TALN 2010, Montréal, 19–23 juillet 2010

Fouille de données séquentielles d'itemsets pour l'apprentissage de patrons linguistiques

Peggy Cellier Thierry Charnois
GREYC – CNRS UMR 6072
Université de Caen – Bd Mal Juin 14032 Caen, France
prénom.nom@info.unicaen.fr

Résumé. Dans cet article nous présentons une méthode utilisant l'extraction de motifs séquentiels d'itemsets pour l'apprentissage automatique de patrons linguistiques. De plus, nous proposons de nous appuyer sur l'ordre partiel existant entre les motifs pour les énumérer de façon structurée et ainsi faciliter leur validation en tant que patrons linguistiques.

Abstract. In this paper, we present a method based on the extraction of itemset sequential patterns in order to automatically generate linguistic patterns. In addition, we propose to use the partial ordering between sequential patterns to enumerate and validate them.

Mots-clés : Fouille de données, motifs séquentiels, extraction d'information, apprentissage de patrons linguistiques.

Keywords: Data mining, sequential patterns, information extraction, linguistic pattern learning.

1 Introduction

Face à l'abondance et à la prolifération des données textuelles, l'accès à l'information pertinente dans les corpus est devenu un enjeu majeur avec des besoins dans différents domaines. On peut citer l'analyse du discours évaluatif qui connaît un intérêt croissant pour des applications telles que la veille d'opinions, l'analyse de tendances ou de marchés (Pang & Lee, 2007). Les approches symboliques du traitement automatique des langues dédiées à l'extraction d'information reposent sur des ressources élaborées manuellement dont le coût est important en temps de développement, voire prohibitif lorsqu'il s'agit de les adapter à un nouveau domaine (Poibeau, 2003). C'est pourquoi les méthodes permettant d'apprendre automatiquement les ressources connaissent un essor important. Certaines approches s'appuient sur des corpus annotés (Califf & Mooney, 2003) souvent difficiles à obtenir. D'autres méthodes utilisent des corpus bruts (Riloff, 1996) mais elles reposent sur une analyse syntaxique qui impacte la qualité des résultats. On peut aussi citer les approches dans la lignée de (Hearst, 1992) qui visent à acquérir des relations sémantiques d'un type particulier (hyperonymie) pour enrichir automatiquement des lexiques ou des ontologies. Dans (Charnois *et al.*, 2009) une approche a été proposée pour apprendre automatiquement des patrons linguistiques pour la découverte de relations entre entités nommées. Cette approche s'appuie sur l'utilisation de motifs séquentiels d'items, qui permettent de générer automatiquement des patrons linguistiques, et sur la fouille de données récursive des motifs qui permet de gérer la quantité de patrons extraits. Cette

Identifiant	Séquence
1	$\langle(homme\ homme\ NOM)(de\ de\ PRP)(culture\ culture\ NOM)\rangle$
2	$\langle(en\ en\ PRP)(vieux\ vieux\ ADJ)(farceur\ farceur\ NOM)(misanthrope\ misanthrope\ ADJ)\rangle$
3	$\langle(réputé\ réputer\ VER\ pper)(pour\ pour\ PRP)(sa\ son\ DET\ POS)(cruauté\ cruauté\ NOM)\rangle$

TAB. 1 – Extrait d'une base de séquences pour les textes : "homme de culture, en vieux farceur misanthrope, réputé pour sa cruauté".

approche a l'avantage de ne pas nécessiter d'analyse syntaxique ni de ressource extérieure autre qu'un corpus d'apprentissage brut, et n'est pas dédiée à l'apprentissage d'un type de patrons spécifiques.

Dans cet article, nous proposons une méthode basée sur les motifs séquentiels d'itemsets. Cela signifie qu'au lieu de décrire un mot par un seul item (son lemme ou sa catégorie grammaticale ou la conjonction des deux) comme dans (Charnois *et al.*, 2009), un mot est décrit par un ensemble d'items. L'avantage de cette description plus riche est de pouvoir générer automatiquement des patrons linguistiques sophistiqués contenant à la fois des lemmes et des catégories grammaticales, comme le patron « homme de NOM ». De plus, pour gérer le problème du grand nombre de motifs extraits à valider par un expert, nous proposons de nous appuyer sur l'ordre partiel existant entre ces motifs. Cela permet une énumération structurée des motifs et facilite leur exploration. Nous appliquons notre approche à l'apprentissage de patrons linguistiques pour la découverte de constituants en position détachée, extra-prédicatifs, et porteurs de qualification, voire de jugement, comme illustrée dans la table 1. Toutefois, le processus peut être facilement adapté pour découvrir d'autres types de patrons linguistiques Dans la section 2, la méthode proposée est détaillée. La section 3 présente l'application de la méthode pour la découverte d'expressions qualificatives.

