Clustering Pipelines of large RDF POI Data

Rajjat Dadwal[1⋆], Damien Graux[1], Gezim Sejdiu[2], Hajira Jabeen[2], and Jens Lehmann[1,2]

[1] Fraunhofer Institute for Intelligent Analysis and Information Systems, Germany
[2] Smart Data Analytics, University of Bonn, Germany
{rajjat.dadwal,damien.graux}@iais.fraunhofer.de
{sejdiu,jabeen,jens.lehmann}@cs.uni-bonn.de

Abstract. Among the various domains using large RDF graphs, applications often rely on geographical information which is often represented via Points Of Interests. In particular, one challenge is to extract patterns from POI sets to discover Areas Of Interest (AOIs). To tackle this challenge, a typical method is to aggregate various points according to specific distances (e.g. geographical) via clustering algorithms. In this study, we present a flexible architecture to design pipelines able to aggregate POIs from contextual to geographical dimensions in a single run. This solution allows any kind of clustering algorithm combinations to compute AOIs and is built on top of a Semantic Web stack which allows multiple-source querying and filtering through SPARQL.

1 Introduction

Various organizations like DBpedia [3], Wikidata [7] etc. are constantly working for gathering information from different sources and storing it in structured form, e.g. RDF[1]. RDF data allow to model various domains and this characteristic helps to solve problems in different areas i.e., from the medical domain to the geographical domain. In this study, we are focusing on Points Of Interests (POIs). POIs are generally characterized by their geospatial coordinates along with their thematic/contextual attributes. A common POI use-case is to find hot zones according to specific topics: i.e. discovering Areas of Interest (AOIs) as a result of aggregation of POIs. With the assistance of AOIs, one can identify other similar areas in the same or a different city, recognize the distinguishing characteristics of this area, and determine potential types of users (or customers) that would be interested in that area.

In this paper, we propose a flexible architecture to design clustering pipelines for POI semantic datasets at once. Indeed, using large and detailed RDF vocabularies allow richer POI descriptions. For example, one POI related to a restaurant might be described by its latitude, longitude, food specialty, reviews, address, phone number etc. which could represent up to 50 distinct triples[2] leading then

⋆ This research was supported by the European project SLIPO (number 731581).

[1] https://www.w3.org/TR/rdf11-primer/

[2] See e.g. the SLIPO ontology: https://github.com/SLIPO-EU/poi-data-model/

to billions of RDF records overall. As a consequence, we require scalability and build our solution on top of the distributed semantic stack SANSA [4] which benefits from Apache Spark [8]. The proposed architecture then enables any kind of clustering algorithm combinations on POI RDF data.

2 Architecture overview

In order to process RDF (containing POIs) datasets in an efficient and scalable way, we first have to adopt a convenient processing framework. SANSA [4] is a data-flow engine for distributed computing of large-scale RDF datasets. It provides APIs for faster reading, querying, inferencing and apply analytics at scale. It uses Apache Spark [8] as an underlying engine. SANSA contains features which are utilized for processing RDF data with thematic and spatial information.

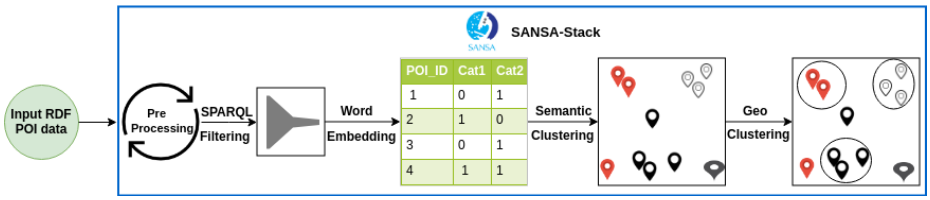

Fig. 1. A Semantic-Geo Clustering flow.

Our proposed approach contains up to five main components (which could be enabled/disabled if necessary) namely: data pre-processing, SPARQL filtering, word embedding, semantic clustering and geo-clustering. In particular, in Figure 1, we present an example of Semantic-Geospatial clustering pipeline. Indeed, we consider two types of clustering algorithms: the semantic-based ones and the geo-based ones.

In semantic based clustering algorithms (which do not consider POI locations but rather aim at grouping POIs according to shared labels), there is a need to transform the POIs categorical values to numerical vectors to find the distance between them. So far, we can select any word embedding technique among the three available ones namely one-hot encoding, Word2Vec and Multi-Dimensional Scaling. All the above mentioned methods converts categorical variables into a form that could be provided to semantic clustering algorithms to form groups of non-location-based similarities. For example, all restaurants are in one cluster whereas all the ATMs in another one. On the other hand, the geo-clustering methods help to group the spatially closed coordinates with in each semantic cluster.

More generically, our architecture and implementation allow users to design any kind of clustering combinations they would like. Actually, the solution is flexible enough to pipe together more than two clustering "blocks" and even to add additional RDF datasets into the process after several clustering rounds.

