Graph-based strategies for matching points-of-interests from different VGI sources

Tessio Novack
Heidelberg University
GIScience Group
Im Neuerheimer Feld 348
Heidelberg
novack@uni-heidelberg.de

Robin Peters
Heidelberg University
GIScience Group
Im Neuerheimer Feld 348
Heidelberg
r.peters@uni-heidelberg.de

Alexander Zipf
Heidelberg University
GIScience Group
Im Neuerheimer Feld 348
Heidelberg
zipf@uni-heidelberg.de

Abstract

Several urban studies have been relying on spatial data provided by Volunteered Geographic Information (VGI) sources. The matching of features from different VGI platforms may serve to assess and improve the reliability and completeness of VGI data. In this paper, we propose two strategies for matching points-of-interests (POIs) based on a graph whose nodes and edges represent the POIs and their possible matching pairs, respectively. We called these strategies *Best-Best Match* and *Combinatorial Match*. The former drastically reduces the amount of ambiguous edges in comparison to the baseline method, which matches each POI with its best matching candidate. The later completely rules out ambiguous matches. Both strategies are able to produce 1:0 matchings, thus tackling the issue that sometimes a POI from a reference dataset is not represented in a second dataset. The *Combinatorial Matching* strategy consists in extracting all possible subsets of edges from the graph in which no node occurs more than once. It then selects the subset with the highest sum of edge weights. As a first evaluation of these strategies, we conducted an experiment for matching POIs from OpenStreetMap and Foursquare. The results show that our two proposed strategies perform as comparatively good as the baseline method.

Keywords: Volunteered Geographic Information, points-of-interest, matching, data conflation, graph theory.

1 Introduction

Several urban studies and applications have been increasingly relying on spatial data provided by Volunteered Geographic Information (VGI) projects. However, VGI data is produced by people with different interests, perceptions and expertise. This has motivated many researches to investigate different ways for assessing the quality of VGI data, as reviewed by e.g. Degrossi et al. (2017). Frequently, these investigations take two important aspects into account, namely, the reliability and the completeness of the data (Ballatore and Zipf, 2015).

Both these aspects can be assessed and improved by matching corresponding representations of the same feature across different VGI sources. The reliability, i.e. the trueness, about the existence of a venue gains strength by the fact that it is represented in more than one VGI project. At the same time, the matching enables us to gather complementary and reconfirm common information about the venue (e.g. address, open time, accepted credit cards etc.), thus improving the completeness aspect of the data.

The matching of road networks and building footprints from VGI and authoritative sources are already well-investigated topics (Fan et al. 2014; Abdolmajidi et al. 2015). On the other hand, strategies for matching points-of-interests (POIs) from different geo-datasets have not yet been fully explored. Scheffler et al. (2012) proposed a simple approach for matching Qype and Facebook places to their OSM counterparts based on their spatial distance and string and topic similarities. McKenzie et al. (2013) proposed a weighted regression model for matching Foursquare and Yelp POIs based on their distance and string similarities. They also considered the POIs textual similarity by means of a Latent Dirichlet Allocation model (Blei et al, 2003). Li et al. (2016) also focused on a strategy for defining the weights of different POIs similarity measures. They proposed defining these weights based on the entropy of the respective similarity measure. Jiang et al. (2015) used a robust string similarity measure developed by Cohen (2003) for matching POIs from Yahoo! and proprietary sources with the ultimate goal of estimating spatially detailed land-use information. These works focus mainly on the evaluation of matching candidates, but they lack on suggestions on how to tackle the issues of (1) ambiguous matching, i.e. cases when two POIs from one dataset are matched to the same POI from a second dataset, and (2) of 1:0 matches, i.e. cases when a POI does not have a corresponding representation on a second dataset.

In this short paper, we propose two graph-based strategies for matching corresponding POIs from two different VGI sources with the intention of avoiding ambiguous matching. Furthermore, the two strategies enable the detection of 1:0 cases. Due to lack of space though, we focus the performance analysis on the matching accuracy levels and, partially, the number of ambiguous matches produced.

