Improving Geo-spatial Linked Data with the Wisdom of the Crowds

Roula Karam
Dipartimento di Elettronica e Informazione
Politecnico di Milano
Milano, Italy
karam@elet.polimi.it

Michele Melchiori
Dipartimento di Ingegneria dell’Informazione
Università degli Studi di Brescia
Brescia, Italy
melchior@ing.unibs.it

ABSTRACT
Currently, there is more and more interest in geo-spatial data sources providing rich information about a huge number of interconnected geo-entities and points of interest located in the real world. Moreover, this kind of sources is one of the first to be published as linked open data. Noteworthy examples are the Geonames and GeoLinkedData initiatives. On the one hand, making available more data sources as linked open data allows querying the sources in an integrated way. On the other hand, it is known that content of geo-spatial data sources suffers from various drawbacks, mainly concerning data quality and conflicts. In this context, relevant feedbacks from users with specific experience and knowledge about POIs in a certain spatial region are considered valuable contributions to improve data quality and solve description conflicts. In this context, we propose a conceptual framework called M-PREGeD (Multi-Providers cRowd-Enhanced Geo linked Data) devoted to collect, organize and rank user-generated corrections and completions to improve accuracy and completeness of Geo-spatial Linked Data from different data sources. Metrics have been defined for both contributors and contents. In the framework, validated and ranked corrections and completions are stored as linked open data in a separate repository but linked to the original data sources. The repository can be queried in a combined way with the original data sources.

Categories and Subject Descriptors
Information Systems [Geographic information systems, World Wide Web]: Web searching and information discovery, Web applications, Crowdsourcing

Keywords
Geospatial Web, Linked Data, Location based Applications, Model-driven approach, Human computation, Crowdsourcing

1. INTRODUCTION
The current popularity of Location Based Services (LBS) on mobile devices is creating a growing interest for Web geo-spatial data sources. These sources typically own very large datasets of geo-spatial data and points of interests (POIs). Well known examples are Foursquare and Google Places that are specialized in describing and searching POIs. Geonames is a geographical database that defines an ontology of geographical concepts and relationships. Another noteworthy initiative in this domain is OpenStreetMap ¹, a community-based project to build and maintain a detailed map of the world. Moreover, some geo-spatial data sources have recently started to publish their content as linked open data (LOD). For example, the LinkedGeoData initiative uses the information collected by OpenStreetMap to make it available as RDF knowledge base according to the linked data (LD) principles. Geonames provides a RDF dump of its content including more than 8.3 million Geonames toponyms and about 156 millions of RDF triples. In the Spanish initiative GeoLinkedData, geo-spatial data produced by the public administration and concerning the Spanish territory have been transformed and published as LOD. Publishing geographical data as LD brings relevant advantages in terms of accessibility and interoperability of the sources, such as:

- Making datasets accessible by non-proprietary languages and tools (e.g., SPARQL, GeoSPARQL ², RDF Browsers, Geo LD Browser [2]);
- Introducing formal semantics to make data machine processable;
- Enriching data with links to external resources (e.g., DBpedia) to set cooperating relationships among data sources;
- Allowing to query/browse multiple data sources in a combined way.

However, geo-spatial data may suffer from two known drawbacks: variable quality and description conflicts. The first one concerns updating, completeness and accuracy of the data. The second one concerns inconsistent descriptions provided by different sources for the same POI. These problems are especially relevant and frequent in geographic data dealing with urban environments. In fact, these are featured by a high pace of changing and development; every day new POIs can appear or some of their features can

¹http://www.openstreetmap.org
²http://www.opengeospatial.org/standards/geosparql
change. With reference to the quality issue, geographic data providers as Geonames and OpenStreetMap make available simple models for improving data by users’ feedbacks. For example, moderator-based approval of data corrections and insertion of new POIs as proposed by users. Concerning description conflicts, techniques have been studied to estimate when conflictual descriptions are referring to the same POI and how to unify them [14, 19]. However, these automatic techniques are not completely effective (usually they cannot solve 20-30% of the conflicts) and cannot be helpful if the information is incorrect or missing in all the sources. Therefore, feedbacks and contributions from users that have a specific knowledge or physical presence in the requested environment are considered valuable contributions to increase data quality. In general, we can say that putting the user in the life cycle of geographic data is fundamental for describing the urban environments.