2 Apprentissage des patrons linguistiques

L'extraction automatique des patrons linguistiques et leur validation s'effectuent en deux étapes : 1) les motifs séquentiels fréquents sont extraits de la base d'exemples ; 2) ces motifs extraits sont ensuite structurés dans un *diagramme de Hasse* afin de faciliter leur sélection en tant que patrons linguistiques.

2.1 Extraction de motifs séquentiels d'itemsets pour la découverte de patrons

L'extraction de motifs séquentiels d'itemsets a été introduite dans (Agrawal & Srikant, 1995; Srikant & Agrawal, 1996). L'extraction de motifs séquentiels se fait à partir d'une base de séquences, BDD, où chaque séquence est décrite par une liste ordonnée d'ensembles de littéraux appelés *items*. Un ensemble d'items est communément appelé *itemset*, noté $(i_1 i_2 ... i_m)$ où les i_j sont des items. Une séquence est donc une liste ordonnée d'itemsets, notée $\langle s_1 ... s_n \rangle$ où les s_j sont des itemsets. Dans le cas de la découverte de patrons linguistiques, nous constituons une base de séquences (la base d'exemples) à partir de morceaux de texte. Chaque morceau de texte représente donc une séquence de la base et est décrit par l'ensemble des mots qui le composent. Les mots[1]sont représentés par des itemsets qui décrivent le mot par sa forme fléchie, son lemme et sa catégorie grammaticale [2]. Un extrait de base contenant 3 constituants en position

[1]En fonction de l'application il est tout à fait possible de choisir un grain différent, par exemple : syllabe, phrase, paragraphe.
[2]Notons que d'autres informations pourraient être ajoutées comme des traits sémantiques.

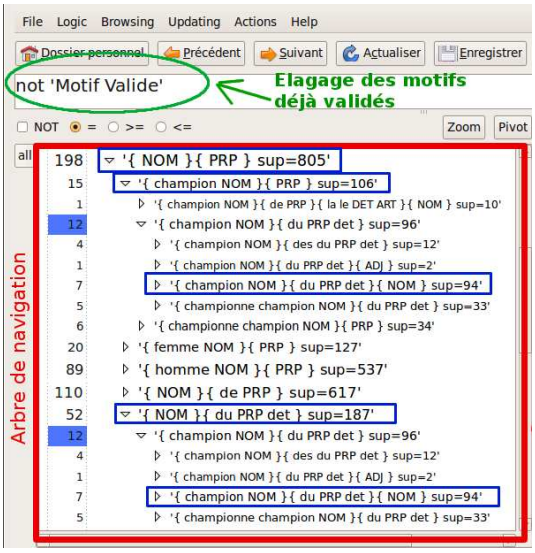

FIG. 1 – Navigation dans les patrons linguistiques pour les valider.

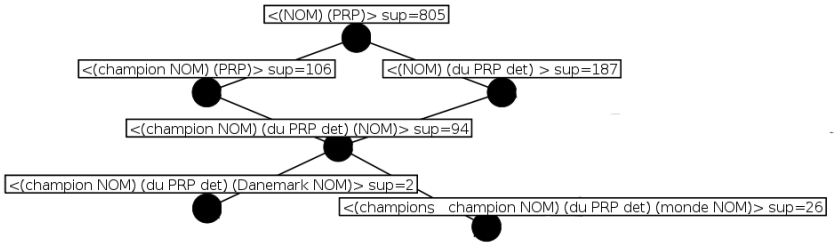

FIG. 2 – Extrait du diagramme de Hasse contenant des motifs séquentiels.

