Tagging automatique de données dans le cadre d'objets connectés, Fiabilité et impact sur la confiance

Antonin PENON, Laurence DEVILLERS
IRT-SystemX / LIMSI-CNRS, LIMSI-CNRS/Paris-Sorbonne 4
antonin.penon@ens-cachan.fr, devil@limsi.fr

ABSTRACT

Ce papier décrit un projet de thèse commencé en octobre 2016 avec la collaboration du LIMSI et de l'IRT SystemX. Il s'inscrit dans le projet TNC (Territoire Numérique de Confiance). L'apparition d'un grand nombre d'objets connectés a motivé la création de la plateforme TNC, centralisant ces derniers, tout en assurant un contrôle plus éthique des données utilisateur. Face à la quantité croissante de données extraites, une problématique importante est celle d'aider l'utilisateur à choisir lesquelles partager ou non. Nous proposons ici une solution basée sur l'étiquetage automatique des données (i.e. les tagger automatiquement): l'enrichissement des données est obtenu avec des outils de fouille de contenu: détection de topics, de sentiments, etc. Un rapide état de l'art sur les méthodes de classification pour divers types de données est effectué, et les premiers objectifs de la thèse sont posés. L'objectif des tags peut être double, à la fois de vérifier les données produites par les algorithmes des IoTs et d'enrichir les données de l'utilisateur pour qu'il décide de les diffuser ou non. Il s'agit de travailler sur l'explication et la transparence des données et des algorithmes, ainsi que sur la confiance des utilisateurs.

Keywords: self data, user confidence, document tagging.

1 INTRODUCTION

L'émergence actuelle d'une multitude d'objets connectés au réseau Internet (Internet des Objets: IDO) pose de nombreux enjeux, tant techniques qu'éthiques. Un grand nombre de services existent afin de traiter et d'ajouter de la valeur aux données récoltées par ces objets, en vue d'améliorer l'expérience utilisateur. Cependant, la collecte massive de telles informations pose plusieurs problèmes, notamment celui de la vie privée. En effet, la plupart des applications actuelles ne donnent que peu (voire pas du tout) de contrôle à l'utilisateur sur ses propres données.

Cette thèse s'inscrit dans le cadre du projet TNC (Territoire Numérique de Confiance), développé à l'IRT SystemX en collaboration avec plusieurs industriels. Un des objectifs de ce projet est de remettre l'utilisateur au centre du processus de collecte des données. La plateforme développée au sein de TNC rend le contrôle à l'utilisateur, en lui permettant de sélectionner les données qu'il souhaite partager ou non, et avec quel service.

Une des problématique liée à cet objectif est le développement d'outil aidant l'utilisateur à sélectionner éfficacement les données à partager. La possibilité de tagger systématiquement les données générées permettra à l'utilisateur d'avoir une idée du contenu de ses données sans les traiter une à une. Il lui serait ensuite offert la possibilité d'établir des règles de partages (eg: pas de partage de photos prises à l'intérieur de sa maison, pas de partage de documents écrits lié à son activité professionelle, etc…).

Plusieurs cas d'études seront menés, sur les types de données les plus courants (image, vidéo, texte, …). Outre le développement d'algorithmes adaptés à ce problème, une attention particulière sera portée sur la précision des tags générés et sur l'impact qu'il peuvent avoir sur la relation de confiance entre l'utilisateur et les services. La confiance est un sujet complexe, nous étudierons les différents niveaux mis en jeu dans la relation personne/données, personne/IOT.

2 TAGGING DE DOCUMENTS : ETAT DE L'ART

Il existe une grande variété de méthodes de tagging automatique, différentes suivant le type de donnée considéré. La première tâche consistera à établir un état de l'art de ces méthodes dans les types de document les plus fréquents. La présente section n'a pas pour vocation d'être exhaustive, tant le nombre d'algorithmes permettant la classification est élevé. On se concentrera sur des méthodes d'apprentissage supervisé, l'idée étant de permettre à l'utilisateur de définir des tags par l'exemple.

2.1 Dans le texte

Il existe plusieurs méthodes pour la classification de texte supervisée, dont les plus fréquentes sont résumées et comparées dans [1].

