My web intelligence : un outil pour l'analyse du web et des réseaux

Amar Lakel

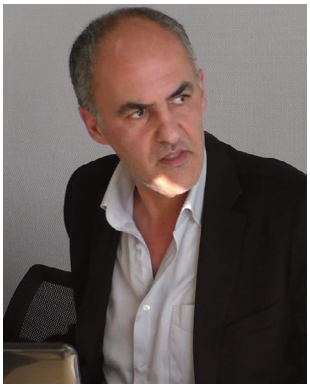
Dans I2D - Information, données & documents **2021/1 n° 1** , pages 96 à 103
Éditions A.D.B.S.

ISSN 2428-2111
ISBN 9782242821111
DOI 10.3917/i2d.211.0096
Date de mise en ligne : 24/05/2021

Mots-clés : Analyse réseaux, Cartographie web, Corpus web Crawler, Digital Studies, *Viralité informationnelle*

MY WEB INTELLIGENCE : UN OUTIL POUR L'ANALYSE DU WEB ET DES RÉSEAUX

L'analyse des sources ouvertes nécessite des outils qui soient capables d'effectuer des *crawls* de sites web pour mieux les catégoriser et faciliter leurs analyses sous des formes notamment cartographiques. Basé sur l'analyse des communautés en ligne et des controverses, *My Web Intelligence* est un outil pour les *digital studies* dont l'intérêt dépasse les seuls intérêts de la recherche pour faciliter l'étude et l'analyse des réseaux d'influence et des stratégies de viralité de l'information.



Amar LAKEL

Amar Lakel est maître de conférences en sciences de l'information et de la communication à l'Université Bordeaux Montaigne. Spécialisé en humanités digitales, en gouvernance politique et en analyse des controverses, il mène ses recherches au sein du laboratoire MICA. Il est le concepteur du logiciel *My Web Intelligence*.

⊕ Amar.Lakel@u-bordeaux-montaigne.fr

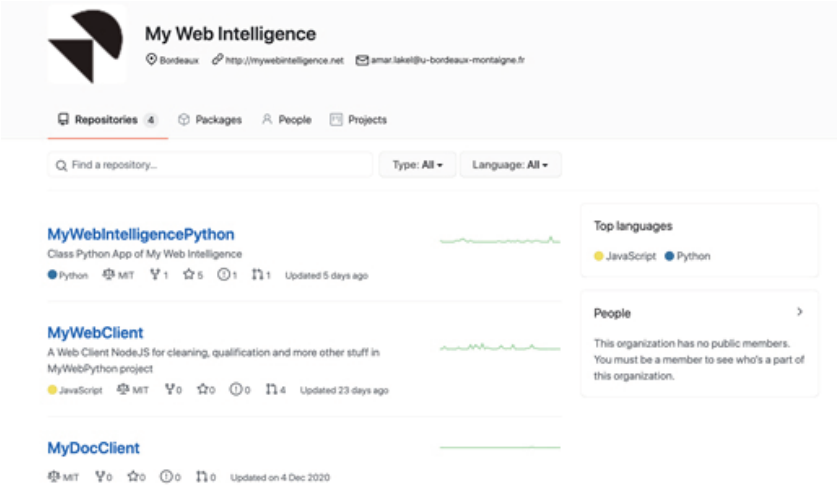
My Web Intelligence est un programme que je dirige au sein de l'équipe E3D du Laboratoire MICA (MICA) de l'Université Bordeaux Montaigne[1]. Le programme vise à développer un outil d'extraction (*crawl*), d'archivage, de qualification et de visualisation du Web au service des *digital methods*. L'objectif est de fournir, à tous les experts et chercheurs qui souhaitent développer des études dans le domaine de l'intelligence numérique et des humanités digitales, un dispositif basé sur l'analyse des prises de parole en ligne.

[1] https://mywebintelligence.net/en/my-web-intelligence-mapping-web-controversies/

Crawler le web : un échantillonnage de proche en proche

My Web Intelligence s'appuie sur les moteurs de recherche web pour obtenir un premier corpus de documents pour démarrer la collecte d'informations. En croisant les sources des différents types d'infomédiaires, on multiplie les rationalités algorithmiques qui nous font entrer dans notre espace public numérique. À partir de ce corpus de premier niveau, l'exploration continue des liens sortants qualifiés permet de s'enfoncer dans les couches profondes du web pour obtenir le territoire numérique le plus complet possible au cœur de nos préoccupations. Ainsi, entre *crawl* profond et évaluation progressive des informations les plus pertinentes au regard de son dictionnaire projet, la plate-forme, travaillant en tâche de fond, finit par constituer un territoire d'informations traitant d'un sujet donné. Le *crawler* est donc une machine à échantillonner sur une méthode de proche en proche. Mais il faut nécessairement l'associer à des algorithmes d'approbation qui se doivent de rejeter le bruit et de classer les documents dans un ordre de priorité. L'extracteur de corpus en charge de la constitution des archives numériques embarque un navigateur web en charge d'absorber les ressources numériques qui est doté de la capacité d'extraire le contenu éditorial de la page (en mode *readable*) et d'isoler les documents multimédias de ce contenu (détections des liens hypertextes, détection des médias, etc.). Si le document est jugé pertinent, les liens hypertextes sont explorés pour récupérer les documents cités. De proche en proche, le *crawler* extrait un échantillon semi-représentatif du web.

