Weighted Multi-Attribute Matching of User-Generated Points of Interest

Grant McKenzie
University of California, Santa Barbara
grant.mckenzie@geog.ucsb.edu

Krzysztof Janowicz
University of California, Santa Barbara
jano@geog.ucsb.edu

Benjamin Adams
University of California, Santa Barbara
adams@nceas.ucsb.edu

ABSTRACT

To a large degree, the attraction of Big Data lies in the variety of its heterogeneous multi-thematic and multi-dimensional data sources and not merely its volume. To fully exploit this variety, however, requires conflation. This is a two step process. First, one has to establish identity relations between information entities across the different data sources; and second, attribute values have to be merged according to certain procedures which avoid logical contradictions. The first step, also called matching, can be thought of as a weighted combination of common attributes according to some similarity measures. In this work, we propose such a matching based on multiple attributes of Points of Interests (POI) from the Location-based Social Network Foursquare and the Yelp local directory service. While both contain overlapping attributes that can be use for matching, they have specific strengths and weaknesses which makes their conflation desirable. We present a weighted multi-attribute matching strategy and evaluate its performance. Our strategy can automatically match **97%** of randomly selected Yelp POI to their corresponding Foursquare entities.

Categories and Subject Descriptors

H.4 [**Information Systems**]: Information Systems Applications; H.2.8 [**Information Systems**]: Database Management—*Spatial databases and GIS*

Keywords

Volunteered Geographic Information, Location-based services, Point of interest, POI, Conflation, Similarity

1. INTRODUCTION

Recently, an economy of data and service providers has evolved around Points Of Interest (POI).[1] This includes map-centric applications such as Google Maps, local directory services such as Yelp, several location-based social networks, e.g., Foursquare, as well as numerous spatially-enabled sharing services such as Path or Flickr. Each of these services specializes on certain kinds of place-related information.

From a research perspective, the combination of these data sources is desirable for multiple reasons. First, by conflating Points of Interest, we can exploit complementary attributes to arrive at a more holistic understanding of places. For instance, one can combine user reviews from different communities to study sentiment, compare the place categorization hierarchies and match them using ontology alignment techniques, mine check-in behavior for patterns, compare pictures from tourists versus locals, and so on. Second, we can increase data quality by comparing the same attributes across data sets. One potential application is to remove typos in place names contributed by volunteers.

The process of conflating Points Of Interest involves two steps. First, identity has to be established between them. That is, it has to be determined whether both information entities correspond to the same place in the physical world. We refer to this part as *matching* throughout the paper. To do so, one would usually compare the values of attributes common to both datasets using a particular similarity measure. For example, if two datasets both contain a name attribute for their Points of Interest, the Levenshtein distance can be used to match them. Simply comparing names alone, however, will only work for certain cases. Thus, other attributes such as geographic locations can be compared using appropriate measures as well. In practice, these measures will rarely return exact matches, and we have to combine them and define a matching threshold.

The second step involves conflating Points of Interest. For example, while a place may have multiple names and one can be chosen to be canonical, this is not feasible for geographic locations. Understanding how to proceed with different attributes is an ontological question. It is beneficial to understand the process by which attribute values are recorded as well as the resulting types of errors. For example, one could naively assume that because POI locations from location-based social networks (LBSN) are recorded by GPS positioning via smartphones, they may be inaccurate to about 5-30 meters and averaging positions from two LBSN would improve accuracy. We will later discuss why this is not the case.

In this work we will focus on the first step of conflation and show how to match POI from the LBSN Foursquare and

[1]We use the term Points Of Interest here instead of the more general Places Of Interest as the used data sources only contain point-like feature geometries.

This is a draft version, not the final submission.
ACM X-XXXXX-XX-X/XX/XX
http://dx.doi.org/10.1145/2525314.2525455.

the local directory services Yelp. Foursquare specializes on user check-ins and, thus, social and temporal aspects. While it also provides user tips, those are typically short personal statements. In contrast, Yelp focuses on detailed user reviews and a wide range of semi-structured place attributes such as the ambiance, prices, noise level, and wifi availability. For example, this would enable queries for *places visited by friends, that have a low noise level, friendly staff, and free wifi.*

The contributions of this work are as follows:

- We will test whether a syntactic string measure can successfully match at least 80% of our sample data. It is important to keep in mind that for an automatic matcher a success rate of 80-90% is not sufficient. In our case, given the > 30 million POI in the USA alone, at least 3 million POI would still have to be corrected and matched manually.
- Following the Pareto principle, we assume that matching the remaining (less than) 20% of POI will require a weighted combination of matchers which exploit additional POI attributes. To do so, we will use *Double Metaphone* to match for phonetic similarity as well as matchers based on textual user reviews and geographic distance.
- Subsequently, we will use binomial probit regression to arrive at a weighted combination of all matchers and evaluate the results.
- Finally, we will discuss some interesting insights made during our work. For instance, we will try to explain why the geographic coordinates of POI clearly differ between Yelp and Foursquare.