In this context and motivated by the above considerations, we propose a provider independent conceptual framework, called M-PREGeD (Multi-Providers cRowd-Enhanced Geo linked Data), to collect, organize and rank user-generated corrections and completions to improve accuracy and completeness of geo-spatial data in the context of LD. Within such framework, the user can play different roles as contributor. Metrics have been defined for both contributors and user-generated content. Moreover, validated and ranked user-provided corrections and completions are stored as LOD. A repository for such data input is maintained separately from the original data sources but linked to them. The repository therefore can be queried together with the original data sources.

The paper outline is the following. In Section 2, we analyze the problem of conflicts introduced by a motivating example. In Section 3, some relevant literature and approaches about conflicts in Geo data, data quality and users contribution are discussed; Section 4 presents the requirements and the main features of the M-PREGeD framework and its data model; Section 5 describes the metrics used to evaluate users and their activities as contributors. Finally, in Section 6 we give concluding remarks and future research directions.

2. PROBLEM STATEMENT

The problems of data quality and specifically of conflicts in description may concern different aspects of a POI or a geo entity description. Let us consider a user’s request to find the nearest restaurant to her current position, submitted to two different geo data providers. The answers obtained by the providers may present various conflicts, as depicted in Figure 1.

First of all, both the answers describe an Italian restaurant but they provide slightly different positions for it on the maps (e.g., with a difference of about 100 meters). The restaurant is named “Carlo’s Pizza” in the first answer and “Da Carlo Trattoria” in the second one, and there are differences in the details of the descriptions. The restaurant is represented with two different cartographic symbols (icons) on different base-maps. Different levels of heterogeneities (both in the underlying data and in the representation), can be identified:

- Syntax level. The data is extracted from different sources that adopt different representation formats (e.g., ESRI Files, MapInfo files, Spatial Oracle DB, PostGIS, GRASS files, etc.). These differences may have an impact on how the data is presented to the user;
- Structure level. The geographic characteristics are visualized using different geometries or basic forms. Very often, the same entity is represented by different icons or geometric entities (e.g., roads can be represented either as lines or polylines);
- Semantic level. It concerns heterogeneities in data schemes or instances concerning class names and structures, attribute names and values. For example, name conflicts are present when values of attributes of two different representations are homonyms (e.g., Congo river and Congo country) or when values of attributes representing the same geo entity are named differently (synonymy). The latter problem is depicted in our example where the same restaurant is named differently due to lack of data updates in one or both the providers;
- Cartographic level. The considered geo maps are developed independently so there are diverse map formats and legends.

Another type of heterogeneity occurs when different providers specify different positions for the same POI, as seen in the example. Usually, these differences are related to the used positioning techniques that can have different levels of precision. As a consequence, this may lead to imprecise results for the geocoding functions. Differences in place names are often related to spatial databases that lack of frequent updating. Details about POIs can be different among providers due to lack of update, as well as common agreements about representation rules for email addresses, websites/URL, facilities, etc.
3. RELATED WORK

Literature and approaches about conflict management in data integration and managing user contributions to improve quality of data are relevant to our discussion.

Conflict management in Geo data. Data integration of (non spatial) data sources in presence of semantic heterogeneity has been largely studied in literature both for the traditional [3, 20] and Web [5, 6] data sources. Inspired by these approaches and by the works of Fonseca et al. [10], Kavouras et al. [15], Laurini [16] and by the wide literature in the domain of GIS and LBS interoperability (e.g., [12]), we proposed a general solution to integrate spatial data (more precisely POIs) in presence of representation conflicts [13]. More in general, the detection/resolution of conflicts is an important aspect in the problem of data integration. Different types of conflicts at structural, constraint, schematic or semantic levels (naming, scaling and confounding conflicts) has been discussed in [22]. In the Semantic Web, conflicts usually arise in processes of knowledge merging and ontology matching, as mentioned in [18] where case-based resolution are proposed. The authors in [17] had classified different types of conflicts (conspicuous/ in-conspicuous) and present detection/resolution strategies to assist humans for image annotation. They describe the different situations that may lead to conflicts, such as:

1. Conversion phase: when an annotation conflict happens from the translation of users' representation to the internal representation of the system;
2. Merging phase: when two annotations (one from a human and another one from a software agent) for the same image have been merged thus can lead to a different result.
3. Consistency of annotations: two annotations are correct individually but when considered together this may lead to a wrong conclusion or to a situation that a system cannot understand.

