Extraction de clés de liage de données (résumé étendu)

Jérôme Euzenat

INRIA & Univ. Grenoble Alpes
F-38000 Grenoble, France
Jerome.Euzenat@inria.fr
http://exmo.inria.fr/~euzenat/

Résumé. De grandes quantités de données sont publiées sur le web des données. Les lier consiste à identifier les mêmes ressources dans deux jeux de données permettant l'exploitation conjointe des données publiées. Mais l'extraction de liens n'est pas une tâche facile. Nous avons développé une approche qui extrait des clés de liage (link keys). Les clés de liage étendent la notion de clé de l'algèbre relationnelle à plusieurs sources de données. Elles sont fondées sur des ensembles de couples de propriétés identifiant les objets lorsqu'ils ont les mêmes valeurs, ou des valeurs communes, pour ces propriétés. On présentera une manière d'extraire automatiquement les clés de liage candidates à partir de données. Cette opération peut être exprimée dans l'analyse formelle de concepts. La qualité des clés candidates peut-être évaluée en fonction de la disponibilité (cas supervisé) ou non (cas non supervisé) d'un échantillon de liens. La pertinence et de la robustesse de telles clés seront illustrées sur un exemple réel.

1 Web des données

Le web des données est l'utilisation des technologies du web sémantique pour publier des données sur le web (Heath et Bizer, 2011). Les données sont publiées sous forme de graphe dans le langage RDF et le vocabulaire de ce graphe décrit dans une ontologie. De grandes quantités de données sont publiées de la sorte.

2 Lier les données sur le web

L'intérêt de RDF est de pouvoir lier les données provenant de différentes sources de manière à pouvoir les exploiter conjointement. Cela est souvent obtenu en connectant les ressources représentant la même entité à l'aide du prédicat `owl:sameAs`. Au regard des importantes quantités de données publiées sur le web, il est utile de pouvoir les lier automatiquement. Le *liage de données* (data interlinking) consiste à identifier les mêmes ressources dans deux jeux de données. Cette tâche est apparentée à des tâches connues de reconnaissance d'entités ou de déduplication.

Nous distinguons deux approches du liage de données :