Visualizing The Semantic Similarity of Geographic Features

Gengchen Mai
STKO Lab, UCSB, CA, USA
gengchen_mai@geog.ucsb.edu

Krzysztof Janowicz
STKO Lab, UCSB, CA, USA
janowicz@ucsb.edu

Sathya Prasad
Esri Inc, Redlands, CA, USA
sprasad@esri.com

Bo Yan
STKO Lab, UCSB, CA, USA
boyan@geog.ucsb.edu

Abstract

Topographic and thematic maps have been extensively used to visualize geographic information and spatial relationships. However, it is rather difficult to directly express non-spatial relationships such semantic similarity among features using such maps. Analogies to these classical maps are necessary to visualize the distribution of geographic features in a semantic space in which semantically similar entities are clusters within the same region and the distance between geographic features represents how similar they are. In this work, we discuss one approach for such a semantically enriched geospatial data visualization and searching framework and evaluated it using a subset of places from DBpedia. The resulting map, as a representation of the semantic distribution of these geographic features, is produced by using multiple techniques including paragraph vector and clustering. Next, an information retrieval (IR) model is developed based on the vector embedding of each geographic feature. The results are visualized using the semantic similarity-based map as well as a regular map. We believe such visualization can help users to understand latent relationships between geographic features that may otherwise seem unrelated.

Keywords: Semantic similarity; visualization; geographic information retrieval.

1 Introduction

Information visualization supports humans in understanding large amounts of data, e.g., in finding patterns, that would otherwise not be apparent. With the increase of geographic information, the higher levels of detail and complexity of such data, and the heterogeneous sources these data originated from, effective visualization techniques are becoming a key element of data analytics. They also play a role in fields such as geographic information retrieval (GIR) in which they foster the interpretation of search results beyond simple rankings. As the core component of GIR and general information retrieval systems more broadly, semantic similarity can be computed from different semantic relationships between (geographic) features as well as by comparing shared and unique characteristics of such features(Schwering 2008, Janowicz et al. 2011). To give an intuitive example, two historic places may be similar because they are from the same epoch, i.e., they share a common attribute, or because both were part of the same empire, thereby being related by a common place hierarchy. Instead of ranking such places or returning numerical values for their similarity, one can help the user to understand and evaluate the retrieved results by putting them into context, e.g., by showing the distribution of all geographic features (of a given type) in a semantic search space. The visualization of such space draws from the fact that humans intuitively understand the analogy between distance in visualization space, e.g., a 2D screen space, and semantic similarity, although humans have a tendency to over interpret some visual patterns (Montello et al. 2003).

Semantic web technologies have been widely used to explicitly encode relationships between different entities in knowledge graphs. Although these technologies have been identified as promising ways for addressing many longstanding problems in GISciences (Kuhn et al. 2014) and visualization techniques have been proposed to interact with these geographic knowledge graphs (Mai et al. 2016), users might feel lost when exploring large graphs. Hence, an overview of the distribution of these geographic knowledge graphs can provide guidance to the users. The question is how to produce such a semantic similarity map in ways that resemble familiar cartographic layouts, e.g., by depicting clusters as regions (continents). Figure 1 shows such a visualization of one specific geographic dataset, namely all historic places in the DBpedia dataset.

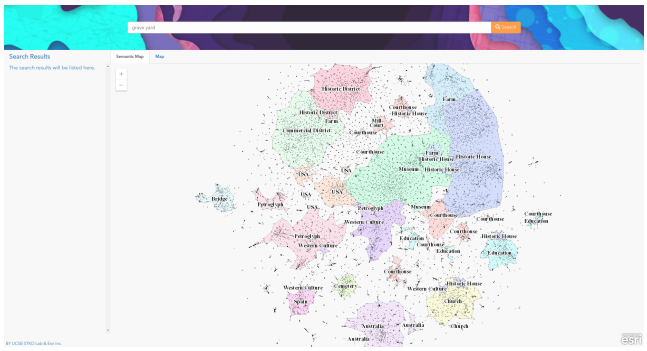


Figure 1: A map rendering of the similarities among DBpedia historic places.

