Visualizing The Semantic Similarity of Geographic Features

Gengchen Mai [1] Krzysztof Janowicz [1] Sathya Prasad [2]
Bo Yan [1]

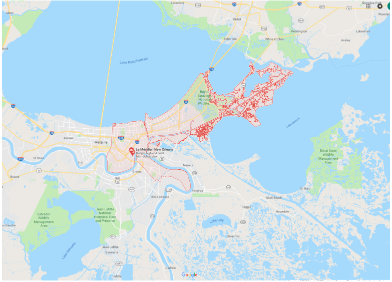
[1]STKO Lab, UC Santa Barbara, CA, USA

[2]Esri Inc, Redlands, CA, USA

April, 2018

A Semantically Enriched Visualization

- **Maps**: extensively used to visualize GI and spatial relationships.



  - Difficult to directly express non-spatial relationships (semantic similarity) using such maps.

A Semantically Enriched Visualization

A Semantically Enriched Visualization: An **analogy** of thematic maps to visualize the distribution of geographic features in a semantic space.

- Points: Geographic Coordinates $\rightarrow$ Locations in the Semantic Space
- Polygons: Administration Regions/Continents $\rightarrow$ Semantic Continents

A Semantically Enriched Visualization

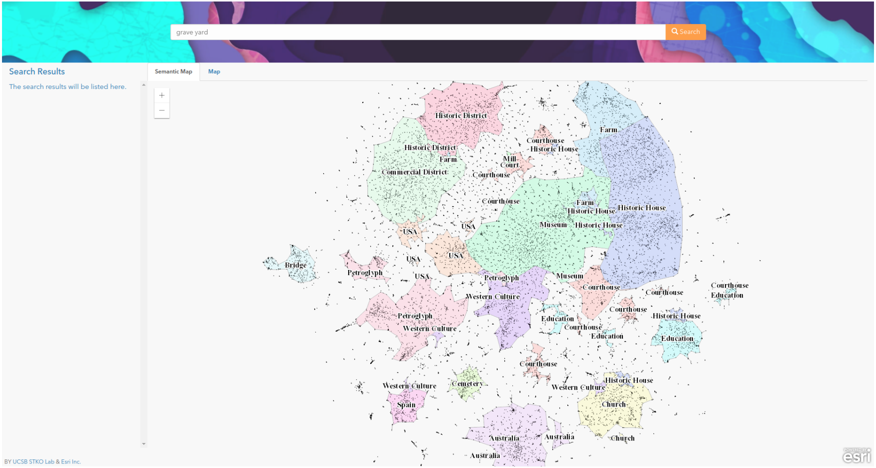
A Semantically Enriched Visualization:

- Semantically similar entities are clusters within the same region;
- The distance between geographic features represents how similar they are.

A Semantically Enriched Visualization

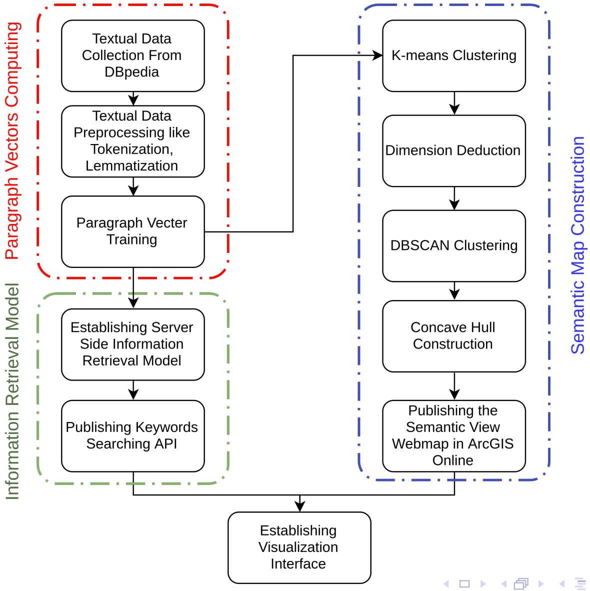
- In this work, a **semantically enriched** geospatial data **visualization** and **searching** framework are presented.
- We evaluate it using a subset of places from DBpedia.
- Multiple techniques:
  - Paragraph Vector
  - Spatial Clustering
  - Concave Hull Construction
  - Information Retrieval (IR) Model

