China Scholarship Council / Université de Lyon Scholarships for doctoral mobility

Call for Thesis subjects for 2019/2020

RESEARCH SUBJECT TITLE:
Automatic metadata extraction from unstructured data lakes

Name of the laboratory: ERIC EA 3083
Website: https://eric.msh-lse.fr/en/

Name of the research team: Decision Information Systems (DIS)
Website: https://eric.msh-lse.fr/en/recherche/equipe-sid/

Name of the supervisor: Jérôme Darmont et Sabine Loudcher
University / Institution: Université Lumière Lyon 2
E-mail adresse: jerome.darmont@univ-lyon2.fr, sabine.loudcher@univ-lyon2.fr

Doctoral School: Infomaths ED 512

Lab Language: English, French

Minimum language level required:

- English: C1
- French: B2
- Other:

Abstract:

Data lakes, a term coined by Dixon (2010), were born with big data with the aim of storing voluminous, varied and diversely structured data in their native format, for the sake of various analyses (reporting, dataviz, machine learning…). However, in the absence of predefined data schemas, an efficient metadata system is required to allow querying the

data and prevent the lake to become an unexploitable so-called data swamp (Hai et al., 2016).

Up to now, research dedicated to design metadata systems for data lakes essentially focused on structured and semi-structured data (Halevy et al., 2016 ; Quix et al., 2016 ; Beheshti et al., 2017 ; Maccioni & Torlone, 2017). Very few address unstructured data such as textual documents, images, sounds and videos, while they represent the majority of big data (Miloslavskaya & Tolstoy, 2016).

The aim of this PhD thesis is thus to design a process allowing to automatically extract metadata from unstructured data, particularly with the help of machine learning methods, and to contribute to the design of a metadata system allowing to jointly querying structured, semi-structured and unstructured data from a data lake.

Beheshti et al. (2017). CoreDB: a Data Lake Service. *2017 ACM on Conference on Information and Knowledge Management (CIKM 2017)*, Singapore. 2451-2454.

Dixon J. (2010). Pentaho, Hadoop, and Data Lakes. James Dixon's Blog. https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/.

Hai et al. (2016). Constance: An Intelligent Data Lake System. *2016 International Conference on Management of Data (SIGMOD 2016)*, San Francisco, USA. 2097-2100.

Halevy et al. (2016). Managing Google's data lake: an overview of the GOODS system. *2016 International Conference on Management of Data (SIGMOD 2016)*, San Francisco, USA. 795-806.

Maccioni & Torlone (2017). Crossing the finish line faster when paddling the data lake with KAYAK. *Proceedings of the VLDB Endowment*, vol. 10, no. 12. 1853-1856.

Miloslavskaya & Tolstoy (2016). Big Data, Fast Data and Data Lake Concepts. *7th Annual International Conference on Biologically Inspired Cognitive Architectures (BICA 2016)*, New-York, USA. 1-6.

Quix et al. (2016). Metadata Extraction and Management in Data Lakes With GEMMS. *Complex Systems Informatics and Modeling Quarterly*, no. 9. 289-293.

Expected duration of the thesis: 36 months

Keywords: Data lakes, Unstructured data, Metadata, Machine learning