Using the proposed framework, geographic features can be searched and visualized in a semantic space as well as a geographic space. Representation learning techniques, e.g., Paragraph Vector (Le & Mikolov 2014), are used to encode the semantics of each geographic feature into a high-dimensional real number

vector based on its textual description. These embeddings of geographic features are used in both the semantic visualization and the information retrieval model. In order to provide an overview of the semantic distribution of geographic features, several analysis techniques are applied to these embeddings such as dimension reduction, clustering, and concave hull construction.

2 Related Work

As a key to geographic information retrieval, semantic similarity has been studied from different perspectives across multiple domains and the proposed similarity measures depend on different data models. Similarity measures often compute the similarity between two concepts based on the distance between them in a concept taxonomy or an ontology. Ballatore and Bertolotto (Ballatore et al. 2013) extend these ideas to geographic domain and compute semantic similarity between OpenStreetMap geographic classes based on OSM Semantic Network. However, these semantic similarity measures typically focus on the type-level similarity.

In contrast to these top-down approaches, representation learning techniques provide another possibility to compute semantic similarity in a data-driven way. Popular word embedding techniques (Mikolov et al. 2013) compute distributional representations of each word and phrase based on shallow neural network models. By scanning through corpuses and predicting center words from context words, high-dimensional embeddings are learned which encode the semantics of each word and phrase. The semantic similarity between two words can be computed by calculating the cosine similarity between corresponding embeddings. Following the same idea, paragraph vectors (Le & Mikolov 2014) learn fixed dimension distributional representations for variable-length pieces of texts. Our work is inspired by these techniques to learn embeddings of each geographic feature based on its textual description. The textual description of geographic features can be obtained from their Wikipedia pages or gazetteers. In this work, we use the textual description of each geographic feature from DBpedia based on `dbo:abstract` and `rdfs:comment` predicates. Vector embeddings for geographic feature types have also recently been proposed by Yan et al. (Yan et al. 2017).

In terms of visualizing the semantic distribution of entities, the Cartograph tool (Sen et al. 2017) is a great example. In many ways, our work can be seen as a continuation of these ideas and *spatialization* as a visual support methods more generally (Skupin & Fabrikant 2003).

3 Method

In this section we will explain the methods used to build a semantically enriched information retrieval and visualization interface for geographic data. Figure 2 shows the workflow. As can be seen, the workflow can be divided into four major steps: 1) data preprocessing and computation of paragraph vectors; 2) establishing the information retrieval model; 3) constructing the semantic similarity map from the dataset; and 4) setting up the visualization and searching interface. The most important and interesting step is the map construction step. We will explain this step in detail and briefly describe the others steps.

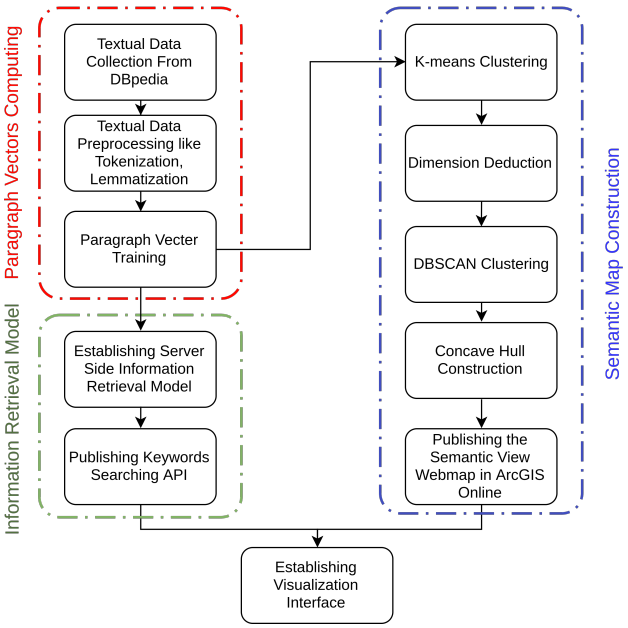


Figure 2: The workflow of the semantically enriched visualization and retrieval interface.

