Extraction automatique et massive de données
Présentation d'une méthodologie pour les sciences sociales

Thomas Delclite
Université Lille 1 – Laboratoire Clersé

Journée d'étude « Données numériques en Sciences Sociales. Collecte, structuration, analyse et diffusion ».

Avril 2015

Afin d'effectuer ses recherches, le chercheur en sciences sociales est régulièrement amené à faire de l'analyse quantitative, et donc à étudier des bases de données, quel que puisse être l'objet de ses recherches. Cette volonté peut être bloquée par un constat simple : les données peuvent ne pas être disponibles. En France, par exemple, la statistique publique fournit un large choix de données, souvent fiables. Ces données permettent de mener des études sur de nombreux sujets, mais en excluent quasi automatiquement certains autres. Pour citer quelques exemples, les entreprises n'ont aucune obligation de fournir des données historiques sur le tarif de leurs produits, ou sur leur stratégie de communication. Ceci peut limiter certaines études économiques, sociologiques, en gestion, portant justement sur l'analyse de certaines entreprises, de leur discours, ou de certains secteurs d'activités. Concernant la recherche académique, les éditeurs de revues n'ont aucune obligation de fournir des données rétrospectives sur les publications de leur(s) revue(s), notamment leur contenu.

Néanmoins, de telles informations sont disponibles pour l'utilisateur lambda : les entreprises disposent de plus en plus fréquemment de services d'achat en ligne, les chercheurs affiliés au CNRS peuvent obtenir une grande partie des articles de recherche publiés sur les 100 dernières années. L'information est donc partiellement disponible et accessible, mais non directement exploitable. Cette disponibilité n'est ici que partielle, car si l'usager d'un site peut "voir" les informations, le chercheur ne peut pas les traiter en grande quantité. Théoriquement, il serait possible de naviguer sur le(s) site(s) web et extraire manuellement l'information (tarifs, méta-données, ..), la limitation tient au fait que ce travail est long et sujet à erreur de report.

Cette communication présente, en ce sens, une méthodologie d'extraction massive de données. Il ne s'agit pas d'un produit fini, mais d'une méthodologie pouvant s'appliquer à une large variété de sujets. Pour les sciences sociales, cette méthodologie permet la création de bases de données originales, et ainsi la production de connaissances autour de nombreux sujets. La massification de l'usage d'Internet permet de mener des études empiriques en évitant l'écueil de la disponibilité statistique. Cette méthodologie s'inscrit également dans les sciences du discours, en s'intéressant justement à ce que les acteurs produisent comme données et informations.

La communication présentera dans un premier temps la méthodologie d'extraction massive de données. Nous détaillerons ensuite les manières d'analyser l'information obtenue. Pour cela, nous nous appuierons sur deux travaux menés à terme avec cette méthodologie : l'analyse des usages du critère de Pareto par les économistes durant les 100 dernières années et l'analyse de l'évolution des tarifs des billets de TGV proposés par la SNCF. Ces deux travaux, très différents dans leur objet d'étude, permettront de mettre en lumière les possibilités de cette méthodologie.