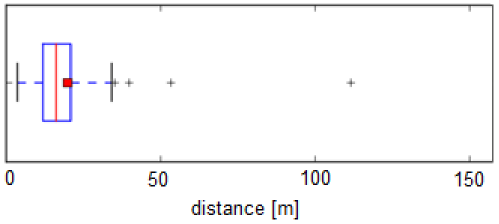
2 Similarity measures for matching POIs

The matching of corresponding POIs from different VGI sources has to rely on measures or metrics of similarity between the points. Based on the literature, one is able to distinguish the following types of measures:

- Temporal gap,
- Spatial distance,
- String similarity,
- Semantic similarity,
- Topologic consistency and
- Topic similarity.

The temporal gap can be useful as an auxiliary measure for detecting no longer existing venues and venues with obsolete attributes.

Figure 1: Boxplot of the spatial distance between one hundred corresponding pairs of POIs from OSM and Foursquare.

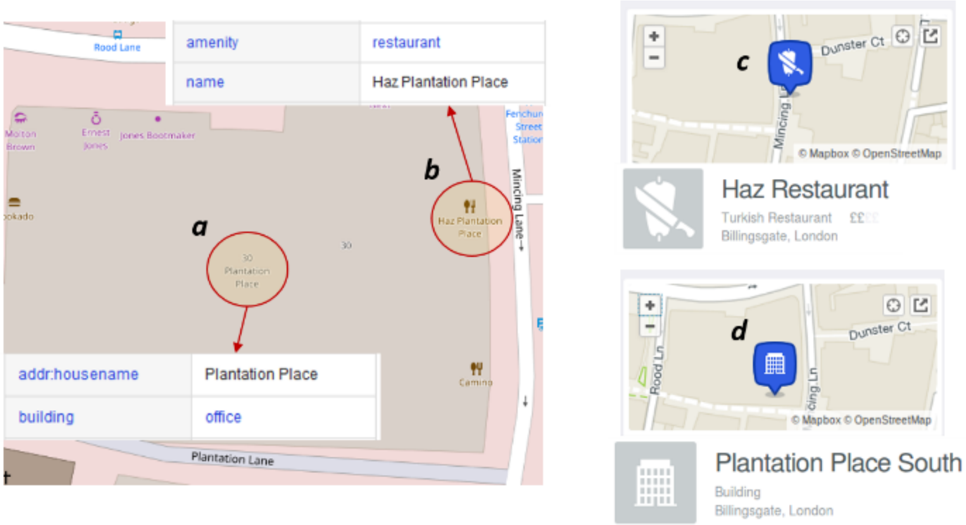


The spatial distance refers to the distance in physical space between the positions of the POIs. Because one expects to find corresponding POIs close to each other, it is the most obvious measure to consider in the matching. However, the possibilities to geocode the POIs position at different scales (i.e. city level, city district, street, street and number) as well as to define its position manually by clicking on the screen may lead to differing positions of corresponding POIs. Figure 1 shows a box plot of the spatial distance between pairs of corresponding POIs from OSM and Foursquare. These matching pairs were collected by randomly selecting one hundred OSM POIs and then finding their corresponding Foursquare venues. It indicates that relying solely on distance is likely to produce inaccurate matchings.

String similarity measures express how similar two sequences of letters are. Stores, restaurants, banks and gyms have names and they provide us valuable evidence for finding corresponding POIs. One should note however that chains like 'McDonald's' or 'Starbucks's' may have several stores in a city. Hence, it is advisable to consider the string similarity in conjunction with the POIs spatial distance. Even then, differences in writing, misspellings and prepositions may lead to wrong matches. In particular, the lack of consensus on the name of a same venue registered by different users in different VGI sources may hamper the correct matching. Figure 2 exemplifies this by depicting two venues as represented on OSM and on Foursquare. The former were coded as 'a' and 'b' and the latter as 'c' and 'd'. Based on the POIs name similarities, 'a' as well as 'b' would be matched with 'd', where in fact 'b' should match 'c'.

The venues on Figure 2 might be correctly matched if along with the spatial distance and the string similarity, the semantic similarity of the POIs is considered. For instance, a POI may belong to a general category 'restaurant' and to a more specific one like 'fast-food restaurant'. Comparing the semantic similarity between the categories from the POIs in Figure 2, would strengthen the potential of matching 'b' with 'c', since they belong to categories 'restaurant' and 'Turkish restaurant' respectively. By the same token, it would weaken the attraction between 'b' (restaurant) and 'd' (building) while strengthening the correct match between 'a' (office building) and 'd'.

