Journées pédagogiques SIF

23 juin 2015

au CNAM

Apports de l'informatique aux SHS… et réciproquement

Nathalie Denos - Univ. Grenoble Alpes

Informatique et SHS : le cas du traitement de données massives

- données massives
  - les exploiter
  - pour quoi faire ?
  - quelle valeur scientifique ?
  - quelles limites à l'automatisation des traitements ?
  - quels liens avec la société, les droits fondamentaux, l'économie numérique ?

Observation : journées scientifiques « regards croisés… »

Objet

- nouveau matériau : millions de messages publiés sur les différents espaces participatifs du web
- indices de pratiques, d'attitudes et d'opinions pour analyser le monde social

Questions

- A partir de quelles méthodes et de quels outils informatiques étudier ces messages ?
- Comment penser la relation entre les données online disponibles sur le Web et les (dis)positions "réelles" des individus ?
- Comment informaticiens et chercheurs en SHS peuvent-ils travailler ensemble sur ces questions ?

Disciplines

- informatique
- sciences politiques
- sociologie
- linguistique
- sémiologie
- info – com
- …

Programme : http://mediamining.univ-lyon2.fr/velcin/webpol/#programme

Instantané 1 : enjeux méthodologiques, épistémologiques, déontologiques

- Dominique Cardon
  - se définit comme sociologue qualitativiste (pas très stat)
- Enjeux méthodologique d'une enquête sur Facebook : le projet ALGOPOL

- éclairer les enjeux méthodologiques et épistémologiques d'une enquête de sciences sociales à partir d'une extraction massive de données numériques sur Facebook.
- retracer les difficultés et les choix effectués pour conduire l'expérimentation
- réfléchir à la nature des connaissances produites à partir des données du web et aux enjeux méthodologiques et déontologiques du projet

Instantané 1 : enjeux méthodologiques, épistémologiques, déontologiques (suite)

- questions préalables
  - est-ce que le web, c'est la société ?
  - les données du web sont accessibles, donc publiques ?
  - l'enquêté a consenti, tout va bien ?
- le travail des données
  - 1 an à comprendre les données fournies par Facebook
    - sale, mal catégorisé, faux, ...
    - tout re-décrire avec un système de règles (le code python sera publié, article écrit à ce sujet)
  - culture du big data / massif
    - entre en conflit avec la qualité des données exigée par les sociologues
- les données du web sont auto-suffisantes ?
  - données externes requises
  - enquêtes hyper-qualitatives (entretiens très approfondis avec un petit nombre de personnes)
- collecter le profil, le réseau social, et le mur de l'enquêté
  - contrat assez compliqué : on prend tout, mais en retour on te rend une vision de ton réseau
  - on récupère des éléments sur les amis qui eux n'ont pas consenti ; c'est interdit
  - tentative de mise en place d'un semi-consentement, idée suggérée par la CNIL
  - aujourd'hui données anonymisées et disponibles dans un cadre contraignant (engagement à détruire les données)

Instantané 1 : éléments de réflexion

- qualité des données, exigence méthodologique
- lien au cadre juridique et aux droits fondamentaux
- automatisation du pré-traitement
- publication du code
- lien au monde économique

Instantané 2 : archivage, constitution d’un corpus, valeur historique et patrimoine

- Jean-Marc Francony et Peter Stirling
  - info-com
- Fouille de données du web dans un contexte d'archivage institutionnel : une collaboration autour des campagnes électorales 2012

- collaboration entre le laboratoire PACTE et la BnF
- mettre en place un accès aux données du dépôt légal de l'internet en vue de les traiter dans une perspective de recherche, à travers les collections sur les élections présidentielle et législatives de 2012

Instantané 2 : archivage, constitution d'un corpus, valeur historique et patrimoine (suite)

- évolution du code du patrimoine
  - tout peut être considéré comme relevant de l'archivage
  - production de documents sur le territoire national ou liés à la France
  - la BNF n'a pas besoin d'avertir, ni d'autorisation particulière pour archiver
- acquis méthodologiques
  - méthode empirique : collecte de sources
    - fixes
    - mouvantes (réajustement en cours de campagne)
    - manuelles (opportunistes)
  - pas une visée d'exhaustivité mais de représentativité
    - logique de collection
- approche technique et documentaire d'archivage de la BNF
- partition entre la BNF et l'INA, qui partagent la mission d'archivage du web
- approche BNF = modèle intégré (BcWeb)
  - collectes larges : annuelle, d'un très grand nombre de sites
  - collectes ciblées : plus fréquemment, sur des sites sélectionnés par des bibliothécaires ou des partenaires
- robot Heritrix open source