Multiple annotators for the same POI or image can be also source of further conflicts even if their intervention was initially aimed to solve a pre-existing conflict. In these cases, strategies are used like majority voting, top-k ranking, reputation or confidence degree of each annotator and its level of expertise.

Improving data quality by involving the users. Generally speaking, data quality has serious implications for the users of every information system [25, 4]. Research on data quality is devoted to identify appropriate and relevant dimensions that characterize it in a given context. In fact, data accuracy is only one of the possible dimensions. Other significant ones are completeness, consistency and accessibility [25]. As emphasized in [7, 11] the LD practices give a contribution to the quality of the published data because: (i) the adoption of shared vocabularies makes explicit the semantics of the data and therefore increases the possibilities to understand and check it, and (ii) the presence of links makes it possible to check the consistency of data across different sources. In [9] an attempt is described to define a classification of data quality dimensions suitable and relevant for LD sources.

4. THE M-PREGeD FRAMEWORK

4.1 Requirements

The main idea of the M-PREGeD framework is to maintain a user-generated collection of data about geo linked data published by LOD sources. The collection includes corrections generated and validated by users, complementary information and conflict resolutions about POIs. From a conceptual perspective the data maintained by M-PREGeD introduces an additional linked corrections layer in the architecture of geo linked data, as shown in Figure 2. Two main motivations have been identified for the proposed framework: (i) introducing more sophisticated and provider-independent models for collecting and evaluating crowd-generated feedbacks than the ones currently offered by geo linked data providers, (ii) dealing with conflicts in the POI
descriptions published by different providers. The main requirements for the M-PREGeD framework keep into account that the data collection has to be complementary to the existing sources and compliant to the LD practices:

- The linked corrections layer is published as LOD with reference to the original geo data. For example, each POI description includes the original URI(s) that can be resolved for accessing the original POI description(s).

- The linked corrections layer constitutes a mediation layer built over the original sources and can be used in a joint way with them.

- Applications and users can browse the layer by using RDF data browser and querying it with the SPARQL, GeoSPARQL languages.

Corrections, complementary information and conflict resolutions stored in the M-PREGeD data model (discussed in the following Section 4.2) are generated and evaluated by the activities of users (voluntary and paid ones). The two kinds of users have different roles. A voluntary user is a user that has a direct knowledge of the POI or geo entity she is providing information about, and she acts under the motivation of sharing her knowledge and improve the quality of the datasets. A paid user performs human computation tasks on the data, specifically she verifies the consistency of information loaded by voluntary users. The human computation tasks can be performed through a game with a purpose (GWAP) engagement and rewarding model [1] or a crowdsourcing model where the user is rewarded monetarily, like in Amazon Mechanical Turk 3.

3https://www.mturk.com

Conflicts in POIs descriptions are detected by automatic procedures, as cited in Section 3, and in some cases can be solved automatically. Otherwise, they are submitted to users that are competent for the POI geographical area if the correction procedure is not able to produce a unified coherent description. This happens for example, when every provider describes a different position for a given POI in a given city and the distances among the positions exceed a given threshold of acceptability.

4.2 Data Model

The M-PREGeD data model is shown in Figure 3. This model describes data requirements for the M-PREGeD evaluation process. A User is specialized as VoluntaryUser and PaidUser. Both have a CompetencyArea that allows to establish competency about POIs. The competency that can also be declared directly by a user (KnowsAbout association). A POI has a URI and keeps reference to the original URI of the POI in the data source. A Description is a set of RDF triples forming the POI description. A VoluntaryUser provides a Completion or a Correction as RDF triple to complete/correct a POI Description. A Correction can also be introduced to correct another Correction considered incorrect by the user. The same kind of user has an important role in evaluating, through the Assesses associations, as positive or negative, a change (i.e., a Completion or a Correction) provided by another user. These evaluations are used to calculate an overall evaluation of the change and of the users, as explained in the Section 5. A change is also analyzed by a PaidUser that expressed its validity or not by performing a CompletionAnalysis or a CorrectionAnaly-
sis. Concerning a DetectedConflict that cannot be solved automatically, it is described by the pair of conflicting Descriptions and it is analyzed by at least a PaidUser that performs a ConflictAnalysis with the purpose to produce a Correction to the POI description.