2. RELATED WORK

The matching and conflation of geographic datasets has a long history in the field of geographic information science. Historically, two related areas of research have emerged, one focusing on the geometric or geographic properties of the data [5, 11] and another centered on the descriptive attributes [8, 2].

A number of methods have been developed for analyzing text and assessing similarities between strings for duplication detection [7, 10], and information retrieval [9]. Though name matching is a common technique used in matching and conflating POI, the geographic coordinates of the POI also play a significant role. Research by Wu and Winter [16] focused on the semantic issues involved in matching place names in a gazetteer. They found that the spatial properties of an entity could be engaged as a supplemental source for matching. Similarly, Mülligann et al. [15], utilized the spatial-semantic interaction of point features to determine duplicates in OpenStreetMap.

The comparison of documents based on unstructured text has been an area of significant research in the past few years. Recent advancements in probabilistic topic models [4] have made it feasible to infer and measure similarities between documents. These topic-based approaches have emerged in the geospatial science literature as well with researchers geolocating individuals based on the content of their social contributions [6, 12] and building location recommendation systems [14, 3], to name a few. It has been shown in previous work that individual words and topics in place descriptions are indicative of geospatial location [1]. However, this effect only becomes present at a coarser spatial resolution; e.g., when comparing meso-level features like cities and national parks.

3. METHODOLOGY

In this section, we describe the used dataset and developed methodology.

3.1 Venues Dataset

A random sample of 200 POI were collected from the continental United States through the public Yelp API. Selected businesses consisted of a name, geographic coordinates, at least one category tag, and a minimum of five user-contributed reviews. The 200 POI were manually compared to venues accessed through the Foursquare API returning a positive matched set of 140 Foursquare venues. Again, a positive match required that the above attributes also be present in the matched Foursquare POI. In place of reviews, this other source of user-generated content includes *Tips*, which are similar in nature to *Reviews* except shorter in length, mostly recommending items or offering advice. For a set of POI to be chosen as a match, a minimum of 3 tips associated with the Foursquare venue were required. We call this matched set of POI, V_M.

The mean great circle distance between the two venue sources equated to 62.8 meters. The largest discrepancy in distance between two matched venues in V_M was found to be 869.3 meters. Using this distance as a rough upper-bound, each known Yelp business was buffered to return all Foursquare venues within a 1000 meter radius. This resulted in a test set, V_T, of 73,304 POI averaging 505 per known Yelp location. All POI in V_T were comprised of some textual name attribute and geographic coordinates, 82.1% of POI were tagged with a minimum of one category and 34.2% listed at least one user-contributed *tip*.

3.2 Venue properties & Measures

The methods used to match POI are grouped by the attribute of the venue used as input.

Venue Name: Levenshtein Distance

The Levenshtein (edit) distance between each Yelp business in V_M and the nearest Foursquare venues in V_T are calculated and ranked based on smallest edit distance. In this case, edit operations are defined as addition, deletion and substitution with each operation given a weight of 1. The fewer edits needed, the smaller the edit-distance and the more similar the two venue names are to be gauged.

Venue Name: Phonetic Similarity

Using the Double Metaphone algorithm, two phonetic codes (primary and alternate) are generated for each Yelp business in V_M and two for each of the venues in V_T. Using the Levenshtein distance metric, each pair of codes is compared, producing four phonetic distance values. As was the case with the previous metric, venues are ranked by distance value from smallest to largest, the smallest value indicating the best estimated match given this similarity measurement method.

Geographic Location

In POI matching, there is an assumption that the geographic distance between two locations is a strong indicator of match accuracy. Though this is highly dependent on the contributing source of the data, in this case, the location of each venue offered by both POI sources is subject to the same contribution errors present in any of the other attributes. Many users either enter an address or cross-street for a venue (which is then geocoded) or, more likely, they rely on the geographic positioning method employed by their mobile device. Given the uncertainty of mobile positioning systems and systematic errors inherent to GPS and wireless positioning, it is not uncommon to find a significant discrepancy in the geographic coordinates of the same location sourced from two applications. The mean distance between two POI in our matched set V_M is 62.8 meters with a maximum difference of 869.3 meters.

