée de la forme $\mathtt{A} \longrightarrow \mathtt{B}$ où $\mathtt{A} = \{\mathtt{t_1}, \mathtt{t_2}, \ldots, \mathtt{t_p}\}$ et $\mathtt{B} = \{\mathtt{t_{p+1}}, \mathtt{t_{p+2}}, \ldots, \mathtt{t_q}\}$. La règle $\mathtt{A} \longrightarrow \mathtt{B}$ s'interprète comme le fait que tous les textes contenant les termes-clés $\{\mathtt{t_1}, \mathtt{t_2}, \ldots, \mathtt{t_p}\}$ ont tendance à contenir aussi les termes-clés $\{\mathtt{t_{p+1}}, \mathtt{t_{p+2}}, \ldots, \mathtt{t_q}\}$, avec une certaine probabilité donnée par la *confiance* de la règle. Plusieurs algorithmes permettent de mettre en œuvre un tel processus d'extraction de règles à partir de motifs fréquents, par exemple *Close* et *Pascal* [1].

Le *support* et la *confiance* sont deux mesures associées aux règles d'association utilisées pour réduire le nombre de règles extraites. Le support d'une règle est donné par le nombre de textes contenant à la fois les termes-clés de $\mathtt{A}$ et de $\mathtt{B}$ — la réunion des termes-clés est notée $\mathtt{A} \sqcap \mathtt{B}$ — tandis que la confiance d'une règle est le rapport entre le nombre de textes contenant $\mathtt{A} \sqcap \mathtt{B}$ et le nombre de textes contenant $\mathtt{A}$ (ce qui reflète la probabilité conditionnelle $\mathtt{P(B/A)}$). La confiance donne une mesure du pourcentage d'exemples et de contre-exemples de la règle. La présence d'un contre-exemple montre qu'il existe des textes possédant les termes de $\mathtt{A}$ mais pas nécessairement tous les termes de $\mathtt{B}$. Lorsque la confiance vaut $\mathtt{1}$, la règle est dite *exacte*, sinon elle est *approximative*. Deux valeurs de seuil sont définies, $\sigma_{\mathtt{s}}$ pour le support minimal et $\sigma_{\mathtt{c}}$ pour la confiance minimale.

Partant d'une règle comme $\mathtt{A} \longrightarrow \mathtt{B}$, si la règle est constituée de motifs $\mathtt{A}$ et $\mathtt{B}$ très fréquents, alors ces motifs sont partagés par presque tous les textes et les probabilités associées, $\mathtt{P(A)}$, $\mathtt{P(B)}$ et $\mathtt{P(A \sqcap B)}$, sont fortes ou très fortes ; inversement, l'intérêt des connaissances impliquées par les motifs $\mathtt{A}$ et $\mathtt{B}$, du point de vue de la découverte de connaissances, est faible. Si la règle est constituée de motifs $\mathtt{A}$ et $\mathtt{B}$ rares, alors ces motifs sont partagés par un petit nombre de textes mais apparaissent conjointement : ils sont vraisemblablement liés dans le contexte du domaine des textes étudiés et peuvent présenter un intérêt du point de vue de la découverte de connaissances. Les mesures de support et de confiance ne permettent pas toujours de discerner à elles seules les règles porteuses de sens et à fournir une classification adéquate des règles d'association extraites. C'est pourquoi d'autres mesures sont également utilisées, comme l'intérêt, la conviction, la dépendance, la nouveauté et la satisfaction (pour des détails voir [4, 2]).

3 La vraisemblance d'une règle d'association

3.1 Le modèle de connaissances

Le modèle de connaissances noté $(\mathcal{K}, \sqsubseteq)$ est caractérisé par un ensemble de termes $\mathcal{K}$ muni d'une relation de *spécialisation* $\sqsubseteq$ (qui est un ordre partiel). Le principe de la classification par la vraisemblance est le suivant : il faut rechercher pour les écarter toute règle d'association $\mathtt{A} \longrightarrow \mathtt{B}$ qui ne fait que traduire une relation $\mathtt{A} \sqsubseteq \mathtt{B}$. Par exemple, si $\mathtt{pomme} \sqsubseteq \mathtt{fruit}$, alors la règle d'association $\mathtt{pomme} \longrightarrow \mathtt{fruit}$ va être rejetée car elle est connue dans le modèle $(\mathcal{K}, \sqsubseteq)$. En revanche, la règle $\mathtt{tarte\text{-}cerise} \longrightarrow \mathtt{chocolat}$ exprime potentiellement une relation intéressante entre les termes $\mathtt{tarte\text{-}cerise}$ et $\mathtt{chocolat}$, entre lesquels il n'existe *a priori* pas de relation de spécialisation $\sqsubseteq$.