3.1 Paragraph Vectors Computing

We collected all geographic features that belong to the `dbo:HistoricPlace` class from *DBpedia* as the dataset for our experiment. Overall, there are 21010 historic places in *DBpedia*. Each historic place has an abstract, comments, images, and geographic coordinates. The Wikipedia page ID for each historic place is also stored. The abstracts and comments of each entity are combined and treated as the description of the current geographic feature. Some textual data preprocessing works like tokenization and lemmalization are applied to these descriptions. Next, the paragraph vector of each historic place is learned based on its descriptions. The paragraph vector model is a two-layer neural network which learns high-dimentional continuous vectors for each document (historical place). The cosine similarity between these vectors represents the semantic similarity between the corresponding places. Previous work has studied the affect of different hyperparameters on the performance of paragraph vectors (Le & Mikolov 2014, Sen et al. 2017). We use a grid search to determine the optimal parameters and manually set the dimension of the paragraph vectors to 300, 10 for the window size, and 0.025 for the learning rate.

3.2 Information Retrieval Model

The information retrieval model is established based on the learned embedding of each historic place from the paragraph vector model. Since these high-dimensional embeddings encode the terms associated with a places (and thus a *semantic context*), the embeddings of semantically similar places will cluster in the high-dimensional space and the cosine similarity between them will be high. The idea behind our information retrieval model is that when the user enters a query, an embedding with the same dimension as the entities' will be dynamically computed based on this query. Then this embedding will be used to search for the top 20 nearest entity embeddings based on the cosine similarity. The corresponding geographic entities will be returned as the search result.

In this work, we explored three ways to compute an embedding from a user's query.

1. The first way utilizes the *Doc2Vec.infer_vector()* function from gensim's Doc2Vec package.
2. The second model computes the embedding of the user query as the weighted average of each word token's embeddings in the query. The weight of each word token is computed as its TF-IDF score (term frequency-inverse document frequency) based on the current corpus which contains the descriptions of all *DBpedia* historic places.
3. The third model computes the embedding as the simple average of the query tokens' embeddings after stop words removal.

Based on our experiments, the simplest solution, i.e., the third model, gives the best results. For example, when we search for "grave yard", the third model will return some cemeteries and grave houses as the retrieved results while the other two models may also return libraries and other features which are not as strongly semantically related to our query. Many quantitative evaluation metrics of IR models can be used to evaluate these three models such Normalized Discounted Cumulative Gain (NDCG). We leave this for future work.

An API [1] is provided for the semantic searching functionality among *DBpedia* historic places.

3.3 Semantic Similarity Map Construction

The next interesting research question of our work is how to construct an overview of the semantic distribution of geographic entities such that it follows a cartographic tradition, e.g., clusters semantic similar entities in the same region – 'continent', if you like.

The learned embeddings of places are in a high-dimensional space. The first step is to extract semantic structures from their semantic distribution in this space. K-means (MacQueen et al. 1967) clustering is used to group these high-dimensional embeddings into different clusters. As the only parameter of k-means, the number of clusters are decided according to the silhouette coefficient (Rousseeuw 1987). We experimented the number of clusters from 2 to 30. As the number of clusters increases, the silhouette coefficient first increases with some fluctuations and then decreases after 16. So we decided not to try larger number of clusters. The highest silhouette coefficient is 0.0525 with 16 as the number of clusters which is used to get the final clustering results. In order to understand the meaning associated to each cluster, we produced a word cloud for each cluster. The descriptions of DBpedia historic places in each cluster are combined and treated as one document. For each cluster/document, 10 words with the highest TF-IDF score are selected which indicates the *topics* of the current cluster. The ration behind this is that the more frequent the current term occurs in the current document and the less documents contains this term, the more *indicative* this term is for the current document. According to the top 10 words of each cluster, cluster names have been assigned to these 16 clusters to indicate their thematic scope, e.g., *Mill*, *Bridge*, *Petroglyph*, and *Museum*; see Figure 3.

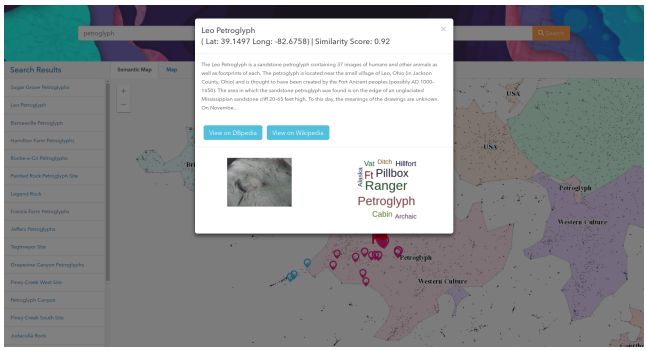


Figure 3: The pop-up window shows some basic information for `dbo:Leo_Petroglyph`.