A Semantically Enriched Visualization



A semantically enriched visualization resembles cartographic layouts

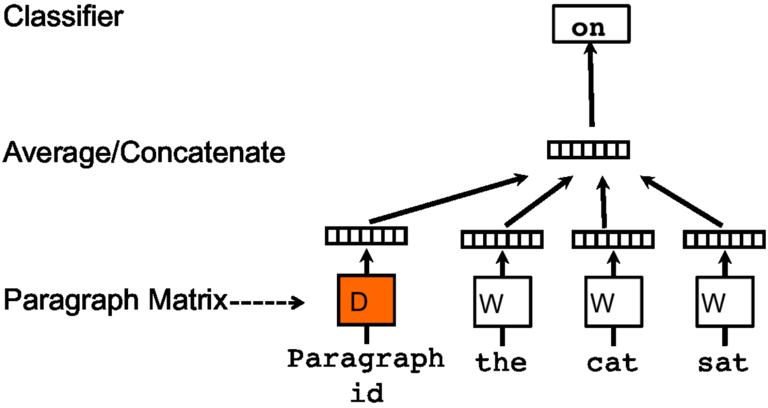
The Workflow



Paragraph Vectors Computing

- **Paragraph Vector** (or called **Doc2Vec**) is a representation learning method proposed by Natural Language Processing community.
- **Idea**: Give a collection of documents, Doc2Vec learns a high-dimensional continuous vector (embedding) for each document.
- The **cosine similarity** between the learned document vectors represents the **semantic similarity** between their corresponding documents.

Paragraph Vectors Computing



The two-layer neural network architectural of Doc2Vec

Paragraph Vectors Computing

Outputs of Doc2Vec:

- Embeddings of documents;
- Embeddings of word tokens in the document corpus.

Paragraph Vectors Computing

- **Data Source:** All entities typed dbo:HistoricalPlace in DBpedia (21010 places)
  - Each historic place has an abstract, comments, images, and geographic coordinates.
- **Method:** Doc2Vec Model (PVDM [4])
  - **Textual data collection**: Treat each place as a document whose content is its abstract and comments
  - **Textual data preprocessing**: tokenization and lemmalization
  - **Paragraph vector training**: embedding dimention $K = 300$; window size $N = 10$; learning rate $\alpha = 0.025$

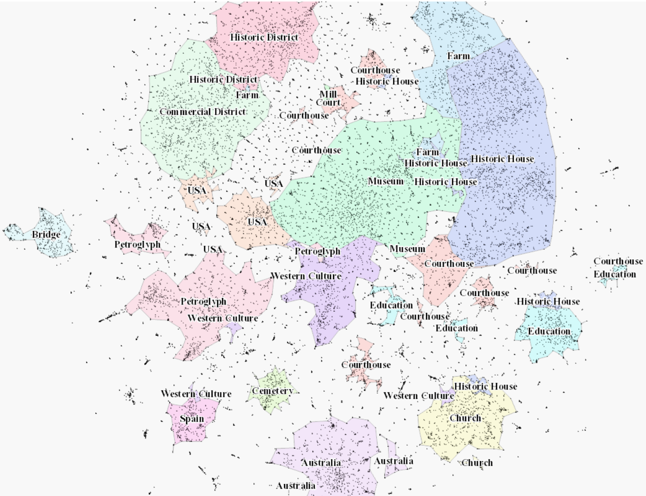
Information Retrieval Model

- **Place Embeddings**: the learned embedding of each historic place from Doc2Vec.
- **Query Embeddings**:
  - Utilize the *Doc2Vec.infer_vector()* function from gensim's Doc2Vec package
  - The TF-IDF score weighted embedding based on word embeddings of query word tokens
  - **The simple average of the query tokens' embeddings after stop words removal**
- **Semantic Similarity Score Function**: the cosine similarity between the query embedding and place embeddings
- An API [1] is provided for the semantic searching functionality among *DBpedia* historic places.

[1] http://stko-testing.geog.ucsb.edu:3050/semantic/search?searchText=grave%20yard.

Semantic Similarity Map Construction

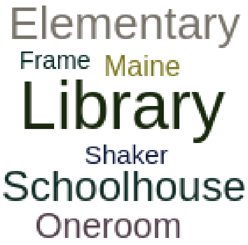
Spatialization: how to construct an overview of the semantic distribution of geographic entities such that it follows a cartographic tradition (*semantic continent*).



Semantic Similarity Map Construction

K-means clustering: group these place embeddings into different clusters;

- Try #(*clusters*) from 2 to 30 and compute silhouette coefficient [5] of the clustering results;
- #(*clusters*) = 16 gives the highest silhouette coefficient;
- The descriptions of places in each cluster are combined as one document;
- Word clouds are produced from 10 word with highest TF-IDF score;
- Each cluster is named according to the its top 10 words.



