Projet ANR-2012-CORD-002-01

ImagiWeb

Programme CONTINT 2011

Mémoire scientifique

Acronyme du projet	ImagiWeb
Titre du projet	Images sur le Web : Analyse de la dynamique des images sur le Web 2.0
Coordinateur du projet (société/organisme)	Laboratoire ERIC, Université de Lyon
Période du projet (date de début – date de fin)	1er Avril 2012 – 30 Septembre 2015
Site web du projet, le cas échéant	http://mediamining.univ-lyon2.fr/velcin/imagiweb

Rédacteur de ce rapport	
Civilité, prénom, nom	M. Julien VELCIN
Téléphone	04 78 77 24 14
Adresse électronique	Julien.Velcin@univ-lyon2.fr
Date de rédaction	27 Novembre 2015

Liste des partenaires présents à la fin du projet (société/organisme et responsable scientifique)	- AMI Software (E. Fourboul) - CEPEL, Université de Montpellier (J.Y. Dormagen) - EDF France (A. Peradotto) - ERIC, Université de Lyon (J. Velcin) - LIA, SFR Agorantic, Université d'Avignon et des Pays du Vaucluse (M. El-bèze) - Xerox Research Center Europe (C. Brun)

MEMOIRE SCIENTIFIQUE

A.1 RESUME DU MEMOIRE

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 novatrice. A cette approche résolument pluridisciplinaire s'ajoute la volonté d'apprécier l'opinion de manière fine au sujet 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.

A.2 ENJEUX ET PROBLEMATIQUE, ETAT DE L'ART

Internet joue un rôle très important sur la manière dont nous percevons le monde qui nous entoure. Ainsi, de nombreuses entités nous parviennent principalement par l'intermédiaire de l'existence virtuelle qu'elles mènent sur la toile et dans les médias, qu'il s'agisse d'un film, d'une personnalité, d'une entreprise, d'une marque. L'enjeu du projet ImagiWeb est de dévoiler les mécanismes qui procèdent à la production, la diffusion, l'évolution des opinions des internautes relativement à ces entités, ce que nous appelons leur *image*. Nous avons choisi en particulier de travailler sur deux cas d'étude : l'image de deux hommes politiques français (en l'occurrence N. Sarkozy et F. Hollande entre 2012 et 2014) et l'image de l'entreprise EDF sur les thématiques liées au nucléaire. Pour étudier ces images, l'idée est de mêler une approche informatique, mettant notamment en œuvre des techniques d'analyse automatique des textes d'expressions postés sur Internet (blogs, tweets), et une approche sociologique afin de déterminer l'identité des producteurs d'opinion. Ce choix différencie nettement ImagiWeb de projets précédents tels que Blogoscopie et Doxa au niveau national, ou ARCOMEM, TrendMiner et Limosine au niveau européen.

D'un **point de vue informatique**, la problématique consiste à récupérer les données textuelles issues du Web et à les analyser de manière automatique afin d'en extraire l'opinion, d'abord au niveau individuel des messages puis des groupes d'internautes. Cela nécessite le recours à deux approches différentes d'apprentissage automatique.

Premièrement, nous développons des techniques d'apprentissage supervisé afin de distinguer la polarité de l'opinion exprimée dans un texte (est-elle positive, négative, neutre ?) ainsi que l'aspect ciblé par l'opinion (est-ce le *bilan* d'un homme politique ou la *politique tarifaire* d'une entreprise ?). Cette analyse conjointe ajoutée à la nature des messages traités (courts, politiques, en français) pose des problèmes sérieux aux algorithmes habituels de classification d'opinion [Liu B., 2015] qui demandent à être adaptés. Deuxièmement, nous développons des techniques d'apprentissage non supervisé afin de regrouper les messages et *a fortiori* les internautes qui affichent des opinions similaires vis-à-vis de l'entité étudiée. Les algorithmes doivent être en mesure de prendre en compte le caractère dynamique des opinions. Pour cela, les algorithmes habituels de clustering [Aggarwal and Reddy, 2014] doivent être revisités.

D'un **point de vue sociologique**, l'objectif est de mieux caractériser les auteurs de messages sur le réseau social Twitter et dans la blogosphère. Il s'agit d'établir, par exemple, leur niveau d'études, leurs profils professionnels, leurs positions dans l'espace social mais aussi leurs pratiques culturelles et leurs orientations politiques. Cette approche sociologique de la Twittosphère et de la blogosphère est très rarement mise en œuvre, entre autres parce qu'elle

se heurte à des verrous majeurs résultant en particulier de la quasi-absence d'informations directement disponibles *online* [Boullier and Lohard, 2013]. Elle est, pourtant, essentielle pour identifier les logiques sociales qui régissent la circulation des messages.

