Geospatial Semantics for Spatial Prediction

Marvin Mc Cutchan
Vienna University of Technology, Austria
marvin.mccutchan@geo.tuwien.ac.at

Ioannis Giannopoulos
Vienna University of Technology, Austria
igiannopoulos@geo.tuwien.ac.at

Abstract

In this paper the potential of geospatial semantics for spatial predictions is explored. Therefore data from the LinkedGeoData platform is used to predict landcover classes described by the CORINE dataset. Geo-objects obtained from LinkedGeoData are described by an OWL ontology, which is utilized for the purpose of spatial prediction within this paper. This prediction is based on an association analysis which computes the collocations between the landcover classes and the semantically described geo-objects. The paper provides an analysis of the learned association rules and finally concludes with a discussion on the promising potential of geospatial semantics for spatial predictions, as well as potentially fruitful future research within this domain.

2012 ACM Subject Classification Information systems → Association rules, Information systems → Geographic information systems, Software and its engineering → Semantics

Keywords and phrases Geospatial semantics, spatial prediction, machine learning, Linked Data

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.45

Category Short Paper

1 Introduction and Related Research

This paper investigates the potential of geospatial semantics for spatial predictions. For this purpose data is obtained from the LinkedGeoData platform [6], which maintains geospatial data with semantic annotations, provided as linked data. This data is then used to predict CORINE landcover classes within a defined region of interest (ROI). Predictions are carried out by computing association rules using the FP-Growth algorithm. Descriptive statistics are calculated for the corresponding association rules and are used for an evaluation of the potential of geospatial semantics for spatial predictions. For the purpose of this research, spatial prediction is defined as the prediction of a CORINE landcover class for a defined region, based on the classes of geo-objects which fall within that region. Examples for such classes of geo-objects are given: tree, restaurant, river and bar. The proposed methodology ultimately enables to predict landcover classes in areas, where no classifications are yet available.

Spatial predictions are identified as one of the use-cases of Digital Earth [5] as well as a key feature for tackling global problems such as urbanization and climate change [3]. Association analysis can spatially predict and has traditionally been utilized as spatial association rule mining [8] or co-location mining [7] within the domain of Geoinformation science. Spatial association rule mining aims at detecting geo-ojects with a frequent spatial relationship. Other papers focus on increasing the performance of co-location mining in terms of computational complexity [1, 10] . Further approaches use contextual information as an auxiliary data source in order to achieve better predictions [11, 14]. Nevertheless, no work

uses extensive semantic information for association analysis or researches on the contribution of geospatial semantic information for spatial predictions. Thus, this work explores the potential, geospatial semantics hold for spatial predictions. The contribution of this paper is as follows: It demonstrates that data with geospatial semantics enable to score meaningful association rules and are therefore a promising data source for this purpose. Additionally, it is shown that geospatial semantics predict association rules with a high conviction in urban areas as well as that a higher number of distinct classes provide better results.
The paper is structured as follows: Firstly the methodology is outlined, followed by a presentation of the results, including an analysis. The paper finalizes with a discussion of the results as well as its fruitfulness to future research and applications.

2 Methodology

Two major steps are performed within the methodology of this work. First, the data is derived and preprocessed. Second, the association analysis is carried out using the FP-Growth algorithm [2]. The FP-Growth algorithm generates association rules describing which set of classes have a relevant association.

