Proposition de recherche doctorale

Quality of Data Interlinking

Mots clés :
- Directeur de thèse : samira SI-SAID CHERFI
- Co-encadrant(s) :
- Unité de recherche : Centre d'Etude et de Recherche en Informatique et Communications
- Ecole doctorale : Ecole Doctorale Informatique, Télécommunications, Électronique de Paris
- Domaine scientifique principal : Divers

Résumé du projet de recherche (Langue 1)

Many organizations and individuals publish a variety of data on the web to enhance transparency. The Open Data initiative that consists in sharing non-sensitive information in a raw and machine-readable format aims to stimulating innovation and creating new business opportunities [1]. A plethora of sources are now available but that suffer from heterogeneity of contents, structures and formats as well as unequal quality. The linking of all kinds of open data sets provides potentially many benefits. It could at least help to infer new data and possibly meaningful knowledge. Consequently, such data users require linked data to meet high quality standards in order to develop applications that can produce trustworthy results. Unfortunately, the LOD (Linked Open Data) does not necessarily meet the expected quality. The process of publication and linkage of resources in the LOD walks through i) data cleaning and transformation into existing RDF formats, ii) storage of the data into RDF storage systems, and iii) data interlinking. The challenge, from the quality viewpoint is how to insure the quality of the resulting datasets. There has been an important research effort on data interlinking. These efforts have produced integrated datasets from several domains such as DBpedia or Yago and a variety of methods and frameworks to support interlinking such as SILK [2], RuleMiner [3] or KnoFuss [4]. However, even if the interlinking approaches ensure or try to ensure the quality of produced links, the continuous evolution and growth of data hampers the preservation of the initial quality. More and more researchers and business actors rely on multiscource data analysis to try to understand or explain observed phenomena. The accuracy of the deduced conclusions relies, however, on the quality of interlinkage between the sources. The intended research work will follow the underlying agenda: - The identification of quality defects related to data interlinking: Many of the links established between data sets lack quality. In [5] authors investigate the incorrect use of owl:sameAs links. They pointed out the fact that the correctness of such links is not insured. Many other situations could happen such as incompleteness of links or incoherence. The challenge is to investigate several interlinked data sources from several domains and try to qualify the underlying quality defects. The difficulty of such work is the definition of efficient algorithms for data exploration and defects detection. The subsumed defects need to be confirmed by human intervention. Once quality criteria identified, it is necessary to associate to each criterion a set of assessment methods and algorithms; - The detection of quality defaults raises the problem of correction that is not sufficiently addressed in literature where quality stands more for evaluation than correction; - The proposed solutions should be implemented through a prototype or a platform to support Data Interlinking evaluation and improvement; - Finally, the problem of scalability should be addressed as Open Data is also Big Data and developed solutions should be scalable.

Résumé du projet de recherche (Langue 2)

- The identification of quality defects related to data interlinking: Many of the links established between data sets lack quality. In [5] authors investigate the incorrect use of owl:sameAs links. They pointed out the fact that the correctness of such links is not insured. Many other situations could happen such as incompleteness of links or incoherence. The challenge is to investigate several interlinked data sources from several domains and try to qualify the underlying quality defects. The difficulty of such work is the definition of efficient algorithms for data exploration and defects detection. The subsumed defects need to be confirmed by human intervention. Once quality criteria identified, it is necessary to associate to each criterion a set of assessment methods and algorithms; - The detection of quality defaults raises the problem of correction that is not sufficiently addressed in literature where quality stands more for evaluation than correction; - The proposed solutions should be implemented through a prototype or a platform to support Data Interlinking evaluation and improvement; - Finally, the problem of scalability should be addressed as Open Data is also Big Data and developed solutions should be scalable.