Towards Knowledge-Enriched Path Computation

Georgios Skoumas[◇] Klaus Arthur Schmid[♠] Gregor Jossé[♠]
Andreas Züfle[♠] Mario A. Nascimento[♡] Matthias Renz[♠] Dieter Pfoser[♣]
[◇]National Technical University of Athens, Greece
gskoumas@dblab.ece.ntua.gr
[♠]Ludwig-Maximilians-Universität München
{schmid,josse,zuefle,renz}@dbs.ifi.lmu.de
[♡]University of Alberta, Canada
nascimento@ualberta.ca
[♣]George Mason University, USA
dpfoser@gmu.edu

ABSTRACT

Directions and paths, as commonly provided by navigation systems, are usually derived considering absolute metrics, e.g., finding the shortest path within an underlying road network. With the aid of crowdsourced geospatial data we aim at obtaining paths that do not only minimize distance but also lead through more popular areas using knowledge generated by users. We extract spatial relations such as "nearby" or "next to" from geo-textual travel blogs, that define closeness between pairs of points of interest (POIs) and quantify each of these relations using a probabilistic model. Using Bayesian inference, we obtain a probabilistic measure of spatial closeness according to the crowd. Applying this measure to the corresponding road network, we derive an altered cost function taking crowdsourced spatial relations into account. We propose two routing algorithms on the enriched road networks. To evaluate our approach, we use Flickr photo data as a ground truth for popularity. Our experimental results – based on real world datasets – show that the computed paths yield competitive solutions in terms of path length while also providing more "popular" paths, making routing easier and more informative for the user.

Categories and Subject Descriptors

H.2.8 [**DATABASE MANAGEMENT**]: Database Applications—*Spatial databases and GIS*

General Terms

Algorithms, Experimentation, Performance

Keywords

Crowdsourced geospatial data, Enriched graphs, Routing

1. INTRODUCTION

User-contributed content has benefited many scientific disciplines by providing a wealth of new data. Technological progress, especially smartphones and GPS receivers, have facilitated contributing to the plethora of available information. OpenStreetMap[1] constitutes the standard example and reference in the area of volunteered geographic information. Authoring geospatial information typically implies coordinate-based *quantitative data*. The broad mass of users contributing content, however, are much more comfortable using *qualitative information*. People typically do not use geographic coordinates to describe their spatial motion, for instance when travelling or roaming. Instead, they use qualitative information in the form of toponyms (landmarks) and spatial relationships ("near", "next to", "close by", etc.). Hence, there is an abundance of qualitative geospatial information (freely) available on the internet, e.g., in travel blogs, largely unused. In contrast to quantitative information, which is mathematically measurable (although sometimes flawed by measurement errors), qualitative information is based on personal cognition. This is of particular interest when considering the "routing problem" (equivalent to "path finding"). Traditional routing queries use directions from systems that only take the structure of the underlying road network into account. In human interaction such information is usually enhanced with qualitative information (e.g. "the bridge next to the Eiffel tower"). Combining traditional routing algorithms with crowdsourced geospatial references we aim to more properly represent human perception while making it mathematically measurable. For that purpose, we enrich a road network with information about spatial relations between pairs of points of interest (POIs) extracted from textual travel blog data. Using these relations, we obtain routes that are easier to interpret and to follow, much rather resembling a route that a person would provide.

As an example, consider the routing scenario in Figure 1 which is set in the city of Paris, France. The continuous line represents the conventional shortest path from starting point "Gare du Nord" to the target at "Quai de la Rapée" –while the dot and dashed lines represent alternative paths computed by the algorithms introduced in this paper. The triangles in this example denote touristic landmarks. For instance, the dashed path on the bottom right passing recognizable locations such as "Place de la République", "Cirque d'hiver" and "la Bastille", as proposed by our algorithms, is considerably easier to describe and follow, and might yield more interesting sights for tourists than the shortest path.

The challenge of this work is to extract crowdsourced information from textual data and incorporate it, in an existing road network. This *enriched* road network is subsequently used to provide paths, between given arbitrary pairs of start and target nodes, that

SIGSPATIAL '14:, November 04 - 07 2014, Dallas/Fort Worth, TX, USA.

ACM 978-1-4503-3131-9/14/11 ...$15.00. http://dx.doi.org/10.1145/2666310.2666485.

