



HAL
open science

VIPE : un outil interactif de classification multilabel de messages courts

Frank Meyer, Sylvie Tricot, Pascale Kuntz, Wissam Siblini

► **To cite this version:**

Frank Meyer, Sylvie Tricot, Pascale Kuntz, Wissam Siblini. VIPE : un outil interactif de classification multilabel de messages courts. Extraction et Gestion des Connaissances. EGC 2017, Jan 2017, Grenoble, France. hal-01624614

HAL Id: hal-01624614

<https://hal.science/hal-01624614v1>

Submitted on 26 Oct 2017

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.

VIPE : un outil interactif de classification multilabel de messages courts

Frank Meyer *, Sylvie Tricot *
Pascale Kuntz **, Wissam Siblini *, **

*Orange Labs - 2 av. Pierre Marzin - 22 300 Lannion, France
prenom.nom@orange.com,

**Laboratoire d'Informatique de Nantes Atlantique - Site Polytech 44300 Nantes, France
prenom.nom@univ-nantes.fr

Résumé. Nous présentons un outil interactif de classification multilabel développé au sein du groupe Orange et utilisé pour l'analyse d'opinions. Basé sur un algorithme de factorisation rapide de matrice, il permet à un utilisateur d'importer des textes courts (tweets, mails, enquêtes, ...), de définir des labels d'intérêts (« client globalement satisfait », « évoque la rapidité du débit »,...) et de proposer pour chaque texte des recommandations de labels et pour chaque label des recommandations de textes.

1 Introduction

L'analyse d'opinions est un enjeu majeur pour les entreprises qui visent à améliorer en permanence leur relation client. Aux enquêtes par sondage s'ajoutent pour l'analyse les informations extraites sur les médias sociaux. Ces informations contribuent à déterminer le degré d'engouement suscité par les offres d'entreprises, à identifier les différents points de vue et les points de convergence entre les clients, et à recueillir de l'information « fraîche » (Gauzente et al. (2012)). Cependant, l'acquisition des informations utiles est une tâche difficile car les sources complémentaires dont elles sont extraites sont hétérogènes et contiennent des données volumineuses, bruitées et non structurées. Les problèmes associés à l'analyse de ces données rendent le traitement automatique délicat en pratique et l'implication de l'utilisateur est cruciale (Keim et al. (2013)).

L'intégration de l'humain dans la boucle d'apprentissage connaît en effet un essor croissant et des systèmes de classification interactifs ont été développés pour des applications variées : e.g. classification d'images (cueFlick), sélection de fichiers (Smart Selection), classification de gestes (Wekinator), classification de documents (iCluster), tri d'alarmes (CueT). Dans ce cadre, l'utilisateur annote, via une interface adaptée, un nombre limité d'exemples et, à partir de ces quelques exemples, un algorithme d'apprentissage tente de capturer l'expertise pour apprendre un premier modèle prédictif. En fonction de sa satisfaction, l'utilisateur peut arrêter l'apprentissage ou continuer à entraîner le modèle. Les retours expérimentaux menés sur des petits échantillons d'utilisateurs semblent très prometteurs. Cependant, la plupart des systèmes existants se limitent à une classification monolabel où un seul label peut être affecté à la fois à un exemple ; ce qui est peu expressif d'autant plus que les données sont très souvent de

VIPE : un outil interactif de classification multilabel de messages courts

nature multi-label. Dans le cadre de l'analyse d'opinions, il s'agit effectivement de dépasser le cadre positif/négatif pour prendre en compte des comportements plus subtils. L'apprentissage multi-label a suscité une grande attention cette dernière décennie (Zhang et Zhou (2014)) et a conduit au développement de nombreuses approches. Mais comme l'ont montré récemment Nair-Benrekia et al. (2015), peu de ces approches résistent aux contraintes d'interactivité.

Dans cette communication, nous présentons un nouvel outil VIPE (« Visual Interactive and Personalized Exploration - of data ») permettant une classification interactive multi-label de textes courts provenant de transcriptions de résultats d'enquêtes d'opinions sur le Web ou par centres d'appels, de forums spécialisés ou de Twitter. L'utilisateur définit initialement un ensemble de labels (par exemple : Efficacité, Innovation, Couverture réseau, Négatif, Positif) puis il procède à la classification manuelle d'un ensemble restreint de textes. Par exemple, il annote le tweet "ça c'est de la #4G !" en positif avec les deux labels : Efficacité et Innovation. En se basant sur l'ensemble de textes étiquetés, un algorithme d'apprentissage assiste ensuite l'agent en lui prédisant les labels les plus probables pour un ensemble de textes ou les textes les plus probables pour un label ou une combinaison sélectionnés.