A.3 APPROCHE SCIENTIFIQUE ET TECHNIQUE

L'un des principaux atouts du projet est de combiner une **analyse automatique des messages** textuels produits sur le Web pour en extraire les images avec une **étude sociologique** pour caractériser au plus près les émetteurs des opinions qui constituent ces images.

La représentativité de ces images et de leurs émetteurs est un verrou important. Pour aborder ce problème, dans le **premier cas d'étude** sur l'image des hommes politiques véhiculée par Twitter, nous avons choisi de découper notre ensemble de messages selon plusieurs échantillons grâce à une logique de **panélisation**. Ces panels (internautes « répondant », « non répondant » et « de forte audience ») résultent de la mise en œuvre d'une approche hybride entre les méthodes d'enquêtes traditionnelles par questionnaires et les méthodes d'observation directe de l'activité numérique via le logiciel AMI Opinion Tracker, réalisées par les experts en sciences politiques. Ils permettent de collecter les publications émises par les individus et de les analyser au fil du temps, tout en connaissant les caractéristiques sociales de leurs auteurs. Au final, ont été collectés et analysés, sur une période de 24 mois (soit du 1er janvier 2012 au 31 décembre 2013) 27 975 tweets provenant du panel représentatif (individus répondants et non répondants, n= 1228) et 38 574 tweets provenant du panel Twitter de forte audience (n=1116). A ces tweets publiés par les individus composant les panels s'ajoutent l'analyse de tweets tirés au sort sur l'ensemble du réseau social, soit, sur la même période d'étude, un échantillon de 128 932 tweets.

Le **deuxième cas d'étude** porte sur l'image d'EDF sur les thématiques liées au nucléaire, telle que véhiculée dans la blogosphère. Le jeu de données étudié comporte 3300 messages de blogs, dont 600 ont été annotés manuellement par des experts dans le cadre du projet afin d'entraîner les algorithmes d'apprentissage automatique.

L'analyse automatique des textes contenant des opinions est un verrou qui peut tirer parti d'annotations manuelles reflétant la connaissance des experts. Un sous-ensemble des messages récoltés a été l'objet d'une première campagne d'annotation manuelle. Pour les messages courts comme les tweets, on précise une polarité d'opinion et une cible extraite à partir d'une grille d'annotation mise en place par le consortium. Ces annotations servent de base aux algorithmes d'apprentissage, puis de nouvelles étapes d'annotation peuvent être ajoutées par la suite afin de corriger le paramétrage des algorithmes et leurs sorties (**approche active**). Pour les messages plus longs, tels que ceux extraits des blogs, plusieurs annotations (cible, polarité) peuvent être indiquées par les humains et une difficulté supplémentaire consiste pour les algorithmes à prédire non pas une seule annotation mais plusieurs. Pour cela, les algorithmes s'appuient sur un découpage en phrase des extraits de blogs auxquelles les annotations sont associées puis une fusion de ces annotations est réalisée *a posteriori* en fonction de leur poids statistique. Les techniques employées sont issues du traitement automatique des langues, autorisant une analyse fine des textes (syntaxe, négations, coréférences…), mais également des techniques basées sur la cooccurrence statistique, afin de proposer des nouvelles **méthodes hybrides** pour l'identification des opinions. Ces techniques explorent notamment l'approche dite active qui permet de déterminer les messages les plus pertinents pour l'apprentissage et de demander à l'expert de nouvelles annotations afin d'améliorer les performances finales.

A partir de l'ensemble des messages annotés automatiquement, nous les agrégeons pour **former les images**. L'agrégation se base sur des algorithmes originaux d'apprentissage

automatique non supervisé s'inspirant du clustering évolutionnaire afin de prendre en compte leur dimension temporelle. Contrairement à l'existant, les modèles proposés dans le projet doivent être capables de prendre en compte des données lacunaires (beaucoup d'opinions restent inconnues) correspondant à des objets différents au cours du temps, tout en essayant d'aider les utilisateurs à **comprendre** l'évolution des catégories (clusters) détectées.

Une fois ces images (re)constituées, différentes techniques sont mises en place pour les **interroger** de manière conviviale *via* une interface. Dans le projet, nous développons trois entrées possibles pour l'utilisateur : un système de type OLAP afin d'aider l'utilisateur à naviguer dans les données et les annotations, un système automatique pour résumer le contenu d'un ensemble de messages et une visualisation interactive de l'évolution des catégories d'opinion dans le temps.

