HAL
archives-ouvertes.fr

Vectorisation paramétrée des données textuelles

Célia Da Costa Pereira, Mathieu Lafourcade, Patrick Lloret, Cédric Lopez, Mathieu Roche

▶ **To cite this version:**

Célia Da Costa Pereira, Mathieu Lafourcade, Patrick Lloret, Cédric Lopez, Mathieu Roche. Vectorisation paramétrée des données textuelles. EGC: Extraction et Gestion des Connaissances, Jan 2014, Rennes, France. 14èmes Journées Internationales Francophones sur l'Extraction et la Gestion des Connaissances, RNTI-E-26, pp.593-596, 2014. <hal-01333661>

HAL Id: hal-01333661

https://hal.archives-ouvertes.fr/hal-01333661

Submitted on 18 Jun 2016

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.

Vectorisation paramétrée des données textuelles

Célia da Costa Pereira [*], Mathieu Lafourcade [**],
Patrick Lloret [***], Cédric Lopez [****], Mathieu Roche [**,♯]

[*] I3S, UMR 7271, Sophia Antipolis - France – celia.pereira@unice.fr
[**] LIRMM, UMR 5506, Montpellier - France – {prénom.nom}@lirmm.fr
[***] Succeed Together, Paris - France – plloret@succeed-together.eu
[****] Objet Direct, Grenoble - France – clopez@objetdirect.com
[♯] UMR TETIS, Montpellier - France – mathieu.roche@cirad.fr

1 Introduction

L'expression en langage naturel recèle des informations riches que les analystes souhaitent souvent explorer. Dans le cadre de l'activité de la Société *Succeed Together* qui consiste, entre autres, à recueillir et analyser des informations produites lors de séminaires interactifs, les animateurs développent et structurent les discussions établies avec les participants. Les réponses ou remarques apportées par les participants peuvent alors être consignées puis traitées, une phase de regroupent est au préalable nécessaire. Le but est ainsi de mettre en exergue des sentiments partagés par les participants selon une thématique donnée. Dans ce cadre, les travaux menés par le LIRMM (Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier) liés au traitement automatique des données textuelles, permettent aux experts de Succeed Together d'analyser semi-automatiquement et à plus grande échelle les données. Ainsi, nous avons focalisé notre étude sur la représentation des données textuelles par des méthodes de TAL (Traitement Automatique du Langage Naturel). Ceci permet, en particulier, d'améliorer les méthodes de classification et/ou regroupement effectuées par le deuxième collaborateur académique du projet (I3S / Université de Nice).

Dans un premier temps, en section 2, nous décrivons les méthodes de représentation des descripteurs textuels. Une application spécifiquement dédiée au projet a été développée. Cette application est décrite en section 3. Les résultats issus des données fournies par la société sont décrits et analysés en section 4. Enfin, quelques perspectives sont données en sections 5.

2 Descripteurs textuels pour les tâches de Clustering

La sélection de descripteurs pertinents à partir de textes est une étape indispensable pour une tâche de clustering (regroupement) qui consiste à regrouper les documents ayant des contenus sémantiques proches. Pour appliquer les algorithmes de regroupement, il est dans un premier temps nécessaire d'établir une représentation pertinente des documents (Béchet (2009)). Dans cet article, nous nous concentrons sur la représentation vectorielle de Salton et al. (1975).

Nous nous appuierons sur le principe *sac de mots* appliqué aux textes des séminaires de la société *Succeed Together*. Plusieurs types de représentations sont alors possibles :

- **Représentation booléenne :** Le vecteur booléen donne des informations liées à la présence ou l'absence d'un descripteur dans un document.
- **Représentation fréquentielle :** La représentation fréquentielle de base revient à considérer le nombre d'occurrences d'un terme i dans un document j, la normalisation par rapport aux nombre de mots dans un document peut être aussi appliquée.
- **Pondération TF-IDF :** La mesure TF-IDF consiste à calculer l'importance et la discriminance d'un mot dans un document relativement à une collection (Salton et al. (1975)).

