SDMC : un outil en ligne d'extraction de motifs séquentiels pour la fouille de textes

Nicolas Béchet, Peggy Cellier, Thierry Charnois, Bruno Crémilleux, Solen Quiniou

SDMC : un outil en ligne d'extraction de motifs séquentiels pour la fouille de textes

Nicolas Béchet*, Peggy Cellier**, Thierry Charnois*,***, Bruno Cremilleux*,
Solen Quiniou****

*GREYC, UMR 6072, CNRS, Université de Caen Basse-Normandie,
14032 Caen Cedex, France, {prenom.nom}@unicaen.fr,
**IRISA, UMR 6074, INSA Rennes,
35042 Rennes cedex, France, peggy.cellier@irisa.fr
***MoDyCO, UMR 7114, CNRS, Univ. Paris-Ouest Nanterre La Défense,
92 001 Nanterre Cedex
****LINA, UMR 6241, CNRS, Université de Nantes,
44322 Nantes Cedex 3, solen.quiniou@univ-nantes.fr

1 Introduction

Introduite par Srikant et Agrawal (1996), la fouille de données séquentielles permet de découvrir des corrélations entre des événements selon une relation d'ordre (e.g. le temps). Ce domaine est devenu au fil des années un champ actif de la fouille de données avec de nombreuses applications comme l'analyse de séquences biologiques, le web mining, ou encore la fouille de textes. L'information découverte se présente usuellement sous la forme de motifs séquentiels. Deux défis majeurs du domaine sont d'une part la définition de méthodes et d'outils permettant d'appréhender de très grands volumes de données et d'autre part la sélection de motifs potentiellement intéressants. Bien que de nombreux outils permettant d'extraire des motifs séquentiels existent dans la littérature (Srikant et Agrawal (1996); Yan et al. (2003); Wang et Han (2004); Zaki (2001); Nanni et Rigotti (2007)), il n'existe à notre connaissance pas d'outil en ligne permettant d'extraire des motifs séquentiels propres aux données textuelles.

Notre objectif est de permettre à des non spécialistes d'extraire des motifs séquentiels sans connaissance a priori en fouille de données. Dans ce contexte, un motif séquentiel est une suite ordonnée d'itemsets. Un itemset peut alors être composé d'informations multiples comme le mot lui même, son lemme, sa catégorie grammaticale. Par exemple, le mot *champions* peut être représenté par l'itemset $\langle(champions\ champion\ NN)\rangle$. Ainsi, le motif séquentiel $\langle(Champions\ champion\ NN)(monde\ NN)\rangle$ signifie que les mots "champions" et "monde" apparaissent souvent ensemble dans des phrases de notre corpus.

L'outil SDMC (Sequential Data Mining under Constraints) présenté fait suite à un intérêt pour la fouille de motifs séquentiels de la part de la communauté de chercheurs linguistes en statistique textuelle, ou encore en linguistique de corpus dans le cadre de projets CPER [1].

1. Contrat Projet État Région, projets « Outils et méthodes pour l'exploration des textes en sciences humaines » et « Hybridation des méthodes de traitement automatique des langues avec la fouille de données » et financés par la région Basse-Normandie.

SDMC : un outil en ligne d'extraction de motifs séquentiels pour la fouille de textes

Cet outil permet d'extraire des motifs séquentiels d'itemsets adaptés à la fouille de textes, en proposant à l'utilisateur de fixer certains critères.