détachée de qualification est donné à la table 1. Une séquence, $S_1 = \langle a_1 ... a_n \rangle$ est contenue dans une séquence $S_2 = \langle b_1 ... b_m \rangle$ s'il existe des entiers $i_1 < ... < i_n$ tels que $a_1 \subseteq b_{i_1}, ..., a_n \subseteq b_{i_n}$. S_1 est alors appelée *sous-séquence* de S_2, noté $S_1 \leq S_2$. Le support d'une séquence[3], S, dans une base, BDD, est le nombre de séquences de BDD qui contiennent S. Par exemple, pour le motif $\langle (PRP)(ADJ) \rangle$ son support dans BDD est de 1 (Séquence 2). Les motifs séquentiels ne sont pas nécessairement des suites contigues. Par exemple, le motif $\langle (PRP)(ADJ) \rangle$ couvre aussi l'expression « en homme courageux ». Un motif séquentiel *fréquent* dans BDD est une séquence dont le support est supérieur à un seuil fixé : $minsup$. Notons que l'utilisation des itemsets permet une description plus riche des mots que l'utilisation de simples items. Cet ajout d'information sur les mots donne la possibilité de découvrir des patrons linguistiques plus sophistiqués, composés d'information mixte comme à la fois les lemmes mais aussi les catégories grammaticales. Par exemple, on trouve des patrons linguistiques de la forme : $\langle (en\ PRP)\ (homme\ NOM)(de\ PRP)(NOM) \rangle$. Pour diriger l'extraction des motifs séquentiels vers l'objectif de l'utilisateur, on définit des contraintes sur les motifs à extraitre. Par exemple, la contrainte *gap* (Dong & Pei, 2007) impose que pour que S_1 soit contenue dans S_2 il faut que chaque couple d'itemsets adjacents de S_1 ne soit pas séparé dans S_2 par plus d'un certain nombre d'itemsets. Ce nombre est appelé $maxgap$. Les contraintes linguistiques permettent de définir le type de patrons linguistiques recherchés.

2.2 Maîtrise du nombre de motifs

Beaucoup de motifs séquentiels sont générés. Pour pallier à ce problème nous utilisons une représentation condensée des motifs qui élimine les redondances entre motifs : les *motifs fermés* (Yan *et al.*, 2003), et un ordre partiel qui permet d'avoir une énumération des motifs structurée et de faciliter leur exploration.

Un motif fréquent, S, est un motif *fermé* fréquent, s'il n'existe pas de motif fréquent S' tel que $S < S'$ et $sup(S) = sup(S')$. Par exemple, soient $S = \langle (homme\ NOM)\ (de)\ (NOM) \rangle$ et $S' = \langle (homme\ NOM)$

[3]Parfois le support normalisé est utilisé : $sup(S) = \frac{nombre\ de\ sequences\ de\ la\ base\ qui\ contiennent\ S}{nombre\ de\ sequences\ de\ la\ base}$

$(de)\ (culture\ NOM)\rangle$ deux séquences telles que $sup(S) = sup(S') = 10$, alors S n'est pas un motif fermé. En effet les 10 exemples couverts par le motif S sont aussi couverts par le motif S'.

Les motifs fermés fréquents extraits sont partiellement ordonnés entre eux. Afin de mieux visualiser cette relation d'ordre on peut afficher les motifs dans un *diagramme de Hasse* qui est une représentation graphique d'un ordre partiel (Davey & Priestley, 1990). Un exemple de diagramme de Hasse est présenté à la figure 2. La taille du diagramme de Hasse peut être trop grande pour que le diagramme soit affiché entièrement. Nous proposons donc d'utiliser un outil de navigation permettant de naviguer dans le diagramme des motifs les plus généraux aux plus spécifiques afin de les valider en tant que patrons linguistiques. À la figure 1 nous donnons un exemple de navigation avec l'outil Camelis[4] (Ferré, 2009). La navigation se fait via « l'arbre de navigation » se trouvant à gauche de l'outil. Notons que lorsqu'un motif, M, est sélectionné par l'expert comme patron linguistique, tous les motifs dont M est une sous-séquence ne sont plus à examiner. En effet, lors de l'application des patrons, les morceaux de texte qu'ils reconnaissent sont inclus dans les morceaux de texte reconnus par M. Camelis offre la possibilité de marquer les motifs validés comme patrons linguistiques et donc de ne plus les afficher ainsi que les motifs qui les contiennent (cf « not 'Motif Valide' » dans figure 1), réduisant l'espace des motifs à vérifier et facilitant l'exploration.

3 Cas d'étude : la reconnaissance d'expressions qualificatives

Nous avons appliqué notre méthode à l'apprentissage de patrons linguistiques dénotant des expressions porteuses de qualification et en position détachée comme décrites dans (Jackiewicz *et al.*, 2009a). Les expressions en gras dans les exemples (1) et (2) illustrent le type d'expressions qui nous intéressent :

(1) **Ni trop sentimental**, **ni trop énergique**, *il maîtrise*, **avec une finesse quasi mozartienne**, *un [...]*

(2) **Figure légendaire de l'opposition au régime communiste**, **éminent professeur d'histoire médiévale**, **ministre des affaires étrangères de la Pologne de 1997 à 2000**, *Bronislaw Geremek avait [...]*

Nous souhaitons apprendre des patrons linquistiques comme ceux définis manuellement par (Jackiewicz *et al.*, 2009a), par exemple :
– Groupes nominaux (GN) : [det] N de GN (*Femme de tête,X* ; *X, le maestro de la désinflation*).
– Adverbes (*courageusement, X*) ; groupes prépositionnels (*en mauvaise posture, X*).
– Constructions détachées : groupes adjectivaux (*imprévisible et fantasque, X*) ; constructions absolues (*l'oeil vigilant, X*) ; participes (*réputé pour son caractère bourru, X*).

