

Searching for Data Sources for the Semantic Enrichment of Trajectories

Luiz André P. Paes Leme¹, Chiara Renso², Bernardo P. Nunes^{3,4},
Giseli Rabello Lopes⁵, Marco A. Casanova³, and Vânia P. Vidal⁶

¹ Fluminense University, Niterói/RJ, Brazil
lpaesleme@ic.uff.br

² ISTI-CNR, Pisa/PI, Italy
chiara.renso@isti.cnr.it

³ PUC-Rio, Rio de Janeiro/RJ, Brasil
{bnunes,casanova}@inf.puc-rio.br

⁴ Federal University of the State of Rio de Janeiro, RJ, Brazil
bernardo.nunes@uniriotec.br

⁵ Federal University of Rio de Janeiro, RJ, Brazil
giseli@dcc.ufrj.br

⁶ Federal University of Ceará, Fortaleza/CE, Brasil
vvidal@lia.ufc.br

Abstract. The fast growing number of datasets available on the Web inspired researchers to propose innovative techniques to combine spatio-temporal data with contextual data. However, as the number of datasets has increased relatively fast, finding the most appropriate datasets for enrichment also became extremely difficult. This paper proposes an innovative approach to rank a set of datasets according to the likelihood that they contain relevant enrichments. The approach is based on the intuition that the sequence of places visited during a trajectory can induce the best datasets to enrich the trajectory. It relies on a supervised approach to learn rules of association between visited places and meaningful datasets.

Keywords: trajectories, semantic enrichment, movement data

1 Introduction

The personal position-enabled mobile devices are becoming our companions in everyday life, leaving tracks of our movements during our daily routine. The tracks collected by mobile devices describe the so-called *raw trajectories* that represent the geometric facets of movement data. Social media have also been proposed as complementary sources of mobility data. Georeferenced social media can be used as sparse and freely annotated movement traces [2,12] or, possibly, can be used to enrich raw GPS data thus getting semantically richer data with high positional accuracy [5].

The approach presented in this paper tackles the problem of searching the most appropriate datasets to enrich mobility data. It is based on the intuition

that the sequence of places visited during a movement, i.e., the sequence of stops, can induce the purpose of the movement and hence suggest the set of datasets for enrichment. For example, assume that a traveler visits the sequence of places [hotel, stadium, restaurant, hotel] in Rio de Janeiro. Also assume that the dataset of tourist attractions available in the Open Data Portal of the government of the city of Rio de Janeiro contains data about the Maracanã stadium. The sequence of places suggests that the person can be a tourist because tourists frequently stay in hotels and visit the Maracanã Stadium in Rio de Janeiro. Therefore, one could attempt to match the place labeled as stadium with the entry *Maracanã stadium* in the dataset. It is important to notice that this is not a deterministic problem that could be solved with an a priori rule such as `if a person visited a stadium then search for enrichments in the dataset of attractions` since there is no obvious evidence, for someone who doesn't know the content of the dataset, that the dataset of attractions would contain an entry that could be matched with the place stadium. However, this can be learnt from previous trajectory enrichments: if most trajectories similar to this one, in terms of the places visited, are enriched with the dataset of attractions then one can select that dataset as a potential source of enrichment for the new trajectory.

In this paper we take advantage of social media traces of movement and their user annotations to propose a technique for searching potentially useful datasets for the enrichment of trajectories. As for related work, the process of semantic enrichment of spatial and spatiotemporal data can be automatic [2,5,10] or semi-automatic [8]. Automatic approaches can use machine learning techniques such as Hidden Markov Models [10,12], probabilistic models [7], similarity measures [2] or simple proximity heuristics [2] to attach annotations. Recent techniques have also stressed the relevant role of the emerging and fast growing Web of Data [3] in the enrichment process. All existing works have used predefined sets of sources. Developers have favored popular sources such as DBpedia, Open Street Map, Open Weather Map, etc. and neglected less popular ones such as government open data and domain specific datasets. The fundamental reason for that is the lack of techniques to crawl and search for potentially useful datasets for enrichment.

