Motive-based Search
Computing Regions from Large Knowledge Bases using Geospatial Coordinates

Liliya Avdiyenko[1], Martin Nettling[1], Christiane Lemke[1],
Matthias Wauer[1], Axel-Cyrille Ngonga Ngomo[2] and Andreas Both[1]
[1]*R&D, Unister GmbH, Leipzig, Germany*
[2]*Universität Leipzig, IFI/AKSW, Leipzig, Germany*

Keywords: Information Search and Retrieval, Online Information Services.

Abstract: To create a better search experience for end users and to satisfy their actual intents even for vaguely formulated queries, a contemporary search engine has to go beyond simple keyword-based retrieval concepts. For a geospatial search, where user queries can be quite complex such as "places for winter sport holidays and culture in Central Europe", we introduce the notion of geospatial motifs denoting traits of geographical regions. Defining a motif by a set of geospatial entities with certain characteristics, we present an approach to inferring important regions for the motif based on density of these entities. The evaluation of the approach for several motifs showed that the inferred regions are among the most popular places for a motif of interest according to the opinion of several experts and official rankings. Thus, we claim that the presented semi-automatic process of detecting regions for geospatial motifs can contribute to more powerful and flexible search applications which are able to answer user queries containing complex geospatial concepts.

1 INTRODUCTION

The evolution of web search engines was described in (Broder, 2006). In their first generation (1994-1997, e.g., Excite, Lycos, AltaVista), retrieval was based on "on-page" textual data only. Search engines of the second generation using "off-page" web-specific data for scoring brought the much needed push in result quality from 1998 onwards, e.g., the PageRank algorithm (Page et al., 1998). The third generation of search engines integrates multiple data sources and focuses on answering the user need behind a query, which remains an active research area until now.

Online search systems now have more information at their disposal than ever. This explosive growth of available data and the described paradigm shift on what web search should offer creates a strong potential for future industrial applications, such as geospatial search in the e-commerce travel vertical. When searching for holiday destinations, users often use concepts from their everyday language in so-called *motive-based searches*, which are hard to process correctly for first and second generation search engines. A user will probably issue a search for "*places **ideal** for winter holidays*" as opposed to "*places in the mountains with at least three ski lifts, snow from December to March and ski rental facilities*" which could be a common interpretation of ***ideal***.

Such search scenarios can be tackled efficiently only by conducting a deeper analysis of the data available and matching it to the user's intent. For this, we introduce the notion of *motifs* to describe characteristics of geospatial regions, which could be any concept a user is looking for, e.g., *winter holiday* or *culture*. These motifs then serve as supporting data objects in general geospatial search scenarios: one can think about a touristic search engine delivering hotels located in regions corresponding to a certain motif.

The regions might well be created by experts. However, as a number of motifs gets larger, hiring experts becomes problematic due to availability, time and cost constraints. Therefore, we propose a semi-automatic approach. Defining a motif by a set of geospatial entities with motif-related characteristics, the approach looks for regions with a high density of these entities using Voronoi tessellation (Okabe et al., 1992). In addition, the inferred regions have the following desirable properties. They are of an arbitrary size and shape, they are scored, allowing retrieval of the most relevant search results on any geospatial

Avdiyenko, L., Nettling, M., Lemke, C., Wauer, M., Ngomo, A. and Both, A..
Motive-based Search - Computing Regions from Large Knowledge Bases using Geospatial Coordinates.
In *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015) - Volume 1: KDIR*, pages 469-474
ISBN: 978-989-758-158-8

scale, and they are robust with respect to the completeness of entities used to build them.

On the data level, many challenges need to be addressed. Though the number of publicly available data sets is high, they might contain only a limited amount of information pertaining to a given motif. While knowledge bases of geospatial entities such as LinkedGeoData and NaturalEarth do exist, the quality of the contained entities differs with respect to the number and the quality of annotations or links to other data sets. Consequently, given available annotated entities describing a certain motif and their coordinates, our approach infers motif-specific regions.

Section 2 reviews the related work. The approach based on Voronoi tessellation is explained in Section 3. Experiments are described in Section 4, whereas Section 5 presents a discussion. Section 6 concludes.