2 Description de l'outil

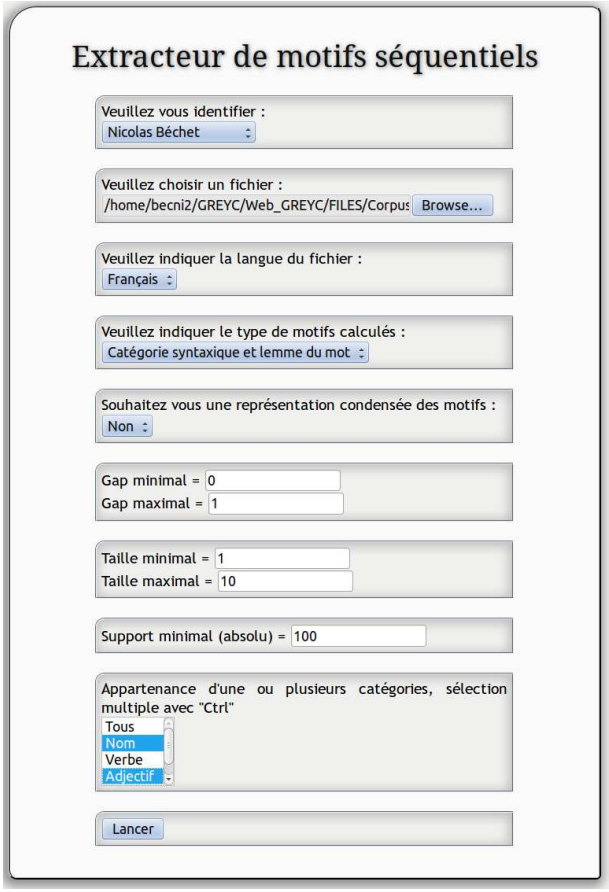

FIG. 1 – *Capture d'écran du site Web de l'extracteur de motifs séquentiels*

Notre outil se présente sous la forme d'une page Web [2] (cf. figure 1). Destinée à des non spécialistes, l'interface à été simplifiée pour en faciliter l'usage. L'utilisateur a alors la possibilité de s'authentifier et de soumettre un fichier à partir duquel les motifs séquentiels seront extraits. Ce corpus peut être rédigé en français ou en anglais, et doit comporter du texte brut. Une phase d'étiquetage grammaticale est réalisée en utilisant l'outil TreeTagger [3].

2. L'outil est accessible à l'adresse https ://sdmc.greyc.fr/, login et mot de passe sont à demander aux auteurs
3. http ://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

La base de séquences, à partir de laquelle seront extraits les motifs, est constituée en fonction des choix de l'utilisateur. Il est ainsi possible d'utiliser les lemmes du mot seul, la forme seule, la catégorie syntaxique seule, ou finalement la catégorie syntaxique et le lemme du mot. Prenons par exemple l'expression "*Champions du monde*". Une fois l'extracteur grammatical appliqué elle devient " *(Champions champion NN) (de de PRP) (le le DET) (monde monde NN)*". Si l'utilisateur choisi de conserver les lemmes uniquement, la séquence, qui est la représentation d'une phrase où chaque mot est un itemset, va être : $\langle(champion)(de)(le)(monde)\rangle$. En revanche, en conservant les lemmes et catégories, la séquence devient :
$\langle(champion\ NN)(de\ PRP)(le\ DET)(monde\ NN)\rangle$.

L'un des principaux avantages de notre outil est la possibilité d'appliquer un certain nombre de contraintes au processus d'extraction de motifs. Ces contraintes sont particulièrement adaptées à la fouille de textes, afin de modéliser des connaissances linguistiques et de filtrer les motifs les plus pertinents en fonction de la problématique.

Nous proposons ainsi à l'utilisateur différentes contraintes :

- La contrainte de **support minimal**. Cette dernière repose sur la notion de support d'un motif qui peut être défini dans ce cas précis comme le nombre de phrases contenant le motif extrait. Ainsi, le support minimal est le nombre minimal de phrases dans lequel ce motif est observé. Cette contrainte traduit une certaine régularité des motifs produits.
- Une autre contrainte intéressante est la contrainte de **gap**. Un motif séquentiel avec contrainte de gap $[M, N]$, noté $P_{[M,N]}$ est un motif tel qu'au minimum M itemsets et au maximum N itemsets sont présents entre chaque itemset voisin du motif dans les séquences à partir desquelles il est extrait.
- La contrainte de **longueur**, qui indique le nombre minimal et maximal d'itemsets dans un motif.
- Une dernière contrainte actuellement proposée dans notre outil est l'**appartenance** qui est particulièrement utile avec des données textuelles. Cette contrainte permet à l'utilisateur d'obtenir des motifs contenant au moins un verbe et/ou un nom et/ou un adjectif et/ou un adverbe.