Parmis elles ont peut citer la représentation en modèle vectoriel du texte (représentation d'un document par vecteur de fréquences, et mesures de similarités dans cet espace) qui fut une des premières techniques automatiques dans le traitement du texte.

L'analyse sémantique latente [3] est une méthode utilisant une autre représentation, permettant de réduire la dimensionnalité du problème (en tronquant les valeurs singulières faibles dans la décomposition SVD de la matrice de fréquence des termes).

Enfin l'utilisation de machines à supports de vecteurs [4], dans lesquelles les données sont mappées vers un espace où elles pourront être séparées linéairement, sera également considérée.

2.2 Dans l'audio

Les techniques de tagging audio reposent dans un premier temps sur une phase d'extraction de caractéristiques audio (features) dans le signal. Ces caractéristiques peuvent être des données spectrales, temporelles, ou toute autre transformation effectuée sur le signal, et leur choix est très important. La deuxième phase concerne l'apprentissage d'un modèle basé sur ces caractéristiques. Plusieurs méthodes sont possibles, suivant le type de classification souhaité.

Dans le cas de l'identification d'interlocuteur, ou de la reconnaissance d'émotions, on peut utiliser des mélanges de gaussiennes [5]. La classification content-based peut être effectuée grâce à des arbres de décisions [6], ou bien encore des machines à supports de vecteurs.

2.3 Dans l'image

Parmis la multitude d'algorithmes de classification d'images, nous nous pencherons sur des algorithmes reposant sur l'extraction de features et l'apprentissage d'un modèle en résultant (eg: eigen

faces, regression logistique), ainsi que sur des approches de type réseaux de neurones et réseaux de neurones à convolution.

3 Objectifs

Un système de tagging automatique sera mis en place, gérant plusieurs types de données, et intégré à la plateforme TNC. Des tags généraux seront pré-entrainés, et une solution sera apportée pour la définition de tags personnalisés par l'utilisateur. Une étude sera menée sur la précision et la fiabilité des tags obtenus (notamment sur la différence entre ceux pré-entrainés et ceux définis par l'utilisateur), ainsi que sur le choix des algorithmes à utiliser. L'outil développé sera testé sur des non-experts, afin de s'assurer de sa simplicité d'utilisation. On en profitera pour évaluer l'apport de ce système de tags automatique par à rapport à un système sans, notamment au niveau de la confiance ressentie. Pour ce faire, l'outil sera testé sur une multitude de corpus (avec ou sans système de tags, variations d'autres paramètres à préciser), et la détermination des tags influençant le plus la confiance pourrait être intéressante. Une telle étude nécessitera également d'élaborer un peu plus ce que peut être une mesure de la confiance.

4 Conclusion

Le contrôle des données est une problématique clé dans le monde numérique d'aujourd'hui. Le projet TNC a pour objectif une centralisation des IoTs, tout en rendant la main à l'utilisateur sur sa vie privée. Cependant, la multitude d'objets connectés et la quantité importante de données en résultant peut rendre sa gestion difficile. L'efficacité de la protection des données dépend aussi de la simplicté d'utilisation de notre solution par un non-expert. Un système de tag automatique peut être un bon moyen de faciliter les choix de l'utilisateur, tout en lui laissant le contrôle sur ses données, et sans compromettre sa vie privée.

References

[1] Cardoso-Cachopo, A., & Oliveira, A. L. (2003, October). An empirical comparison of text categorization methods. In *International Symposium on String Processing and Information Retrieval* (pp. 183-196). Springer Berlin Heidelberg.

[2] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*(1), 1-47.

[3] Dennis, S., Landauer, T., Kintsch, W., & Quesada, J. (2003). Introduction to latent semantic analysis. In *Slides from the tutorial given at the 25th Annual Meeting of the Cognitive Science Society, Boston*.

[4] Christiannini, N., & Shawe-Taylor, J. (2000). Support vector machines and other kernel-based learning methods.

[5] Devillers, L., Tahon, M., Sehili, M. A., & Delaborde, A. (2015). Inference of human beings' emotional states from speech in human–robot interactions. *International Journal of Social Robotics*, *7*(4), 451-463.

[6] Homburg, H., Mierswa, I., Möller, B., Morik, K., & Wurst, M. (2005, September). A Benchmark Dataset for Audio Classification and Clustering. In *ISMIR* (Vol. 2005, pp. 528-31).