2.1 Data acquisition and preparation

In order to perform an association analysis between linked data of the LinkedGeoData dataset and the CORINE landcover dataset (see table 1), a series of steps are performed for deriving and preparing the data: OpenStreetMap data is downloaded for a ROI. SPARQLIFY[6] is then used to load the OpenStreetMap data into a new local triplestore as linked data in the LinkedGeoData structure. Thus, a local copy of the LinkedGeoData endpoint is created for a specific ROI. This enables to access semantic information of geo-objects of the OpenStreetMap dataset, such as its OWL classes. The OWL classes are defined by the LinkedGeoData ontology and are ultimately used to predict the CORINE landcover classes. The ROI is covered by Austria. Geo-objects are then loaded into a PostGIS database. Each object contains three attributes: A unique identifier, the name of its OWL class as well as a geometry encoded as a well-known binary. Thus, every geometry is enhanced with semantics. The dataset contains 3 different types of geometries, namely, point, polygon and linestring. There are 1.080.819 point-objects, 4.024.536 polygon-objects and 1.893.309 line-object which can have one of the 768 OWL classes. Furthermore, the CORINE dataset is transformed to a polygon dataset where each pixel is presented by a square polygon, called a grid-cell. Each grid-cell contains two attributes: An unique identifier, as well as the class number of the corresponding landcover class. There are 44 classes in the CORINE landcover dataset, however, only 28 of them are present in Austria [13]. There are 56667188 grid-cells within the ROI. Finally, a transaction table is generated in the PostGIS database. Each row of this transaction table contains the identifier of a grid-cell and a class of a geo-object which intersects with the corresponding grid-cell. Thus, the table enables to query which classes appear within a certain grid-cell. The distinct set of classes which intersect with a grid-cell is defined as a transaction t_i. All transactions form a set, denoted as $\mathbf{T}$. Thus, t_i is a subset of $\mathbf{T}$.

2.2 Association Analysis

After the preparation of the data, the association analysis is performed. Therefore the FP-Growth algorithm is utilized which computes association rules based on the frequencies of transactions and the number of all available transactions. An example of a computed association rule is given:

$$\{\textbf{Building, Tree, Tramway}\} \rightarrow \{\textbf{Continuous urban fabric }\}$$

This association rule suggests, that the class "Continuous urban fabric" is likely to appear, if the classes "Building", "Tree" and "Tramway" are present. Association rules can have different confidences. The confidence of an association rule can be calculated by equation 1 [12].

$$conf(\mathbf{X} \rightarrow \mathbf{Y}) = \frac{supp(\mathbf{X} \cup \mathbf{Y})}{supp(\mathbf{X})} \tag{1}$$

$$supp(\mathbf{X}) = \frac{|\{t_i | \mathbf{X} \subseteq t_i, t_i, t_i \in \mathbf{T}\}|}{|\mathbf{T}|} \tag{2}$$

The support function $supp(\mathbf{X})$ describes the proportion of a transaction t_i, which contains $\mathbf{X}$, in the set $\mathbf{T}$. Its numerator denotes the number of times t_i (which contain $\mathbf{X}$) is observed among all transactions in $\mathbf{T}$. Whereas the denominator is defined by the number of all transactions within $\mathbf{T}$. The confidence can range from [0,1] and states how often a rule has been found in the transaction database. A confidence of 0 corresponds to no confidence that a given rule is true. In contrast, a confidence of 1 states the maximum confidence that an association rule is correct. An association rule can be additionally described by the Conviction [12]:

$$conv(\mathbf{X} \rightarrow \mathbf{Y}) = \frac{1 - supp(\mathbf{Y})}{1 - conf(\mathbf{X} \rightarrow \mathbf{Y})} \tag{3}$$

The conviction is as a measurement of the degree of implication of an association rule. An association rule can be confident merely because $\mathbf{Y}$ appears with a high frequency and $\mathbf{X}$ with a low frequency. A high conviction corresponds to a high degree of implication of an association rule, whereas a low conviction corresponds to a low degree of implication of a rule. Confidence and conviction are going to be used to validate the generated association rules. The FP-Growth algorithm computes rules based on a defined minimum support. The lower the support is set, the more association rules are computed. However, defining the value too low will yield a long runtime. The support is set as low as possible within this study to compute as much association rules as possible in order to gain more insights on the impact of geospatial semantics on spatial predictions. For this purpose, an optimal value of 0.1 was found in an interative manner. In addition, association rules having a conviction lower than 1 were pruned, as a conviction below that value suggests no significant implication. For running the FP-growth algorithm, rapidminer [9] was used. There are 56.667.188 grid-cells and consequently 56.667.188 potential transactions. Due to computational limitations not all of these transactions were used within this study. Therefore 5000 randomly selected transactions per landcover class were chosen. This balanced selection was made in order to avoid a bias in the association rules.

