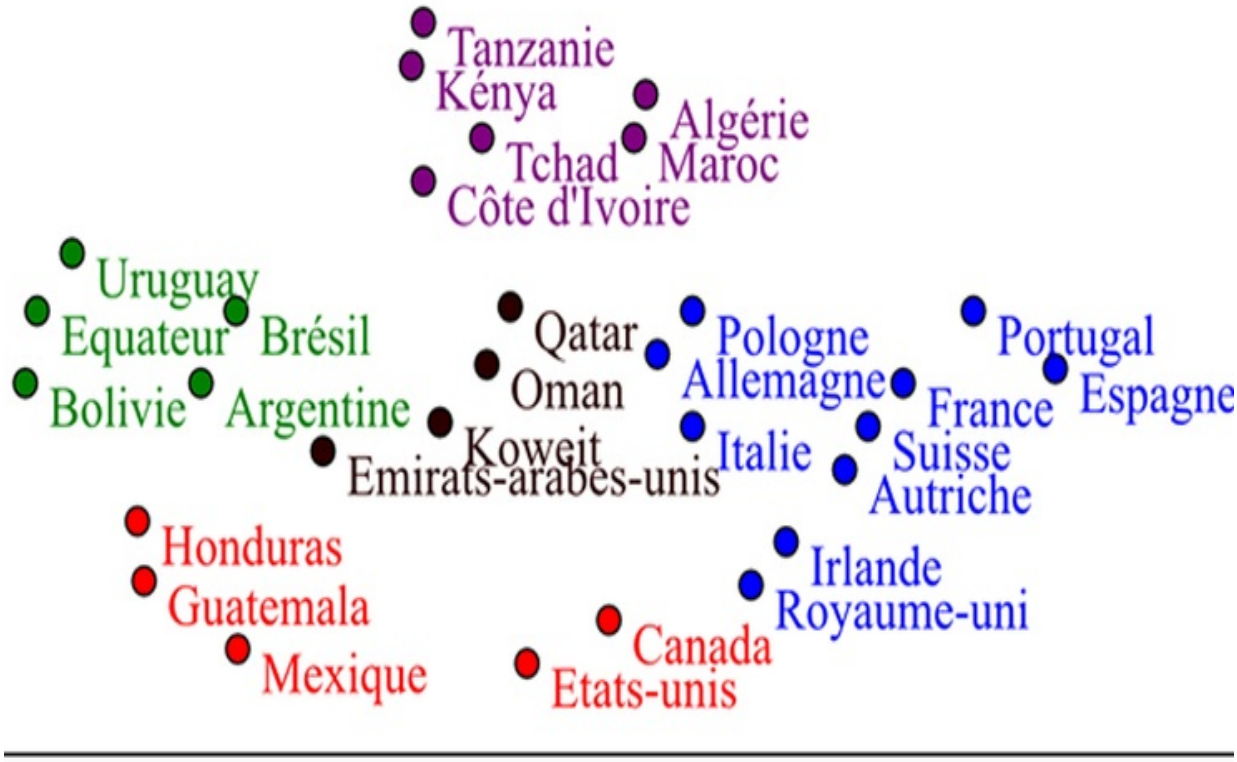
Analyse des bases de connaissances

Ces dernières années, les bases de connaissances, telles que Wikipedia, Allociné, ou encore les réseaux sociaux comme Facebook, se sont multipliées sur Internet. Enrichies par les internautes ou créées automatiquement par des ordinateurs, elles deviennent de plus en plus importantes, et contiennent de plus en plus d'informations. Problème : leur taille les rend extrêmement difficiles à manipuler et à étudier. L'analyse de ces bases de données est donc un domaine d'étude très porteur à l'heure actuelle.

