Application de mise en correspondance d'entités avec pré-sélection par Map Reduce

Contact : Nicolas Lumineau (nicolas.lumineau@univ-lyon1.fr) et
Fabien Duchateau (fabien.duchateau@univ-lyon1.fr)

Type de sujet : recherche

Ce TER se place dans le cadre de travaux de recherche effectués au sein de l'équipe BD du LIRIS sur l'intégration de données à large échelle. Plus précisément, des pairs (e.g., des bibliothèques, musées, des sources de données du Linked Data) possèdent des collections d'entités. Pour enrichir ces collections, il est d'abord nécessaire de détecter quelles sont les entités équivalentes entre ces collections. Etant donné le volume important d'entités dans chaque collection, il n'est pas possible de comparer chaque entité deux à deux. Une étape dite de « blocking » consiste à sélectionner les entités qui seront comparées avec des méthodes plus sophistiquées. Nous proposons d'étudier la faisabilité d'une étape de « blocking » avec le paradigme « Map Reduce ». Les objectifs de ce TER sont les suivants :

- lire plusieurs articles sur « entity matching » et le « Map Reduce » afin de comprendre la mise en correspondance d'entités en utilisant un paradigme à la Map Reduce;
- de développer un prototype simple permettant de prendre deux collections d'entités et de sélectionner les entités à « matcher » par un processus « à la Map Reduce ».

Pré-requis : MIF18

Bibliographie :

Load Balancing for MapReduce-based Entity Resolution.
Lars Kolb, Andreas Thor, and Erhard Rahm. 2012. In *Proceedings of the 2012 IEEE 28th International Conference on Data Engineering* (ICDE '12)
http://arxiv.org/pdf/1108.1631

Pairwise document similarity in large collections with MapReduce.
Tamer Elsayed, Jimmy Lin, and Douglas W. Oard. 2008. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*
www.csee.ogi.edu/~zak/cs506-pslc/docsim.pdf

Linked Data
http://linkeddata.org