In order to visualize the semantic distribution of geographic entities in a 2-dimensional space, dimension reduction techniques are applied to these high-dimensional embeddings. We experiment with different dimension deduction methods including PCA and t-SNE. It turns out that t-SNE performs best and the clusters derived from k-means are still well separated despite the dimensions being reduced from 300 to 2 while the clusters overlapped with each other in the results from PCA. This results align well with other work (Maaten & Hinton 2008, Sen et al. 2017).

After we project all the embeddings of geographic entities into the 2D space, *semantic continents* can be constructed from the k-means clusters. Although t-SNE produces a good dimension reduction result in which many projected features are still clustered together, some points are far away from their cluster centroids and scattered in the 2D space. Hence, we apply DBSCAN (Ester et al. 1996) to each projected k-

[1] An example query can be accessed through `http://stko-testing.geog.ucsb.edu:3050/semantic/search?searchText=grave\%20yard`.

means cluster to extract the "core" parts of them. DBSCAN has two parameters: *Eps* and *MinPts*. Different parameter combinations give different definitions of the result clusters. So we decide to use the same parameter combination when we apply DBSCAN to all 16 k-means clusters' points. In terms of the criteria for selecting the appropriate parameters for DBSCAN, several clustering quality evaluation metrics are available (Mai et al. 2018) such as normalized mutual information (Strehl & Ghosh 2002) and the rand index (Rand 1971). These metrics evaluate the clustering results extrinsically which means that the *ground truth* of the clustering task should be available to compare it with the clustering results. Such kind of *ground truth* data does not exist for many cases such as ours. Hence, we use visual interpretation to select a parameter combination for DBSCAN, namely 1.1 for *Eps* and 6 for *MinPts*.

After using DBSCAN to get the core points of each k-means clusters, we construct polygons to represent region, here *semantic continents*. A common way to do so would be to compute the convex hull of each point set. Using a convex hull to approximate the shape of a point cluster may lead to large empty space which would not be very adequate for a human user. Instead, we use concave hulls by making use of the *Chi-shape* algorithm (Duckham et al. 2008). *Chi-shape* first constructs a Delaunay triangulation based on the current point cluster. The initial boundary is the convex hull of the current point cluster. Then it erodes the outside boundary by deleting the edges of the boundary in the order of edge length until the longest edge of the boundary is shorter than a threshold. A normalized length parameter $\lambda_p \in [1, 100]$ controls this threshold and decides the complexity of the final constructed hull. In order to get an optimal λ_p, a fitness score function (Akdag et al. 2014) is used to balance the *complexity* and *emptyness* of the resulting concave hull (See Equ. 1). Here ϕ is the fitness function/score which is used to balance the *complexity* (Brinkhoff et al. 1995) and the *emptiness* of the constructed polygons. P and D represent the derived simple polygon and the Delaunay triangulation of the corresponding point cluster.

$$\phi(P, D) = Emptiness(P, D) + C * Complexity(P) \quad (1)$$

We iterate λ_p from 1 to 100. For each λ_p, we compute the average fitness score of all point clusters produced by DBSCAN among all the 16 k-means clusters. Figure 4 shows the average fitness score for different λ_p. The optimal λ_p with the lowest average fitness score is 30. The *semantic continents* of each k-means cluster are finally constructed by using this optimal λ_p. Note that applying DBSCAN to some k-means cluster like *USA* and *Australia* will produce more than one clusters which will later on result in multiple concave hulls (See Figure 1). And the points outside of these concave hulls are not noise. They are just not within the major parts of the k-means clusters they belong to. Please also note that a cluster labeled *Australia* should not be confused with the country and its spatial footprint, it is simply a cluster of historical places that are similar to each others, e.g., because they are Aboriginal sacred sites.

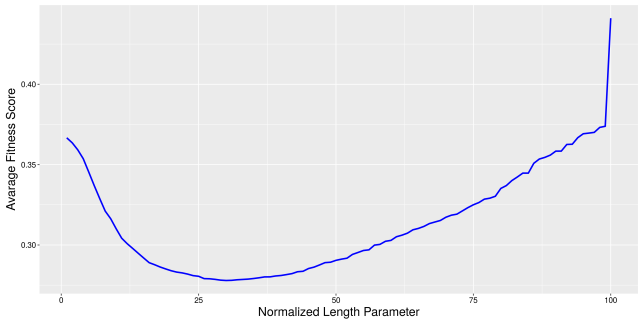


Figure 4: The average fitness score for different λ_p among all DBSCAN clusters.

