Geo Data Annotator: A Web Framework for Collaborative Annotation of Geographical Datasets

Stefano Cresci, Davide Gazzè, Angelica Lo Duca, Andrea Marchetti, Maurizio Tesconi
Institute for Informatics and Telematics (IIT)
National Research Council (CNR), Pisa, Italy
[name].[surname]@iit.cnr.it

ABSTRACT

In this paper we illustrate the Geo Data Annotator (GDA), a framework which helps a user to build a ground-truth dataset, starting from two separate geographical datasets. GDA exploits two kinds of indices to ease the task of manual annotation: geographical-based and string-based. GDA provides also a mechanism to evaluate the quality of the built ground-truth dataset. This is achieved through a collaborative platform, which allows many users to work to the same project. The quality evaluation is based on annotator agreement, which exploits the Fleiss' kappa statistic.

Categories and Subject Descriptors

D.2.11 [**Software Engineering**]: Software Architecture—*Domain-specific architectures*

Keywords

data matching; Web application; manual annotation; ground-truth

1. INTRODUCTION

Most of the data available on the Web is often fragmented and duplicated over many different platforms. This situation clearly represents a problem which poses serious limitations to the possible exploitation of such information. For this reason, considerable effort has already been devoted to research challenges such as *data matching* [1, 4, 3], namely the task of identifying and linking those records which represent the same entity over one or more datasets [1]. Although data matching algorithms define automatic or semi-automatic procedures to compare two or more records, they still need a manual annotation process during the learning or the evaluation phase. In fact, such phases are typically based on a so-called *ground-truth* dataset, which represents the verified situation of links between the records. In this paper we propose the Geo Data Annotator (GDA), an interactive collaborative Web framework which allows human operators to build and annotate a ground-truth dataset, containing the exact matched entries of two geographical datasets provided as input. Many datasets could benefit of GDA, e.g. the commercial places databases offered by Foursquare, Google Places and Facebook thus to create a comprehensive dataset of all touristic places in a city. This is the case of Tourpedia [2], an enciclopedia of tourism, built on datasets provided by social media.

2. METHODOLOGY

GDA faces with to challenges (i) reduce the complexity of the annotation process; and (ii) improve the overall quality of the annotated dataset. In this paper we focus on geographical datasets. A geographical dataset is defined as a collection of records geographically located somewhere in the world. More formally, let $x \in X$ be a specific place within the dataset X. Every $x \in X$ is described by the following tuple: $\langle id, \phi, \lambda, name, address, other \rangle$ where *id* represents the unique identifier of the place; ϕ and λ represent the geographical information about the place, being respectively, its latitude and longitude; *name* and *address* the name and the mail address associated to the place; *other* represents possible optional fields.

2.1 Reducing Matching Complexity

Matching complexity is one of the biggest issues faced when annotating a geographical dataset for data matching purposes. This is mainly due to the low scalability of the annotation process. In fact in data matching, the annotation is related to the links between the records, rather than the records themselves. This results in massive numbers of links potentially requiring a manual annotation. In GDA we mitigate the issue of matching complexity by reducing the total number of links to annotate. Building on the geographical information, as well as on the name and address fields of places, GDA exploits two different indices: geographical-based and string-based. A *geographical index* is built computing the geographical distance between any two places of the given datasets. This $I_{geo}(X, Y, \delta_{geo})$ index consists in listing all the pairs (x, y) such that their geographical distance is less than a given threshold δ_{geo}. We exploit the ϕ and λ attributes to compute the geographical distance $D_{geo}(x, y)$, according to the following formula:

$$D_{geo}(x,y) = \arccos(\ \sin\phi_x \sin\phi_y + \\ + \cos\phi_x \cos\phi_y \cos\Delta\lambda\ )\ R \qquad (1)$$

where $\Delta\lambda = \lambda_x - \lambda_y$ and R is the ray of Earth. The *string index* is based on the *name* and *address* fields. In literature,

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742723.