The rest of this paper is organized as follows. Section 2 introduces the basic concepts used throughout the paper and describes the proposed ranking technique. Section 3 addresses the preparation of the test dataset. Section 4 presents the experiments for assessing the technique and Section 5 contains the conclusions.

2 The problem of searching for sources of enrichments

A *raw trajectory* of a moving object o is a sequence $\rho_o = (p_1, p_2, \dots, p_n)$ of spatio-temporal points such that the timestamp of p_i is earlier than the timestamp of p_{i+1} . A *segment* g of a raw trajectory ρ_o is a continuous subsequence of ρ_o . A *segmented trajectory* of a raw trajectory ρ_o is a sequence $\sigma_o = (g_1, g_2, \dots, g_n)$ of

segments of ρ_o such that $s = g_1 \parallel \dots \parallel g_n$, that is, s is the concatenation of g_1, \dots, g_n . A segment of a raw trajectory is a fragment of the whole raw trajectory where a given property holds.

The notion of semantic trajectory goes further and enriches a segmented trajectory with contextual information retrieved from external datasets. A *contextual resource* r of a dataset d is a pair (r, d) with $r \in d$. We use the notion r^d rather than (r, d) . A *contextual information* of a segment g of a segmented trajectory σ_o , denoted by c , is a set, of contextual resources $c = \{r_1^{d_1}, \dots, r_n^{d_n}\}$. In this way, we say that c *enriches* g . Intuitively, a contextual information is a set of resources that can be used to describe a trajectory. A *semantic trajectory* for a segmented trajectory σ_o is a sequence $\tau_o = (\langle g_1, c_1 \rangle, \dots, \langle g_n, c_n \rangle)$, such that $\langle g_i, c_i \rangle$ is a pair indicating that g_i is enriched with contextual information c_i .

We also define a particular kind of enriched trajectory, called *labeled trajectories*. Labeled trajectories arise from mobility data captured from social media. We define labeled trajectories as follows. A *labeled trajectory* for a segmented trajectory σ_o is a sequence $\lambda_o = (\langle g_1, l_1 \rangle, \dots, \langle g_n, l_n \rangle)$, such that $\langle g_i, l_i \rangle$ is a pair indicating that segment g_i is enriched with a set l_i of labels.

Given a labeled trajectory $\lambda_o \in \Lambda$ of a segmented trajectory σ_o and a set D of available datasets, generate a list $R = [d_1, \dots, d_n]$ of datasets such that $d_i \in D$ and d_i likely contains the resources for the semantic enrichment of σ_o . The list should be ranked according to the likelihood that a dataset contains semantic enrichments for σ_o . More formally, let

- i. Σ be a set of segmented trajectories
- ii. Λ be a set of labeled trajectories of the trajectories in Σ
- iii. T be a set of semantic trajectories of the trajectories in Σ
- iv. Δ be the set of datasets of the contextual resources of the trajectories in T
- v. P be an assessment function that estimates the likelihood that a dataset d_i contains enrichments for $\sigma_o \in \Sigma$ with respect to $\lambda_o \in \Lambda$.