[1] `https://www.openstreetmap.org/`

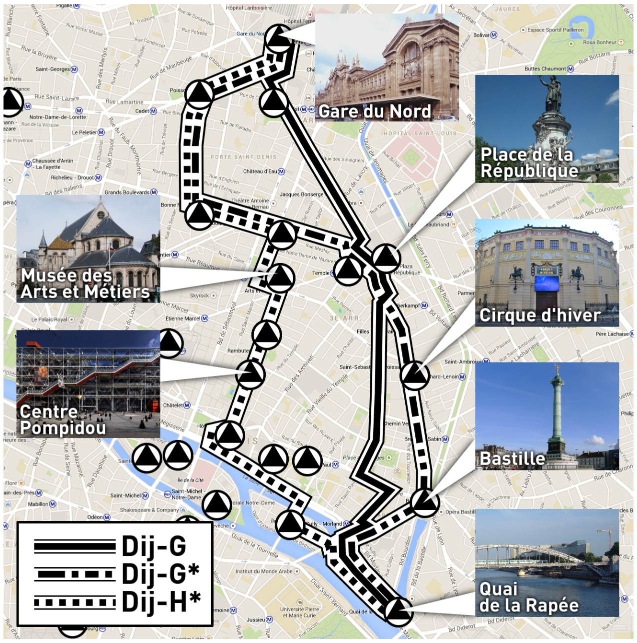

Figure 1: Short and popular paths in Paris.

satisfy the claim of higher popularity[2] while only incurring a minor additional spatial distance.

2. RELATED WORK

Research areas relevant to this work include: (i) qualitative routing and (ii) mining of semantic information from moving object trajectories. The authors in [1] try to tackle the problem of efficient routing by using cost functions that trade off between minimizing the length of a provided path while also minimizing the complexity of the provided path in terms of instructions given. The discovery of semantic places through the analysis of raw trajectory data has been investigated thoroughly over the course of the last years. The authors in [4] provide solutions for the semantic place recognition problem and categorize the extracted PoIs into pre-defined types. Finally, the authors in [6] propose an approach for route recommendation based on a user opinion mining platform about trajectories, providing longer but more pleasant routes. The major drawback of these approaches is that they focus almost exclusively on road network data without taking into account any kind of qualitative information, i.e., information coming from the user. Additionally, they do not integrate the extracted semantic information into the road network. Instead, they use the extracted information to only enrich specific trajectories.

3. SPATIAL RELATION EXTRACTION

In this work, we choose travel blogs as a rich potential source for (crowdsourced) geospatial data. This selection is based on the fact that people tend to describe their experiences in relation to their trips and places they have visited, which results in "spatial" narratives. To gather such data, we use classical Web crawling techniques and compile a database consisting of 120,000 texts, obtained from travel blogs[3]. Obtaining qualitative spatial relations from text involves the detection of (i) POIs (or toponyms) and (ii) spatial relationships linking the POIs. The employed approach involves geoparsing, i.e., the detection of candidate phrases, and geocoding, i.e., linking these phrases to actual coordinate information. Using the Natural Language Processing Toolkit (NLTK) (cf. [3]), a leading platform for analyzing raw natural language data, we managed to extract 500,000 POIs from the text corpus. For the geocoding of the POIs, we rely on the GeoNames[4] geographic gazetteer data, which contains over ten million POI names worldwide and their coordinates. This procedure associates (whenever possible) POIs found in the travel blogs with geographic coordinates. Using the GeoNames gazetteer we were able to geocode about 480,000 out of the 500,000 extracted POIs.

Having identified and geocoded the spatial objects, the next step is the extraction of qualitative spatial relationships. The extraction of spatial relations between entities in text is a hard Natural Language Processing (NLP) problem, especially when applied to a noisy crowdsourced dataset. We address this NLP challenge by implementing a spatial relation extraction algorithm based on NLTK [3] components in combination with predefined strings and syntactical patterns. More specifically, we define a set of language expressions that are typically used to express a spatial relation in combination with a set of syntactical rules. The use of both syntactical and string matching reduces the number of false positives considerably. A formal algorithm describing the architecture of our proposed information extraction system can be found in [8] and [9].