Figure 2: Example where a matching based solely on the venues name similarities would cause mistaken matches. The POI coded as 'b' would be matched with 'd', instead of 'c'. To correct that, we suggest considering the semantic similarity of the POIs.



amenity	restaurant
name	Haz Plantation Place

addr:housename	Plantation Place
building	office

Haz Restaurant
Turkish Restaurant ££
Billingsgate, London

Plantation Place South
Building
Billingsgate, London

Source: Foursquare and OpenStreetMap.

Another factor that one may consider is the topologic consistency of the matching candidates. Two POIs are topologically consistent if they are located inside the same building footprint or on the same side of the river, for example. Due to their frequently inaccurate positions, however, the topologic relations of VGI POIs might not provide valuable evidence for their correct matching.

3 Measures considered and matching strategies

In this section, we describe the similarity measures considered in an initial experiment as well as the matching strategies we propose.

3.1 Spatial similarity

The spatial similarity measure used in our matching strategy is based on the Euclidean distance between the POIs. It decreases from 1 to 0 according to their distance in space:

$$Spt(p_i, p_j) = \begin{cases} 1-\left(\dfrac{d(p_i, p_j)}{\varepsilon}\right) & \text{if } d(p_i, p_j) < \varepsilon \\ 0 & \text{if otherwise} \end{cases} \quad (1),$$

where $d(p_i,p_j)$ is the Euclidean distance between the POIs and ε is a distance threshold, above which the spatial similarity between the POIs is zero.

3.2 String similarity

The string similarity measures we used are two normalized Levenshtein distances (Levenshtein, 1966) available in the FuzzyWuzzy Python library. These are the Token Sort Ratio and the Token Set Ratio measures. These two measures express the similarity of two strings as a percentage value. The metric Token Sort Ratio is more conservative and outputs a score of 83 when comparing “This is a paper” with “This is a short paper”, whereas Token Set Ratio outputs a score of 100.

3.3 Semantic similarity

In order to evaluate the semantic similarity between the POIs categories, the large English semantic network WordNet® (Miller, 1995) was used. In WordNet®, each word is associated to a group of one or more synsets, which are synonyms or definitions from that word. Different measures are available for computing the semantic similarity between synsets in WordNet®. In this initial stage, we have been evaluating the *shortest path* distance, which considers the number of edges separating the two synsets in the semantic network. We have adapted it to the following equation:

$$dist(S_i, S_j) = 1-\left(\frac{\min\limits_{s_i \in S_i, s_j \in S_j}(shortest\_path(s_i, s_j))}{20}\right) \quad (2),$$

where s_i is one from possibly multiple synsets from the word S_i. We iterate through all synsets of words S_i and S_j (i.e. the POIs categories) and extract the minimum value. This is then divided by twenty, which is the maximum depth of the network, i.e. the longest possible path between two synset. This fraction is then subtracted from 1 so that similar synsets are assigned a value closer to one.

3.4 Graph definition and POIs matching strategies

We define a graph where nodes represent POIs and edges represent possible matchings between connected nodes. The matching potential of each pair of nodes is represented by an edge weight. The weight of each edge may be computed by any function of the similarity measures presented above. The graph creation obeys two constraints: (1) two nodes are linked only if they are less than a certain distance threshold apart and (2) two nodes from the same VGI source are never linked. Figure 4(a) depicts an example of such a graph. Based on it, three matching algorithms are evaluated.

The first algorithm, our baseline method called *Best-Candidate*, evaluates the matching candidates of each POI and naively matches it with the best candidate. It ignores the fact that two POIs from one dataset may be matched to the same POI from a second dataset, as depicted in Figure 3(b). It will also match every POI from a reference dataset, thus assuming that every node from the reference dataset has a corresponding one on the second dataset. In other words, it assumes only the existence of 1:1 matches and does not tackle the possible existence of 1:0 matches.

