Insee
Mesurer pour comprendre

Big Data et Statistique Publique

Lettre d'information
N°5 – mars 2017

Enrichir nos données grâce à Internet

De très nombreuses informations, disponibles sur différents sites internet, peuvent enrichir nos données. Si le relevé de ces informations peut se faire manuellement (c'est par exemple le cas de certains tarifs entrant dans le calcul de l'indice des prix), des techniques se sont développées pour permettre une extraction automatique de l'information sur internet : on parle de « webscraping ». Ces techniques sont déjà utilisées pour enrichir l'indice des prix dans certains pays (voir la lettre BDSP n°2) ou pour des projets expérimentaux portant sur l'exploitation d'offres d'emploi, qui font l'objet d'un volet de l'ESSnet Big Data en cours. Les données issues d'internet peuvent plus directement être mobilisées pour enrichir et préciser des données existantes. L'expérimentation décrite dans cette lettre en fournit un exemple. Il faut cependant avoir en tête que l'extraction des données peut soulever des problèmes juridiques, le propriétaire du site conservant le droit de propriété intellectuelle sur les données qu'il met à disposition. Un accord avec le propriétaire d'un site constitue donc un préalable indispensable à une extraction à grande échelle. Par ailleurs, dans un contexte où de plus en plus de données sont mises à disposition librement sur internet, un enjeu important est de pouvoir obtenir simplement des informations précises, sans nécessairement récupérer des données très volumineuses : c'est ce que permet le web sémantique présenté dans le dernier paragraphe.. Les projets décrits ici sont multithématiques et comportent des dimensions techniques poussées. C'est pourquoi les expérimentations en cours à l'Insee associent de très nombreux acteurs (informaticiens, méthodologues, statisticiens…), l'association de compétences complémentaires étant indispensable.

Pauline Givord (Division Méthodes Appliquées de l'Économétrie et de l'Évaluation)

Collecter des données sur internet : le webscraping

Internet : du texte dans des balises

Comment repérer de l'information sur des pages web ? Les pages web sont écrites en langage HTML ou XHTML : en langage courant, cela signifie que s'y entremêlent des **champs d'information textuelle** et des **« balises »**. Les balises structurent les pages web en délimitant les différents éléments sémantiques comme les titres, les listes ou les liens vers d'autres pages. Ainsi en parcourant le code HTML d'une page web on peut **repérer** puis **extraire** l'information pertinente. Par exemple, supposons qu'on souhaite compléter une base de produits et de prix avec des produits vendus sur internet : il s'agit donc de relever les noms et prix de ces produits sur les pages web des sites marchands correspondants. En général, si le site est bien fait, les balises encadrant les noms et prix des produits auront des identifiants permettant de les retrouver automatiquement au sein d'une page contenant beaucoup d'autres informations (par exemple ceinture ou 20€). On utilise pour cela des "robots", c'est-à-dire des programmes conçus pour récupérer automatiquement l'information contenue entre des balises dont les caractéristiques auront été ciblées au préalable (ici les classes 'produit' et 'prix'). Selon le format dans lequel sont stockées les informations sur le site, cette extraction peut être plus ou moins complexe. S'il s'agit déjà d'une table de données, il est assez aisé de repérer les balises pertinentes et d'extraire leur contenu. À l'extrême opposé, il est parfois nécessaire de naviguer dans l'arborescence d'un site, en passant d'une page à une autre en suivant les liens qui y figurent, ou de sélectionner un item dans un menu déroulant.

Des outils clé-en-main existent et permettent de scraper très simplement des données bien structurées (tables en particulier), en fournissant seulement l'adresse de la page. Import.io est l'un d'entre eux. Il est gratuit tant que le nombre de requêtes par mois reste inférieur à 500 requêtes. D'autres outils sont plus personnalisables, notamment Scrapy : avec quelques lignes de code en Python, cet environnement permet de naviguer dans les sites et d'en extraire l'information aisément. Des librairies équivalentes sont également disponibles en R.

En pratique, il faut prendre garde que l'extraction des données à grand volume peut soulever des problèmes pour le propriétaire du site, par exemple parce qu'une extraction massive peut ralentir l'utilisation de son site. Pour s'en prémunir, plusieurs sites mettent en place des dispositifs pour bloquer le scraping : par exemple par . des « CAPTCHA » (qui demandent de cocher la case "je ne suis pas un robot"), ou par le blocage d'adresses IP se connectant trop souvent. Il faut en particulier vérifier que le site autorise la récupération des données à grande échelle, ce qui est en général indiqué sur le site.