3.1 Corpus d'apprentissage

Deux corpus d'apprentissage ont été générés automatiquement pour pallier l'absence de corpus disponible. Le corpus **AXIOLO** est issu des expériences de (Jackiewicz *et al.*, 2009a). Il est constitué d'expressions reconnues par application d'une dizaine de patrons élaborés manuellement sur des articles issus du journal Le Monde, catégorie « Portraits », de la période juillet à décembre 2002 (soit 884 articles), ainsi que sur la période 2003 à 2006 de l'ensemble des articles du Monde pour deux autres patrons spécifiques : *en Adj <expansion>*[5] et *en N <expansion>*[6] (Jackiewicz *et al.*, 2009b). Ce corpus d'expressions qualificatives

[4] http ://www.irisa.fr/LIS/ferre/camelis/index.html
[5] avec Adj appartenant à une liste d'adjectifs fixés (bon, vrai, authentique, ...) : *en vrai professionnel*
[6] avec N appartenant à une liste de noms (homme, femme, virtuose, ...) : *en homme sensible et généreux*

contient $4\,063$ expressions (i.e. séquences), ce qui représente $12\,257$ mots. Il est très peu bruité. Le corpus **ARTS** a été généré à partir de règles heuristiques sur les articles de la rubrique "Art" du journal Le Monde, année 2006, soit 3 539 articles. Ces heuristiques sont destinées à filtrer parmi les constituants périphériques ceux qui ne sont a priori pas porteurs de qualification (exemples : proposition de la phrase contenant un verbe conjugué, groupe circonstanciel de temps, espace, but, causalité). Ce corpus est constitué de $13\,576$ expressions (i.e. séquences), ce qui représente $85\,153$ mots. Il contient des exemples négatifs. Nous estimons à environ 32% le pourcentage d'expressions non qualificatives (bruit).

3.2 Apprentissage des patrons de qualification

Paramètres de l'apprentissage. Pour le calcul des motifs fermés d'itemsets, nous avons utilisé l'implémentation de Clospan (Yan *et al.*, 2003) proposée dans Illimine[7]. Nous cherchons des patrons linguistiques décrivant des expressions de qualification en position détachée. Pour cela, nous fixons deux contraintes sur les motifs à extraire : 1) ils débutent l'expression ; 2) ils sont formés d'éléments contigüs (maxgap=0). Pour chacun des deux corpus nous avons calculé les motifs séquentiels en faisant varier les valeurs du seuil de support, $minsup$, de 50% à $0,05\%$. On constate que des seuils de support élevés (50% ou 25%) ne fournissent que des motifs très généraux où seuls les catégories grammaticales apparaissent. Il est donc plus intéressant de choisir des seuils de support très bas pour obtenir des motifs spécifiques et capturer des expressions, ou phénomènes, linguistiques peu fréquents[8].

Résultats quantitatifs. Avec le corpus AXIOLO, pour $minsup = 0,05$ (2 expressions), $8\,264$ motifs fermés fréquents sont extraits en moins d'une seconde. Après application des deux contraintes il reste $1\,789$ motifs. Avec le corpus ARTS, pour $minsup = 0,05$ (6 expressions), environ 8 millions de motifs fermés fréquents sont extraits en 7h. Après application des contraintes il reste $7\,818$ motifs[9].