A.4 RESULTATS OBTENUS

Deux **bases de données** ont été constituées : une première issue de Twitter pour le cas d'étude sur les hommes politiques et une seconde de billets de blog pour le cas d'étude EDF. L'originalité principale de l'approche est que les tweets ont été sélectionnés sur la base de panels d'internautes sociologiquement identifiés afin de pouvoir répondre à des questions jusqu'alors impossible à aborder à cause de l'anonymat des pseudonymes (livrable L2.1).
Une **procédure complète d'annotation** des données d'opinion a été mise en place afin d'acquérir une base suffisante pour entraîner les algorithmes d'analyse. Cette procédure se manifeste par plusieurs livrables (L2.3). Tout d'abord, une plateforme permet de réaliser des annotations fines en associant des polarités et des cibles d'opinion à des passages d'un texte. Cette plateforme, aisément extensible et *open-source*, est disponible pour la communauté scientifique[1]. Il s'agit également d'un jeu de 7283 tweets anonymisés annotés (sous-ensemble de tous les tweets réellement annotés) mis à disposition de la communauté[2]. Il s'agit enfin d'un échantillon de 600 messages de blogs annotés pour le deuxième cas d'étude. La procédure d'acquisition et d'annotation a donné lieu à une publication commune à la conférence LREC [Velcin et al., 2014].
Les recherches traitant de l'**annotation automatique** des messages d'opinion ont permis de mettre au point des algorithmes hybrides capables de prédire la polarité et la cible des opinions exprimées dans les textes, qu'il s'agisse de tweets ou de messages de blog (livrables L3.2). En moyenne, les taux de succès obtenus sur ces tâches, soit environ 80% d'exactitude pour la polarité et 70% pour les cibles (cf. livrable L6.1), atteignent les résultats consignés dans la littérature récente voire les dépassent dans certains cas : pour la détection de la polarité sur des tweets français, les meilleurs systèmes de l'édition 2015 du défi fouille de texte (DEFT2015[3]) présentent des résultats de l'ordre de 73% d'exactitude. Ces algorithmes ont pu être utilisés pour annoter l'ensemble des données à notre disposition, étape indispensable à la suite du processus. Des approches hybrides combinant information linguistique riche à des modèles statistiques de classification ont été entre autres expérimentées [Stavrianou et al, 2014]. Notons que la fusion des approches a fait l'objet d'une attention toute particulière, tout comme le choix d'une méthode appropriée pour obtenir les meilleurs résultats [Cossu et al., 2013a].
Les travaux sur l'**extraction et le suivi des images** proprement dites ont conduit aux livrables L3.4. D'une part, nous avons commencé par définir l'image en nous inspirant d'éléments issus des technologies du Web sémantique car elle est employée ensuite dans la mise en place des scénarios d'utilisation (cf. tâche T4). Ensuite, nous avons utilisé une

[1] Voir : http://dev.termwatch.es/~molina/imagiweb2/static/systeme_description.html
[2] Voir : http://mediamining.univ-lyon2.fr/velcin/imagiweb/dataset.html
[3] Voir : http://www.atala.org/taln_archives/ateliers/2015/DEFT/deft-2015-long-001.pdf

formalisation plus mathématique dans laquelle l'image exprimée par un groupe de personnes est une distribution sur des couples (cible,polarité). Cette définition sert de base à deux modèles originaux de clustering évolutionnaire : *Temporal Mixture Model* (TMM) et *Parametric-Link Multinomial Mixture* (PLMM). Pour notre tâche, ces deux modèles se sont avérés plus performants que d'autres algorithmes de la littérature, tel que le modèle *Dynamic Topic Model* [Blei and Lafferty, 2006]. Ces travaux ont conduit à des publications dans les conférences ECIR [Kim et al., 2015] et IDA [Hasnat et al., 2015]. Un article de revue internationale est en cours d'évaluation [Hasnat et al., submitted].

Le **prototype de démonstration** correspond aux livrables de la tâche T5. La version finale (livrable L5.3) permet de réaliser et tester plusieurs des scénarios développés dans les livrables de la tâche T4. Le prototype est une application Web intégrée à la plateforme de veille éditée par AMI Software. Il permet une synchronisation avec la plateforme centrale d'annotation ainsi que la mise en œuvre d'une anonymisation des données assurant une conformité du prototype vis-à-vis de règles de sécurité des données personnelles. Le prototype de démonstration propose un ensemble d'outils d'affichage et de visualisations adéquates, pour une exploration aisée des annotations, une synthèse de la dynamique temporelle de l'image selon les thématiques abordées et la comparaison de plusieurs images. Le prototype a été présenté dans le cadre de la conférence TALN [Khouas et al., 2015].