3 Logiciel de Vectorisation

Dans le cadre du projet, nous avons développé un logiciel dédié à l'extraction des descripteurs utiles pour l'étape de clustering afin de représenter les textes sous forme d'une matrice. Le format ARFF choisi pour cette représentation est typique des logiciels de clustering tel que Weka [1]. De manière plus précise, la matrice de sortie est constituée d'un nombre de lignes correspondant au nombre de réponses des participants pour une question donnée lors des séminaires. La capture d'écran du logiciel (cf. Figure 1) montre les différentes fonctionnalités détaillées ci-dessous :

- **Fichiers d'entrée.** Deux types de données peuvent être fournies en entrée du logiciel. Le premier correspond aux données textuelles fournies par Succeed Together (format XML). Cette liste est pour l'instant figée. Mais l'utilisateur peut aussi introduire de nouvelles données textuelles via le panneau dédié.
- **Traitement de type n-grammes et par patrons grammaticaux.** Différents types de filtres peuvent être appliqués. Les premiers filtres que nous proposons sont de types n-grammes de mots. L'utilisateur peut choisir le nombre n par l'intermédiaire de notre logiciel. Par ailleurs, l'utilisateur peut effectuer une recherche terminologique selon des patrons définis reposant sur un étiqueteur grammatical [2] (Nom-Préposition-Nom, Nom-Nom, Nom-Adjectif, Adjectif-Nom, etc). Les syntagmes extraits sur la base de patrons morpho-syntaxiques se révèlent en général plus pertinents.
- **Traitement statistique**. Le logiciel permet de ne conserver que les n-grammes présents au moins N_{occ} fois, ce seuil étant déterminé par l'utilisateur. Précisons qu'un élagage trop strict peut engendrer un résultat nul si la taille du corpus est insuffisante.
- **Génération de différentes sorties**. Le logiciel développé propose différents types de sorties (cf. Figure 1).

4 Expérimentations

Dans cette section, nous discutons les résultats obtenus en utilisant différents paramètres du logiciel développé dans le cadre du projet.

Filtres de n-grammes. La Figure 2 indique le nombre de n-grammes moyen extrait en fonction de n et en fixant le nombre d'occurrences $N_{occ} = 3$. Les résultats indiquent que

1. http ://www.cs.waikato.ac.nz/ml/weka/
2. Sygfran : http ://www.lirmm.fr/∼chauche/ExempleAnl.html

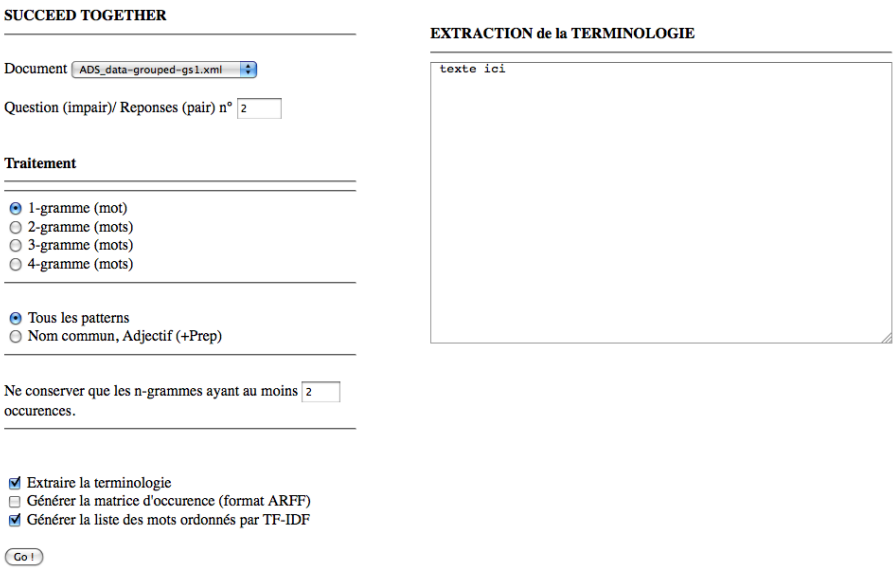

Fig. 1: Vue globale du logiciel d'extraction des descripteurs linguistiques

pour le corpus donné, de nombreux n-grammes sont extraits pour $n = 1$. Avec un nombre n supérieur ou égal à 2, le nombre de n-grammes extraits est significativement plus faible. Cette même constatation a été relevée quel que soit N_{occ}.