3 Results and Analysis

There are 28 predictable CORINE landcover classes within the ROI (see table 1). Each landcover class is a subclass of a more general parentclass. Tables 2, 3 and 4 summarize the number of learned association rules, the descriptive statistics for each class as well as the corresponding parent class, according to the definition of the CORINE dataset [13].

Table 1 All available 28 CORINE classes in Austria and their description.

1	2	3	4	5	6	7	10
Continous urban fabric	Discontinous urban fabric	Industrial and commercial units	Road and rail network and associated land	Port areas	Airports	Mineral extraction sites	Green urban areas
11	12	14	15	18	20	21	23
Sport and leisure facilities	Non-irrigated arable land	Rice fields	Vine yards	Pastures	Complex cultivation patterns	Agriculture with natural vegetation	Broad leaved forest
24	25	26	27	29	31	32	34
Coniferous forest	Mixed forest	Natural grassland	Moors and heathland	Transitional woodland shrubs	Bare rock	Sparsely vegetated area	Glaciers and perpetual snow
35	36	40	41				
Inland marshes	Peatbogs	Water courses	Waterbodies				

Table 2 Predictions for CORINE landcover classes 1-11.

CLASS number	1	2	3	4	5	6	7	10	11
Number of association rules	5752	157	345	226	-	1525	45	496	11
Min (Confidence)	0.11	0.11	0.11	0.11	-	0.11	0.11	0.11	0.11
Max (Confidence)	0.87	0.36	0.58	0.72	-	0.98	0.90	0.71	0.86
Mean (Confidence)	0.32	0.17	0.23	0.25	-	0.60	0.51	0.24	0.52
Standarddeviation (Confidence)	0.17	0.05	0.12	0.14	-	0.25	0.28	0.12	0.28
Min (Conviction)	1.08	1.09	1.08	1.08	-	1.10	1.08	1.08	1.08
Max (Conviction)	7.23	1.50	2.27	3.48	-	61.47	9.23	3.40	7.06
Mean (Conviction)	1.59	1.17	1.31	1.37	-	6.17	3.25	1.34	2.87
Standarddeviation (Conviction)	0.64	0.07	0.27	0.41	-	8.58	2.50	0.34	1.89
CORINE Parentclass	Artificial surfaces								

Observing tables 2, 3 and 4, several trends can be observed: Most rules were computed for class 1(Continuous urban fabric), followed by class 6 (Airports). Association rules predicting class 1 exhibit a relatively high confidence, up to 87%, as well as a relatively high conviction, 7.23. An association rule which exhibits both, a high conviction and high confidence can be considered a meaningful rule. Generally, it can be observed that all subclasses of “Artificial surfaces” yield the most promising results. In contrast, confidence and conviction decline for classes which are a subclass of “Agricultural areas”, with one exception, i.e. class 15 (vine yards), which was predicted with exceptional conviction and confidence. However, no predictions could be made for class 14 (rice fields). The lowest confidence as well as conviction can be observed among subclasses of “Forests and semi-natural areas” and “Wetlands” with one exception, class 34 (Glaciers and perceptional snow). No association rules were computed for classes 27 (Sclerophyllous vegetation), class 29 (Transitional woodland shrub), class 32 (Sparsely vegetated areas), as well as class 36 (Peatbogs). The confidence as well as conviction inclines for subclasses of water bodies, as specially for subclass 41 (water bodies).