The spatial relation extraction procedure results in a *relationship graph*, where edges represent one or more spatial relations and vertices represent POIs. For the scope of this work, we have collected data for the region of Paris which will be our main dataset during the experimental evaluation of the proposed approach. Here, we should point out that in the context of this work, i.e., computation of a combination of short and enriched routes, we only consider distance and topological relations that denote closeness ("near", "close", "next to", "at", "in" etc). The use of relations that denote direction, e.g., "north", "south", "east" etc. or remoteness, e.g., "away from", "far" etc., is an open direction for future work.

4. MODELING SPATIAL RELATIONS

In this section we briefly describe the probabilistic modeling we follow in order to quantify qualitative spatial relations between POIs. We model a spatial relation between two POIs P_i and P_j based on *distance* and *orientation* features as presented in [7]. For each spatial relation occurrence, we create a two-dimensional spatial feature vector $D = (D_d, D_o)^\intercal$ where D_d denotes the distance and D_o denotes the orientation between P_i and P_j. This way, we end up with a set of two-dimensional spatial feature vectors $\mathcal{D}_{\text{rel}} = \{D_1, D_2, \ldots, D_n\}$ for each spatial relation. We employ Gaussian Mixture Models (GMMs) and we train a probabilistic model for each spatial relation based on the spatial feature vectors. In general, a GMM is a weighted sum of M-component Gaussian densities as $p(d|\lambda) = \sum_{i=1}^{M} w_i g(d; \mu_i, \Sigma_i)$ where d is a l-dimensional data vector (in our case $l = 2$), w_i are the mixture weights, and $g(d; \mu_i, \Sigma_i)$ is a Gaussian density function with mean vector $\mu_i \in \mathbb{R}^l$ and covariance matrix $\Sigma_i \in \mathbb{R}^{l \times l}$. To fully characterize the probability density function $p(d|\lambda)$, one requires the mean vectors, the covariance matrices and the mixture weights. These parameters are collectively represented in $\lambda = \{w_i, \mu_i, \Sigma_i\}$ for $i = 1, \ldots, M$. For the parameter estimation of spatial relation GMM, we use Expectation Maximization (EM). This approach is extensively analyzed in [7] and [8].

[2] Formally introduced in Section 6

[3] http://www.travelblog.com/
http://www.traveljournal.com/
http://www.travelpod.com/

[4] http://www.geonames.org/

5. ROAD NETWORK ENRICHMENT

In this section we describe how we transform a *relationship graph*, as presented in Section 3, into a weighted graph, and how we use the edge weights of the weighted graph in order to modify the edge costs of a real world road network. As presented in Section 3, the spatial relation extraction procedure results in a relationship graph between POIs. Let $\mathcal{P} = \{P_1, \dots, P_m\}$ denote the set of nodes representing the POIs and let $\mathcal{R} = \{R^1, \dots, R^n\}$ denote the predefined set of spatial closeness relations, represented by spatial NLP expressions like “next to” or “close by”. Furthermore, let $R_{i,j} \subseteq \mathcal{R}$ denote the set of relations observed (i.e., extracted from the text) between two distinct nodes P_i and P_j. Note that R^k denotes an abstract relation, while $R_{i,j}$ denotes a set of occurrences of relations between a pair of nodes. Let $D_{i,j}$ denote the spatial feature vector (distance and orientation), between two distinct POIs P_i and P_j. Finally, let $\mathcal{D} := \bigcup_{i \neq j \wedge R_{i,j} \neq \emptyset} D_{i,j}$ denote the set of all spatial feature vectors between all pairs of POIs which have non-empty sets of relations. We want to estimate the posterior probability of a class $R^k \in R_{i,j}$ based on the spatial feature data $D_{i,j}$ between two POIs P_i and P_j. This is given by $P(R^k|D_{i,j}) = \frac{P(D_{i,j}|R^k)P(R^k)}{\sum\limits_{l=1}^{n} P(D_{i,j}|R^l)P(R^l)}$. $P(D_{i,j}|R^k)$ denotes the likelihood of $D_{i,j}$ given relation R^k based on the trained GMM, while $P(R^k)$ denotes the prior probability of relation R^k given only the observed relations $R_{i,j}$.