One wants to find a ranking function $rank : \Lambda \mapsto \bigcup_{n=1}^{\infty} \Delta^n$ such that if $rank(\lambda_o) = [d_1, \dots, d_n]$ then $P(\lambda_o, d_i) > P(\lambda_o, d_{i+1})$, for $i = 1, \dots, n - 1$. We segment trajectories with the stop-and-move strategy [11] and label each segment with taxonomic classifications of the place visited at the end of the segment. We cast the problem as a supervised multi-class classification problem. If one takes the set of available datasets as classes of trajectories, one can induce a ranking function as follows. A *classification model* is a function $C : \Lambda \mapsto \bigcup_{n=1}^{\infty} (\Delta \times \mathbb{R})^n$ that assigns each labeled trajectory λ_o to a list with n pairs (d, s) , where $d \in \Delta$ is a dataset and s is the *assessment score* of d , represented by a Real number. Let \mathcal{C} be the set of all classification models. Let $2^{\Lambda \times T}$ be the set of sets of pairs (λ_o, σ_o) . Intuitively, $\Theta \in 2^{\Lambda \times T}$ is a set of pairs (λ_o, σ_o) , where λ_o is a labeled trajectory and σ_o is a semantic trajectory, such that the pairs in Θ will be used for training a classification model. Then, we introduce the function *Modeling* : $2^{\Lambda \times T} \mapsto \mathcal{C}$ to represent a machine-learning-based process that takes as input sets of pairs of labeled trajectories and that corresponding semantic trajectories, called a *training set*, and outputs a classification model. Finally, the *ranking function induced by a classification model* C is defined as

$rank(\lambda_o) = sortDescending(C(\lambda_o))$, where *sortDescending* sorts pairs by the second coordinate in descending order.

As for the features of trajectories, we tested four types of sets: the set of labels of the places visited in a trajectory (W_{λ_o}), e.g. {Residence, Law School, Pizza Place}, boolean model of the set of labels (X_{λ_o}), as used by Information Retrieval (IR) techniques, the set of all valid sequences of labels visited by a trajectory (Y_{λ_o}), e.g. {(Residence, Law School, Pizza Place), (Residence, Pizza Place), (Residence, Law School), ...}, and the boolean model of the sequences of labels (Z_{λ_o}), also as in IR.

3 Dataset preparation

This section describes the preparation of the dataset used as training data to validate the proposed technique, i.e., the set of labeled and semantic enriched trajectories. We used a set of 9,594,421 geolocated tweets, between June and July 2014 generated in the city of Rio de Janeiro, as trails of movement of people during the FIFA World Cup 2014. A trajectory is defined as the movement of one person between 4:00 AM and 4:00 AM of the next day. There were 912,643 trajectories with 11 samples (tweets) on the average. Each trajectory was segmented using a stop/move heuristic, labeled with place check-ins and semantically enriched with entities from a set of datasets available on the Web.

Labeled Trajectories - Highly dense sampled trajectories are usually segmented using the *speed* and *minimal stop time* criteria. However, low density trajectories, like social media tracks, are not suitable for this kind of segmentation due to the impossibility of correctly computing the speed. We adopted a simpler heuristic for segmenting low density trajectories, yet following the stop-and-move strategy.

The segmentation is based on the intuition that if the time interval between two consecutive tweets is above a given threshold, the user might have moved from one position to another and, therefore, there would be a move segment from the position of the first tweet to the position of the second tweet. On the other hand, if the time interval is short, the user might be stopped or on the move. This last condition is justified because some mobile applications checks-in users automatically. In some cases, it was observed a series of consecutive tweets with short intervals of time and space, giving the idea that the user was on the move. All consecutive tweets considered to be part of the same move segment can be merged into a single segment, while the tweets in the same static position can be merged with the previously identified segment.

Recall from Section 2 that the modeling process receives as input a set of labeled trajectories. The trajectories were labeled with the categories of the places visited by users and enriched with the entities from the datasets of Table 1. The places visited by users were captured from Foursquare check-ins made available through tweets. Each Foursquare check-in contains metadata about the place which includes its classification according to the Foursquare taxonomy. This tax-

onomy is a three-level hierarchy that has, at the highest level, general categories, such as **Food** and **Event**. On the other hand, the lowest level is a very specific classification that contain, for example, the categories **Preschool** and **Private School**. Both levels, however, would lead to a poor discrimination regarding to the class association. The classification model would either over-associate trajectories with classes or discard some associations. Therefore, the trajectories were enriched with the intermediary level of classification, such as **Breakfast Spot**, **Coffee Shop** and **Beach**. The labeling procedure labels segments with information about the place at the end of the segment.