The final semantic similarity map can be seen in Figure 1. This map serves as a base map which indicates the locations of the search results in the semantic space. We publish this map as a webmap service in ArcGIS Online [2].

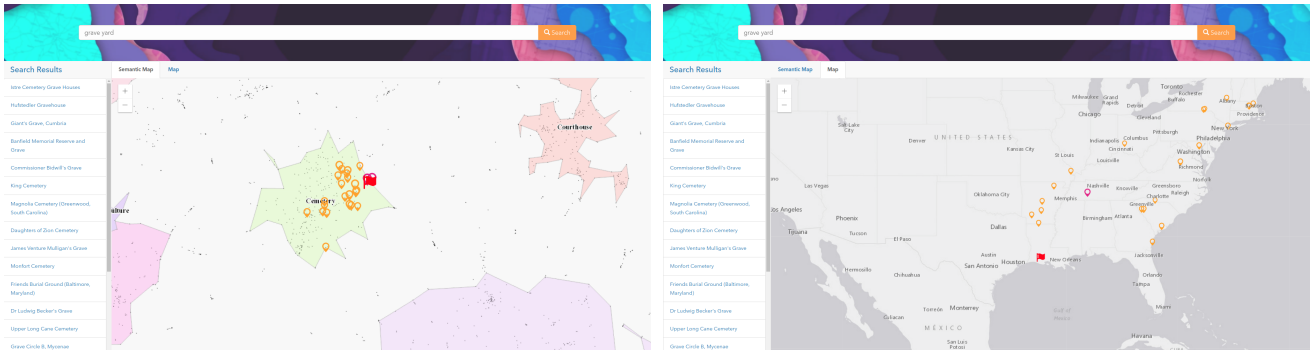
3.4 Visualization and Retrieval Interface

We have deployed a web-based user interface[3] to showcase the functionality using the historical places dataset. Figure 5a and 5b shows how this interface visualizes the search result of "grave yard" in the semantic space and the geographic space. Since the semantic similarity map is made in a way that it follows the tradition of thematic maps and the same symbol style is used to represent the geographic features in both of these maps, we believe that the user will have a better understanding of the retrieved results by switching back and forth between these two visualizations. The retrieved geographic features are represented as pin symbols with their size being proportional to the similarity score between these entities and the user query given by the IR model, while the colors of the symbols indicate the k-means cluster membership information. In both maps, the results are listed in the left side panel in descending order of the similarity scores. Clicking on one entity in this list will make both maps zoom to the location of this entity and change the symbol to a red flag. As shown in Figure 3, when users click on one of the symbols on both maps, a pop-up window will show up which contains some information about the currently selected entity including the name, geographic coordinates, description, an image, and the similarity score.

From Figure 5a and 5b, some interesting observations can be made: 1) by searching for "grave yard", the retrieved geographic entities are cemeteries, grave houses, and graves. Many of them do not contain the search terms in their names like Monfort Cemetery.

[2] http://www.arcgis.com/home/item.html?id=7e15f98399ff4788a502fd04320bdafc

[3] http://stko-testing.geog.ucsb.edu:3050/



(a) data a (b) data b

Figure 5: An example to show how the interface visualizes the searching results in the semantic space and geographic space: (a) searching results in the semantic space; (b) searching results in the geographic space.

But all of them are semantically similar to the concept "grave yard"; 2) In the semantic similarity map, almost all the search results are clustered within the "Cemetery" semantic region/continent which indicates that all retrieved places are within the semantic scope of the "Cemetery" cluster and the IR model successfully retrieves semantically similar geographic features.

4 Conclusion

In this work, we presented a framework and interface to support the interpretation of geographic information retrieval results and similarity scores more generally and tested it on a subset of places from DBpedia. Our experiments mainly focused on studying the different approaches that can be taken to reinforce human cognition by visualizing results in a semantic and in a geographic space. A lot of work remains to be done on the end-user interface level. Similarly, the proposed methods have to be calibrated, e.g., by setting the hyperparameters, based on results of human participants testing. Hence, while a lot of interesting work remains to be done in the future, we believe that the presented experiments showcase how similarity between geographic features can be learned bottom-up and how these similarities can be presented to the user.