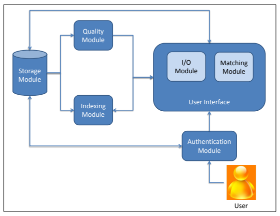

Figure 1: GDA architecture.

many different algorithms have been proposed to implement the concept of string/text similarity. Among those are the cosine similarity measure and its dual metric the cosine distance, the Hamming distance, the Jaccard index and the Levenshtein distance [6]. For the sake of simplicity, we report on the cosine distance measure, being our implementation choice, and one of the most widely used string similarity metrics. However, GDA could be easily extended with other algorithms. The cosine distance $D_{cos}(x, y)$ is defined as the complement of the cosine similarity $S_{cos}(x, y)$:

$$D_{cos}(x, y) = 1 - S_{cos}(x, y) = 1 - \frac{T_x \cdot T_y}{\|T_x\| \|T_y\|} \qquad (2)$$

In Equation (2), T_x and T_y represent the term frequency vectors of the *name* and *address* fields for records x and y respectively. Cosine similarity ranges from -1 meaning exactly opposite, to 1 meaning exactly the same, with 0 usually indicating independence: $S_{cos}(x, y) \in [-1, 1]$. Cosine distance values therefore range in $D_{cos}(x, y) \in [0, 2]$, where 0 indicates exact similarity. The $I_{str}(X, Y, \delta_{str})$ string index compares the values of cosine distance for records of X and Y and returns only those record pairs (x, y) whose distance is less than a given δ_{str} threshold.

2.2 Assessing Matching Quality

Another crucial challenge of all annotation systems lies in the detection of wrong annotations. This issue is strictly related to the assessment and improvement of the annotation process. In GDA, we build on the collaborative nature of the platform and exploit the concept of *annotator agreement*. What typically happens in data matching tasks, is that more than one human annotator is tasked with the annotation of all the links between records of one or more given datasets. The same data is therefore annotated by all human annotators who perform their task blindly. The GDA framework implements the well-known Fleiss' kappa statistic [5] as in the following:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \qquad (3)$$

where $\bar{P}$ is the relative observed agreement among the annotators and $\bar{P}_e$ is the hypothetical probability of random chance agreement. If the annotators are in complete agreement then $\kappa = 1$. If there is no agreement, other than what would be expected by chance, then $\kappa \leq 0$. Concordance metrics, such as Fleiss' kappa, can be exploited to assess the reliability of the annotation process by measuring the extent of agreement among the annotators, thus contributing to the assessment, and possibly to the improvement, of the annotation process.

3. THE FRAMEWORK

The Geo Data Annotator (GDA) is implemented as a Web framework which allows registered users to manage their projects (i.e. collections of datasets). Once logged in the platform, users may execute the following operations: a) create a new project and import the datasets to be matched; b) share the project with other users; c) perform the matching annotation on one or more datasets, this is done blindly with respect to other users annotating the same datasets; d) evaluate the results of the annotation process by analyzing the metrics computed by the framework; e) export the annotated datasets. Figure 1 shows the architecture of GDA. Once registered to the framework through the Authentication Module, a user can interact with the User Interface, in order to perform the manual matching, upload a new dataset or download the results. The User Interface is composed of two submodules: the *I/O module* and the *Matching Module*. The I/O module allows a user to manage projects, while the Matching Module provides the different views and interfaces with which users can browse and annotate their datasets. Specifically, two different views are defined: (i) a geographical view which exploits the geographical index, and (ii) a tabular view which exploits the string index. Such views are implemented as interactive Web interfaces, thus allowing the manual annotation of the datasets. All datasets and results are stored in a relational database (Storage Module). GDA was used in Tourpedia[1], a tourism dataset derived from some popular social media (Facebook, Google Places, Booking.com and Foursquare). All information extracted from those social media were integrated through a matching algorithm and verified through GDA manually. In details, three annotators merged the Amsterdam accommodations of Booking.com and Google Places manually through GDA. The final dataset contained 604 accommodations.

4. REFERENCES

[1] P. Christen. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Data Centric Systems and Applications. Springer-Verlag, Berlin, 2012.

[2] Cresci, S. and D'Errico, A. and Gazzè, D. and Lo Duca, A. and Marchetti, A. and Tesconi, M. . Towards a DBpedia of Tourism: the case of Tourpedia. In *Proceedings of the 2014 International Conference on Semantic Web - Poster and Demo Track*, ISWC2014, pages 129–132, 2014.

[3] N. Dalvi, M. Olteanu, M. Raghavan, and P. Bohannon. Deduplicating a places database. In *Proceedings of the 23rd international conference on World wide web*, pages 409–418. International World Wide Web Conferences Steering Committee, 2014.

[4] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios. Duplicate record detection: A survey. *Knowledge and Data Engineering, IEEE Transactions on*, 19(1):1–16, 2007.

[5] K. L. Gwet. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters*. Advanced Analytics Press, 2012.

[6] I. Oliver. *Programming classics: implementing the world's best algorithms*. Prentice-Hall, Inc., 1994.

[1] http://www.tour-pedia.org