The word cloud for *Education* cluster

Semantic Similarity Map Construction

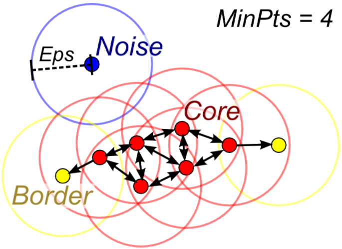
Dimension reduction: to visualize the semantic distribution of geographic entities in a 2-dimensional space

- Different dimension deduction methods including PCA and t-SNE are experimented;
- t-SNE performs best and the clusters derived from k-means are still well separated.

Semantic Similarity Map Construction

DBSCAN:

- Although t-SNE produces a good dimension reduction result, some points are far away from their cluster centroids and scattered in the 2D space.
- We apply DBSCAN [3] to each projected k-means cluster to extract the "core" parts of them.
- Visual interpretation are used to select the parameter combination for DBSCAN. ($Eps = 1.1$ and $MinPts = 6$)



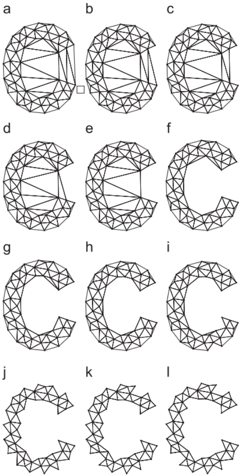
Semantic Similarity Map Construction

Concave Hull Construction: *Chi-shape* algorithm [2]

- It first constructs a Delaunay triangulation;
- It erodes the boundary by deleting boundary's edges until the longest edge less than a threshold.
- A normalized length parameter $\lambda_p \in [1, 100]$ controls this threshold;
- To get optimal λ_p, a fitness score function [1] is used to balance the *complexity* and *emptyness* of the resulting concave hull.

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

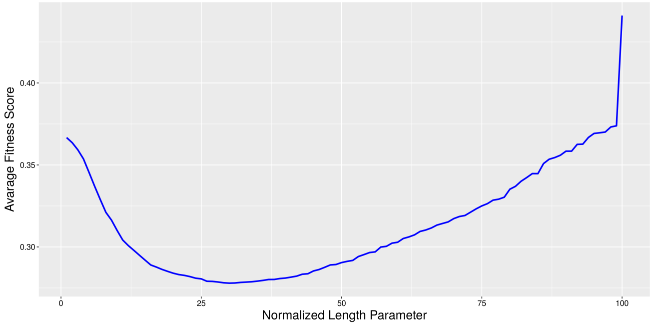
P: the derived simple polygon; D: the Delaunay triangulation of the corresponding point cluster.



Semantic Similarity Map Construction

Concave Hull Construction:

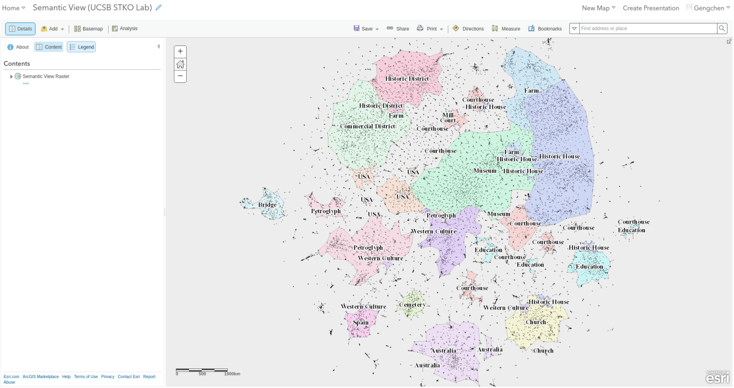
- We iterate λ_p from 1 to 100 and compute the average fitness score of all point clusters produced by DBSCAN;
- The optimal λ_p with the lowest average fitness score is 30.



