TER - Coûts et gains du polymorphisme des données dans un contexte de données textuelles

Pegdwendé Nicolas Sawadogo
pegdwende.sawadogo@univ-lyon2.fr

October 2, 2019

1 Contexte et objectifs

A l'ère de l'intelligence artificielle et des *big data*, les organisations exploitent une grande diversité d'algorithmes pour analyser automatiquement les données dont-elles disposent. Pour prendre en charge ces analyses, les données doivent préalablement être transformées dans un format ou un autre, selon le besoin d'analyse.

Dans le but d'accélérer les analyses, certains auteurs proposent d'anticiper le pré-traitement des données en générant systématiquement des représentations nettoyées et transformées des données brutes [Laskowski, 2016, Leclercq and Savonnet, 2018]. Cela permet ainsi de réduire le temps nécessaire aux futures analyses en partant de données prêtes à l'emploi. On parle de polymorphisme des données [Sawadogo et al., 2019].

Toutefois, il convient de noter que le gain de temps offert par cette approche est au prix d'un stockage plus coûteux, en ce sens qu'il implique un stockage multiple des mêmes données. Nous proposons donc dans ce TER d'étudier et de chiffrer de façon expérimentale les gains et les compromis qu'impliquent l'application du polymorphisme des données (en termes d'espace de stockage et de temps de traitements). Nous nous focalisons principalement sur le cas des données textuelles.

2 Étapes de réalisation

Nous proposons dans un premier temps de procéder de façon itérative pour identifier pour plusieurs types d'analyses (extraction de thématiques, extraction de mots clés, etc.) les coûts et gains chiffrés induits par le procédé basé sur la pré-transformation des données. Dans un second temps (si le temps le permet), vous pourrez proposer des stratégies de compression des données transformées, qui permettraient de réduire le coût de stockage des représentations de données.

Les traitements pourront se faire en Python et le stockage des données via MongoDB ou PostgreSQL.

References

[Laskowski, 2016] Laskowski, N. (2016). Data lake governance: A big data do or die. https://searchcio.techtarget.com/feature/Data-lake-governance-A-big-data-do-or-die.

[Leclercq and Savonnet, 2018] Leclercq, E. and Savonnet, M. (2018). A Tensor Based Data Model for Polystore: An Application to Social Networks Data. In *Proceedings of the 22nd International Database Engineering & Applications Symposium (IDEAS 2018), Villa San Giovanni, Italy*, pages 110–118.

[Sawadogo et al., 2019] Sawadogo, P. N., Scholly, E., Favre, C., Ferey, É., Loudcher, S., and Darmont, J. (2019). Metadata Systems for Data Lakes: Models and Features. In *BI Big Data Applications - ADBIS 2019 Short Papers and Workshop, Bled, Slovenia.*