2 RELATED WORK

There are many ideas on how to enhance information contained in large linked knowledge bases for increasing its applicability for search applications. Several sources providing types for DBpedia resources are available, e.g. DBpedia ontology (DBPO), YAGO (Suchanek et al., 2008). However, being based on Wikipedia categories and infoboxes, they neither assign types to all DBpedia entities, nor can they cover all existing concepts. One of the recent works demonstrated the automatic type inference within one large knowledge base (Paulheim and Bizer, 2013). Additional external information sources can be used as well as shown in (Gangemi et al., 2012), where the natural language definitions of Wikipedia pages were extracted and parsed to produce an OWL representation. Applying heuristics on the output graph and performing word sense disambiguation, types were identified and linked to other ontologies. While these approaches focus on completing the assignment of existing types to entities, we are looking for more flexible higher level characteristics (motifs). Such motifs could be defined by possibly unrelated entity types and be treated as new types or categories themselves.

Finding abstractions and grouping entities is an area closer to our work. Topic modeling based on the DBpedia graph, which was introduced in (Hulpus et al., 2013), identifies topics with the most promising DBpedia concept. In (Titze et al., 2014), an approach to identifying a common label in the Wikipedia category tree using clusters based on finer-grained category annotations of each entity was presented. Membership of a DBpedia entity to a given freely defined concept was predicted using machine learning algorithms based on features from several knowledge bases in (Both et al., 2015). In contrast, our method is designed to be maximally independent of potentially error-prone annotations within a knowledge base. It relies only on coordinates of motif-related entities and their primary types allowing to assign them to a motif.

A common use of geospatial information for search engines providing top-N search results is to influence result ranking for personalization depending on user context and location, e.g., as described in (Bennett et al., 2011). Spatialization can also be used to enhance a user experience in exploratory search systems, e.g., (Adams et al., 2015). The idea of clustering geospatial entities for information analysis was pursued in (Wang et al., 2010), where different clustering methods were modeled in an ontology and applied to Canadian population data. In (De Jonge et al., 2012), a phone call activity data set was used to infer dynamic population density. To estimate areas each mobile phone tower serves, the authors used Voronoi tessellation (Okabe et al., 1992), the plane partitioning, which is also used here to compute motif regions.

3 APPROACH

Our goal is to determine geospatial regions $R = \{R_1, R_2, \ldots, R_n\}$ for a geospatial motif $\mathcal{M}$ and to assign a value I_i for each region R_i to quantify its importance. Let $\mathcal{M}$ be characterized by a set of geospatial entities $E = \{E_1, E_2, \ldots, E_n\}$ of a certain type or several types, e.g., museums and theatres for a cultural motif. Then, we say that a region for a motif $\mathcal{M}$ should have a high density of entities defining $\mathcal{M}$.

The approach is based on Voronoi tessellation. Given a set of n entities as seeds, a spatial plane is partitioned into n tiles in a way that every fictional point of the plane is associated with its closest entity and therefore lies in the corresponding tile (Okabe et al., 1992). Consequently, plane subregions where the density of the entities is high contain many smaller tiles, whereas sparse subregions are represented by a few tiles of larger size. This is a key idea of the tessellation density estimation technique that assigns a constant density value to all points of a cell, which is inversely proportional to its area (Browne, 2007). In the following, we describe how we adapt Voronoi tessellation and the related density estimation technique to determine geospatial regions and their importance.

First, the Voronoi tessellation for the set of n geospatial entities E is calculated resulting in n enclosing tiles $T = \{T_1, T_2, \ldots, T_n\}$. However, the tiles can be very large for regions with few entities, thus, we limit their size. Similarly to the idea of (Browne,

2007) for solving the problem of unbounded tiles, the following steps are done:

- A circle C_i of the fixed radius is calculated around each entity E_i.
- The intersection of a tile T_i with a circle C_i is computed, giving the final polygon[1] $R_i = T_i \cap C_i$.

An importance score I_i of every region R_i is calculated as proportional to its density, i.e., inversely proportional to its area $A(R_i)$: $\hat{I}_i = \frac{1}{A(R_i)}$.