It is important to remark that only trajectories that visited three or more places were selected. It seems not make sense to classify trajectories with small sets of visited places since the induced purpose of the trajectories might be hidden. Therefore, we empirically considered trajectories with three or more places. So, the total number of trajectories was reduced from 912,643 to 8,730.

Semantic Trajectories - Regarding semantic enrichment, we used datasets (Table 1) made available through the open data portals <http://dados.gov.br> (Portal Brasileiro de Dados Abertos - ODBr) and <http://data.rio.rj.gov.br> (Portal de Dados abertos da Prefeitura do Rio - ODRio). The enrichment process was semi-automatic and matched the metadata of the places visited by the users (captured from the Foursquare check-ins) with the metadata of entities contained in each dataset.

Table 1. Datasets used for semantic enrichments of the trajectories.

dataset URI	source	alias
http://dados.gov.br/dataset/instituicoes-de-ensino-basico	ODBr	schools
http://dados.gov.br/dataset/instituicoes-de-ensino-superior	ODBr	universities
http://data.rio.rj.gov.br/dataset/pontos-turisticos-e-culturais	ODRio	attractions
http://data.rio.rj.gov.br/dataset/hoteis	ODRio	hotels
http://data.rio.rj.gov.br/dataset/museus	ODRio	museums
http://data.rio.rj.gov.br/dataset/teatros	ODRio	theaters
http://data.rio.rj.gov.br/dataset/estabelecimentos-de-saude	ODRio	hospitals
http://data.rio.rj.gov.br/dataset/unidades-administrativas	ODRio	offices

The matching process consisted in computing a similarity measure between places p and entities e and manually deciding the matchings. The similarity function used is defined as follows.

$$sim_N(p, e) = 1 - \frac{levenshteinDistance(p[name], e[name])}{p[name].length + e[name].length} \quad (1)$$

$$sim_G(p, e) = \frac{1}{(0.19 \cdot geoDistance(p[position], e[position]) + 1)} \quad (2)$$

$$sim(p, e) = harmonicMean(sim_N(p, e), sim_G(p, e)) \quad (3)$$

The sim_G is defined such that $sim_G = 1$ for $distance = 0$, $sim_G = 0$ for $distance \rightarrow \infty$ and $sim_G = 0.05$ for $distance = 100m$. A place and an entity are matching candidates iff $sim(p, e)$ is greater than 0.95. The matching task only aimed at providing a Gold Standard.

4 Experiments

The main goal of the experiments was to assess the performance of the ranking function. As for the performance measure, the experiments computed the Mean Average Precision (MAP) of the ranking. The next subsections describe the creation of the classification model, the ranking function and the ranking assessment.

Classification model - We investigated different classification algorithms and concluded that the best classification function is a combination of binary classifiers using the JRip algorithm [1]. Table 2.a compares the F-Measure of different classification algorithms, JRip, J48 [6], OneR [4], ConjunctiveRule [9] and DecisionStump [9], with respect to a positive classification for the classes **offices**, **theaters**, **hotels**, **hospitals**, **museums**, **attractions**, **ies**, **schools**. This experiment used the set of valid sequences of places (Z_{λ_o}) for the features of trajectories. As we show, none of the algorithms statistically improves the performance of the reference algorithm (JRip). The statistic significance was determined by the paired T-Test method using a set of 10 randomly partitions of the test dataset of the type 2/3 for training set + 1/3 for test set.