Résultats qualitatifs et discussion. Des expériences ont été conduites en ne considérant qu'un seul item pour décrire un mot comme dans (Charnois *et al.*, 2009). Sans surprise, on constate que les motifs découverts sont soit très spécifiques (par exemple : $\langle homme\ de\ conviction \rangle$ pour les séquences représentées par les lemmes seuls), soit très génériques ($\langle NOM\ PRP\ NOM \rangle$) lorsque les séquences ne sont formées que des catégories grammaticales. Un motif comme $\langle (homme)(de\ PRP)(NOM) \rangle$ ne peut être appris qu'à partir de séquences d'itemsets. De plus, l'analyse des motifs séquentiels d'itemsets extraits du corpus AXIOLO montre la complétude de la méthode. On retrouve en effet tous les motifs présentés précédemment et qui ont servi à générer ce corpus. Enfin, les expériences réalisées sur le corpus ARTS ont permis de tester la méthode à une échelle relativement importante sur un corpus généré automatiquement et non annoté. L'ensemble des motifs produits comporte des motifs inintéressants dûs au bruit présent dans le corpus. Face à ce problème, la navigation au sein de la hiérarchie des motifs est un point fort de la méthode permettant aisément d'élaguer des groupes de motifs inintéressants. Enfin, de nouveaux patrons sont extraits au regard de ceux déjà conçus manuellement dans (Jackiewicz *et al.*, 2009a). C'est le cas du motif $\langle (ADJ\ pour\ DET\ NOM) \rangle$ ("célèbre pour son monastère", "baroque pour une histoire d'amour", ...) et de ses variantes ou extensions : $\langle (ADV)(ADJ)(pour) \rangle$ ("très célèbre pour..."), $\langle (ADJ)(pour)(VER) \rangle$ ("indispensable pour assurer...").

[7] http ://illimine.cs.uiuc.edu/

[8] Les motifs obtenus sont consultables : `http://users.info.unicaen.fr/~pcellier/taln2010/`

[9] Soulignons que les indices classiques de rappel et de précision ne sont pas adaptés pour cette approche qui vise à organiser et non à élaguer les motifs extraits. Cette méthode aurait donc un très bon rappel et une précision médiocre.

4 Conclusion

Dans cet article nous présentons une méthode utilisant l'extraction de motifs séquentiels d'itemsets pour la génération automatique de patrons linguistiques. Cette approche a l'avantage d'éviter un recoupement manuel d'expressions pour déterminer des patrons. De plus, les patrons extraits sont compréhensibles par un humain. L'avantage de décrire les mots non plus par un seul lemme comme dans (Charnois *et al.*, 2009), mais par un ensemble d'items, est une plus grande expressivité des patrons linguistiques découverts (combinant des informations hétérogènes). Nous proposons aussi une solution pour gérer le nombre de motifs extraits en s'appuyant sur un ordre partiel qui existe entre ces motifs. Un utilisateur humain peut ainsi facilement naviguer dans les motifs et les valider en tant que patrons linguistiques. Nous avons utilisé notre méthode pour l'apprentissage de patrons linguistiques pour la reconnaissance de constituants de qualification en position détachée. Le processus peut être utilisé pour apprendre d'autres types de patrons linguistiques, comme les relations entre entités nommées, en définissant des contraintes appropriées.

Références

AGRAWAL R. & SRIKANT R. (1995). Mining sequential patterns. In *Int. Conf. on Data Engineering* : IEEE.

CALIFF M. E. & MOONEY R. J. (2003). Bottom-up relational learning of pattern matching rules for information extraction. *J. Mach. Learn. Res.*, **4**, 177–210.

CHARNOIS T., PLANTEVIT M., RIGOTTI C. & CRÉMILLEUX B. (2009). Fouille de données séquentielles pour l'extraction d'information. *Traitement Automatique des Langues*, **50**(3).

DAVEY B. A. & PRIESTLEY H. A. (1990). *Introduction To Lattices And Order*. Cambridge University Press.

DONG G. & PEI J. (2007). *Sequence Data Mining*. Springer.

FERRÉ S. (2009). Camelis : a logical information system to organize and browse a collection of documents. *Int. J. General Systems*, **38**(4).

HEARST M. (1992). Automatic acquisition of hyponyms from large text corpora. In *Int. Conf. on Computational Linguistics*.

JACKIEWICZ A., CHARNOIS T. & FERRARI S. (2009a). Jugements d'évaluation et constituants périphériques. In *Conférence sur le traitement automatique des langues naturelles*.

JACKIEWICZ A., VIGIER D., CHARNOIS T. & FERRARI S. (2009b). Vers une analyse automatique des discours évaluatifs. Le cas des constituants détachés "en N <exp>". In *Linguistic and Psycholinguistic Approaches to Text Structuring* : ENS.

PANG B. & LEE L. (2007). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, **2**(1-2), 1–135.

POIBEAU T. (2003). *Extraction automatique d'information : Du texte brut au web sémantique*. Lavoisier.

RILOFF E. (1996). Automatically generating extraction patterns from untagged text. In *AAAI/IAAI'96*.

SRIKANT R. & AGRAWAL R. (1996). Mining sequential patterns : Generalizations and performance improvements. In *Int. Conf. on Extending Database Technology (EDBT)*, LNCS, p. 3–17 : Springer.

YAN X., HAN J. & AFSHAR R. (2003). Clospan : Mining closed sequential patterns in large databases. In *Int. Conf. on Data Mining* : SIAM.