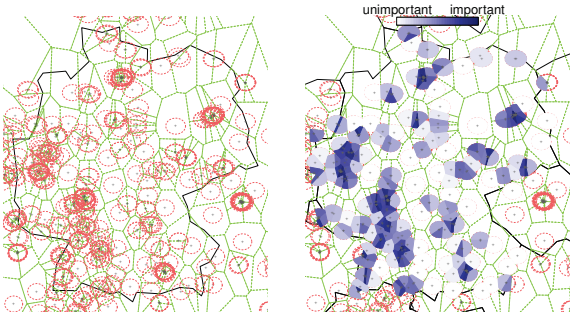
4 EXPERIMENTS

To evaluate the proposed approach to inferring spatial motif-specific regions, we picked three motifs: education, culture and winter sport holidays. In our opinion, the educational and the cultural motif are characterized not only by geospatial components, but also by different qualitative aspects of both motif-specific entities and regions, e.g., the academic and employer reputation of an educational entity. Therefore, we considered them to be more complex. At the same time, the winter sport motif could be defined only by quantity of entities specific to winter sport holidays such as ski lifts. Thus, we evaluate two scenarios: how much can be inferred about a region only having coordinates of the motif-related entities, and how good our approach is for purely geospatial motifs.

4.1 Educational Motif

First, we consider an educational motif in Germany. This motif is defined by entities of educational institutions, such as schools and universities, extracted from the DBpedia[2] as belonging to the type dbpo:EducationalInstitution[3]. There were 646 entities located within the bounding box of Germany, which were used as the seeds for Voronoi tessellation. Figure 1 illustrates the steps of our approach: a) partitioning the bounding region into Voronoi cells and introducing restricting circles of the chosen radius[4], and b) defining the final regions as intersections of tiles and circles as well as their importance scores.

As it is difficult to assess the quality of abstract regions, the evaluation was done on German cities extracted from the Natural Earth map data sets (NaturalEarth). This data was used to make a comparison



(a) (b)

Figure 1: Steps of the approach for the educational motif. (a) Tiles of the Voronoi tessellation and bounding circles around entities. (b) Final polygons as intersection of tiles and circles colored w.r.t. to their importance scores.

with the Breiman's random forest (Breiman, 2001) predicting the importance of populated places for the educational motif (Both et al., 2015). The classifier was trained using features constructed from DBpedia, Natural Earth and GeoNames (GeoNames) data sets, such as population of a place, its area as well as features based on the number of educational entities in its proximity. Note that our approach uses only entity coordinates to give the same answer, i.e., whether a place lies in the region important for education.

For the first evaluation, 3 experts rated 47 German cities pairwise and indicated whether the first city in the pair is more important for education than the second one. The result consists of 3 lists with 181[5] comparisons each, which can be presented as binary rating vectors. Further, we picked only pairs where all experts agreed, resulting in 91 pairs. The classifier's rating vector is based on comparisons of the probabilities of the places being important for the motif, that are produced by the classifier. The rating vector for our approach is constructed by comparing the importance scores of the regions containing the considered cities. Table 1 presents Cohen's κ coefficients for the classifier and the Voronoi approach indicating their agreement with the experts. Despite using only entities' coordinates, our approach shows better results.

Table 1: Evaluation metrics for the educational and the cultural motifs: Cohen's κ coefficients for the classifier and the Voronoi approach illustrating agreement with experts; precision at 10 w.r.t. the official ranking lists of German cities.

	κ w.r.t. experts		P@10 w.r.t. RL	
	class.	Voronoi	class.	Voronoi
education	0.47	0.55	0.7	0.6
culture	-	0.54	-	0.7

The second evaluation is based on the official

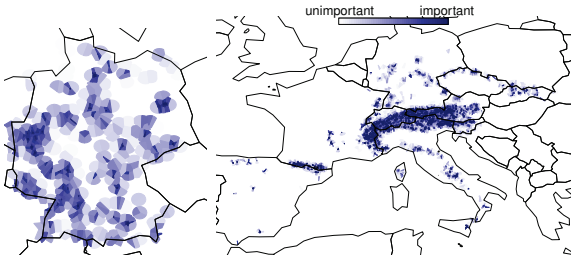
[1] A "polygon" and a "region" are used interchangeably.

[2] The latest revision was used which captures Wikipedia data extracted in May 2014.

[3] dbpo: stands for http://dbpedia.org/ontology/.

[4] $r = 25$ km was chosen for the considered travel use case. Thus, a touristic search engine does not offer hotels located further than 25 km from a motif-relevant entity.