L'algorithme que nous avons proposé pour extraire les motifs séquentiels se fonde sur la notion de *pattern growth* (Pei et al. (2001)) et est brièvement discuté dans Béchet et al. (2013). D'autres contraintes sont implémentées dans l'extracteur de motifs mais n'ont pas encore été portées sur l'outil en ligne comme la contrainte d'**association** ou de **commence par** (cf. Béchet et al. (2012)).

L'utilisateur a la possibilité d'utiliser une représentation condensée des motifs afin d'en réduire le nombre. Une fois l'extraction lancée, il peut récupérer les motifs extraits par un lien téléchargeable transmis par courrier électronique.

3 Conclusion

Nous avons présenté un outil d'extraction de motifs séquentiels adapté à la fouille de textes. Ce dernier est déjà expérimenté par des chercheurs en sciences humaines, comme le CRISCO à Caen ou le LASLA à Liège. En fonction des usages dans ce domaine, l'outil sera amené à évoluer. Nos allons ainsi ajouter prochainement de nouvelles contraintes à notre outil, comme la contrainte d'association. Cette contrainte porte sur les itemsets des motifs extraits et permet d'associer systématiquement à une catégorie syntaxique donnée (*e.g.* la catégorie verbale), le

lemme associé. De même, le calcul du taux d'émergence (ou spécificité) des motifs extraits d'un corpus vis à vis de motifs extraits d'un autre corpus sera intégré ; cette fonctionnalité est intéressante pour l'analyse stylistique (par exemple de textes littéraires). De plus, nous souhaitons permettre l'importation de fichiers XML, pouvant être déjà étiquetés. Nous proposerons aussi différents formats d'exportation de nos motifs, afin de permettre leur édition dans d'autre outils comme Camelis [4] (Ferré (2009)) ou TXM [5] (Heiden (2010)).

Enfin, l'outil sera aussi adapté pour un usage générique, c'est-à-dire pour la fouille de séquences quelconques sous un format numérique standard. Une page web va être développée pour permettre un tel usage.

Références

Béchet, N., P. Cellier, T. Charnois, et B. Crémilleux (2012). Discovering linguistic patterns using sequence mining. In *proceedings of CICLing'2012*, pp. 154–165.

Béchet, N., P. Cellier, T. Charnois, et B. Cremilleux (2013). Extraction de motifs séquentiels sous contraintes multiples. In *proceedings of EGC'2013, to appear*.

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

Heiden, S. (2010). The TXM Platform : Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In *proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, Sendai, Japon, pp. 389–398.

Nanni, M. et C. Rigotti (2007). Extracting trees of quantitative serial episodes. In *Proc. of KDID'07*, pp. 170–188.

Pei, J., J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, et M. Hsu (2001). Prefixspan : Mining sequential patterns by prefix-projected growth. In *ICDE*, pp. 215–224.

Srikant, R. et R. Agrawal (1996). Mining sequential patterns : Generalizations and performance improvements. In *EDBT*, pp. 3–17.

Wang, J. et J. Han (2004). Bide : Efficient mining of frequent closed sequences. In *Int. Conf. on Data Engineering*, pp. 79–90.

Yan, X., J. Han, et R. Afshar (2003). Clospan : Mining closed sequential patterns in large databases. In *SDM*.

Zaki, M. J. (2001). SPADE : An efficient algorithm for mining frequent sequences. *Machine Learning Journal 42*(1/2), 31–60. special issue on Unsupervised Learning.

Summary

We present a tool in order to extract sequential pattern. This tool is specially adapted to text mining, allowing part-of-speech extraction and linguistic constraints like gap and membership. Already used by other institutions, our tool can be useful for many text mining tasks like clustering or named entity recognition.

4. http ://www.irisa.fr/LIS/ferre/camelis/index.html
5. http ://textometrie.ens-lyon.fr/