Filtres morpho-syntaxiques. Sans filtrage syntaxique (FS), de nombreux n-grammes retournés se révèlent non pertinents. Par exemple, les 2-grammes *des collaborateurs* ou *notre présence*. À partir d'un des fichiers caractéristiques fourni par la société, 437 n-grammes sont extraits (avec $N_{occ} = 2$). En appliquant les filtres syntaxiques, seulement 5% des 2-grammes sont conservés par rapport à l'extraction sans FS. Par exemple, nous retenons les syntagmes *libre service*, *nouvelles technologies*, *service client*, *satisfaction client*, *contact humain*, *relation client*. De tels termes se révèlent sémantiquement plus pertinents.

Filtres fréquentiels. La Figure 2 indique le nombre moyen de 1-grammes extraits en fonction du seuil N_{occ}. Par exemple, 96 n-grammes sont extraits avec un seuil égal à 2. Il est notable qu'à partir de $N_{occ} = 6$, très peu de 1-grammes sont extraits.

Par ailleurs, nous avons comparé un classement de type fréquentiel avec un classement sur la base d'une pondération TF-IDF. Nous avons calculé le recouvrement moyen des termes selon ces classements. Par exemple, avec $N_{occ} = 2$, environ 44% des 1-grammes sont communs aux deux mesures (en tenant compte des 10 premiers termes de plus haut score).

5 Conclusion et perspectives

Dans le cadre de ce projet, nous nous sommes concentrés sur l'extraction des descripteurs linguistiques à partir des données textuelles fournies par la société *Succeed Together*. Nous avons évalué l'utilisation de différentes méthodes de Traitement Automatique du Langage à

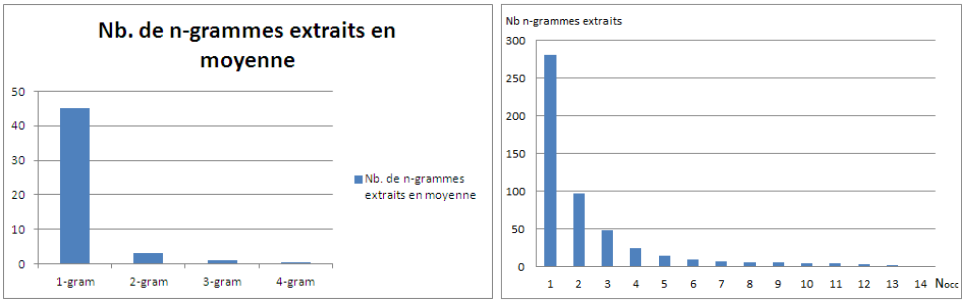

Fig. 2: Figure de gauche : Nombre de n-grammes extraits en fonction de n (N_{occ}=3).
Figure de droite : Nombre moyen de 1-grammes extraits en fonction de N_{occ} ($1 \leq N_{occ} \leq 14$).

partir de ces données. Par la suite, il semble essentiel d'évaluer l'utilisation de ces différents descripteurs pour la tâche de clustering menée par le laboratoire I3S de l'Université de Nice.

De plus, nous souhaitons intégrer des informations sémantiques issues de *Jeux de Mots (JDM)* (Lafourcade et Joubert (2009)) et son réseau lexical d'envergure. La base lexicale de JDM a été amorcée avec environ 150.000 termes et aucune relation entre ces termes. En 5 ans, plus de 2.200.000 relations entre 260.000 termes ont été collectées de manière contributive. Le réseau lexical peut être exploré via une interface dictionnaire (nommée Diko) accessible à l'adresse : http ://www.jeuxdemots.org/diko.php. L'expérience menée avec *Succeed Together* a permis, sur plusieurs séances, d'introduire de nouveaux termes liés au management d'entreprise et de valider environ 300 relations. Ces nouveaux termes permettent de consolider et d'enrichir les descripteurs linguistiques issus des textes. Ceci constitue une perspective particulièrement intéressante dans nos futurs travaux.

Références

Béchet, N. (2009). *Extraction et regroupement de descripteurs morpho-syntaxiques pour des processus de Fouille de Textes*. Ph. D. thesis, Université Montpellier 2.

Lafourcade, M. et A. Joubert (2009). Similarity between term senses in a lexical network. *TAL 50*(1), 177–200.

Salton, G., A. Wong, et C. S. Yang (1975). A vector space model for automatic indexing. *Commun. ACM 18*(11), 613–620.

Summary

Automatic processing of textual data enables users to analyze semi-automatically and on a large scale the data. This analysis is based on two successive processes: (i) representation of texts, (ii) gathering of textual data (clustering). The software described in this paper focuses on the first step of the process by offering expert a parameterized representation of textual data.