Data Fusion for Data Quality Assessment

Research Context

Recently, with the initiative of « Linked Open Data cloud (LOD) », the number of sources of structured data made available on the Web has lead to an explosive growth of the global data space with billions of assertions (61 billions in January 2014). In this data space, semantic links can be established between data. These links allow crawlers, browsers or applications to navigate through the data sources and combine information from different sources. However, in an open environment like the Web, different URIs are regularly created to identify the same object. Generating identity links (owl:sameAs) between resources is crucial to allow applications to exploit the richness of the LOD. To thereby, these application should be able to fuse the resources linked using owl:sameAs links in order to obtain a unified representation. This is the data fusion problem arising after data linking problem. It is this data fusion problem that interests us in this PhD project.

This PhD project is part of the ANR project Qualinca (www.lirmm.fr/qualinca) aiming at the improvements, the evaluation and the represention of data quality of bibliographical knowledge bases managed by ABES (Agence Bibliographique de l'Enseignement Supérieur) and Ina (Institut National de l'Audiovisuel). This PhD project also will seek to define a generic approach for which it will rely on case studies of very different fields, such as agrifood data sources.

The data fusion problem

When we study the problem of data fusion, the main difficulty concerns the conflicts in property values, that is, several possible values for the same property. These conflicts are mainly due to the heterogeneity of the data where different vocabularies and conventions are used to describe the data. Poor data quality (freshness, errors and incomplete information) may contribute to the amplification of conflicts between values.

This aim of this PhD project is to develop a data fusion approach where: (i) it is possible to choose and combine several criteria (e.g, freshness, frequency, reliability of sources, functional and semantic dependencies) to choose the right values (ii) the schema constraints should be checked in the merged data and (iii) provenance information on the original data sources but also mappings that are applied during schema integration of the data sources should be exploited. In addition, the richness of Web of Data will be used by navigating the graph of owl: sameAs links, known to have more precise information (e.g., dates of birth on dbpedia) and sometimes more reliable information (e.g., geographic coordinates on geonames).

Objectives:

Study the theoretical and practical aspects of data fusion approach in order to: (i) avoid data redundancy and (ii) summarize information from a point of view.

Work directions:

1. A first work direction concerns the annotation of linked data. Indeed, once the linking achieved, an essential element for data quality is the interpretability and the explanation of data

fusion result. It is important to be able to identify and keep track of the criteria that are used to choose the "best value" or to rank the possible values of a property during the fusion step. It is possible to achieve such a complex annotation for storing: excluded values, synonymy relations, specialization/generalization relations, mereology relations, and so on. For each criteria, a quality score is computed. These scores should also be kept and represented in the annotations.

2. A second direction of work concerns the case where several data can be grouped into the goal of creating a more generic entity with common characteristics. A case characteristic concerns the concept of "work" in the bibliographical data sources. This concept could be embodied in various "events" (instances of the work in different collections, editions, translations, etc.) that are listed in the data source. This is the case, without materializing the notion of work in the data sources. This fusion approach within a data summary objective needs to be developed.

State of the art

Some approaches for data fusion exist in the literature (see [J. Bleiholder and F. Naumann, 2008] for a state of the art), developed in the domain of relational databases where conflicts are handled at the querying step using predefined operators, such as Max, Min, the value of the most reliable source, the most recent value, etc. More specifically, in [F. Saïs et R. Thomopoulos 2008] we developed a fusion method for RDF data where all the possible values of a property are conserved and stored in an ordered way according to a confidence degree associated with each value and computed using several criteria such as data source reliability, value frequency, value age, etc. The fuzzy set formalism was used to represent the fused data. The limits of this approach are twofold : the importance level is the same for the all the criteria used to choose the values. Moreover, neither schema constraints (e.g., the cardinalities of properties) nor mappings - which are possibly complex - between elements of the schema (e.g., [s1.address = (s2.street, s2.city, s2.zip-code, s2.country)]) are taken into account.

Expected results

The thesis will produce theoretical as well as methodological developments that will have to be validated on benchmarks and applied to 2 domains: on the one hand, documentary databases using the INA and ABES datasets, on the other hand, agrifood data using IATE datasets.

Valorization

From an academic point of view, data integration as well as data quality assessment are both growing research domains, in relation with the works pursued with other international teams. Their application to the documentary and agronomical fields, together with their genericity, will also allow their diffusion in several application areas. The deep insertion of the project partners in the concerned communities, and the participation in several international conferences, will ensure the visibility of the work produced.
A particular effort will focus on the achievement of open-source software, so as to permit the methodological results of the project to be reused and to ensure an efficient diffusion of the results within the international communities in computer science, documentation and agronomy.

Partnership

- Computer Science Research Laboratory (LRI), Orsay

- IATE Joint Research Unit (Ingénierie des Agropolymères et Technologies Emergentes), Montpellier
- INRIA Project Team « GraphIK », Montpellier
- Institut National de l'Audiovisuel (INA), Paris
- Agence Bibliographique de l'Enseignement Supérieur (ABES), Montpellier

Contacts

Fatiha.Sais@lri.fr
Rallou.Thomopoulos@supagro.inra.fr

References

[J. Bleiholder and F. Naumann, 2008] Jens Bleiholder and Felix Naumann. Data fusion. ACM Comput. Surv. 41(1), December 2008, 41 pages

[F. Saïs and R. Thomopoulos 2008] Fatiha Saïs, Rallou Thomopoulos. Reference Fusion and Flexible Querying. In Proceedings of OTM Conferences 2008, Mexico, November 2008, Lecture Notes in Computer Science #5332, Springer, pp. 1541-1549

[P. N. Mendes et al. 2012] Pablo N. Mendes, Hannes Mühleisen, Christian Bizer: Sieve: linked data quality assessment and fusion. EDBT/ICDT Workshops 2012: 116-123

[V. Bryl and C. Bizer 2014 Volha Bryl and Christian Bizer. Learning Conflict Resolution Strategies for Cross-Language Wikipedia Data Fusion. In WebQuality 2014 workshop of WWW (Companion Volume) 2014: 1129-1134