In a traditional classification problem, the spatial relation R^k between a pair of POIs would be classified to the spatial relation model with the highest posterior. In contrast to this approach, we consider each posterior probability $P(R^k|D_{i,j})$ as a measure of confidence of the existence of relation R^k between P_i and P_j. Remember that all the relations we consider reflect terms of spatial closeness. We combine all these posteriors into one measure which we refer to as *closeness score* $\mathcal{W}_{i,j}$ of the pair of POIs P_i and P_j, defined as $\mathcal{W}_{i,j} = \frac{1}{|\mathcal{R}|} \cdot \sum\limits_{k=1}^{|R_{i,j}|} \frac{P(R^k|D_{i,j})}{\max_k\{P(R^k|\mathcal{D})\}}$. We refer to these pairs as *close* since at least one of our relations reflecting closeness exists. Assigning the respective weights $\mathcal{W}_{i,j} \in [0,1]$ to the edges of the relationship graph, we obtain a weighted graph. In this weighted relationship graph, denoted by H^*, there exists a vertex for each POI and an edge (P_i, P_j).

Now that we have extracted and statistically condensed the crowd-sourced data into a closeness score, we need to apply the obtained closeness scores to the underlying network. We have investigated several strategies and have decided upon a compromise between simplicity and effectiveness.

Let $G = (V, E, d)$ denote the graph representing the underlying road network, i.e., the vertices $v \in V$ correspond to crossroads, dead ends, etc., the edges $e \in E \subseteq V \times V$ represent roads connecting vertices. Furthermore, let $d : E \rightarrow \mathbb{R}_0^+$ denote the function which maps every edge onto its distance. We assume that $\mathcal{P} \subseteq V$, i.e., each POI is also a vertex in the graph. This is only a minor constraint since we can easily map each POI to the nearest node of the graph or introduce pseudo-nodes.

Now, for each pair of spatially connected POIs, P_i, P_j, we compute the shortest path connecting P_i and P_j in G, which we denote by $p(i,j)$. We then define a new cost function $c : E \rightarrow \mathbb{R}_0^+$ which modifies the previous cost $d(e)$ of an edge as $c(e) := d(e) \cdot \prod_{e \in p(i,j)} (1 - \alpha \mathcal{W}_{i,j})$. Here, $e \in p(i,j)$ iff e is an edge within the shortest path from P_i to P_j and where $\alpha \in [0,1]$ is a weight scaling factor to control the balance between the spatial distance $d(e)$ and the modification caused by the closeness score $\mathcal{W}_{i,j}$. In the case of $\alpha = 0$, we obtain the unadapted edge weight $c(e) = d(e)$. Summarizing, the more shortest paths between POI pairs run through e, the lower its adjusted cost $c(e)$ is. We now define the *enriched graph* $G^* = (V, E, c)$. It consists of the original vertices and edges and is equipped with the new cost function which implies the reweighting of edges. When computing the cost of a whole path within G^*, as before, we sum the respective edge weights which differ from the original edge weights (because of the altered cost function). In order to measure the influence of the adjusted cost values along a path $p = (e_1, \dots, e_r)$, we introduce the *enrichment ratio* (ER) function er defined as $er(p) = \frac{1}{d(p)} \sum_{i=1}^{r} c(e_i)$. By normalizing with the total length of the path, we are able to compare the spatial connectivity of paths independent of length as well as start and target nodes. Here, a lower ratio implies higher closeness score values along the edges of the path. If none of the edges of a path is part of any shortest path between POIs, its enrichment ratio is 1, while the (highly unlikely) optimal enrichment ratio is 0. On the enriched graph G^* we may now define our path finding algorithms.

In this work, all path computations are based on Dijkstra’s algorithm, because our main focus is not the routing itself but the incorporation of textual information into existing road networks. If desired, speed-up techniques, such as preprocessing steps and/or other search algorithms, could easily be employed. The first approach is denoted as Dij-G*. Given start and target nodes, it executes a Dijkstra search in the enriched road network graph G^* using the adjusted cost function. Our second approach, referred to as Dij-H*, uses the enriched road network graph G^* as well as the relationship graph H^*. Given start and target nodes within G^*, entry and exit nodes within H^* are determined. Subsequently, we route within H^*, i.e., from POI to POI. In Section 6, both algorithms are compared against shortest paths as obtained with Dijkstra’s algorithm, which we denote by Dij-G. An extended analysis of both algorithms can be found in [9].

6. EXPERIMENTAL EVALUATION

We compare the results of the conventional Dijkstra search, i.e., Dij-G, to the results of Dij-G*, which uses the enriched graph G^*, and the results of Dij-H*, which mainly relies on the relationship graph H^*. All the text processing parts were implemented in Python while modeling parts were implemented in Matlab. Network enrichment and path computation tasks were conducted using the Java-based MARiO Framework [2] on an Intel(R) Core(TM) i7-3770 CPU at 3.40GHz and 32 GB RAM running Linux (64 bit).