La **comparaison** entre les sondages d'opinion traditionnels par questionnaires et les mesures d'opinion produites sur Twitter, dans le cas d'étude des hommes politiques, est présentée dans le livrable L6.3. Elle permet d'établir deux résultats principaux : 1) Le réseau social Twitter surreprésente massivement les opinions négatives et, ce, quels que soient les contextes, avec pour conséquence de rendre plus difficile le repérage des évolutions en matière de jugements portés sur les hommes politiques ; 2) On note cependant une chute plus rapide de la popularité de F. Hollande sur le réseau Twitter que dans les enquêtes d'opinion traditionnelles. Le réseau social semble ainsi livrer certains indices permettant d'anticiper des dynamiques de mobilisation et démobilisation des électorats et des publics militants (dont nos résultats permettent d'affirmer qu'ils sont surreprésentés sur le réseau social). Tout semble indiquer que ces dynamiques de (dé)mobilisation en ligne constituent un facteur à l'origine des dynamiques d'opinions enregistrés par les sondages. Les premiers résultats font l'objet d'une publication en cours [Velcin and Boyadjian, 2016].

Les résultats de l'**étude sémiologique** qui concerne l'image (émise et perçue) de l'entreprise EDF constituent le livrable L6.2. Comparés à l'évaluation du prototype fournie dans le livrable L6.4, ils permettent de mettre en lumière l'apport de la quantification apportée par les outils d'analyse automatique et surtout l'intérêt d'une détection de la polarité.

A.5 EXPLOITATION DES RESULTATS

AMI Software

L'ensemble des outils développés dans le cadre du projet ImagiWeb ont été intégrés avec succès au sein de la plateforme de veille éditée par AMI Software autour de deux cas d'études, à savoir l'image des hommes politiques et l'image de marque d'EDF concernant le nucléaire. Les évaluations réalisées, au travers notamment du prototype de démonstration, permettent de valider la pertinence de l'approche pour les cas d'études considérés et sa généralisation à d'autres types d'entités. Ces outils constituent une avancée notable en termes d'analyse d'opinion permettant à l'entreprise d'améliorer son offre, notamment à destination des études marketing et d'analyse d'image de marque. Ceci ouvre de nouvelles perspectives commerciales pour l'entreprise et de nouveaux marchés.

EDF

Pour EDF, comme pour d'autres grandes entreprises, Internet représente une source de données importante pour évaluer son image. Dans le cadre de ce projet, c'est l'expression des internautes de la blogosphère autour du thème du nucléaire qui a été exploitée afin de valider l'intérêt d'une approche d'analyse automatique. Le prototype issu d'ImagiWeb apporte une vue globale quantitative des prises de position des internautes autour de ce thème en couplant volumétrie, thématique, polarité et également chronologie, permettant ainsi de hiérarchiser le poids des différents sujets du corpus tels que prédéfinis ainsi que leur polarité. De plus, l'aspect chronologique permet de croiser les thématiques abordées avec l'actualité pour une interprétation des résultats plus pertinente. Afin d'aider l'utilisateur final, une voie d'amélioration serait maintenant d'ajouter un modèle de résumé automatique et/ou une visualisation du contenu des ciblages, car le nombre de documents relatifs à un thème et une polarité spécifiques peut être très important.

XRCE

Les algorithmes d'annotation développés dans le cadre du projet ImagiWeb ont permis à la fois d'améliorer la couverture des outils linguistiques de XRCE mais également de construire des outils statistiques hybrides, c'est-à-dire utilisant de l'information linguistique riche combinée à des modèles statistiques. Ces outils sont d'un intérêt tout particulier dans un ensemble de projets conduits par Xerox. En effet, ces outils hybrides peuvent être adaptés à d'autres langues et à des problématiques telles que l'analyse d'émotion, ou la détection des traits de personnalité, qui sont essentielles dans le contexte de projets de recherche et de recherche appliquée menés par Xerox autour du « Customer Care » (« satisfaction du client), et ce à des fins de transfert vers les unités commerciales.