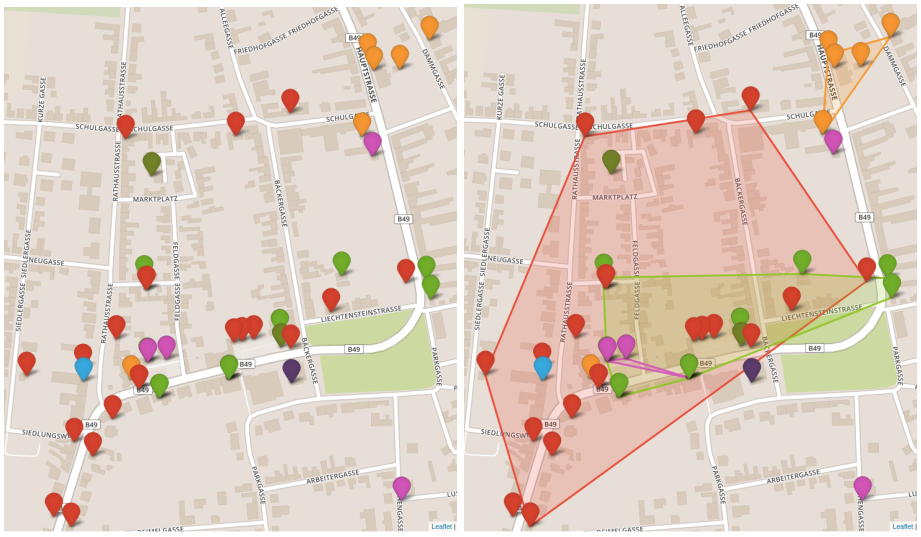

Fig. 2. Visualizations (on a map) of the Semantic-Geo clustering pipeline steps.

In addition, we directly embedded the state-of-the-art clustering algorithms into the SANSA Machine Learning layer[3] so that these pipelines are prone to be built out of the box.

3 Achieved Results

To illustrate the feasibility of our approach and demonstrate the potential of the RDF POI clustering library we developed in SANSA, we present –as an example– in this section the implementation results of the specific architecture presented in Figure 1 i.e. a Semantic-Geo clustering pipeline.

In order to test the process, we used an RDF POI dataset which follows the ontology described in [1] containing around 18 000 triples which represent information on 623 POIs (i.e. around 28 triples per POI). We then chose Word2Vec [6] as embedding for the K-means [5] semantic-clustering algorithm, before running DBSCAN [2] as geo-clustering method. In details, we gave the following parameters to the algorithms: 8 clusters within 5 iterations for K-means and $\epsilon = 0.002$ with at least 2 points per cluster for DBSCAN. The complete process took around 20 seconds using a 8GB-memory laptop running a single-node SANSA & Spark stack.

We present the results obtained at the various steps in Figure 2 on a map, the figure presents a zoom over a particular Austrian region. The figure is twofold, we first display (left side) the only result of the K-means where POIs are pinned on a map and where each color corresponds to a specific cluster. As expected,

[3] `https://github.com/SANSA-Stack/SANSA-ML`

the semantic clusters are distributed over the entire country since POIs of a color are sharing common "sense" with regards to the categories in the ontology. As a consequence, the geographical step of aggregation allows then to break those country-spread clusters into pieces and obtain (right side of Figure 2) relevant AOIs. In particular, four AOIs are visible: an orange one in the corner, a large red one which also embeds a green one and a little magenta.

4 Conclusion

In this article, we presented a solution to extract AOIs from big POI data while considering several dimensions at the same time. The architecture is embedded inside a state-of-the-art Semantic Web stack (i.e. SANSA [4]) and then benefits from the advantages of it. For instance, it allows source aggregation or datasets filtering via SPARQL to only focus on some interesting regions, e.g., a specific country can be selected. Moreover, even if we restricted our description in this study to a Semantic-Geo clustering pipeline, our architecture allows any kind of clustering combinations. Finally, the above-presented pipeline is also openly available from a demonstrating notebook[4] on the SANSA repository.

References

1. Athanasiou, S., Giannopoulos, G., Graux, D., Karagiannakis, N., Lehmann, J., Ngonga Ngomo, A.C., Patroumpas, K., Sherif, M.A., Skoutas, D.: Big POI data integration with linked data technologies. In: 22nd International Conference on Extending Database Technology, Lisbon, Portugal. pp. 477–488 (2019)
2. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd **96**(34), 226–231 (1996)
3. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., et al.: Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web **6**(2), 167–195 (2015)
4. Lehmann, J., Sejdiu, G., Bühmann, L., Westphal, P., Stadler, C., Ermilov, I., Bin, S., Chakraborty, N., Saleem, M., Ngonga Ngomo, A.C., Jabeen, H.: Distributed semantic analytics using the SANSA stack. In: ISWC Resources Track (2017)
5. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: 5th Berkeley Symposium on Mathematical Statistics and Probability. pp. 281–297 (1967)
6. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
7. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledge base. (2014)
8. Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., Franklin, M.J., Shenker, S., Stoica, I.: Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In: Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation. pp. 2–2. USENIX Association (2012)

[4] `https://github.com/SANSA-Stack/SANSA-Notebooks`