[5] 181 comparisons ($k \log k$) are needed to reconstruct the true order of $k = 47$ items



(a) (b)

Figure 2: Regions for (a) the cultural motif in Germany and (b) the winter sport motif in several European countries.

ranking of German universities (U.S. News&World Report). Using the classifier probabilities and the Voronoi importance scores, we ranked 47 German cities and compared P@10 considering the official ranking as the true ranking, see Table 1. As a result, our approach performs slightly better than the classifier w.r.t. the expert rating. Considering the official ranking, our approach is not as good as the classifier. However, when randomly drawing 10 cities, the probability to get at least 6 cities correct out of 47 is only 0.0029, thus the achieved precision is acceptable.

4.2 Cultural Motif

Similarly, we evaluated the quality of the approach for cultural regions in Germany. The cultural entities were extracted from DBpedia and YAGO knowledge bases. The entities included, for example, museums and operas corresponding to the types dbpo:Museum, dbpo:Opera (see the full list in the online appendix). There were 1113 cultural entities within the bounding box of Germany used for Voronoi tessellation. Here, we treated all entities equally. Alternatively, one can assign weights to every entity type influencing the importance score of the resulting tiles.

On Figure 2(a), the final polygons colored w.r.t. their importance scores are plotted on the map of Germany. Though cultural highlights such as Berlin or Cologne are identified correctly, we performed a formal evaluation similar to the one done for the educational motif. In the first evaluation block, 3 experts rated pairs of 47 German cities. For further evaluation, we used 108 pairs where they agreed on whether the first city is more important for culture than the second one in the pair. As we do not have another algorithm for comparison such as the above-described classifier for the educational motif, here the approach was compared to expert ratings only. Its agreement with the experts is given in Table 1.

The precision of the approach for ranking the first 10 cities from the official ranking list (HWWI/Berenberg, 2014) can be seen in Table 1, which can be considered as a good result.

4.3 Motif for Winter Sport Holidays

Finally, the approach was evaluated for the more abstract "winter sport holidays" motif. Thinking about winter sport regions, one imagines sparsely populated areas in mountains equipped with skiing facilities. Thus, this motif was defined by entities like ski lifts and ski rentals extracted from the LinkedGeoData knowledge base (Stadler et al., 2012; LinkedGeoData) as entities of the types lgdo:ChairLift[6], lgdo:SkiRental etc (see the online appendix).

In order to investigate the approach for larger regions, we changed the "scale" and considered several countries of Central Europe, i.e., Germany, France, Italy, Switzerland, Austria, Slovenia, the Czech Republic, Slovakia and Poland. The Voronoi tessellation was run using 16521 winter sport entities within the bounding box of all considered countries. Figure 2 illustrates the inferred regions. Well-known mountain ranges such as Alps or Pyrenees can be easily identified validating our approach. In addition, we evaluated formally the precision of the detected regions on a list of places from the corresponding countries.

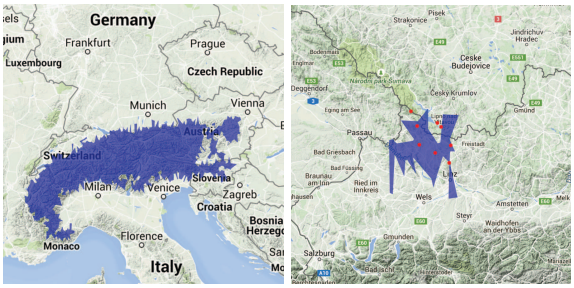
325 ski resorts were extracted from DBpedia as entities of the types like yago:SkiAreasAndResortsInGermany[7] etc. 166 places not appropriate for skiing in the considered countries were selected manually. Thus, the intention was to check whether the detected winter sport regions contain the test ski resorts and do not include the negative examples. The experiment consisted of 100 iterations where a balanced set of 200 places was sampled from the original list. Our approach considered a place as a ski resort if it is located in the polygon with an importance score above a certain threshold[8]. The true positive and true negative rates averaged over 100 iterations are 0.88 ± 0.042 and 0.91 ± 0.014, respectively.

