Matching Points of Interest from Different Social Networking Sites

Tatjana Scheffler[1], Rafael Schirru[1,2], and Paul Lehmann[3]

[1] DFKI GmbH, Alt-Moabit 91c, 10559 Berlin, Germany
{tatjana.scheffler,rafael.schirru}@dfki.de
[2] University of Kaiserslautern, Gottlieb-Daimler-Strasse, 67663 Kaiserslautern, Germany
[3] Brandenburg University of Applied Sciences, Magdeburger Str. 50, 14770 Brandenburg an der Havel, Germany

Abstract. Valuable user-generated information about locations (points of interest, POIs) is stored in various online social media platforms. Merging the data associated with one POI is hard because the platforms lack common identifiers. In addition, user-generated data is commonly faulty or contradictory. Here we present an approach matching POIs from Qype and Facebook Places to their counterparts in OpenStreetMap. The algorithm uses different similarity measures taking the geographic distance of POIs into account as well as the string similarity of selected metadata fields, showing good results.

Keywords: Data Integration, Social Networks, User-Generated Content, Points of Interest.

1 Introduction

In recent years, users have contributed valuable information about locations (points of interest, POIs) in community projects such as OpenStreetMap[1] (OSM) as well as in commercial social networks like Yelp or its German variant, Qype.[2] These platforms often provide different types of information for the same objects, for example ratings (Qype), check-ins (Facebook Places[3]), descriptions, categories, etc. For researchers and application developers it is often necessary to merge these distinct representations of POIs in order to obtain rich and complete information about the associated locations. Unfortunately the records representing the POIs do not share a common identifier across platforms thus making their matching a difficult task. In this paper we present an approach matching POIs from Qype and Facebook Places to their counterparts in OSM. The algorithm uses different similarity measures taking the geographic distance of POIs into account as well as the string similarity of selected metadata fields.

[1] http://www.openstreetmap.org/
[2] http://www.qype.com/
[3] http://www.facebook.com/facebookplaces

B. Glimm and A. Krüger (Eds.): KI 2012, LNCS 7526, pp. 245–248, 2012.

2 Related Work

Elmagarmid et al. survey methods proposed in the literature tackling the issue of lexical heterogeneity, i. e., records have fields that are identically structured across databases, but different representations of the data are used to refer to the same real-word objects (e. g., *44 West Fourth Street* vs. *44 W. 4th St.*) [3]. A data integration approach used for similarity joins in data bases is presented by Cohen [1]. Dozier et al. present Concord, a generic tool for constructing record resolution systems [2]. The tool streamlines the matching task into several steps, including finding a correspondence between fields in the two data bases, defining similarity functions between the fields, and setting up a machine learner to train and use a model for distinguishing good and bad matches. To our knowledge, all of these previous approaches do not specifically deal with POI data.

3 Approach

Our approach integrating POIs from different social platforms is a three staged process. First we apply a geo filter restricting the search space to a smaller number of candidate POIs. For the POIs in the candidate list we apply string preprocessing on their titles and then conduct a two phase matching process.

We use the geographic coordinates (latitude and longitude) of the POIs to localize the search space and reduce the number of comparisons that are required to find the counterpart of a query POI in the OSM data base. For that purpose we determine a bounding box of configurable size d ($d = 0.01°$ in our experiments) around the query POI and add all POIs from OSM that lie within the borders of the bounding box to a list of matching candidates. This list is used as the basis for further processing.

In order to match POI title strings the titles are normalized by removing non-alphanumeric characters, lowercasing and filtering stop words. The string matching phase is itself divided into a two phase process. In the first phase we check whether the title of a candidate POI is within a 10% edit distance of the title of the query POI, i. e., the number of required edit operations is less or equal than 10% of the length of the title of the query POI. The edit distance between two titles of POIs s_1 and s_2 is the minimum number of required edit operations (insertion, deletion, substitution) to transform s_1 into s_2. The measure is often also referred to as Levenshtein distance (e.g., [4] p. 58). If this condition is met, the candidate POI is counted as a match.

In case that no match can be found in phase one, our approach calculates the cosine similarity between the TF-IDF weighted term vectors representing the query POI and the candidates. TF-IDF is a term weighting measure that is widely used in the field of information retrieval [5]. It assigns a higher weight to terms that are supposed to be more discriminative, i. e., terms that appear frequently in one document but rarely in the whole document corpus. In our system the titles of the matching candidates and query POI constitute the corpus for the TF-IDF measure. We represent each document as a bag of words.