Topic Similarity: Descriptive Reviews

An unsupervised topic model approach is taken to measure the textual similarities across locations in order to determine a match. *Latent Dirichlet allocation (LDA)* is an unsupervised, generative probabilistic model used to infer the latent topics in a textual corpus [4]. Here we train LDA by treating the text associated with each venue as a single document. LDA "discovers" topics, represented as multinomial distributions over words. The words that compose the topics emerge from the training set of documents based on co-occurrences of words within and across documents.

As input to the LDA model, all Yelp reviews in V_M are merged based on venue and the LDA model is run with 40 topics. After removing venues with less than 40 characters, the percentage of POI on which topic modeling can performed is reduced to 26.1% of V_T. The MALLET toolkit [13] provided the LDA implementation used in this work.

Once the V_M and V_T POI are represented as topic distributions, the *Jensen-Shannon divergence (JSD)* (Equation 1) is employed to compute a dissimilarity value between two places. V_{Yelp} and V_{FS} represent the topic signatures for a Yelp business and Foursquare venue respectively, $M = \frac{1}{2}(V_{Yelp} + V_{FS})$ and $KLD(V_{Yelp} \parallel M)$ and $KLD(V_{FS} \parallel M)$ are *Kullback-Leibler divergences* as shown in Equation 2.

$$JSD(V_{Yelp} \parallel V_{FS}) = \frac{1}{2}KLD(V_{Yelp} \parallel M) + \frac{1}{2}KLD(V_{FS} \parallel M) \quad (1)$$

$$KLD(P \parallel Q) = \sum_i P(i) \log_2 \frac{P(i)}{Q(i)} \quad (2)$$

The *JSD metric* is calculated by taking the square root of the value resulting from the divergence. Given the inclusion of the logarithm base 2, the resulting metric is bound between 0 and 1 with 0 indicating that the two venue topic distributions are identical and 1 representing complete dissimilarity. Computing the dissimilarity between each known Yelp business and its nearby Foursquare venues in V_T, produces a ranked set of POI from which a match can be extracted.

3.3 Weighted Multi-attribute Model

The above methods for matching POI are based on single attributes. In order to build a weighted model based on measured outcomes, a binomial probit regression model is used to estimate the overall contribution each attribute makes in correctly determining a match. The positioned rank for each distinct attribute, measured across all 140 venues are entered as independent variables to the model. The dependent "correct match" variable consists of either a 1 (match) or 0 (no match) for each pair of venues. The coefficients resulting from the model are normalized and applied as weights to a *regression-based weighted multi-attribute model.*

4. EVALUATION

The results shown in Table 1 demonstrate that a regression-based weighted attribute model outperforms each of the independent attribute models. Our approach ranks each Foursquare POI in V_T by its attribute-matching measure to the associated Yelp POI in V_M. Proceeding through the ranked list of items, the position of the actual attribute match is recorded. A perfect match would result in the correct Foursquare location matching to the top ranked POI in the attribute ranked set. A second position rank indicates that the attribute model chose the "correct Foursquare POI" in V_M as its second most likely match, and so on.

As one can see both the Levenshtein and Double Metaphone methods performed quite well with geographic distance producing the least accurate matching and Topic Modeling showing close to 62% accuracy. These results contradict the widely made assumption that proximity of POI is a strong match indicator (i.e., that "geometry trumps semantics" when performing conflation tasks [2]).

Though each independent matching method exhibited excellent results, the model that merged these methods proved superior. This was based on a regression of the four independent attribute methods, producing remarkable match-accuracy of almost 98% correct matches across 140 POI. The regression-based weights are shown in Equation 3.

$$M_{reg} = 0.562Lev + 0.094DM + 0.170Dist + 0.174LDA \quad (3)$$

4.1 Validating the Model

In order to test the validity of this regression-based model, we randomly selected 100 new POI in the Yelp dataset that were manually matched to the same number of POI in the Foursquare venue set. The results of the evaluation of these test POI are shown in Table 2.

The four independent measures illustrate results similar to those seen in Table 1 with both *name matching* methods producing high levels of accuracy and the *Distance* and *LDA* attributes showing comparable accuracy percentages. The match-accuracy of the *Regression-weighted* model correctly matched **97%** of the 100 POI sample, validating the technique and proposed model.