Figure 1. Codes sources sur Github en licence open-source MIT



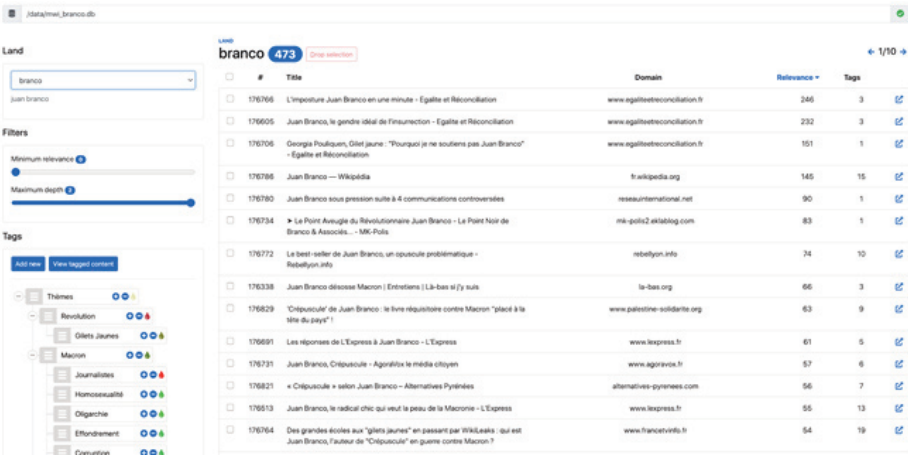
Constituer un corpus web : une logique d'assistance au chercheur

My Web Intelligence est composé de deux briques logicielles[2]. *My Web Intelligence* Python, une brique logicielle en mode console développée sous python et qui permet d'extraire les données du web : c'est l'agent d'enquête du projet. *My Web Client* qui permet de naviguer dans son corpus de recherche pour non seulement nettoyer, mais appréhender son corpus par une interface de navigation web. Après l'ouverture d'un projet de recherche, le professionnel doit compléter un dictionnaire de mots clés qui permettra au *crawl* d'évaluer la pertinence des pages qu'il collecte et auxquels il attribuera une note qui servira, plus tard, aux filtres et exports de corpus. L'entrée et la sortie de données par des fonctions d'import/export, dans le cadre d'un projet, permettent l'export des énoncés en format csv ou gexf (pour l'analyse réseau sous Gephi), des domaines en csv ou gexf (données regroupées à l'échelle du site web), des médias en liste csv (images et vidéo) pour l'analyse visuelle. Les formats csv, gexf et l'utilisation d'une base de données fichier SQLite assurent l'interopérabilité avec tous les logiciels d'analyse du marché (R, iramuteq, etc.). On trouve parmi les variables qui qualifient la page : le titre*, l'URL*, la relevance* (pertinence au regard du dictionnaire projet), depth* (la profondeur d'extraction avec 0 pour les pages ajoutées par l'utilisateur), le domain_id* et le domain_name* (id et nom du domaine d'expression) et son contenu texte*. On ajoute manuellement la date de publication* sur Google et le nombre de partages*, de commentaires*, d'interactions* sur Facebook (obtenus grâce à l'accès à son API). Il faut ajouter les données issues de l'analyse structurale des réseaux que l'on calcule grâce au fichier gexf des pages et le logiciel GEPHI (indegree*, outdegree*, etc.)

Ce sont en tout pas moins de 23 variables qui viennent identifier le texte et son contexte d'énonciation (inscription dans les réseaux de citation des pairs et réception de son lectoratsur les réseaux sociaux). Un second niveau d'analyse opère par regroupement des expressions au niveau du domaine d'expression que l'on qualifie humainement selon la nature sociale du propriétaire du média (secteurs d'activité, le niveau d'institutionnalisation, type de média numérique, etc.). Aux variables qui visent à inscrire sociologiquement l'équipe éditoriale, on ajoutera les indicateurs MOZ (autorité du site web) et l'*Alexa Rank* (indicateur d'audience), mais aussi les données des pages engagées dans le débat (somme des partages, des commentaires et réactions totales sur Facebook, nombre de pages engagées dans la controverse, date de la première publication).