Dans ce qui suit, nous nous restreignons aux règles d'association dites *simples*, où la prémisse et la conclusion sont réduites à un seul et unique terme-clé, comme $\mathtt{a} \longrightarrow \mathtt{b}$ par exemple. Une règle d'association $\mathtt{a} \longrightarrow \mathtt{b}$ est dite *triviale* si la relation $\mathtt{a} \sqsubseteq \mathtt{b}$ existe dans le modèle $(\mathcal{K}, \sqsubseteq)$.

La *vraisemblance* d'une règle $\mathtt{a} \longrightarrow \mathtt{b}$ par rapport à un modèle de connaissances $(\mathcal{K}, \sqsubseteq)$)

comme la probabilité de trouver un chemin allant de $\mathtt{a}$ vers $\mathtt{b}$ dans $(\mathcal{K}, \sqsubseteq)$. Cette probabilité s'appuie sur le principe de la « propagation de l'activation » (*spreading activation theory* [6]) selon laquelle un marqueur d'information part d'un sommet (un concept) du modèle, par exemple $\mathtt{k_1}$, et se propage à travers ce modèle avec une certaine force, à la recherche d'un autre élément, par exemple $\mathtt{k_2}$. La force s'affaiblit de façon proportionnelle à la distance parcourue par le marqueur et au nombre de possibilités offertes à chaque branchement.

3.2 La définition de la vraisemblance d'une règle

Étant donné un modèle de connaissances du domaine $(\mathcal{K}, \sqsubseteq)$, nous définissons une table de probabilités de transitions qui va servir de base à la mesure de vraisemblance entre deux termes-clés. La probabilité de transition entre un terme-clé $\mathtt{k_i}$ et un terme-clé $\mathtt{k_j}$ est calculée sur la base du chemin de longueur minimale qui relie $\mathtt{k_i}$ et $\mathtt{k_j}$ dans le modèle. Il existe deux cas particuliers : (1) par convention, pour tout terme $\mathtt{k_i}$, $\mathtt{d}(\mathtt{k_i}, \mathtt{k_i}) = \mathtt{1}$; ceci pour prendre en compte la réflexivité de la relation de spécialisation et éviter des probabilités anormalement élevées en cas d'absence d'arc sortant ; (2) s'il n'existe pas de chemin entre un terme $\mathtt{k_i}$ et un terme $\mathtt{k_j}$, alors $\mathtt{d}(\mathtt{k_i},\mathtt{k_j}) = \mathtt{2N} + \mathtt{1}$ où $\mathtt{N}$ est le cardinal de $\mathcal{K}$ ($\mathcal{K}$ est *fini*).

La probabilité de transition entre $\mathtt{k_i}$ et $\mathtt{k_j}$ définit la vraisemblance notée $\mathtt{V}(\mathtt{k_i},\mathtt{k_j})$ de la règle $\mathtt{k_i} \longrightarrow \mathtt{k_j}$ et repose sur le produit de deux facteurs : la distance de $\mathtt{k_i}$ à $\mathtt{k_j}$ et le *poids* de $\mathtt{k_i}$ dans le modèle noté $\delta(\mathtt{k_i})$. En outre, il faut encore tenir compte de deux éléments : (1) plus la distance entre deux termes-clés est grande, plus la valeur de la vraisemblance doit être faible, (2) le poids d'un élément $\mathtt{k_i}$ dépend de l'ensemble des termes-clés du modèle, qu'ils soient atteignables ou non depuis $\mathtt{k_i}$. Ainsi, la formule qui calcule la vraisemblance entre $\mathtt{k_i}$ et $\mathtt{k_j}$ est la suivante : $\mathtt{V}(\mathtt{k_i},\mathtt{k_j}) = [\mathtt{d}(\mathtt{k_i},\mathtt{k_j}) \times \delta(\mathtt{k_i})]^{-\mathtt{1}}$, où le poids de $\mathtt{k_i}$ $\delta(\mathtt{k_i}) = \sum_{\mathtt{x} \in \mathcal{K}} \mathtt{1}/\mathtt{d}(\mathtt{k_i},\mathtt{x})$. Le poids $\delta(\mathtt{k_i})$ d'un élément $\mathtt{k_i}$ est dépendant du nombre d'arcs sortants associés à $\mathtt{k_i}$ dans le modèle $\mathcal{K}$: plus ce nombre est élevé plus le poids est élevé. À l'inverse, lorsqu'il n'existe aucun arc sortant, l'élément lui-même devient prépondérant car $\mathtt{d}(\mathtt{k_i},\mathtt{k_i}) = \mathtt{1}$. Il faut encore remarquer que le poids $\delta(\mathtt{k_i})$ est calculé une seule fois pour tout $\mathtt{k_i}$ et que l'équation suivante est vérifiée : $\sum_{\mathtt{x} \in \mathcal{K}} \mathtt{V}(\mathtt{k_i},\mathtt{x}) = \mathtt{1}$. Un exemple est proposé et détaillé dans [5].