01 Apr 2013



Une base de données est un graphe où chaque nœud représente un concept et chaque lien, une relation particulière liant un concept à un autre. Les liens sont donc de différents types. Par exemple, les réseaux sociaux sont des bases de données où les personnes connectées sont des nœuds, et les liens reliant ces nœuds sont les liens entre les utilisateurs. Les bases de données actuelles, comme les réseaux sociaux ou les bases de connaissances (Allociné, Imdb...), comportent des millions de nœuds et souvent une centaine, voire plus, de types de relations différentes. La plupart des bases sont collaboratives, c'est-à-dire qu'elles sont complétées et enrichies par les internautes eux-mêmes. Elles comportent de ce fait souvent des doublons et des erreurs. D'autres bases de connaissances sont créées automatiquement par des robots en collectant des infos à partir du web, et par conséquent, elles comportent aussi des erreurs.

Afin de pouvoir pleinement exploiter ces bases il est donc nécessaire de repérer et de corriger ces erreurs. Mais, selon Antoine Bordes, *" ces bases sont d'une taille tellement grande que ce travail ne peut pas être effectué par des humains. Il est donc nécessaire de concevoir un système pour les gérer, un logiciel qui,*

en trouvant les régularités sous-jacentes aux données, permettrait d'extraire les données qui ne correspondent pas à ces régularités, le plus souvent des données erronées, pour qu'elles soient réétudiées. "

Simplifier et résumer les bases de données

L'objectif du projet ANR, qui a débuté en janvier 2013, piloté par Antoine Bordes, et qui va durer quatre ans, est justement de *" rendre ces bases plus lisibles et plus simples en les résumant "*. Pour cela l'équipe du laboratoire Heudiasyc qui va travailler sur ce projet va projeter ces bases de connaissances dans un espace vectoriel, afin de pouvoir modéliser les liens entre chaque nœud par des probabilités.

Ces probabilités vont permettre d'établir des distances entre les nœuds et donc des similarités entre certains. Le but est de regrouper les millions de nœuds en catégories qui englobent plus ou moins de données. Modéliser la base de données permet d'en saisir les régularités, c'est-à-dire les groupes d'entités exprimant des choses similaires ou des liens exprimant des choses similaires.

Cet espace vectoriel peut ensuite être projeté en deux dimensions afin de visualiser les catégories. Ainsi, explique Antoine Bordes, *" en appliquant ces calculs à la base de données Wordnet, où chaque nœud représente un groupe de synonymes (ainsi le mot manche, qui peut avoir plusieurs sens, sera représenté par plusieurs nœuds), et chaque lien les relations lexicales qui les relient (ainsi, une manche est une partie d'un pull), il est possible, en partant d'un mot, de voir les mots qui en sont les plus proches "*. L'algorithme a permis de regrouper les pays européens, mais également de voir quels sont les pays proches de l'Europe, mais qui n'en font pas partie.

Des applications variées

Cet algorithme permet donc de suggérer des liens manquants dans la base de connaissances, par probabilités. Il serait donc particulièrement intéressant pour suggérer de nouveaux liens dans les réseaux sociaux, par exemple.

Mais il pourrait également être utilisé en génétique, sur des bases de connaissances de protéines et de gènes afin de suggérer des interactions possibles entre un gène et une protéine même si, comme le signale Antoine Bordes, *" cela ne remplacera jamais la recherche classique. Mais cela peut suggérer de nouvelles pistes de recherche. "*

Fusionner les bases existantes

Un objectif à plus long terme de ce projet serait de fusionner plusieurs bases de connaissances complémentaires, en évitant de créer des doublons, dus à des encodages de liens différents. Par exemple, pour fusionner deux bases de connaissances traitant de cinéma, si l'une possède le lien "tel acteur a joué dans tel film" et l'autre "tel film a untel pour acteur", cela va entraîner des doublons, si elles sont fusionnées. Mais l'algorithme permettrait de repérer ces doublons puis de les éliminer.

Une fois fusionnées et débarrassées de leurs erreurs, les bases de connaissances pourraient fournir beaucoup plus d'informations et de bien meilleure qualité.