5. CONCLUSIONS AND OUTLOOK

In this work we addressed the problem of matching Points of Interest from the location-based social network Foursquare and the local directory service Yelp. Conflation of these datasets is very attractive from a research perspective. There is sufficient overlap between the attributes

Position	Levenshtein	D Metaphone	Distance	LDA	Weighted Model
1	86.5	85.8	51.8	61.7	97.9
2	3.5	4.3	13.5	14.2	0.7
3	2.1	2.1	11.3	5.0	0.7
4	1.4	0.0	4.3	2.8	0.0
5	0.7	0.0	5.0	2.1	0.0

Table 1: Independent attribute and weighted method by percentage of top five ranked positions.

Position	Levenshtein	D Metaphone	Distance	LDA	Weighted Model
1	76.0	87.0	51.0	63.0	97.0
2	5.0	7.0	16.0	14.0	1.0
3	3.0	0.0	11.0	6.0	0.0
4	0.0	1.0	4.0	2.0	1.0
5	2.0	0.0	5.0	2.0	0.0

Table 2: The independent attribute matchers by percentage match with the weighted multi-attribute model.

stored by Yelp and Foursquare to support matching, and enough differences to justify the effort of conflation. The presented work focuses on the first step of POI conflation, namely identifying whether two information entities refer to the same place in the physical world. In the presented work, we demonstrated a weighted multi-attribute matching strategy that can successfully match **97%** of randomly selected Yelp POI to their corresponding Foursquare entities. This approach has shown that the distance between matched POI from different providers can be substantial and matching points based on geographic location alone is often imprudent. We touched on some of the reasons why there may be discrepancy in user-generated locations and how this discrepancy varies across providers.

Future work in this area will involve enhancing our model to match additional user-generated and non-user-generated POI datasets. There are an enormous amount of geographically referenced data publicly available online and identifying the same POI in different datasets is a substantial step forward in the aspiration of POI data conflation. While this paper makes use of the more prominent properties of POI, additional attributes can and should be exploited. For example, semi-structured price-range values and user-contributed *star* rankings. Finally, in the future we also plan to look into the second step of conflation, namely how to merge attribute *values*.

6. REFERENCES

[1] B. Adams and K. Janowicz. On the geo-indicativeness of non-georeferenced text. In J. G. Breslin, N. B. Ellison, J. G. Shanahan, and Z. Tufekci, editors, *ICWSM*, pages 375–378. The AAAI Press, 2012.

[2] B. Adams, L. Li, M. Raubal, and M. F. Goodchild. A general framework for conflation. *Extended Abstracts Volume, GIScience 2010*, 2010.

[3] J. Bao, Y. Zheng, and M. F. Mokbel. Location-based & preference-aware recommendation using sparse geo-social networking data. In *ACM SIGSPATIAL*, 2012.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] C.-C. Chen, C. A. Knoblock, and C. Shahabi. Automatically conflating road vector data with orthoimagery. *GeoInformatica*, 10(4):495–530, 2006.

[6] Z. Cheng, J. Caverlee, and K. Lee. A content-driven framework for geolocating microblog users. *ACM Transactions on Intelligent Systems and Technology*, 4(1):2, 2013.

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

[8] J. Hastings. Automated conflation of digital gazetteer data. *International Journal of Geographical Information Science*, 22(10):1109–1127, 2008.

[9] C. B. Jones and R. S. Purves. Geographical information retrieval. *Int. Journal of Geographical Information Science*, 22(3):219–228, 2008.

[10] A. Lait and B. Randell. An assessment of name matching algorithms. *Technical Report Series-University of Newcastle Upon Tyne Computing Science*, 1996.

[11] L. Li and M. F. Goodchild. An optimisation model for linear feature matching in geographical data conflation. *International Journal of Image and Data Fusion*, 2(4):309–328, 2011.

[12] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, and W.-Y. Ma. Mining user similarity based on location history. In *ACM SIGSPATIAL 2008*, page 34. ACM, 2008.

[13] A. K. McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[14] G. McKenzie, B. Adams, and K. Janowicz. A thematic approach to user similarity built on geosocial check-ins. In *Proceedings of the 2013 AGILE Conference*, 2013.

[15] C. Mülligann, K. Janowicz, M. Ye, and W.-C. Lee. Analyzing the spatial-semantic interaction of points of interest in volunteered geographic information. *Spatial Information Theory*, pages 350–370, 2011.

[16] Y. Wu and S. Winter. Inferring relevant gazetteer instances to a placename. In *10th International Conference on GeoComputation. UNSW, Sydney, Australia*, 2009.