2 Architecture de VIPE

VIPE est une application Web composée de 4 modules. Le **module 1** est un gestionnaire d'import des textes courts par chargement de fichiers textes et d'export des résultats de l'apprentissage. Les textes sources sont partagés entre les utilisateurs mais les labels sont du domaine privé de chacun. VIPE apprend un modèle unique sur les différentes sources et labels, mais chaque utilisateur peut uniquement annoter ou consulter les résultats pour ses propres labels. Les résultats peuvent être de trois types : (i) les textes correspondants à des labels donnés - sélectionnés par un ranking sur la matrice des facteurs calculée par l'algorithme de factorisation - ; (ii) les labels prédits pour un texte donné ; (iii) la matrice des scores des labels prédite pour l'ensemble des textes.

Le **module 2** est un gestionnaire de la matrice $X_{m \times n}$ de données qui contient en ligne les textes et en colonne deux informations : (i) une description des textes en sacs de n-grammes (matrice $X_{m \times n_1}$) et (ii) une description binaire de l'affectation des textes aux labels (matrice $X_{m \times n_2}$).

Le **module 3** est composé d'un serveur et d'une interface Web qui permet de manipuler les textes et d'effectuer l'étiquetage interactif (figure 1). VIPE intègre trois actions pour l'apprentissage interactif : (i) par correction manuelle des résultats proposés ; (ii) par sélection d'exemples positifs et négatifs à partir de la base de textes. Cela est fait manuellement par une requête à base de mots-clés ou en parcourant la liste de textes ; (iii) par ajout d'exemples virtuels grâce à une boîte de dialogue qui permet à l'utilisateur de construire un exemple et de l'étiqueter.

Le **module 4** contient l'algorithme d'apprentissage basé sur l'algorithme de factorisation de matrices Molecule. Molecule est une adaptation de l'algorithme Gravity (Takács et al. (2007)) pour les données « positive-only ». Pour rappel, la version de base Gravity factorise une matrice creuse $X_{m \times n}$ en l'approximant par la matrice de rang faible $P^T Q$ (où $P \in \mathbb{R}^{k \times m}$ et $Q \in \mathbb{R}^{k \times n}$). Les matrices P et Q sont calculées par minimisation de l'erreur quadratique moyenne (RMSE) entre $X_{m \times n}$ et $P^T Q$, sur les cellules non nulles de $X_{m \times n}$, par descente de gradient stochastique avec régularisation. Initialement la matrice $X_{m \times n_1}$ (description des

The screenshot shows the VIPE Text interface. On the left, a sidebar lists sentiment labels with their counts and percentages:

- Très content: 20 (21.06%)
- très mécontent: 13 (10.82%)
- Très content usages: 17 (32.73%)
- Très mécontent usages: 13 (2.584%)
- Très content débit: 15 (19.83%)
- très mécontent débit: 13 (9.475%)
- Très content couverture: 13 (4.927%)
- très mécontent couverture: 14 (24.36%)
- je suis sur Paris: 1 (0.179%)

The main panel displays a text snippet from source 'Corpus4GPourEval' with ID 48831687. The text is: "Depuis que j'ai la 4G je peudaire des recherches sur internet à tout moment et rapidement, consulter mon compte bancaire, de regarder des vidéos en HD sans contrainte de temps de téléchargement, me déplacer avec la navigation itinéraire (n'ayant pas de GPS ça dépanne vraiment), profiter pleinement de la ligue 1, et surtout regarder la télé !!". Below the text is a table of predicted scores for various concepts:

concept	threshold	predicted score	Give your appreciation
Très content	0.5	1.0664	true ✓ not applicable false ✗
très mécontent	0.5	-0.3596	true ✓ not applicable false ✗
Très content usages	0.5	1.3172	true ✓ not applicable false ✗
Très mécontent usages	0.5	-0.3032	true ✓ not applicable false ✗
Très content débit	0.5	0.6843	true ✓ not applicable false ✗
très mécontent débit	0.5	0.0209	true ✓ not applicable false ✗
Très content couverture	0.5	0.6432	true ✓ not applicable false ✗
très mécontent couverture	0.5	-0.0754	true ✓ not applicable false ✗
je suis sur Paris	0.5	0.1967	true ✓ not applicable false ✗

