Anonymisation de contributions OSM

Guillaume Touya, Arnaud Le Guilcher

5 janvier 2017

1 Présentation du sujet

1.1 Le projet OSM

OpenStreetMap est un projet visant à collecter des informations géographiques à échelle mondiale de manière collaborative. Dans ce contexte, chaque membre du projet peut créer ou éditer une géométrie, et ajouter ou modifier des méta-données relatives à une géométrie. Sur un même territoire, la qualité et le niveau de détail des géométries peuvent être très hétérogènes, de même que l'exhaustivité et la précision des méta-données.

Pour cartographier et utiliser les données issues du projet OpenStreetMap, une évaluation de la qualité des contributions est donc nécessaire. Une solution pour qualifier les contributions du projet est d'utiliser des méthodes d'apprentissage pour identifier différents types de contributeurs et utiliser cette classification pour évaluer chaque contribution en fonction du contributeur, mais l'application directe d'une telle méthode aux données d'OpenStreetMap se heurte à des problèmes liés à la protection de la vie privée des contributeurs.

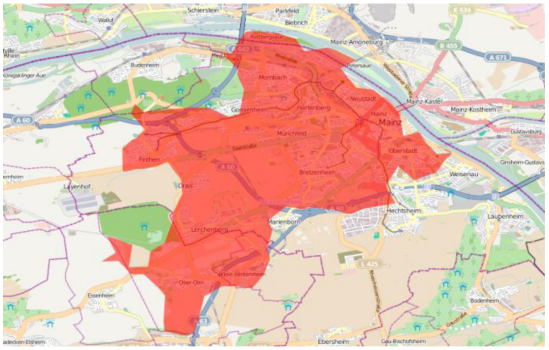

Figure 1: Exemple de zone d'activité d'un contributeur OSM

1.2 Protection de la vie privée

Les recherches en anonymisation partent du constat qu'il peut être possible, en croisant les informations d'une base de données avec des informations externes, de retrouver la personne à laquelle se rapporte une information contenue dans

une base de données, même lorsqu'on a retiré de la base les informations permettant une ré-identification immédiate (nom ou numéro d'identification) ([1]). De nombreuses méthodes d'anonymisation ont été proposées pour les bases de données relationnelles ([2]) et les bases content des données géolocalisées (par exemple [3]).

Quand les bases de données étudiées contiennent des données géo-localisées, les risques liés à une possible ré-identification des informations sont multiples, car les données peuvent indiquer des habitudes de déplacements comme des visites ponctuelles, et révéler de informations relatives à la santé, la religion, la scolarité des enfants, ou aux habitudes alimentaires d'un individu. Ces dangers existent quand on considère l'ensemble des contributions au projet OpenStreetMap dès lors que l'on souhaite regrouper toutes celles qui viennent d'un même contributeur.

1.3 Objectifs du stage

L'objet de ce stage est d'utiliser des méthodes d'anonymisation pour ré-organiser les liens entre contributeurs et contributions sur l'ensemble d'un territoire pour prévenir ces risques pour la vie privée, tout en préservant certaines propriétés garantissant que la base obtenue sera pertinente et pourra être utilisée pour la conception et l'étalonnage d'une méthode de qualification des contributions fondée sur l'apprentissage et la classification des contributeurs en profils. Par exemple, les contributeurs d'OpenStreetMap suivent une configuration classique avec 90% de contributeurs inactifs, 9% de contributeurs peu actifs et 1% de contributeurs très actifs. Nous souhaitons que les données anonymisées suivent toujours ce schéma.

Deux méthodes pourront être envisagées pour parvenir à ce résultat. La première est la permutation d'un certain nombre de contributions entre les contributeurs, la deuxième est la suppression de toute trace des contributeurs réels et la création de contributeurs synthétiques auxquels seront affectés les contributions.

2 Informations pratiques

Profil: Nous cherchons un étudiant en informatique ayant un gout ou une curiosité pour les données géographiques et/ou les projets collaboratifs comme OpenStreetMap.

Niveau: Master 2 ou stage de fin d'étude d'école d'ingénieur.

Compétences: Développement informatique (idéalement Java), maitrise de l'anglais pour la lecture d'articles, base de données, modélisation.

Durée: 5 mois, commençant en mars ou avril 2017.

Lieu: Laboratoire COGIT, IGN, 73, avenue de Paris, Saint Mandé.

Contact:
Arnaud Le Guilcher : Arnaud.Le-Guilcher@ign.fr

Guillaume Touya : Guillaume.Touya@ign.fr
Envoyer CV et lettre de motivation.

Références:

[1] L.Sweeney, *k-anonymity: a model for protecting privacy*, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (2002), no. 5, 557–570.

[2] G. J. Matthews, O. Harel, it Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy, Statistics Surveys, 5 (2011), 1–29 (electronic)

[3] A. E. Cicek, M. E. Nergiz, Y. Saygin, *Ensuring location diversity in privacy-preserving spatio-temporal data publishing*, The VLDB Journal, 23 (2014), no. 4, 609–625.