The document representations are mapped to a vector space where each axis represents a term and the respective value is its weight as determined by TF-IDF. The similarity between the query document and a matching candidate is then obtained by calculating their cosine similarity (cf. [4], pp. 120-123). In order to count a match, we require a minimum cosine similarity between the vectors of two POIs of 0.5. If candidate POIs exceeding this threshold are available, the most similar candidate is selected as a match. Otherwise no match has been detected.

4 Evaluation

To evaluate our algorithm, we manually chose 50 random POIs in the area of Berlin from Facebook Places and 50 POIs from Qype respectively. Then we obtained the detailed metadata of the POIs from the platforms. From Qype only the metadata of 49 POIs could be obtained. When developing the algorithm, we split the data in training (34 instances FB Places, 33 instances Qype) and test data (16 instances). However, as the amount of data is rather small, we chose to present the results based on the complete data set for each platform. To determine the accuracy of our approach, we add the number of correct matches and the number of POIs for which it has been correctly detected that an OSM match does not exist.

We compare the results of our approach (geo filter combined with string preprocessing and a vector space model, GSV) with two baseline algorithms:

Nearest Point of Interest (NP): The first baseline is selection of the nearest POI, within a threshold radius of 0.001°. This baseline only takes the geographic location of a POI into account without considering other metadata. We calculate the Euclidean distance between the query POI and the candidate POIs from the OSM data base, disregarding the curvature of the earth. However as all POIs in

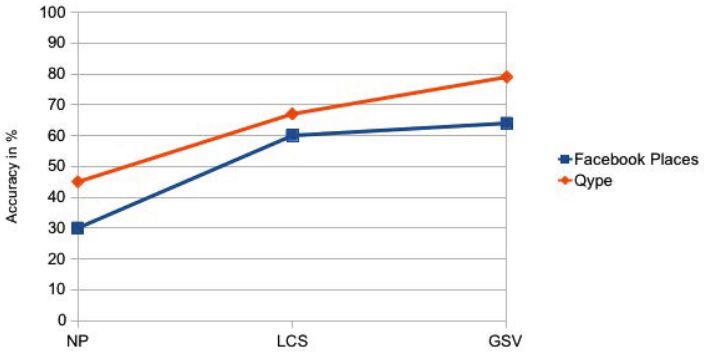

Fig. 1. Accuracy of the baseline approaches nearest POI (NP), longest common substring (LCS) and our method (GSV) for the platforms Qype and FB Places

our data base are restricted to the area of Berlin/Germany this measurement is precise enough for our purposes.

Longest Common Substring (LCS): The second baseline selects the candidate whose title shares the longest common substring with the target POI, independent of location. If several candidates have a LCS of equal length, we select the one with the highest ratio of the length of the LCS to the length of the candidate's title. Minimum ratio of candidate to target POI title is 40%.

In general it can be observed that the geographic information on its own (method NP) does not lead to satisfactory results when integrating POIs from different platforms. Comparing the titles of the POIs (method LCS) results in a higher accuracy. However the best results are obtained when geo data is combined with string similarities in the matching process. Figure 1 shows the overall accuracy of the approaches. For Qype our approach achieves an overall accuracy of 79% compared to 45% for NP and 67% for LCS. For FB Places the accuracy of our method is 64% against 30% for NP and 60% for LCS.

5 Conclusion and Future Work

In this paper we presented an approach matching the representations of POIs from different platforms to obtain rich descriptions about locations. The method combines geographic information with string similarities thus achieving a higher accuracy in the matching process than two baseline approaches that either rely on geographic information or string similarity respectively. In our future work we have to consider further metadata that is often annotated for POIs. For instance, category information is often available which can help to distinguish POIs that lie close around a famous place (e.g., a square) and carry the name of the place.

Acknowledgements. This research has been funded by the Investitionsbank Berlin in the project "Voice2Social", and co-financed by the European Regional Development Fund.

References

1. Cohen, W.W.: Data integration using similarity joins and a word-based information representation language. ACM Trans. Inf. Syst. 18, 288–321 (2000)
2. Dozier, C., Molina-Salgado, H., Thomas, M., Veeramachaneni, S.: Concord - a tool that automates the construction of record resolution systems. In: Proceedings of the Entity 2010 Workshop at LREC 2010, Valetta, Malta (2010)
3. Elmagarmid, A., Ipeirotis, P., Verykios, V.: Duplicate record detection: A survey. IEEE Transactions on Knowledge and Data Engineering 19(1), 1–16 (2007)
4. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval, online edn. Cambridge University Press (April 2009)
5. Sparck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation 28(1), 11–21 (1972)