A Framework for the Continuous Creation of a Knowledge Base System

Mots clés :
- Directeur de thèse : Paolo PAPOTTI
- Co-encadrant(s) :
- Unité de recherche : Laboratoire de recherche d'EURECOM
- École doctorale : École Doctorale Informatique, Télécommunications, Électronique de Paris
- Domaine scientifique principal: Sciences et technologies de l'information et de la communication

Résumé du projet de recherche (Langue 1)

A Knowledge Base System (KBS) is a large-scale integration of information built on two main components. The first is the knowledge base (KB), which is the repository of data. It is composed of triples in the form of (subject, predicate, object), where the entities (e.g., The Mona Lisa painting and Leonardo Da Vinci) are the subject/object and their relationship is an instance of a predicate (e.g., createdBy). The second component is the language used to express knowledge representation. This is made of a logical formalism for expressing facts and rules, and a reasoning engine that uses it. By understanding the data semantics, KBSs enable analytics such as querying with structured languages (e.g. SQL/SPARQL), and a wide array of learning and reasoning tasks. However, creating and maintaining a KBS is difficult. Domain specialists guarantee quality, but, because of the large and increasing number of documents, cannot do it manually. Automatic algorithms scale over large number of documents, but cannot provide high quality in the results, because of process complexity. Creating and curating high quality KBS cannot be done manually at scale and new automatic methods designed around the users need to be developed. We envision new hybrid systems that use the machine to scale over large datasets, and the human experts to teach the algorithms how to solve the hard cases. This vision involves challenges from multiple fields, including data integration, information extraction, data mining, natural language understanding, and probabilistic reasoning.