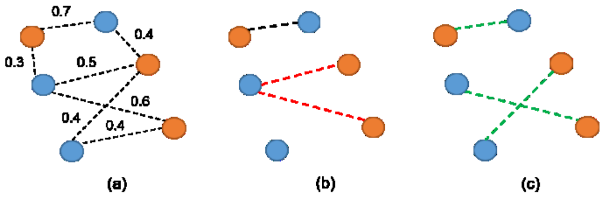
The second algorithm, named *Best-Best Match*, matches node *j* with node *i* only if *j* is the best match of *i* and *i* is the best match of *j*. As opposed to the *Best-Candidate* algorithm, it drastically reduces the possibility of ambiguous matches. Furthermore, it will not match every POI from the reference dataset, thus assuming that not each of its nodes has a corresponding one on the second dataset. Hence, it will hopefully detect 1:0 cases.

The third algorithm, named *Combinatorial Matching*, performs matching in the graph-theory sense of the word. A matching solution is a set of edges where no node occurs more than once. The algorithm exhaustively searches for the subset of edges whose sum of edge weights is the highest (Galil, 1983). An example of a matching solution is depicted in Figure 3 (c). This algorithm completely excludes ambiguous matches such as the one in Figure 3(b).

By excluding ambiguous matches, the *Combinatorial Matching* algorithm is also likely to increase the matching accuracy by avoiding mistaken matches as the one discussed in Section 2 based on Figure 2. The expectation is that linking node ‘b’ with ‘d’ would demand a linkage between ‘a’ and any other node other than ‘d’. Such a subset of edges would hopefully have a lower total final weight sum than a subset containing edges ‘b-c’ and ‘a-d’.

With the intent of not matching nodes from the reference dataset that have no corresponding node on the other dataset (1:0 cases), we implemented the following adjustment in the edge weights. For all nodes, we computed the average value of the weights from the edges connected to it. Note that each edge will have two average values associated to it. Next, we subtracted each edge weight from these two average values. The results of these subtractions are then added and the resulting value defined as the final edge weight. Figure 4 exemplifies the whole process.

Figure 3: Schematic example a network with nodes, edges and edge weights (a). An example of two ambiguous matches (b). An all-valid matching result (c).



4 Data and experiment

In order to test our matching strategy, we collected POIs from Foursquare and OSM from a small bounding-box located at the central area of London (England). All Foursquare POIs located inside the bounding-box were collected, whereas only POIs and building centroids from OSM containing the tags ‘name’, ‘amenity’, ‘shop’, ‘cuisine’, ‘tourism’, ‘office’, ‘leisure’, ‘land-use’, ‘food’, ‘sport’, ‘memorial’ and ‘brewery’ were collected. In total, 824 and 2229 POIs from OSM and Foursquare respectively were considered in the experiment. For assessing the matching accuracy of the three algorithms presented above, we manually collected one hundred matching pairs. Following, we created the graph with a distance threshold of 220 m for linking the POIs. This threshold lies just above the distance between the pair of matching nodes the furthest apart belonging to our test-set.

The final weights of the edges were set in ten different ways. Table 1 shows the codes of the final weights computation functions. The coding serves to a better depiction of the experiments results. Note that C6 to C10 are just the same as C1 to C5 except that the semantic similarity measure is included in the former. In these initial experiments, the final weights are computed simply as the linear addition of the similarity measures presented above.

Figure 5 depicts the accuracy levels obtained with the different edge weights functions (Table 1) for the three matching algorithms presented in Section 4. At a first glance, it seems that the *Best Candidate* algorithm is the most accurate. However, many of its matchings are ambiguous, as presented by Table 2 for the edge weights functions C1 and C2, which take the name similarity into consideration. Also, this algorithm is not able to detect 1:0 matches. On the other hand, the *Best-Best Match* algorithm is able to detect 1:0 matches and has achieved just slightly worse accuracy levels than the *Best Candidate* algorithm. Nonetheless, as shown exemplary by Table 2 for the edge weight functions C1 and C2, this algorithm does not completely eliminate all ambiguous edges. A closer look into Figure 5 leads to the conclusion that the *Combinatorial Matching* algorithm in general delivered a comparable performance as the other algorithms. It even achieved the best overall accuracy when applied with the edge weight function C5 (85%). Furthermore, it has the advantages that it rules out the risk of ambiguous matchings and, due to the edge weight adjustment procedure proposed on Section 3.4, it is capable of producing 1:0 matches.

