GeoMerger: A Tool for Integration of Geographical Datasets by String Similarity

Stefano Abbate, Davide Gazzé, Angelica Lo Duca, Andrea Marchetti, Maurizio Tesconi, Fabio Valsecchi

Institute of Informatics and Telematics - CNR, Pisa, Italy
name.[surname]@iit.cnr.it

Abstract

In this paper we illustrate GeoMerger, a tool that integrates data belonging to two different geographical datasets. The tool is intended to merge the fragmented data of Facebook, FourSquare, Booking and Google Places hotels datasets.

1 Introduction

Nowadays the information available on the Web is incredibly increasing and spread among different sources. For example, Hostelsbase[1], which is a Web portal, hosts more than 500,000 hotels. Moreover, the same hotel has its own Web site as well as its Facebook[2] page and its profile on Foursquare[3]. If on one hand this is extremely positive, because of the variety of data and the different details for each source, on the other hand this leads to data fragmentation. For this reason, data integration is mandatory. The goal of data integration is to merge together different sources, in order to build a common schema.

In this paper, we face with a sub part of the problem of data integration: given two geographical datasets, we assume that they have already the same schema and we propose GeoMerger, a tool, for the integration of geographical entries of the Social Media datasets. GeoMerger merges two geospatial datasets finding the common matching elements. It filters the elements by their geographical coordinates and compares their string labels.

2 GeoMerger

GeoMerger is a tool that takes two geographical datasets as input and produces as output a third dataset containing the matching elements be-

[1] http://hostelsbase.org
[2] http://www.facebook.com
[3] http://foursquare.com

tween the input datasets. Each dataset must be provided as a collection of JavaScript Object Notation (JSON) objects, havin the following stucture.

```
{{"name": "value",
   "lat": value,
   "lng": value,
   "...": ...
},
{ ... }, ... }
```

Each element of both datasets is characterized at least by two attributes: a) a geographical coordinate (*lat* an *lng*), b) a representing string label (*name*), which could be its name or title. Moreover, we assume that the same resource 1) could not be present in both datasets, 2) could have different values of attributes in the two datasets, 3) may have different (or similar) string labels in the two datasets.

The output format is equal to the JSON structure described above. It includes *name*, *lat*, *lng* and the additional attributes of both the initial datasets. In fact, if a match is found by the tool the output dataset must include the additional attributes of both the input datasets for providing more information about the resources (hotel, accomodations, points of interest and restaurants).

The matching algorithm used by GeoMerger is described below. Given two datasets A and B, the algorithm computes for each entry $a \in A$ a set G composed by every entry $b \in B$ such that the geographical distance between a and b is less than a threshold x. Both a and b are composed by at least three attributes: a label, the latitude and the longitude.

For each element $(a, b) \in G$, the tokenization of the label of a and b is firstly computed, then the matching algorithm returns a value for each pair of tokens, representing the percentage of matching between them. If this value is greater than a token threshold (tt), then the two tokens are considered matched. When all the tokens have been compared, the string similarity between the two labels can be calculated as a function of the matching tokens. If the string similarity is greater than a string threshold (ss), the two strings are considered matched.

3 Experiment Results

In order to show the potentiality of GeoMerger, we have set up an experiment, which consisted in merging two datasets containing accommodations

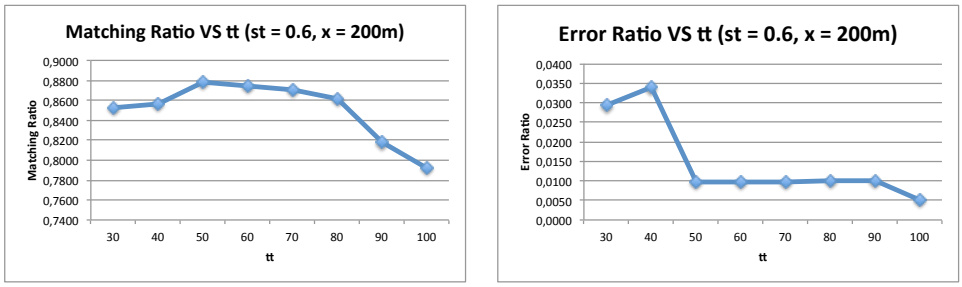

Figure 1: Matching Ratio VS tt. and Error Ratio VS tt.

in Amsterdam. In particular, we have extracted the two datasets from Foursquare and Facebook, by exploiting the APIs they provide. For each accommodation, we have taken the name, the address, the latitude and longitude. It may happen that for the same accommodation, the latitude and the longitude differ in the two datasets.

As a result, we have extracted 394 accommodations from Facebook and 775 from Foursquare. Through a graphical interface we have developed, we have manually merged the two datasets, obtaining 231 matched entries. Then we have run uploaded them in GeoMerger, by executing different experiments with different values of tt, st and x.

We have defined two metrics: matching ratio (MR) and error ratio (ER). The matching ratio is defined as the ratio between the number of matched entries by the tool (excluded errors) and the number of total merged entries by hand. The error ratio, instead, is the ratio between the number of errors committed by GeoMerger and the number of matched entries.

Setting the parameters to 200m, 0.6 and 50 respectively for x, st and tt the algorithm reaches the best case as shown in Figure 1. The best case has a matching ratio of 87.87% and an error ratio less than 1%.

4 Conclusions

In this paper we have illustrated GeoMErger, a tool which performs merging of geographical datasets by string similarity. We have demonstrated that it is able to reach a matching ratio of 87.87% in the best case. However, we would like to propose a challenge on how to improve its performance.