Utiliser les sites d'offres d'emploi pour enrichir les statistiques d'emploi : un projet complexe

Plusieurs instituts de statistique européens s'intéressent à l'exploitation de données issues d'Internet. Un volet de l'Essnet Big Data, auquel participe la Dares, porte sur l'utilisation des portails d'offres d'emploi (par exemple Jobijoba, le Bon Coin...) pour améliorer les statistiques sur les offres d'emploi et les emplois vacants.

Les atouts identifiés de ce type de source sont qu'elles peuvent fournir rapidement des informations sur le stock d'offres disponibles, ainsi que des détails sur les employeurs, leur localisation géographique, les compétences attendues...

L'exploitation de ces données soulève plusieurs difficultés néanmoins. Tout d'abord, l'identification des plateformes intéressantes. Celles-ci peuvent être très nombreuses, (par exemple 1600 portails ont été identifiés en Allemagne et plusieurs centaines en France) et très variées (par exemple spécialisées sur une profession ou une région), et leur nombre évolue régulièrement.

L'option la plus naturelle est soit de se limiter aux sites les plus importants en termes d'offres d'emploi (à condition de disposer d'information fiable et indépendante sur ces taux de fréquentation, ces chiffres étant souvent surestimés par les plate-formes), soit de développer des partenariats avec plusieurs sites diffusant des offres d'emploi pour obtenir les données sans directement procéder au scraping.

Cette dernière option, actuellement expérimentée par Pôle emploi pour ses « sites partenaires », présente l'intérêt de fournir de l'information structurée. En revanche, l'agrégation de plusieurs sites expose au risque de créer des doublons, qui ne sont pas toujours simples à identifier : les employeurs peuvent poster la même offre d'emploi sur plusieurs sites, avec un niveau de description qui peut varier selon le site utilisé. Les classements, les termes utilisés pour décrire les emplois, les nomenclatures, peuvent potentiellement être spécifiques à la plate-forme utilisée.

Dans tous les cas, l'information extraite des sites doit être travaillée et confrontée à des données de référence. En effet, une annonce peut correspondre à plusieurs emplois, plusieurs lieux, et le fait que l'emploi soit effectivement vacant à la date de l'extraction n'est pas toujours assuré.

Plusieurs pays (Royaume-Uni, Slovénie en particulier) ont déjà lancé des travaux exploratoires sur le sujet. La Dares participe à l'ESSnet sur le sujet. Ainsi elle est associée à une expérimentation en cours avec le SGI et la division MAEE pour scraper et analyser des offres du site « Le Bon Coin » : ces projets, complexes, restent encore à un stade exploratoire.

Utiliser les données issues d'internet pour enrichir une base de sondage : un exemple d'expérimentation menée par la division Commerce

Les divisions Commerce et Services réalisent, en alternance, une enquête sur les réseaux d'enseignes du commerce de détail deux années sur trois. La connaissance de ces réseaux est centrale pour décrire l'équipement commercial : ils regroupent la majorité de l'emploi salarié du secteur et les points de ventes qui en sont membres présentent des spécificités marquées. Suivant cet objectif, l'enquête « Contours de réseaux » 2016 vise à établir un recensement des réseaux du commerce alimentaire et la liste exhaustive des points de vente affiliés.

Un des enjeux pour garantir la qualité des données de cette opération est de disposer d'une base de sondage précise et exhaustive. Or, à la différence de la plupart des enquêtes « Entreprises » de l'Insee qui s'appuient sur le répertoire Sirus, il n'existe pas de répertoire des réseaux. La connaissance des réseaux repose ainsi sur des sources diverses et partielles : anciennes enquêtes, fédérations professionnelles et informations sur l'enseigne du répertoire Sirene.

Ces diverses sources permettent d'établir une liste de mots clés correspondant à des enseignes potentielles. La division Commerce a fait appel à l'expertise du département des Méthodes Statistiques pour automatiser partiellement les recherches internet permettant de déterminer qu'un mot clé désigne effectivement un réseau.

Deux procédures ont été mises en place :

- pour vérifier si un mot clé a fait l'objet d'un enregistrement comme marque en interrogeant le site de l'Institut National de la Propriété Industrielle (Inpi) puis, le cas échéant, pour obtenir la raison sociale du déposant ;
- pour identifier les sites internet portant le nom du réseau.

Ces deux procédures ont été menées sur une liste de plus de 4 000 mots clés. Si une vérification manuelle a été ensuite nécessaire, les requêtes automatisées ont accéléré ce travail. Elles ont également permis d'écarter d'emblée certains mots clés n'ayant donné aucun résultat lors des recherches web. Finalement, plus de 400 réseaux d'enseignes ont été identifiés dans le commerce de détail alimentaire, dont la majorité n'était pas connue auparavant.