Figure 4: Original edge weights (a). Average edge weights from each node (b). Subtraction of the average weights from the original edge weights (c). Final edge weights computed as the addition of the subtractions from the previous step (d).

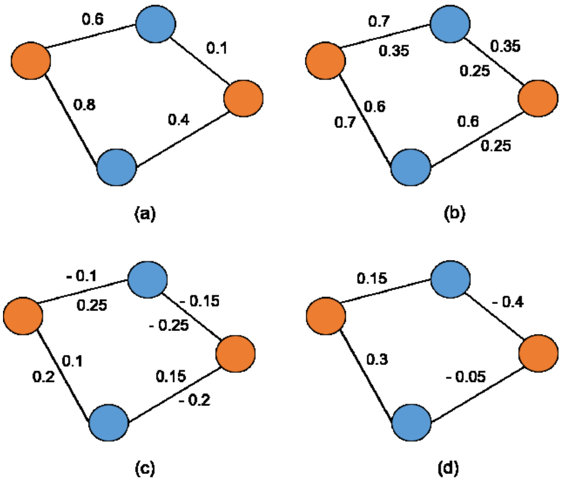


Figure 5 also shows that matchings with edge weight functions involving the semantic similarity measure (C6 to C10) have performed worse than functions C1 to C5. Probably, the semantic similarity adds indecision to the matching, as many POIs close to each other have frequently similar categories (e.g. restaurant, cafés, pubs). This does not mean though that the semantic similarity measure is not relevant, as we suggested above based on a concrete example. Possibly, if the weights of the individual similarity measures are tuned, the accuracy levels of matching experiments conducted with the name, spatial and semantic similarity measures would be higher than when only the name and spatial similarities are considered.

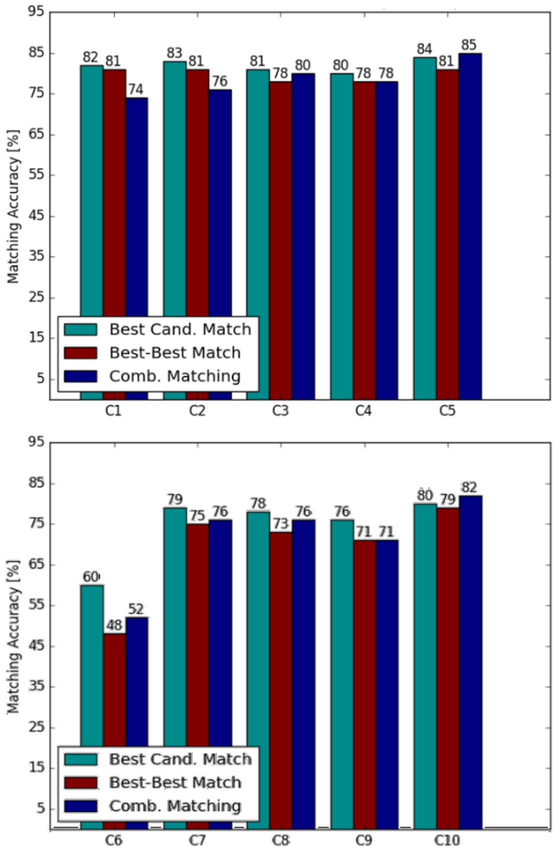
Table 1: Measures, and the respective coding, used for computing the final edge weights. Each measure is simply a linear addition of different similarity measures.

Code	Edge Weights Measures
C1	Token Sort Ratio (TSoR)
C2	TSoR + Token Set Ratio (TSeR)
C3	TSoR + Spatial Similarity (SpS)
C4	TSeR + SpS
C5	TSeR + TSoR + SpS
C6	TSoR + Semantic Similarity (SemS)
C7	TSeR + TSoR + SemS
C8	TSoR + SpS + SemS
C9	TSeR + SpS + SemS
C10	TSeR + TSoR + SpS + SemS

Table 2: Measures, and the respective coding, with which the final weights of the graph's edges were set.

	Best. Cand. Match		Best-Best Match	
	C1	**C2**	**C1**	**C2**
Number of Foursquare nodes with more than one OSM node matched to it	112	113	22	16
Number of ambiguous edges	253	250	50	37
Mean number of edges matched to the same Foursquare node	2.25	2.28	2.2	2.3
Max. number of edges matched to the same Foursquare node	7	7	6	6

Figure 5: Accuracy levels of the ten matching experiments conducted with different similarity measures and combinations of them.