4 Discussion and Future Research

Considering the findings based on the results presented in tables 2, 3 and 4 it can be said that geospatial semantics can be used for spatial predictions and exhibit different qualities depending on the landcover class to be forecast. Classes closely related to urban areas are predicted better than classes which can be found more often in rural areas, such as forests or wetlands. A potential explanation for this effect is given: LinkedGeoData is based on

Table 3 Predictions for CORINE landcover classes 12-25.

CLASS number	12	14	15	18	20	21	23	24	25
Number of association rules	5	-	64	2	4	1	-	-	-
Min (Confidence)	0.12	-	0.11	0.12	0.12	0,15	-	-	-
Max (Confidence)	0.13	-	0.86	0.13	0.17	0.15	-	-	-
Mean (Confidence)	0.13	-	0.48	0.12	0.14	0.15	-	-	-
Standarddeviation (Confidence)	0.01	-	0.25	0.01	0.02	-	-	-	-
Min (Conviction)	1.09	-	1.10	1.09	1.09	1.13	-	-	-
Max (Conviction)	1.11	-	6.73	1.10	1.16	1.13	-	-	-
Mean (Conviction)	1.10	-	2.59	1.10	1.12	1.13	-	-	-
Standarddeviation (Conviction)	0.01	-	1.68	0.01	0.03	-	-	-	-
CORINE Parentclass	Agricultural areas						Forests and semi-natural areas		

Table 4 Predictions for CORINE landcover classes 26-41.

CLASS number	26	27	29	31	32	34	35	36	40	41
Number of association rules	1	-	-	2	-	11	8	-	47	12
Min (Confidence)	0.15	-	-	0.11	-	0.15	0.11	-	0.11	0.11
Max (Confidence)	0.15	-	-	0.27	-	0.86	0.25	-	0.27	0.52
Mean (Confidence)	0.15	-	-	0.19	-	0.44	0.16	-	0.15	0.22
Standarddeviation (Confidence)	-	-	-	0.12	-	0.24	0.05	-	0.04	0.12
Min (Conviction)	1.13	-	-	1.08	-	1.13	1.08	-	1.08	1.08
Max (Conviction)	1.13	-	-	1.32	-	7.11	1.29	-	1.31	1.99
Mean (Conviction)	1.13	-	-	1.20	-	2.49	1.15	-	1.14	1.28
Standarddeviation (Conviction)	-	-	-	0.17	-	2.13	0.07	-	0.05	0.25
CORINE Parentclass	Forests and semi-natural areas						Wetlands		Waterbodies	

OpenStreetMap and therefore relies on volunteers collecting geospatial data. Thus, there is a greater likelihood that a higher coverage of geospatial data is present in urban areas, increasing the number of classes per grid-cell. A higher number of classes per grid-cell enable to compute association rules with a higher conviction as they exhibit a higher distinction and therefore result in a lower support. The same argument could be made for the number of available classes: A higher amount of available classes would increase the chances to get association rules with a higher conviction as it would increase the distinction. However, this aspect is not covered in this study. Future research will focus deeper on the investigation of the potential of geospatial semantics for predictive purposes. Therefore two major aspects will be investigated: (1) The effect of the class hierarchy on the quality of spatial predictions. For this purpose classes will be exchanged with their parent class. (2) Future studies will investigate the impact of adding other geospatial data with different ontologies. The obtained knowledge can be used as input in spatial human-computer interaction [4], for future geosensor networks in order to create better predictions as well as to measure the impact on integrating different geospatial data sources with semantic annotations. This could help to explain yet undiscovered geospatial phenomena and it is therefore argued that further analysis in this domain is paramount to research progress.

References

1 Witold Andrzejewski and Pawel Boinski. GPU-accelerated collocation pattern discovery. In Barbara Catania, Giovanna Guerrini, and Jaroslav Pokorný, editors, *Advances in Databases and Information Systems*, pages 302–315, Berlin, 2013. Springer.