Besides comparing the computed path w.r.t. their enrichment ratio (ER) and length (as presented in Section 5), we introduce a measure of popularity based on Flickr data. Our experimental setup is based on the region of Paris, France. This region has comparatively high density of spatial relations, Flickr photo data, and OSM data, which accounts for an exact representation of the road network. Since ER is a measure introduced in this paper, we use Flickr data as an independent ground truth. We are aware that to cognitive aspects (like the importance of sights or the value of landmarks) there is no absolute truth. However, in order to be able to draw comparisons, we presume that if the dataset is large enough, the bias can be neglected. We use a geotagged Flickr photo dataset, provided by the authors in [5], to assign a number of photos to each vertex of the underlying road network. The number of Flickr photos assigned to each vertex is referred to *popularity*. In our settings, every photo which is within the 20-meter radius of a vertex, contributes to the popularity of that vertex. The popularity of a path is computed by summation of all popularity values along this path.

In our first experiment we examine the influence of different scalings of the closeness score $\mathcal{W}_{i,j}$ in terms of enrichment ratio, path length increase (distance) and popularity. Figure 2 (a)

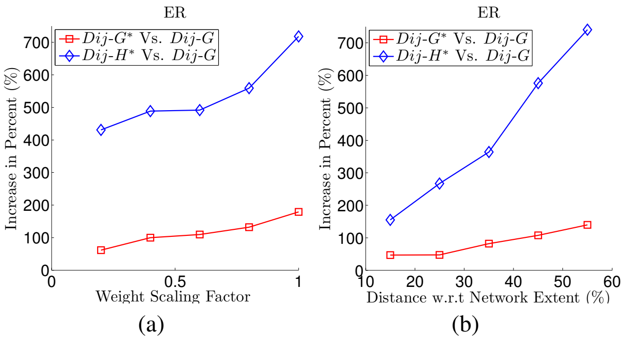

Figure 2: ER increase for algorithms Dij-G* and Dij-H*.

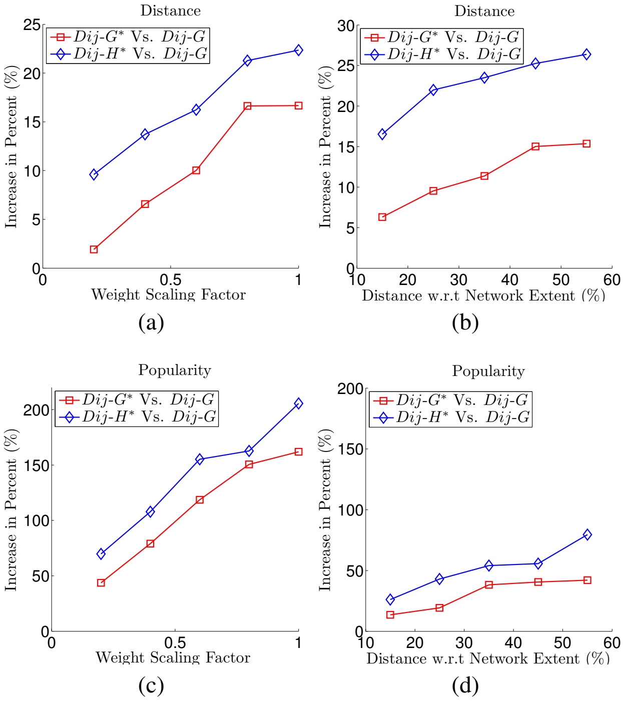

Figure 3: Popularity and distance trade-off.