4.3 Change submission process

The flow of activities performed to collect user’s provided changes are depicted in the process shown in Figure 4. Similarly, a process has been defined to describe the workflow of activities that a paid user performs for conflict analysis and resolution.

5. SCORING MODEL

In the M-PREGeD framework, users and users’ changes are evaluated by scores. A user score is stored in the userScore attribute of the User class (see Figure 3). Score calculation is based on: (i) the history of activities performed by the user and (ii) the history of activities other users performed on the user changes. The user’s score is used to evaluate her trustability and is consequently considered also for evaluating the changes she proposes. A weight is associated with each considered activity based on its relative importance in the evaluation process. An activity provides a positive or negative indication about a change, as shown in the Table 1.

The score of a user $U_j$ is evaluated by adapting a formula proposed for information retrieval queries with positive and negative feedback [21].

$$\text{userScore}_{U_j} = \alpha_U \cdot \frac{POS_j}{n_{posj}} - \beta_U \cdot \frac{NEG_j}{n_{negj}} \in [-2, +2] \quad (1)$$

where $POS_j$ and $NEG_j$ are weighted sums of performed activities $a_i$ concerning the user $U_j$:

$$POS_j = \sum_i \text{count}(a_i) \cdot w_{a_i}$$

where $\text{count}(a_i)$ is the number of occurrences of $a_i$ of positive type. $NEG_j$ is defined similarly. $\alpha_U$ and $\beta_U$ are coefficients to determine the relative importance of negative and positive components of the score. $0 \leq \alpha_U, \beta_U \leq 1$ and $\alpha_U + \beta_U = 1$. $n_{posj}$ and $n_{negj}$ are the total number of occurrences of activities of positive, resp. negative, type concerning the user $U_j$.

A change $C_k$ proposed by the user $U_j$ is evaluated as follows. Firstly, a community score is assigned according to the users’ assessments about $C_k$.

$$\text{communityScore}_{C_k} = 4 \cdot (\alpha_C \cdot \frac{POS_k}{n_{posk}} - \beta_C \cdot \frac{NEG_k}{n_{negk}}) \quad (3)$$

The community score ranges in $[-2, +2]$. $POS_k$ and $NEG_k$ are defined as for the user score but considering only assessment activities concerning the change $C_k$. The coefficients $\alpha_C$ and $\beta_C$ follow the same constraints defined for $\alpha_U$ and $\beta_U$.

Figure 4: Process of change submission and evaluation.
Secondly, a overall score is defined for $C_k$, as:

$$\text{overallScore}_{C_k} = \alpha_0 \ast \text{communityScore}_{C_k} + \beta_0 \ast \text{userScore}_{U_j}$$  \hspace{1cm} (4)$$

The coefficients $\alpha_0$ and $\beta_0$ define the relative weighting of the scores of $C_k$ and $U_j$. Moreover, they follow the same constraints as in the previous definitions.

### 6. CONCLUSIONS AND FUTURE WORK

In this paper we described the M-PREGed conceptual framework aiming at managing and providing a collection of data about geo linked data published by LOD sources. The collection is maintained independently from the original sources and constitutes an additional data layer. The collection is formed by crowd-generated corrections, complementary information and conflict resolutions about POIs. The users who contribute by proposing changes to the POIs descriptions, are evaluated based on assessments given by other users on the changes. A scoring mechanism is introduced for evaluating both users and changes. The final purpose is to make available a collection of proposed changes about a POI that are ranked based on the scores. Improving the data quality of large data sets by calling end-users as proposed in our framework is considered challenging. Most of the existing crowdsourcing platforms require considerable efforts to process large-scale data and do not provide functionalities for data quality analysis and malicious/wrong input detection. One intervention to improve such limitations is discussed in [8] where the authors present a methodology for developing a unified crowdsourcing framework to support high-quality data extraction and complement the current platforms (Amazon Turk, Microtask, etc.) with new functionalities for definition of the user’s tasks and quality control.

Our framework will be completed with the definition of the ontology of the published data. We are adopting for this task the methodology proposed by the authors of GeoLinkedData in [24]. Other future work includes studying the implementation issues of the framework in the context of LD practices and tools with the purpose to define a prototype version and perform experimentation. Finally, we plan to investigate the creation of personalized views on the data (e.g., customizable maps) based on users’ geo-cultural and topical affinities. Urban environments are in continuous development and POIs are nowadays more and more related to the individual culture, profile and preferences.

### 7. REFERENCES


