Analyse des images qui circulent sur Internet : un aperçu du projet ImagiWeb

Julien Velcin*

*Université de Lyon (ERIC, Lyon 2)
julien.velcin@univ-lyon2.fr,
http://mediamining.univ-lyon2.fr/velcin/

Résumé. Dans cette présentation, je donne un aperçu du projet ImagiWeb. L'objectif du projet consiste à capturer l'image d'une entité, au sens de sa représentation, qui circule sur Internet et dans les médias sociaux. Le projet a mobilisé six équipes de chercheurs entre informatique et sciences sociales (ERIC à Lyon, LIA à Avignon, CEPEL à Montpellier, Xerox à Grenoble, EDF à Paris et AMI Software à Montpellier) pendant trois ans et demi.

1 Introduction

L'image de nombreuses entités (par ex. célébrités, entreprises, marques) nous parvient principalement par l'intermédiaire de l'existence virtuelle qu'elles mènent sur le Web et dans les nouveaux médias. L'objectif du projet ImagiWeb est d'analyser l'opinion exprimée dans les messages postés sur Internet au sujet de ces entités, à l'aide de techniques informatiques et statistiques, et de la relier aux caractéristiques sociales des individus qui les ont produits en suivant une logique de panélisation.

A cette approche résolument pluridisciplinaire s'ajoute la volonté d'apprécier l'opinion en regard de cibles qui décrivent l'entité (par ex. : les soutiens de l'homme politique ou la politique tarifaire de l'entreprise) et de la suivre de manière dynamique. Pour cela, il est nécessaire d'aller au-delà des méthodes habituelles de classification d'opinion (Liu, 2015). Ce type de travaux peut avoir un impact aussi bien technique (développer des nouveaux algorithmes et logiciels), stratégique (apporter une nouvelle solution pour la gestion de la réputation en ligne) que sociétal (mieux comprendre la naissance et la diffusion des représentations).

2 Méthodes et technologies utilisées

Pour résoudre cette problématique, après avoir mis en place un cadre complet d'acquisition et d'annotation des données (Velcin et al., 2014), nous avons choisi de combiner des outils avancés de traitement automatique de la langue (approche linguistique), de fouille de textes (approche statistique) et d'apprentissage automatique (supervisé et non supervisé).

La prédiction des cibles et des polarités d'opinion est obtenue à l'aide d'une méthode hybride et active de classification supervisée (Stavrianou et al., 2014; Cossu et al., 2015) tandis

que le regroupement des groupes d'opinion homogène est construit à l'aide de clustering probabiliste évolutionnaire (Kim et al., 2015).

Les producteurs des messages analysés sont identifiés à l'aide d'une stratégie originale de panélisation des internautes mise en place en modernisant l'approche traditionnellement employée en sociologie (Dormagen et al., 2014).

Enfin, ces outils sont intégrés dans un prototype permettant de démontrer l'intérêt de l'approche sur deux cas d'étude et selon plusieurs scénarios d'usage envisagés, tels que la navigation dans les données et les annotations ou la visualisation temporelle des groupes d'opinion (Khouas et al., 2015).

3 Résultats obtenus dans le projet

Le projet a principalement permis de montrer qu'il était possible de capturer les opinions fines au sujet d'une entité en utilisant des outils d'analyse automatique des messages d'expression sur Internet.

Sur le cas des hommes politiques, il a été possible de comparer cette opinion en ligne aux baromètres habituels de l'opinion et d'en tirer des conclusions sur certaines convergences observées mais surtout sur des différences très marquées (Velcin et Boyadjian, 2016).

Sur le cas de l'entreprise EDF, le prototype mis en place a permis de confirmer des conclusions tirées par les sémiologues, et ce de manière exhaustive, mais également de proposer des informations nouvelles.

D'un point de vue industriel, ce projet a permis l'élaboration d'une méthode générale pour étudier l'image de marque (réputation) sur Internet. Celle-ci a été d'ores et déjà intégrée à la plateforme de veille éditée par l'entreprise partenaire.

4 Conclusion

Après trois ans et demi, le projet ImagiWeb a permis de montrer qu'il était possible de capturer l'opinion des groupes d'individus et, ce, à un degré fin d'analyse. Contrairement à l'idée reçue sur l'anonymat des internautes, une logique de panélisation est possible afin d'identifier les producteurs d'opinion mais il ne faut pas se tromper sur la valeur de représentativité des sources d'information, telle que Twitter. Bien sûr, il s'agit d'être vigilant sur les aspects liés à la vie privée des internautes en intégrant des mécanismes d'anonymisation.

L'image des entités peut être ainsi étudiée via l'utilisation de logiciels, comme nous l'avons montré avec le prototype de démonstration sur plusieurs scénarios d'analyse. L'évaluation par un sémiologue de l'apport de ce type d'outil affiche clairement ses avantages (navigation facilitée dans les données, résumé d'un vaste corpus de documents, capacité d'innovation) ainsi que des pistes d'amélioration (par exemple la gestion dynamique des cibles).

Références

Cossu, J.-V., E. SanJuan, J.-M. Torres-Moreno, et M. El-Bèze (2015). Multi-dimensional reputation modeling using micro-blog contents. In *Proceedings of the 22nd International Symposium on Methodologies for Intelligent Systems (ISMIS)*, pp. 452–457. Springer.

Dormagen, J.-Y., J. Boyadjian, et M. Neihouser (2014). Hybrid method for measuring opinion. In *Proceedings of Asia Conference for e-Democracy and Open Government (CeDEM)*, Hong-Kong.

Khouas, L., C. Brun, A. Peradotto, J.-V. Cossu, J. Boyadjian, et J. Velcin (2015). Étude de l'image de marque d'entités dans le cadre d'une plateforme de veille sur le web social. In *Actes de la 22ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Caen.

Kim, Y.-M., J. Velcin, S. Bonnevay, et M.-A. Rizoiu (2015). Temporal multinomial mixture for instance-oriented evolutionary clustering. In *Proceedings of the European Conference on Information Retrieval (ECIR)*, pp. 593–604. Springer.

Liu, B. (2015). *Sentiment Analysis : Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

Stavrianou, A., C. Brun, T. Silander, et C. Roux (2014). NLP-based feature extraction for automated tweet classification. In *Workshop on Interactions between Data Mining and Natural Language (DMNLP), in collocation with ECML-PKDD*, Nancy, France, pp. 15–19.

Velcin, J. et J. Boyadjian (2016). De l' "opinion mining" à la sociologie des opinions en ligne. pour une approche interdisciplinaire de l'étude du web politique. *Question de communication*. Article en cours d'évaluation.

Velcin, J., Y. Kim, C. Brun, J. Dormagen, E. SanJuan, L. Khouas, A. Peradotto, S. Bonnevay, C. Roux, J. Boyadjian, A. Molina, et M. Neihouser (2014). Investigating the image of entities in social media : Dataset design and first results. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, pp. 818–822.

Summary

In this talk, I will give an overview of the ImagiWeb project. The objectif of this project consists in capturing an entity's image (that is, its representation) that circulate through Internet and the social media. The project involved six research teams inbetween computer science and social sciences and humanities (ERIC at Lyon, LIA at Avignon, CEPEL at Montpellier, Xerox at Grenoble, EDF at Paris et AMI Software at Montpellier) for three years and a half.