References

Akdag, F., Eick, C. F. & Chen, G. (2014), Creating polygon models for spatial clusters, *in* 'International Symposium on Methodologies for Intelligent Systems', Springer, pp. 493–499.

Ballatore, A., Bertolotto, M. & Wilson, D. C. (2013), 'Geographic knowledge extraction and semantic similarity in openstreetmap', *Knowledge and Information Systems* **37**(1), 61–81.

Brinkhoff, T., Kriegel, H.-P., Schneider, R. & Braun, A. (1995), Measuring the complexity of polygonal objects., *in* 'ACM-GIS', p. 109.

Duckham, M., Kulik, L., Worboys, M. & Galton, A. (2008), 'Efficient generation of simple polygons for characterizing the shape of a set of points in the plane', *Pattern Recognition* **41**(10), 3224–3236.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. et al. (1996), A density-based algorithm for discovering clusters in large spatial databases with noise., *in* 'Kdd', Vol. 96, pp. 226–231.

Janowicz, K., Raubal, M. & Kuhn, W. (2011), 'The semantics of similarity in geographic information retrieval', *Journal of Spatial Information Science* **2011**(2), 29–57.

Kuhn, W., Kauppinen, T. & Janowicz, K. (2014), Linked data-a paradigm shift for geographic information science, *in* 'Proceedings of The Eighth International Conference on Geographic Information Science (GIScience2014), Berlin.', Springer, pp. 173–186.

Le, Q. & Mikolov, T. (2014), Distributed representations of sentences and documents, *in* 'International Conference on Machine Learning', pp. 1188–1196.

Maaten, L. v. d. & Hinton, G. (2008), 'Visualizing data using t-sne', *Journal of machine learning research* **9**(Nov), 2579–2605.

MacQueen, J. et al. (1967), Some methods for classification and analysis of multivariate observations, *in* 'Proceedings of the fifth Berkeley symposium on mathematical statistics and probability', Vol. 1, Oakland, CA, USA, pp. 281–297.

Mai, G., Janowicz, K., Hu, Y. & Gao, S. (2018), 'Adcn: An anisotropic density-based clustering algorithm for discovering spatial point patterns with noise', *Transactions in GIS* pp. 1–22.

Mai, G., Janowicz, K., Hu, Y. & McKenzie, G. (2016), A linked data driven visual interface for the multi-perspective exploration of data across repositories., *in* 'VOILA@ ISWC', pp. 93–101.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. (2013), Distributed representations of words and phrases and their compositionality, *in* 'Advances in neural information processing systems', pp. 3111–3119.

Montello, D. R., Fabrikant, S. I., Ruocco, M. & Middleton, R. S. (2003), Testing the first law of cognitive

geography on point-display spatializations, *in* 'International Conference on Spatial Information Theory', Springer, pp. 316–331.

Rand, W. M. (1971), 'Objective criteria for the evaluation of clustering methods', *Journal of the American Statistical association* **66**(336), 846–850.

Rousseeuw, P. J. (1987), 'Silhouettes: a graphical aid to the interpretation and validation of cluster analysis', *Journal of computational and applied mathematics* **20**, 53–65.

Schwering, A. (2008), 'Approaches to semantic similarity measurement for geo-spatial data: A survey', *Transactions in GIS* **12**(1), 5–29.

Sen, S., Swoap, A. B., Li, Q., Boatman, B., Dippenaar, I., Gold, R., Ngo, M., Pujol, S., Jackson, B. & Hecht, B. (2017), Cartograph: Unlocking spatial visualization through semantic enhancement, *in* 'Proceedings of the 22nd International Conference on Intelligent User Interfaces', ACM, pp. 179–190.

Skupin, A. & Fabrikant, S. I. (2003), 'Spatialization methods: a cartographic research agenda for non-geographic information visualization', *Cartography and Geographic Information Science* **30**(2), 99–119.

Strehl, A. & Ghosh, J. (2002), 'Cluster ensembles—a knowledge reuse framework for combining multiple partitions', *Journal of machine learning research* **3**(Dec), 583–617.

Yan, B., Janowicz, K., Mai, G. & Gao, S. (2017), 'From itdl to place2vec–reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts', *Proceedings of SIGSPATIAL* **17**, 7–10.