[2] Le logiciel est téléchargeable à cette adresse : https://github.com/MyWebIntelligence

Figure 2. Interface de navigation du corpus My Web Client



Nettoyer les données massives : la gestion collective des corpus

My Web Intelligence est dotée d'un tableau de bord pour gérer les grands corpus à l'aide d'un certain nombre d'indicateurs. Cette interface de nettoyage et de qualification des données permet, non seulement un contrôle et une suppression du bruit, mais aussi une qualification thématique des pages web. Le nettoyage de données est une étape essentielle dans toute recherche. Pour autant face à la taille des corpus, il ne peut se faire sans l'aide d'une part d'agents algorithmiques et d'autre part par la mobilisation de collectif.

L'interface *My Web Client* offre la possibilité à l'utilisateur d'annoter humainement le document. Une gestion des thèmes et des contenus qu'il identifie permet, par la suite, de travailler thème par thème sur l'analyse de contenu soit directement sur l'interface soit en exportant la base de contenu thématisée pour une analyse lexicologique postérieure.

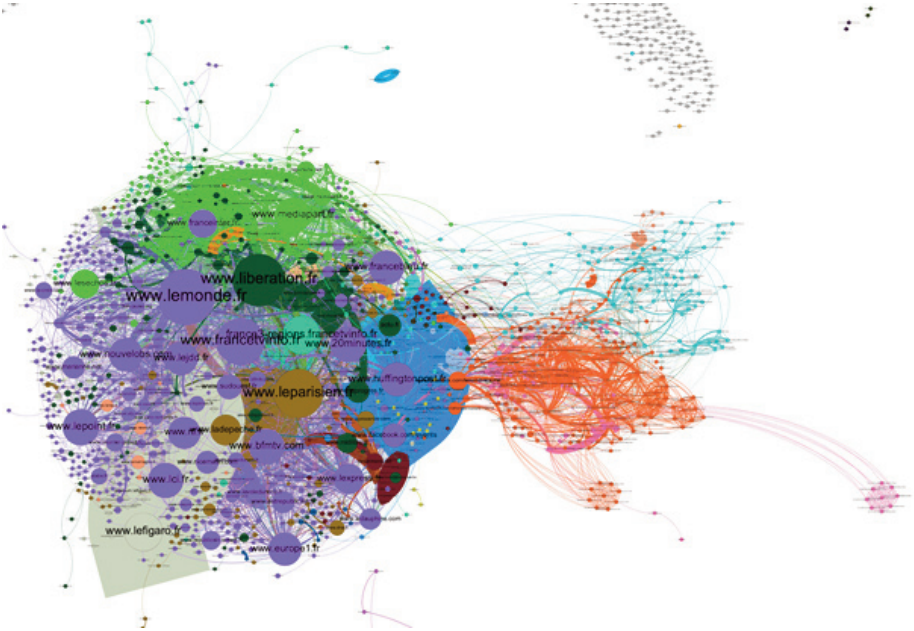
Fig. 3. Interface d'annotation des expressions

Classer, catégoriser et comprendre

Une fois l'extraction et la qualification des données d'une controverse achevées, My Web Intelligence donne accès à un corpus nettoyé qui permet de pouvoir mettre en place un ensemble de traitements d'analyse et de traitements des données pour tirer véritablement une compréhension de l'économie de la discussion en ligne. Le premier travail est d'utiliser la théorie des graphes et l'analyse structurale des réseaux pour générer des cartographies des médias qui sont à l'origine de la controverse. En effet, derrière les mots, il y a des locuteurs aux commandes de supports médiatiques. Des locuteurs situés et engagés dans un espace public numérique. Il faut non seulement pouvoir qualifier ces médias selon leur nature sociale, leurs comportements éditoriaux, mais il faut avant tout révéler à travers la structure de leurs citations qualifiées, le contexte d'alliance et d'adversité qu'ils tissent dans les processus de légitimation, mais aussi d'opposition. « Dis-moi qui tu cites, quelles sont tes références et je te dirai qui tu es. »