Table 2. F-measure of classification algorithms.

a) Using binary vector of sequences of places.						b) Multi-class version of JRip binary vector of sequences of places	
Dataset	(1)	(2)	(3)	(4)	(5)	Class	F-Measure
offices	0.66	0.10	0.00	• 0.00	• 0.00	offices	0.31
theaters	0.82	0.63	0.66	0.66	0.66	theaters	0.67
hotels	0.14	0.03	• 0.03	• 0.00	0.00	hotels	0.06
hospitals	0.53	0.32	0.26	• 0.00	• 0.00	hospitals	0.33
museums	0.58	0.28	• 0.16	0.00	• 0.00	museums	0.63
attractions	0.69	0.69	0.53	• 0.62	0.53	attractions	0.65
universities	0.82	0.81	0.78	0.73	0.78	universities	0.79
schools	0.72	0.71	0.71	0.68	0.71	schools	0.69
Average	0.62	0.45	0.39	0.34	0.34	None	0.90
						Average	0.56

o, • statistically significant improvement or degradation

(1) rules.JRip '-F 3 -N 2.0 -O 6 -S 1'

(2) trees.J48 '-C 0.25 -M 2'

(3) rules.OneR '-B 6'

(4) rules.ConjunctiveRule '-N 3 -M 2.0 -P -1 -S 1'

(5) trees.DecisionStump ''

The low performance of the classification with respect to the dataset **hotels** can be explained by its generality. That is, a **hotel** can be a stop in several different sequences of places, which makes it more difficult to find a pattern of correlation between trajectories and datasets. The average performance of JRip was 62%. Table 2.b shows the performance of the multi-class version of JRip, which has an average performance of 56%. The binary classifiers, therefore, proved to be more efficient. Tables 3.a and 3.b show the performance measures using the set of features in Definitions W_{λ_o} and X_{λ_o} . The best performance was achieved

with the JRip algorithm using the categories of the places visited by a trajectory (Definition X_{λ_o}), which was 61%. These results corroborate the intuition that using sequences of places is more discriminating. For example, a sequence of places (Definition Z_{λ_o}) such as [Residence, School, Residence] could indicate that the person is a Student, while a sequence [Residence, School, School, School, Residence], if the schools are different, could indicate that the person is, for example, a professional delivery boy. Both cases, however, have the same set of features. In the first example, the enrichment with a dataset of schools would make sense, while in the last one it seems not to be the case.

Table 3. F-measure of classification algorithms.

a) Using sets of categories as features.					b) Using binary vector of categories as features.				
Dataset	(1)	(2)	(3)	(4)	Dataset	(1)	(2)	(3)	(4)
offices	0.17	0.07	0.07	0.00	offices	0.00	0.00	0.53	o 0.57
theaters	0.62	0.53	0.53	0.00	theaters	0.67	0.67	0.86	0.86
hotels	0.00	0.10	0.10	0.00	hotels	0.26	0.26	0.12	0.14
hospitals	0.27	0.08	o 0.08	o 0.00	hospitals	0.28	0.28	0.52	o 0.45
museums	0.44	0.27	0.27	0.00	museums	0.24	0.24	0.72	o 0.64
attractions	0.61	0.66	0.66	0.00	attractions	0.68	0.68	0.69	0.74
universities	0.77	0.62	o 0.62	o 0.00	universities	0.69	0.69	0.83	o 0.83
schools	0.71	0.58	0.58	0.00	schools	0.68	0.63	0.69	o 0.69
Average	0.45	0.36	0.36	0	Average	0.43	0.43	0.62	0.62

o, • statistically significant improvement or degradation

- (1) bayes.NaiveBayesUpdateable "
- (2) bayes.NaiveBayes "
- (3) rules.JRip '-F 9 -N 2.0 -O 6 -S 1'
- (4) trees.J48 '-C 0.25 -M 2'

Ranking assessment - Rankings were generated, as before, using a set of 10 randomly partitions of the enriched dataset such that 2/3 were used for training set and 1/3 for test set. We used sets of binary classifiers, one for each dataset, based on JRip algorithm which is a rule-based classifier that, while trained, generates classification rules such as

Rule: if the a person visited a place of type **States & Municipalities** and did not visit a **Residence** and moved from a place of type **States & Municipalities** to a place of type **Food** then classify trajectory as **attraction**

The training step computes, for each rule, its precision, recall and F-measure. We used the F-measure as an estimate of the confidence of the rule, since intuitively the higher the precision and recall are, the higher the confidence on the classification will be. The confidence, therefore, was used as the *assessment score* (Section 2) output by the *classification model*.