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

Semantic Similarity Map Construction

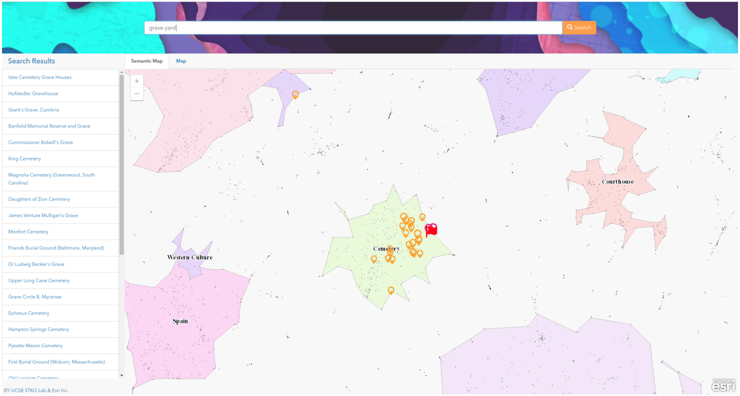
Publishing the Semantic View Webmap in ArcGIS Online: ne [2].



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

Result

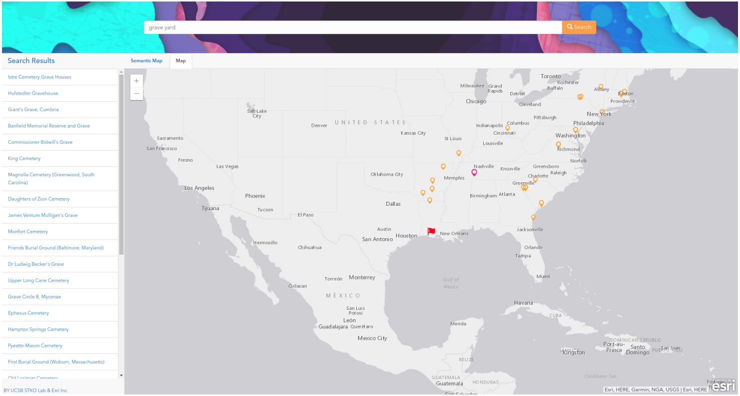
We have deployed a web-based user interface[3] to showcase the functionality using the historical places dataset.



the search result of "grave yard" in the semantic space

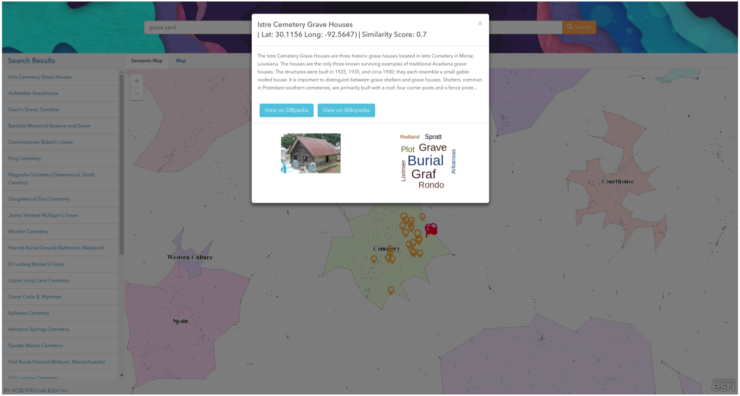
[3] http://stko-testing.geog.ucsb.edu:3050/

Result



the search result of "grave yard" in the geographic space

Result



The pop-up window shows some basic information for `dbo:Istre_Cemetery_Grave_Houses`.

Conclusion & Future Work

- In this work, we presented a semantically enriched geospatial data visualization and search framework.

- In the future, the proposed methods have to be calibrated, e.g., by setting the hyperparameters, based on results of human participants testing.

Paper

This work has been accepted to AGILE 2018.

Gengchen Mai, Krzysztof Janowicz, Sathya Prasad, Bo Yan. Visualizing The Semantic Similarity of Geographic Features, In: *Proceedings of AGILE 2018*, June 12 - 15, 2018, Lund, Sweden.

References I

[1] Akdag, Fatih, Eick, Christoph F, & Chen, Guoning. 2014. Creating polygon models for spatial clusters. *Pages 493–499 of: International Symposium on Methodologies for Intelligent Systems*. Springer.

[2] Duckham, Matt, Kulik, Lars, Worboys, Mike, & Galton, Antony. 2008. Efficient generation of simple polygons for characterizing the shape of a set of points in the plane. *Pattern Recognition*, **41**(10), 3224–3236.

[3] Ester, Martin, Kriegel, Hans-Peter, Sander, Jörg, Xu, Xiaowei, *et al.* . 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *Pages 226–231 of: Kdd*, vol. 96.

References II

[4] Le, Quoc, & Mikolov, Tomas. 2014. Distributed representations of sentences and documents. *Pages 1188–1196 of: International Conference on Machine Learning*.

[5] Rousseeuw, Peter J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, **20**, 53–65.