FIG. 1 – L'interface de VIPE pour l'étiquetage des textes

textes en n-grammes) ne contient que des 1 (« positive-only ») et la matrice $X_{m \times n_2}$ (labels) contient des 1 et des 0. Molecule ajoute des 0 virtuels à la matrice $X_{m \times n_1}$ pour équilibrer l'apprentissage : les 0 sont générés par échantillonnage aléatoire à chaque présence de 1 dans $X_{m \times n_1}$. De nombreuses comparaisons expérimentales avec différentes stratégies ont montré l'efficacité de cet échantillonnage. L'algorithme est anytime et est appliqué selon une boucle sans fin sur les données d'apprentissage (une passe dure quelques secondes pour 1 million de cellules traitées dans la matrice).

3 Usages et performances

Actuellement, VIPE est utilisé par une vingtaine d'utilisateurs au sein d'Orange, principalement dans les services marketing pour analyser les opinions sur les différents produits et services de l'entreprise. En moyenne, la phase interactive dure moins de 2 heures ; une fois la confiance établie dans l'outil, l'utilisateur exporte les résultats. VIPE produit une table avec les scores prédits. Ces derniers peuvent être utilisés pour de la classification mais aussi pour du ranking selon les objectifs de l'analyse. En mai 2016, la volumétrie des données stockées pour

VIPE : un outil interactif de classification multilabel de messages courts

l'analyse d'opinions était de : 446 700 textes dans la base (m), 1 380 000 n-grammes (n_1), 156 labels (n_2), 17 900 000 (resp. 4550) cellules non vides dans la matrice $X_{m \times n_1}$ (resp. $X_{m \times n_2}$).

L'algorithme d'apprentissage Molecule a préalablement été testé sur quatre bases de tailles variées - de 6000 à 330000 lignes (m) et de 3600 à 480 000 colonnes ($n = n_1 + n_2$) - : MovieLens IM, Netflix et deux extraits de catalogue VOD (Video On Demand) IMDB (table films/mot-clés extraite en 2012) et Orange. Les données, codées en « positive only » (binarisées selon la moyenne globale des notes dans les cas de Netflix et MovieLens), ont été partitionnées en 10 ensembles avec un processus classique de validation croisée. Les résultats avec le critère BER multi-label qui évalue le ratio des labels mal classés sont les suivants : 11.3% (Movie Lens), 12.9% (Netflix), 7.4% (IMDB), 20.2% (Orange). L'apprentissage résiste bien à la très forte parcimonie des données (au mieux, plus de 96% de valeurs manquantes).

En conclusion, VIPE est un prototype de système d'aide à la classification interactive multi-label de textes courts qui permet de gérer une volumétrie importante, avec des bonnes performances prédictives. Des améliorations de l'outil sont prévues à court terme comme, par exemple, l'ajout de fonctions d'aide au regroupement automatique pour du clustering interactif.

Références

- Gauzente, C., P. Volle, et al. (2012). Développer l'intelligence client. In *Stratégie clients : Points de vue d'experts sur le management de la relation client*, P. Volle (ed). Pearson.
- Keim, D. A., M. Krstajic, C. Rohrdantz, et T. Schreck (2013). Real-time visual analytics for text streams. *Computer* (7), 47–55.
- Nair-Benrekia, N.-Y., P. Kuntz, et F. Meyer (2015). Learning from multi-label data with interactivity constraints : an extensive experimental study. *Expert Systems with Applications* 42(13), 5723–5736.
- Takács, G., I. Pilászy, B. Németh, et D. Tikk (2007). Major components of the gravity recommendation system. *ACM SIGKDD Explorations Newsletter* 9(2), 80–83.
- Zhang, M.-L. et Z.-H. Zhou (2014). A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering* 26(8), 1819–1837.

Summary

In this article, we present VIPE, an interactive tool for visualization and multi-label classification developed at Orange Labs and used for opinion analysis. Based on a fast matrix factorization algorithm, it allows a user to import texts (tweets, emails, surveys, etc...), to define his labels of interest ("satisfied customer", "likes the 4g", etc...) and then it recommends labels for any text and texts for any label.