As it is not trivial to get "real" polygons of mountain areas to validate our approach, we engaged people in this task. For this, the final neighboring Voronoi tiles were merged so that there were several regions instead of a large set of individual ones. Note that it was done only for visualization purposes as after merging the precision of importance scores on a finer spatial scale deteriorates. Thus, the 10 biggest merged regions were given to 5 experts to decide whether they are suitable for winter sport vacations. With an aver-

[6] lgdo: is http://linkedgeodata.org/ontology/

[7] yago: stands for http://dbpedia.org/class/yago/

[8] It was selected via cross-validation and corresponds to the importance score of a spherical polygon with $r = 7$ km.



(a) (b)

Figure 3: Examples of rated winter sport regions: (a) The Alps rated by all as correct. (b) Austrian side of tri-border region of Germany, Austria and the Czech Republic, it was rated by all as wrong. The red points indicate ski resorts making (b) suitable for the ski holidays.

age agreement between people of $\kappa = 0.71$, on average 8.6 out of 10 regions were rated as correct. While all rated regions can be seen in the online appendix, Figure 3 shows two examples (Kahle and Wickham, 2013): (a) presents the Alps rated as correct by all experts, (b) illustrates a region which in the rater's opinion is not suitable for skiing. However, it contains small ski resorts, see the red points plotted over the detected region. Thus, our method was not wrong in detecting this region but it is not as well-known.

5 DISCUSSION

5.1 Robustness

It is likely that the computed regions will still be valuable if a few of the relevant entities are missing for their calculation. An experiment was performed to assess robustness of the approach. 100 data subsets were randomly sampled without replacement for different sample sizes containing from 10% to 90% of the original data. Based on each of the samples, Voronoi tessellation was recomputed to produce new motif regions. As before, their quality was measured by the agreement with experts for pairwise comparison of cities for the educational and cultural motifs.

Figure 4 shows the results. For the cultural motif, one can see that performance remains similar with increasing standard deviation until using as little as 30% of the data. For the educational motif, performance drops steadily, but can still be considered acceptable when using 80% of the data. This behavior can be explained by the different number of available entities for the different motifs, i.e. 1113 and 646 entities for the cultural and the educational motifs, respectively. Hence, robustness increases with the number of points used for the computation of the regions.

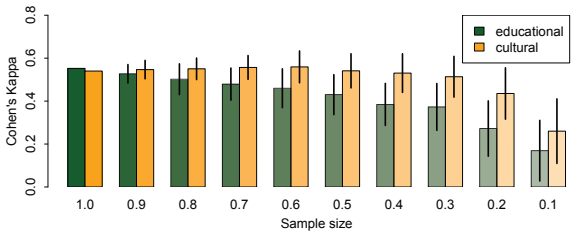


Figure 4: Agreement of the Voronoi approach with experts depending on sample size, averaged over 100 iterations.

5.2 Runtime

Experiments were run on a Virtual machine with Ubuntu 64bit allowed to use 1578Mb of memory and 4 of 8 cores of the AMD-V CPU. Execution time of all steps was 8.27s for 646 entities, 18.9s for 1113 entities and 478s for 16521 entities, which depends solely on the number n of entities. The implementation of the tessellation is an $O(n\log(n))$ algorithm. All other steps run in $O(n)$. Thus, for using the approach in search engines in real time, regions must be preprocessed and stored in databases like PostGIS optimized for geospatial requests. However, note that our approach is not performance optimized, yet.

5.3 Quality of the Approach

We investigated the ability of the approach to rank cities w.r.t. the educational and cultural motifs. Though reproducing official rankings was not our primary intention, the results show that the approach is able to detect important motif-specific regions. Though, the agreement with experts is moderate, it can be explained by the fact that we used only the geospatial information for defining the motifs and ignored other qualitative features. However, this was done on purpose to show that the reasonable accuracy can be achieved using only the entity coordinates.

The quality of the inferred winter sport regions is promising. In the opinion of our experts, they can often resemble parts of the known mountain ranges. The approach remains also precise for small winter sport areas and non-mountain regions. Thus, we expect that integrating inferred regions for different motifs in a retrieval system will increase user satisfaction.

Obviously, there is space for improvement. To decrease the noise influence on the density estimates, one could use bootstrapping. Discretization of importance scores goes in the similar direction but can impair the accuracy on the small scale. Various clustering methods for producing motif regions can be investigated as well.