To assess the ranking function we computed the Mean Average Precision (MAP) of the rankings of a set of trajectories. We assessed the ranking function on 2,508 trajectories out of the 8,730 trajectories available in the dataset. These trajectories had 1.5 relevant datasets on the average and the computed MAP was 66%, which means that one would need, on the average, just the three top most entries of the rank to find two datasets for enrichments.

5 Conclusions

This work proposes a novel approach for finding datasets for semantic enrichment based on the types of places visited. The technique takes advantage of place check-ins available on social networks to identify the sequences of places. It is a supervised approach that uses a set of semi-automatically enriched trajectories to learn correlations between the places visited and the datasets available for enrichment. We investigated different classification algorithms and different sets of features for the trajectories. The best performance was obtained with the JRip algorithm and sets of features that contain all possible sequences of places for a trajectory. The resulting ranks obtained, on the average, a MAP of 66% in the experiments.

Acknowledgements: This work has been funded by CNPq/BR and FAPERJ under grants E-26-170.028/2008, 557128/2009-9, 248743/2013-9, 248987/2013-5, 303332/2013-9, 442338/2014-7, 444976/2014-0 and E-26-201.337/2014.

References

1. Cohen, W.W.: Fast Effective Rule Induction. In: the 12th Int'l. Conf. on Machine Learning. pp. 115–123. Morgan Kaufmann (1995)
2. Fileto, R., Krüger, M., Pelekis, N., Theodoridis, Y., Renso, C.: Baquara: A Holistic Ontological Framework for Movement Analysis Using Linked Data. In: the 32nd Int'l. Conf. on Conceptual Modeling. pp. 342–355. Springer (Nov 2013)
3. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space, vol. 1. Morgan & Claypool, San Rafael, CA (Feb 2011)
4. Holte, R.C.: Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. *Machine Learning* 11(1), 63–91 (1993)
5. Nabo, R.G.B., Fileto, R., Nanni, M., Renso, C.: Annotating Trajectories by Fusing them with Social Media Users Posts. In: the XI Brazilian Symposium on Geoinformatics. pp. 25–36 (2014)
6. Quinlan, J.R.: *C4.5: Programs for Machine Learning*. Morgan Kaufmann (1993)
7. Spinsanti, L., Celli, F., Renso, C.: Where you stop is who you are: understanding peoples' activities. In: the 5th workshop on behaviour monitoring and interpretation—user modelling (2010)
8. Uzun, A.: Linked crowdsourced data - Enabling location analytics in the linking open data cloud. In: 2015 IEEE Int'l. Conf. on Semantic Computing. pp. 40–48. IEEE (2015)
9. Witten, I.H., Frank, E., Hall, M.A.: *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann (Jan 2011)
10. Yan, Z., Chakraborty, D., Parent, C., Spaccapietra, S., Aberer, K.: SeMiTri: a framework for semantic annotation of heterogeneous trajectories. In: The 14th Int'l. Conf. on Extending Database Technology. pp. 259–270 (2011)
11. Yan, Z., Chakraborty, D., Parent, C., Spaccapietra, S., Aberer, K.: Semantic trajectories: Mobility data computation and annotation. *Transactions on Intelligent Systems and Technology* 4(3), 49 (2013)
12. Yuan, J., Liu, X., Zhang, R., Sun, H., Guo, X., Wang, Y.: Discovering Semantic Mobility Pattern from Check-in Data. In: the 15th Int'l. Conf. on Web Information System Engineering. pp. 464–479. Springer, Cham (2014)