Detection, Integration and Correction Methods for Homologous Geographic Objects

Bilal Berjawi, Franck Favetta, Fabien Duchateau, Maryvonne Miquel, Robert Laurini

LIRIS CNRS UMR 5205, INSA de Lyon, Université de Lyon, F-69621 Villeurbanne, France
LABEX IMU ANR-10-LABX-0088/ ANR-11-IDEX-0007, Université de Lyon
{prenom.nom@liris.cnrs.fr}

In the last few years, new technologies have emerged, especially in the field of telecommunications, digital systems and devices. This development allowed companies to offer new services, such as *Cloud Computing* or *Location-Based Services*. The visualization of these services leads to many new issues and research topics. In the market there are several providers for Location-Based Services, and each of them uses its own concepts, models and data. As a consequence, customers obtain different answers from one provider or from another [RK11]. In this context, the UNIMAP project aims at providing Location-Based Services in which data from various suppliers are aggregated. This ensures the completeness and the correction of the results for a given query. However, the main research question is to deal with the detection and the integration of homologous objects from the databases of these suppliers.

Location-Based Services allow people to find nearby places from different categories (Hotel, Restaurant, etc.), based on their geographic location (latitude/longitude or a simple street address) within a given distance. For example, a customer may request the nearby restaurants which offer Lebanese-style food within 500 meters of her current location that is detected by a GPS device. Places which have been found using Location-Based Services' requests are called Points of Interest (POIs).

Google Maps, OpenStreetMap, Microsoft Bing Maps, Nokia Here Maps, MapQuest, Rhône-Alpes Tourist Office SITRA and many others are offering Location-Based Services. Some of them are considered as Leader providers, because they are not offering this service only to the end user, but also they share their data using web services with other companies.

Our state of the art has identified three main methods to detect, integrate and correct homologous objects:

1. **Practitioners method:** It can be divided into two models: 1) Geographic experts visit places on the ground in order to check and validate POIs positioning and attributes. *Michelin, Google,* or *NAVTEQ* promote this model. 2) The second model is specific to tourist companies which propose to all POIs in their area to be added into the database, usually by charging a fee. As *Tourist Offices, Hotel.com, etc.* do it. The accuracy of the data will depend directly on the contributions of members by updating regularly their information.

   This method should produce accurate data, and it is very efficient for a limited region. Otherwise, we can distinct two drawbacks: how to cover a large area? How to guarantee an efficient data updating process?

2. **Databases integration:** Some providers use different data sources (*Yellow Pages*, *Tourist Offices*, etc.) which might have different structures, in order to construct their own database of POIs.

   The heterogeneity between the data sources can be found both at the schema level (e.g. different structure, synonymy) and at the data level (e.g. different values).

   Many works have been proposed in schema matching [BBR11] and ontology matching [ES07]. For instance, COMA++ [EM07] combines various similarity measures to select the correspondences between the schemata. Cupid [MBR01] uses machine learning techniques to discover these correspondences. At the data level, entity matching (or record linkage) is a common process to remove duplicates in databases [KR10].

3. **Collaborative method:** In this proposition, data are opened for everyone. The main source of POIs is users' contributions. People from everywhere may add/edit/delete POIs using many technologies such as web user interface or API. The main benefit of this method is the unlimited contributions because data are offered from people to people. In the other hand, many of non-expert users may participate and this can imply inaccuracies in POIs. The *OpenStreetMap* provider uses this method.

It may happen that some providers use multiple methods at the same time.

In the UNIMAP project, all providers which share their data are considered as data sources suppliers. Since they use ICTs to share POIs, in a dynamic (frequent updates) and large scale environment (many POIs for a given area), the second method is recommended to solve the detection issues, and regarding integration/correction issues the third method seems appropriate to guarantee the quality and to dynamically enrich and improve a POI-based cartography. Result quality may be increased by combining with other methods.

References

[RK11] Roula Karam. Multi-Providers Location Based Services for Mobile-Tourism: a Use Case for Location and Cartographic Integrations on Mobile Devices PhD thesis. Liris-5272, INSA de Lyon, September 2011.

[BBR11] Zohra Bellahsene and Angela Bonifati and Erhard Rahm. Schema Matching and Mapping. In Springer 2011.

[ES07] Jérôme Euzenat and Pavel Shvaiko. Ontology matching. In Springer 2007.

[ADMR05] David Aumueller and Hong Hai Do and Sabine Massmann and Erhard Rahm. Schema and ontology matching with COMA++. In SIGMOD 2005.

[MBR01] Jayant Madhavan and Philip A. Bernstein and Erhard Rahm. Generic Schema Matching with Cupid. In VLDB 2001.

[KR10] Hanna Köpcke, Erhard Rahm. Frameworks for entity matching: A comparison. In Data knowledge Engineering 2010.