Instantané 2 : éléments de réflexion

- problématique de l’archivage
  - volatilité des données du web
  - valeur historique
- en recherche, mais aussi…
  - tirer profit des données

Instantané 3 : faire émerger un sens caché

- Pascal Marchand
  - info-com
- L'économie au FN : programme ou rhétorique ?

  l'euro est au centre du discours économique du FN

  l'immigration est traitée majoritairement comme une question économique plus que comme une question sociale

- analyse textométrique appliquée à un corpus constitué d'extractions d'articles sur les sites Web du Front national indexés par le mot clé "économie"
- classification lexicale et analyse d'une matrice de similitudes
- espaces lexicaux mêlant un discours économique et des thématiques identitaires

Instantané 3 : éléments de réflexion

- cadre théorique et méthodologique
  - capacité à interpréter les données
  - tirer du sens des données « brutes »
- représentation graphique
  - nuage de mots vs. graphe de similitudes

- ex de Graphs, maps, trees
  - que peut-on dire des figures féminines dans la littérature du 18è et 19è siècle à partir des titres ?

Instantané 4 : nécessité et limites du traitement automatique pour une approche sémiotique

- Virginie Julliard
  - info-com, approche sémiotique
- Étudier la controverse sur la "théorie du genre" dans Twitter, Enjeux méthodologiques et épistémologiques d'une recherche outillée

- étudie la façon dont le genre est mis en débat à l'occasion de la controverse sur la "théorie du genre" dans la presse quotidienne nationale et Twitter
- hypothèse : le dispositif technico éditorial qu'est Twitter influerait sur la manière dont se déroule la controverse sur la théorie du genre et sur la mise en débat du genre

Instantané 4 : nécessité et limites du traitement automatique pour une approche sémiotique (suite)

- différence entre les formes d'enregistrement et les formes de restitution
  - le chercheur dispose d'autre chose que ce à quoi le lecteur / l'internaute est confronté
- importance de l'image
  - ex : la théorie du genre c'est CA ! + photo d'un garçon avec une brique de lait scotché sur le corps qui nourrit un bébé
- masse des messages empêche un traitement manuel : quels traitements automatiques ?
- co-conception d'un outil de captation et de traitement des tweets
  - reconstruire les conditions de possibilité d'une démarche sémiotique sur Twitter
    - outils existants, mais les outils ne sont pas neutres
    - travail itératif de conception avec un ingénieur qui a développé
  - point de vue situé et qualitatif vs. outils d'analyse quantitative
- aller-retour quantitatif / qualitatif
  - idées suggérées par l'informaticien de ce qu'il peut proposer

Instantané 4 : éléments de réflexion

- Masse des données : un obstacle
- Traitement automatique vs. aide à l'exploration
- Coopération avec l'informaticien

Instantané 5 : SHS 3.0, la vibration

- Dominique Boullier
  - sociologue
- De l'opinion aux vibrations de la vie politique sur le web

- L'opinion mining a trompé son monde, à la fois sur les performances de ses méthodes mais aussi sur son nom
  - opinion = concept qui doit être réservé à l'entité très bien fabriquée depuis près de 80 ans par les sondages
  - plate-formes et marques produisent avant tout des vibrations à partir des traces
  - bénéficie de la puissance de calcul et de la vélocité du Big Data mais perd toute exhaustivité et toute représentativité
  - construire une nouvelle convention qui ouvre la voie à des sciences sociales de troisième génération

Instantané 5 : éléments de réflexion

- les représentations de la société
  - face à la complexité
- apprivoiser les médias sociaux
  - cadre théorique d’analyse

Importance de …

- la qualité des données
  - représentativité
  - exhaustivité
  - qualification
  - nettoyage
- ce qu'on peut automatiser… ou pas !
  - scripts
  - image
- l'archivage
  - du web…
  - valeur historique
- la représentation graphique
  - rigueur de l'interprétation
- le partage du code
  - un effort
  - un réflexe
  - une culture
- l'exploitation des données
  - cadre juridique
  - accords et conventions

Formes de coopération entre informaticien et spécialiste de SHS

- complémentarité / juxtaposition
  - le technicien informaticien au service du spécialiste de SHS
- croisement fertile
  - l'informaticien voit des limites, des potentialités, ... et les soumet au spécialiste de SHS, enrichissant ainsi sa réflexion et ses pistes de travail
  - le spécialiste de SHS porte des exigences quant à la qualité des données et au sens qu'on peut leur associer (chasse aux interprétations abusives qu'on pourrait en faire avec un regard critique et une méthodologie moins affutés), il cherche à établir des conventions équitables et socialement acceptables avec les fournisseurs de données, ...
- appropriation
  - le spécialiste de SHS "met les mains dans le cambouis"