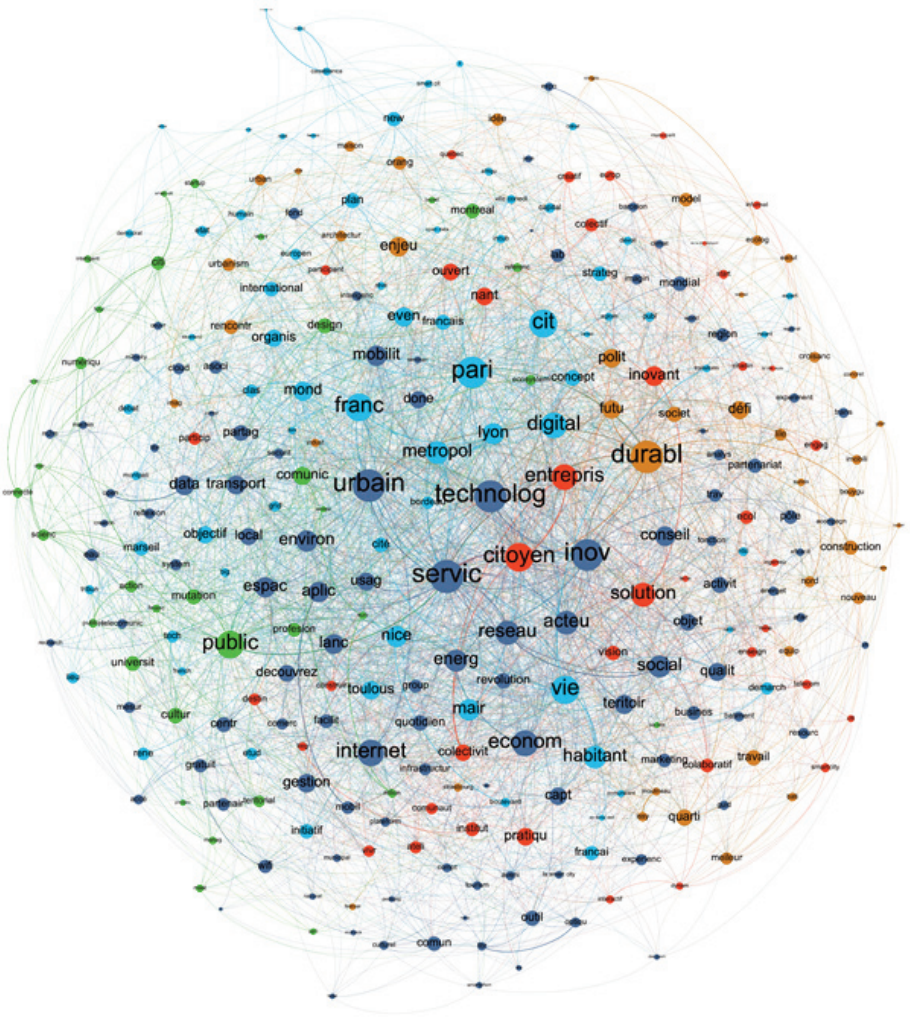
Une vision globale et structurale des acteurs révèle non seulement la structure des alliances et des oppositions, mais elles révèlent les communautés d'intérêts idéologiques et situe chaque média selon un rôle social dans le débat et au sein de sa communauté (leader d'opinion, vigie, marginal sécant, bridge, etc.). Cette recontextualisation du locuteur au cœur de ses « amis » nous informe sur la position sociale du média au sein d'une communauté stratégique.

Fig. 4. Cartographie médias de la couverture des Gilets Jaunes de oct. 2018 à juin 2019



Analyse des graphes et analyse de contenus : des pistes nouvelles ?

Figure 5. Analyse lexicale des pages web traitant de la Smart City en 2020



Enfin, l'analyse des graphes peut être utilisée pour comprendre la structuration argumentaire. La cartographie de mots-clés révèle alors une structure du dictionnaire qui est le produit latent de la construction de la réalité par la prise de parole dans des médias donnés. La cartographie dynamique argumentaire permet de retracer la genèse des stratégies argumentatives. L'utilisation des variables topologiques des graphes nous permet de comprendre aussi le rôle

et la place de chaque concept dans une stratégie argumentaire globale. En réalité les sujets qui prennent position dans une controverse sont dans leur très grande majorité des porte-paroles qui habitent des discours qui leur pré-existent et qu'ils travaillent à la marge. La controverse voit rarement la création innovante d'arguments et bien plus souvent une prise de position sur des arbres argumentaires toujours déjà là dans des énoncés produits comme des mêmes. Elle permet surtout de repérer les émergences et les innovations, la diffusion voire la viralité de certains concepts.

Bibliographie indicative

Lakel, A. 2019. « Prises de positions et influences sur le web : le cas de l'information de santé ». *Revue française des sciences de l'information et de la communication* (18). doi: 10.4000/rfsic.8376.

Lakel, A., et Le Deuff, O. 2017. « À quoi peut bien servir l'analyse du web ? » *Les Cahiers du numérique*, 13(3):39-62.

Des vidéos de formation au niveau de la démarche et de la prise en main de l'outil sont disponibles en vidéo : *My Web Intelligence - Formations* https://www.youtube.com/playlist?list=PLbCMGWVe0gqGjHwqSwz9TT5nhTFWpthQZ