6 CONCLUSIONS

We presented an approach for bridging the gap between expectations users have of search systems and data available in (linked) knowledge bases. We proposed a solution as to how this data can be utilized in applications without having deeper insights into its structure. This is particularly valuable as knowledge bases permanently gain complexity via interlinking or automatic enrichment processes. The experiments showed that motifs based only on entity coordinates can become useful in complex search scenarios. The approach is flexible with regard to the boundaries and size of the initial area and robust with respect to missing relevant entities. Moreover, it assigns scores to regions to allow retrieval of the most relevant regions for any map zoom level.

Future work can continue in many directions. Using additional features like the importance of entities could result in a performance boost, although it would increase the complexity of the approach. The same applies to giving different weights to several entity types and calculating weighted intersections of the different tessellations. In addition, we plan to implement a search system using precalculated regions for a larger number of motifs and perform a user study to measure their impact in an actual search scenario.

ACKNOWLEDGEMENTS

We thank Luise Erfurth and other colleagues at Unister. This work has been supported by the European Union's 7th Framework Programme in the scope of the project GeoKnow (GA no. 318159).

REFERENCES

Adams, B., McKenzie, G., and Gahegan, M. (2015). Frankenplace: Interactive thematic mapping for ad hoc exploratory search. In *WWW'15*, pages 12–22.

Bennett, P. N., Radlinski, F., White, R. W., and Yilmaz, E. (2011). Inferring and using location metadata to personalize web search. In *SIGIR'11*, pages 135–144. ACM.

Both, A., Avdiyenko, L., and Lemke, C. (2015). Computing geo-spatial motives from linked data for search-driven applications. In *Know@LOD at ESWC'15*.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Broder, A. (2006). The future of web search: From information retrieval to information supply. In *Next Generation Information Technologies and Systems*, pages 362–362. Springer.

Browne, M. (2007). A geometric approach to non-parametric density estimation. *Pattern Recognition*, 40(1):134–140.

DBPO. http://www.dbpedia.org/ontology.

De Jonge, E., van Pelt, M., and Roos, M. (2012). Time patterns, geospatial clustering and mobility statistics based on mobile phone network data. In *FCSM'12*.

Gangemi, A., Nuzzolese, A. G., Presutti, V., Draicchio, F., Musetti, A., and Ciancarini, P. (2012). Automatic typing of DBpedia entities. In *The Semantic Web – ISWC 2012*, pages 65–81. Springer.

GeoNames. Geonames geographical database. http://www.geonames.org/. Accessed 2015-03-15.

Hulpus, I., Hayes, C., Karnstedt, M., and Greene, D. (2013). Unsupervised graph-based topic labelling using DBpedia. In *WSDM'13*, pages 465–474. ACM.

HWWI/Berenberg (2014). HWWI/Berenberg Kulturstädteranking 2014. Die 30 groessten Städte Deutschlands im Vergleich.

Kahle, D. and Wickham, H. (2013). ggmap: Spatial visualization with ggplot2. *The R Journal*, 5:144–162.

LinkedGeoData. http://linkedgeodata.org/sparql. Accessed 2015-05-21.

NaturalEarth. http://www.naturalearthdata.com/. Accessed 2015-03-15.

Okabe, A., Boots, B., and Sugihara, K. (1992). *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. John Wiley & Sons, Inc., New York, USA.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University.

Paulheim, H. and Bizer, C. (2013). Type inference on noisy rdf data. In *The Semantic Web – ISWC 2013*, pages 510–525. Springer.

Stadler, C., Lehmann, J., Höffner, K., and Auer, S. (2012). LinkedGeoData: A Core for a Web of Spatial Open Data. *Semantic Web Journal*, 3(4):333–354.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2008). Yago: A large ontology from Wikipedia and WordNet. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(3):203–217.

Titze, G., Bryl, V., Zirn, C., and Ponzetto, S. P. (2014). DBpedia Domains: augmenting DBpedia with domain information. In *LREC'14*.

U.S. News&World Report. Best Global Universities in Germany. http://www.usnews.com/education/best-global-universities/germany. Accessed 2015-05-20.

Wang, X., Gu, W., Ziebelin, D., and Hamilton, H. (2010). An ontology-based framework for geospatial clustering. *International Journal of Geographical Information Science*, 24(11):1601–1630.

APPENDIX

Additional material is accessible online: http://bit.ly/1L2GKIe.