Au-delà de ces premiers résultats encourageants, ces procédures restent à étendre à d'autres secteurs du commerce, ce qui pourrait permettre d'initialiser un répertoire des réseaux et faciliter les prochaines enquêtes. Plusieurs pistes sont envisagées pour rendre cette démarche encore plus efficace : restreindre le nombre de sites internet obtenus en sélectionnant plus finement les résultats pertinents, ou encore étendre la récupération de données internet à l'extraction de la liste des magasins d'un réseau, disponible sur son site.

Enrichir ses données par le Web sémantique

Le web scraping n'est pas toujours indispensable pour récupérer des données sur Internet. Beaucoup d'informations sont librement disponibles sur le web : on parle de données ouvertes. Dans une variante particulièrement aboutie, les Linked Open Data (LOD), la sémantique des données est formalisée, afin de les structurer et de les connecter entre elles. Ce web de données, dit aussi « web sémantique » apporte des technologies qui facilitent l'exploration et la réutilisation des données, en particulier un langage de requête (SPARQL) qui permet de récupérer les données directement sur le web.

Comme les pages HTML sur le web habituel, les LOD sont identifiées par des adresses Internet. Ainsi, l'Insee publie à http://id.insee.fr/geo/region/11 des informations sur l'Île-deFrance : nom, population, liste des départements, etc. D'autres données sur cette région sont fournies dans d'autres bases, par exemple son contour géographique à http://nuts.geovocab.org/id/FR10 ou les informations générales de Dbpedia, la version « web sémantique » de Wikipedia, à l'adresse http://fr.dbpedia.org/resource/Île-de-France. Lier les données entre ces différentes sources, c'est simplement ajouter dans les informations publiées des relations sémantiques entre les différentes ressources. Par exemple, la déclaration :

http://id.insee.fr/geo/region/1"sameAS" http://nuts.geovocab.org/id/FR10

établit que les deux adresses identifient en fait la même ressource : la région Île-de-France. On peut, en suivant les liens sémantiques, rassembler toutes les informations publiées par les différents producteurs de données concernant la région.

"sameAs" n'est qu'un exemple des prédicats sémantiques disponibles pour lier entre elles les ressources du web de données : des relations plus complexes peuvent être également définies (une des entités contient l'autre ou en est une partie, représente un concept plus général que l'autre ou au contraire plus précis, etc.).

Tous ces liens s'expriment toutefois dans un même formalisme très simple : des triplets sujet – prédicat – objet. Ce modèle, connu sous le nom de RDF, est en fait utilisé pour exprimer toutes les propriétés des LOD, par exemple :

http://id.insee.fr/geo/region/11 "a pour nom" "Île-de-France"

Comme l'objet d'un triplet peut être le sujet d'autres triplets, les LOD forment en fait un gigantesque graphe dont la taille est estimée à 150 milliards de triplets. De nombreux autres triplets RDF sont inclus dans des pages HTML et peuvent être extraits pour être connectés aux LOD. Ils obéissent à des vocabulaires normés, notamment schema.org, prôné par les grands moteurs de recherche (Google, Microsoft Bing, etc.). Plus de 10 millions de sites web, en particulier les grands sites de commerce en ligne, utilisent schema.org.

Actualités / Brèves

- une réunion de l'ESSnet BigData auquel le SSP participe a eu lieu le 21 février ; les principaux résultats ont été présentées au workshop de diffusion

- les supports des présentations du Séminaire de Méthodologie Statistique sur le Big Data et machine learning sont disponibles ici.

- Eurostat organise du 13 au 15 mars un Big Data Hackathon pour des équipes des INS volontaires. Benjamin Sakarovitch (DMS), Yves-Laurent Benichou (SGI) et Stéphanie Combes (DMS) représenteront l'Insee.

Agenda

- 14 et 15 mars : New Techniques and Technologies for Statistics à Bruxelles

- 16 et 17 mars : « Science XXL – ce que l'abondance et la diversité des données numériques font aux sciences sociales » conférence organisée par l'Ined,

Ont participé à ce numéro Franck Cotton, Stéphanie Combes, Benjamin Sakarovitch, Romain Tailhurat, Corentin Trévien, Maxime Bergeat, Françoise Dupont, Pauline Givord.

Cette lettre est une occasion d'informer largement et d'échanger. N'hésitez pas à nous transmettre vos réactions et suggestions d'articles à stephanie.combes@insee.fr ***et*** benjamin.sakarovitch@insee.fr

Les archives de cette lettre sont disponibles sur :
http://www.agora.insee.fr/jahia/Jahia/site/dmcsi/SiteDMCSI/DMSaccueil/DMAEEaccueil/BigData

Demande d'inscription individuelle à la lettre : dg75-l101@insee.fr