5 Future work

On future work, we intend to investigate ways of aggregating the different similarity measures other than by simply adding them. This is expected to increase the accuracy of the matchings, in particular when the semantic similarity measure is considered. Different data mining and classification algorithms shall be investigated for defining the weights of the different measures.

The question on how to consider the semantic similarity between POIs is also to be considered. So far, we have experimented with a fairly simple measure. However, the literature is rich in evidence that more effective semantic similarity measures lead to better results (Ballatore et al., 2013).

This work has proposed two strategies for matching POIs from VGI sources that, as opposed to the naive though frequently used best-candidate matching approach, are able to avoid ambiguous matches and to leave POIs from a reference dataset unmatched (hopefully 1:0 cases). The accuracy of these two strategies regarding the detection of 1:0 matches needs to be reported. Unfortunately, we lack space for presenting this analysis in this short paper. Furthermore, the fact that in VGI datasets a same venue is frequently represented by more than one POI has been so far ignored by researchers. These duplicate cases require a matching approach able to perform n:1 and n:m matches. We are currently putting effort to cover these gaps and hence significantly contribute to this very relevant research topic.

Lastly, future experiments will be conducted over larger areas and evaluated based on larger test-sets.

References

Abdolmajidi, E., Mansourian, A., Will, J. and Harrie, L. (2015) Matching authority and VGI road networks using an extended node-based matching algorithm. *Geo-spatial Information Science*, 18, 65-80.

Ballatore, A., Bertolotto, M. and Wilson, D. C. (2013) The semantic similarity ensemble. *Journal of Spatial Information Science*. 7, 27-44.

Ballatore, A. and Zipf, A. (2015) A conceptual quality framework for volunteered geographic information. In: *XII Conference on Spatial Information Theory*, Santa Fe, 2015.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research*. 3, 993-1022.

Cohen, W. W., Ravikumar, P., Fienberg, S. E. (2003) A comparison of string distance metrics for name-matching tasks. In: *Proceedings of 2003 International Joint Conferences on Artificial Intelligence (IJCAI-03)*. San Francisco, 2003.

Fan, H., Zipf, A., Fu, Q. and Neis, P. (2014) Quality assessment for building footprints data on OpenStreetMap. *International Journal of Geographical Information Science*. 28, 700-719.

Galil, Z. (1986) Efficient algorithms for finding maximal matching in graphs. *Journal ACM Computing Surveys*. 18(1), 23-38.

Jiang, S., Alves, A., Rodrigues, F., Ferreira, J. Pereira, F. C. (2015) Mining point-of-interest data from social networks for urban land use classification and disaggregation. *Computers, Environment and urban Systems*. 53, 36-46.

Levenstein, V. I. (1966) Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8), 707-710.

Li, L., Xing, X., Xia, H. and Huang, X. (2016) Entropy-weighted instance matching between different sourcing points of interest. *Entropy*, 18(2), 1-15.

Mckenzie, G., Janowicz, K., Adams, B. (2013) Weighted multi-attribute matching of user-generated points of interest. In: *Proceedings of the 21st International Conference on Advances in Geographic Information Systems*. Orlando, 2013.

Meng, L., Huang R., Gu J. (2013) A review of semantic similarity measures in WordNet. *International Journal of Hybrid Information Technology*, 6(1), 1 – 12.

Miller, G. A. (1995) WorldNet: a lexical database for English. Communications of the ACM 38(11), 39-41.

Scheffer, T., Schirru, R. and Lehmann, P. (2012) Matching points of interest from different social networking sites. In: Glimm, B. and Krüger, A. (eds.) *KL 2012: Advances in Artificial Intelligence*. Springer Berlin Heidelberg, pp. 245-248.

Degrossi, L. C.; Albuquerque, J. P.d.; Rocha, R. d. S.; Zipf, A. (2017) A framework of quality assessment methods for crowdsourced geographic information: a systematic literature review. In: *Proceedings of the 14th International Conference on Information Systems for Crisis Response and Management.* Albi (France), 2017.