4 Discussion, conclusion et perspectives

Beaucoup de travaux en fouille de textes s'intéressent à la façon de gérer le très grand nombre de règles d'association extraites. La plupart de ces travaux abordent le problème du point de vue statistique, sans chercher à tenir compte de connaissances du domaine.

Dans cet article, nous proposons deux méthodes de classification de règles d'association pour la FDT, qui s'appuient pour l'une sur des mesures statistiques et pour l'autre sur une mesure de vraisemblance qui est fonction d'un modèle de connaissances du domaine. Le calcul et l'utilisation de la mesure de vraisemblance sont de nature différente de ce qui se pratique habituellement en FDT avec des mesures statistiques. La mesure de vraisemblance permet de classifier les règles d'association extraites et de focaliser l'attention de l'analyste sur des règles qui ne reflètent pas une relation de spécialisation dans le modèle. Ainsi, le comportement de la mesure de vraisemblance en terme d'apport de nouvelles connaissances pour enrichir une ontologie par exemple est cohérent avec ce qui peut être attendu d'un processus d'extraction de connaissances à partir de textes pour un analyste spécialiste du domaine des textes. Par ailleurs, l'approche présentée ici autorise un enrichissement mutuel entre modèle de connaissances et processus de FDT, ce qui montre qu'une telle approche peut être véritablement qualifiée d'approche pour l'extraction de connaissances à partir de textes guidée par les connaissances du domaine.

Le travail de recherche actuel peut être prolongé dans un certain nombre de directions : définition de la vraisemblance de règles complexes (avec des prémisses et des conclusions comptant plus d'un terme-clé), prise en compte dans le modèle de connaissances de relations de causalité, temporelles ou spatiales ... Il reste encore à approfondir les liens entre les aspects statistiques et les aspects qualitatifs, et à relier de façon plus significative les classifications issues des mesures statistiques et celles qui sont issues du modèle de connaissances : peu de travaux se sont jusqu'à présent intéressés cette tâche ; le domaine est donc ouvert et potentiellement fertile pour la fouille de textes.

Références

[1] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal. Pascal : un algorithme d'extraction des motifs fréquents. *Technique et science informatiques*, 21(1) :65–95, 2002.

[2] H. Cherfi. *Étude et réalisation d'un système d'extraction de connaissances à partir de textes*. Thèse d'université, Université Henri Poincaré (Nancy 1), 2004.

[3] H. Cherfi, D. Janetzko, A. Napoli, and Y. Toussaint. Sélection de règles d'association par un modèle de connaissances pour la fouille de textes. In M. Liquière and M. Sebban, editors, *Actes de la conférence sur l'apprentissage (CAp 2004), Montpellier*, pages 191–206. Presses Universitaires de Grenoble, 2004.

[4] H. Cherfi, A. Napoli, and Y. Toussaint. Vers une méthodologie de fouille de textes s'appuyant sur l'extraction de motifs fréquents et de règles d'association. In R. Gilleron, editor, *Actes de la conférence sur l'apprentissage (CAp-03), Laval*, pages 61–76, 2003.

[5] H. Cherfi, A. Napoli, and Y. Toussaint. Deux méthodes de classification de régles d'association pour la fouille de textes. Rapport de recherche, LORIA, Nancy, 2005.

[6] A. Collins and E. Loftus. A spreading-activation of semantic processing. *Psychological Review*, 82(6) :407–428, 1975.