A.6 Discussion

La plupart des objectifs que nous nous étions fixés ont été atteints puisque nous avons été en mesure de capter l'opinion fine des individus, et ce à des degrés d'exactitude importants vis-à-vis des travaux actuels du domaine, dans les deux cas d'étude considérés. Nous avons pu la confronter, soit à la « vérité terrain » des sondages dans le premier cas, soit à une analyse sémiologique traditionnelle dans le second cas. Certains accomplissements mineurs n'ont pu être atteints, tels que la prise en compte de cibles dynamiques ou l'évaluation qualitative des images issues du *clustering*. Ils constituent des perspectives à court terme de ce projet. Nous n'avons pas non plus été en mesure de combiner l'analyse de différentes sources pour traiter un même cas d'étude, comme par exemple Twitter et les blogs qui ont été utilisés séparément, et nous avons été forcés d'abandonner Facebook pour des raisons de changement de sa politique d'accès aux données. Cela ouvre la perspective de traiter plusieurs sources hétérogènes à la fois pour analyser la même entité, voire analyser plusieurs entités de manière réellement concurrente (deux hommes politiques, voire toute une classe politique). Les travaux réalisés dans le cadre du projet ont montré une résonance importante avec de nombreux problèmes concrets, tels que l'analyse des controverses ou la gestion de la réputation, dont la résolution possède un impact direct sur notre société.

A.7 Conclusions

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. 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).

A.8 References

[Aggarwal and Reddy, 2014] Aggarwal, C. C., & Reddy, C. K. : *Data clustering: algorithms and applications*. CRC Press, 2014.

[Blei and Lafferty, 2006] Blei, D. M., & Lafferty, J. D. : Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning* (ICML), pp. 113-120, ACM.

[Boullier and Lohard, 2012] Boullier, D., and Lohard, A. : *Opinion mining et Sentiment analysis*. OpenEdition Press, 2012.

[Cossu et al., 2013a] Cossu J-V., Torres-Moreno J-M. et El-Bèze M. : Recherche et utilisation d'entités nommées conceptuelles dans une tâche de catégorisation. Actes de la *20ème Conférence sur le Traitement Automatique des Langues Naturelles* (DEFT/TALN). Sables d'Olonne, Juin 2013.

[Hasnat et al., 2015] M.A. Hasnat, J. Velcin, S. Bonnevay and J. Jacques : A Comparative Study of Clustering Methods with Multinomial Distribution. *Proceedings of the 14th International Symposium on Intelligent Data Analysis* (IDA), regular poster, 2015.

[Hasnat et al., submitted] M.A. Hasnat, J. Velcin, J. Jacques and S. Bonnevay : Parametric link among multinomial mixture models: Application to temporal/evolutionary data analysis. Submitted to *Annals of Applied Statistics* en Septembre 2015.

[Khouas et al., 2015] Leila Khouas, Caroline Brun and Anne Peradotto, Jean-Valère Cossu, Julien Boyadjian, Julien Velcin : Étude de l'image de marque d'entités dans le cadre d'une plateforme de veille sur le Web social, *22ème Conférence sur le Traitement Automatique des Langues Naturelles* (TALN), Caen, 22-25 Juin 2015.

[Kim et al., 2015] Young-Min Kim, Julien Velcin, Stéphane Bonnevay and Marian-Andréi Rizoiu : Temporal Multinomial Mixture for Instance-oriented Evolutionary Clustering. *Proceedings of the European Conference on Information Retrieval* (ECIR). Vienna, Austria, 2015.

[Liu B., 2015] B. Liu : *Sentiment Analysis: mining opinions, sentiments, and emotions*. Cambridge University Press, March 2015.

[Stavrianou et al., 2014] Anna Stavrianou, Caroline Brun, Tomi Silander, Claude Roux : NLP-based Feature Extraction for Automated Tweet Classification. *Workshop on Interactions between Data Mining and Natural Language* (DMNLP), in conjunction with ECML/PKDD, Nancy, France, 15-19 September 2014.

[Velcin et al., 2014] Julien Velcin, Young-Min Kim, Caroline Brun, Jean-Yves Dormagen, Eric SanJuan, Leila Khouas, Anne Peradotto, Stéphane Bonnevay, Claude Roux, Julien Boyadjian, Alejandro Molina and Marie Neihouser : Investigating the Image of Entities in Social Media: Dataset Design and First Results. *Proceedings of the 9th International Conference on Language Resources and Evaluation* (LREC), pp.818-822. Reykjavik, Iceland, 2014.

[Velcin and Boyadjian, 2016] J. Velcin et J. Boyadjian : De l'« opinion mining » à la sociologie des opinions en ligne. Pour une approche interdisciplinaire de l'étude du Web politique. Dans: Questionner les humanités numériques, perspectives croisées, numéro spécial de la revue *Question de communication* dirigé par J. Longhi (prévu en 2016).