2 Christian Borgelt. An implementation of the fp-growth algorithm. In *Proceedings of the 1st International Workshop on Open Source Data Mining: Frequent Pattern Mining Implementations*, OSDM '05, pages 1–5, New York, NY, USA, 2005. ACM. doi:10.1145/1133905.1133907.

3 Max Craglia, Kees de Bie, Davina Jackson, Martino Pesaresi, Gábor Remetey-Fülöpp, Changlin Wang, Alessandro Annoni, Ling Bian, Fred Campbell, Manfred Ehlers, John van Genderen, Michael Goodchild, Huadong Guo, Anthony Lewis, Richard Simpson, Andrew Skidmore, and Peter Woodgate. Digital earth 2020: towards the vision for the next decade. *International Journal of Digital Earth*, 5(1):4–21, 2012. doi:10.1080/17538947.2011.638500.

4 Ioannis Giannopoulos, Peter Kiefer, and Martin Raubal. Mobile outdoor gaze-based geo-HCI. In *Geographic Human-Computer Interaction, Workshop at CHI 2013*, pages 12–13, 2013.

5 M. F. Goodchild. The use cases of digital earth. *International Journal of Digital Earth*, 1(1):31–42, 2008. doi:10.1080/17538940701782528.

6 Jon Jay Le Grange, Jens Lehmann, Spiros Athanasiou, Alejandra Garcia Rojas, Giorgos Giannopoulos, Daniel Hladky, Robert Isele, Axel Cyrille Ngonga Ngomo, Mohamed Ahmed Sherif, Claus Stadler, and Matthias Wauer. The geoknow generator: Managing geospatial data in the linked data web. http://jens-lehmann.org/files/2014/lgd_geoknow_generator.pdf. (Accessed on 04/30/2018).

7 Y. Huang, S. Shekhar, and H. Xiong. Discovering colocation patterns from spatial data sets: a general approach. *IEEE Transactions on Knowledge and Data Engineering*, 16(12):1472–1485, 2004. doi:10.1109/TKDE.2004.90.

8 Krzysztof Koperski and Jiawei Han. Discovery of spatial association rules in geographic information databases. In Max J. Egenhofer and John R. Herring, editors, *Advances in Spatial Databases*, pages 47–66, Berlin, 1995. Springer.

9 Rapidminer. Lightning fast data science platform | rapidminer. https://rapidminer.com/. (Accessed on 04/30/2018).

10 Arpan Man Sainju and Zhe Jiang. Grid-based colocation mining algorithms on gpu for big spatial event data: A summary of results. In Michael Gertz, Matthias Renz, Xiaofang Zhou, Erik Hoel, Wei-Shinn Ku, Agnes Voisard, Chengyang Zhang, Haiquan Chen, Liang Tang, Yan Huang, Chang-Tien Lu, and Siva Ravada, editors, *Advances in Spatial and Temporal Databases*, pages 263–280, Cham, 2017. Springer International Publishing.

11 Muhammad Shaheen, Muhammad Shahbaz, and Aziz Guergachi. Context based positive and negative spatio-temporal association rule mining. *Knowledge-Based Systems*, 37:261–273, 2013. doi:10.1016/j.knosys.2012.08.010.

12 Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Addison-Wesley Longman Publishing, Boston, MA, USA, 2005.

13 UBA. CORINE Landcover Nomenklatur (deutsch). http://www.umweltbundesamt.at/fileadmin/site/umweltthemen/raumplanung/1_flaechennutzung/corine/CORINE_Nomenklatur.pdf. (Accessed on 04/30/2018).

14 Cunjin Xue, Wanjiao Song, Lijuan Qin, Qing Dong, and Xiaoyang Wen. A spatiotemporal mining framework for abnormal association patterns in marine environments with a time series of remote sensing images. *International Journal of Applied Earth Observation and Geoinformation*, 38:105–114, 2015. doi:10.1016/j.jag.2014.12.009.