shows the influence of the different scalings of $\mathcal{W}_{i,j}$ on ER. As we increase $\mathcal{W}_{i,j}$ influence, we observe an increase of ER for both Dij-G* and Dij-H* in comparison to Dij-G, with Dij-H* having a much more significant impact. Figure 2 (b) shows the increase of ER, depending on the distance between start and destination nodes. Moreover, Figures 3 ((a), (c)) show the influence of the different scalings of $\mathcal{W}_{i,j}$ on distance and popularity. As we increase the influence of $\mathcal{W}_{i,j}$ from 0.2 to 1.0, we observe an increase of distance and popularity for both Dij-G* and Dij-H* in comparison to Dij-G. It is clear that Dij-G* always performs better than Dij-H* in terms of path length increase, but Dij-H* performs always better in terms of ER and popularity. This is because Dij-H* routes directly through the POIs, causing greater detours, but passing along highly weighted parts of the enriched graph G^*, which mostly coincide with dense Flickr regions. Finally, Figures 3((b), (d)) show that, as we increase the distance between start and destination, we observe an increase of the distance and popularity for Dij-G* as well as Dij-H* in comparison to Dij-G. As in our previous experimental setting, it is clear that Dij-G* always outperforms Dij-H* in terms of path length increase, while Dij-H* always outperforms Dij-G* in terms of enrichment ratio and popularity. The explanation for these results is that, similarly to the first setting, Dij-H* routes directly through the POIs and therefore accumulates greater ER and higher popularity values.

We conclude that both Dij-G* and Dij-H* show convincing results. Both algorithms yield significant increase in terms of ER as well as in terms of the independent Flickr-based popularity, while increasing path length only slightly. Consequently, we can claim that spatial relations, extracted from crowdsourced information, can indeed be used to enrich actual road networks and define an alternative kind of routing which reflects what people perceive as “close”.

7. CONCLUSIONS AND OUTLOOK

In this work, we presented an approach to computing knowledge-enriched paths within road networks. We employed novel methods to extract spatial relations between pairs of points of interest such as “near” or “close by” from crowdsourced textual data, namely travel blogs. We quantified the extracted relations using probabilistic models to handle the inherent uncertainty of user-generated content. Based on these models, we proposed a new cost function to enrich real world road networks, reflecting the closeness aspect according to the crowd. In contrast to existing approaches, we did not enrich previously computed paths with semantical information, but the entire network. Continuingly, two routing algorithms were presented taking this closeness aspect into account. The evaluation on real world datasets showed that both algorithms perform very well providing popular paths, while they increase the path length only slightly. For future work, we are researching alternative methods for aggregating all categories of spatial relations.

Acknowledgements

The research leading to these results has received funding from the European Union 7^{th} Framework Programme-Marie Curie Actions, Initial Training Network GEOCROWD (grant No. FP7-PEOPLE-2010-ITN-264994) as well as the Shared-E-Fleet project by the German Federal Ministry of Economics and Technology (grant No. 01ME12107). Mario A. Nascimento has been partially supported by NSERC Canada. Dieter Pfoser has been partially supported by NGA NURI (grant No. HM02101410004).

8. REFERENCES

[1] M. Duckham and L. Kulik. Simplest paths: Automated route selection for navigation. In *Spatial Information Theory. Foundations of Geographic Information Science*, pages 169–185. 2003.

[2] F. Graf, H.-P. Kriegel, M. Renz, and M. Schubert. Mario: Multi-attribute routing in open street map. In *Advances in Spatial and Temporal Databases*, pages 486–490. 2011.

[3] E. Loper and S. Bird. Nltk: The natural language toolkit. In *Proc. of the ETMTNLP*, pages 63–70, 2002.

[4] M. Lv, L. Chen, and G. Chen. Discovering personally semantic places from gps trajectories. In *Proc. of the 21st ACM CIKM*, pages 1552–1556, 2012.

[5] H. Mousselly-Sergieh, D. Watzinger, B. Huber, M. Döller, E. Egyed-Zsigmond, and H. Kosch. World-wide scale geotagged image dataset for automatic image annotation and reverse geotagging. In *Proc. of the 5th ACM MMSys*, pages 47–52, 2014.

[6] D. Quercia, R. Schifanella, and L. M. Aiello. The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city. *CoRR*, abs/1407.1031, 2014.

[7] G. Skoumas, D. Pfoser, and A. Kyrillidis. On quantifying qualitative geospatial data: A probabilistic approach. In *Proc. of the 2nd ACM GEOCROWD*, pages 71–78, 2013.

[8] G. Skoumas, D. Pfoser, and A. Kyrillidis. Location estimation using crowdsourced geospatial narratives. *CoRR*, abs/1408.5894, 2014.

[9] G. Skoumas, K. A. Schmid, G. Jossé, A. Züfle, M. Nascimento, M. Renz, and D. Pfoser. Towards knowledge-enriched path computation. *CoRR*, abs/1409.2585, 2014.