HOW TO GEOTAG THE WEB?

Alexander Czech

E280 - Center of Regional Science at Vienna University of Technology

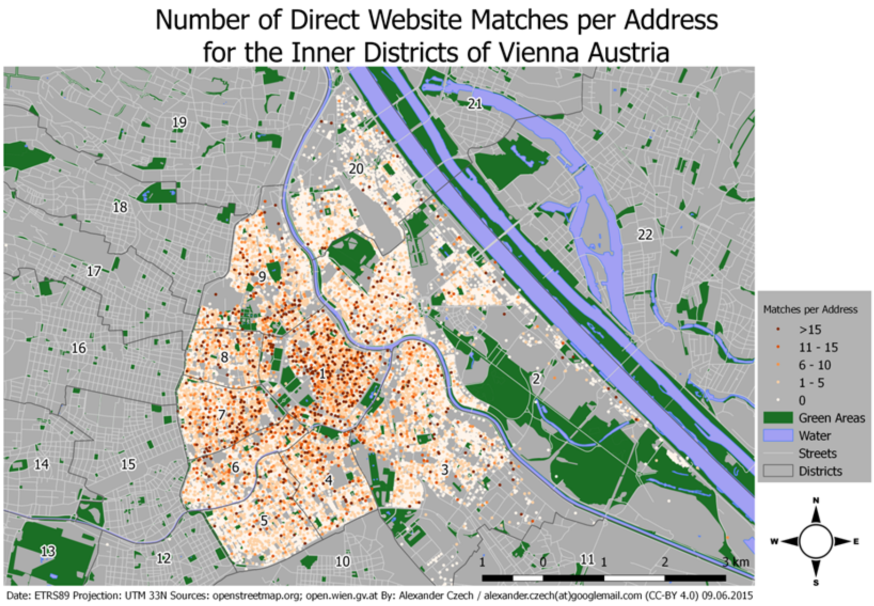
Spatial information guides people’s activities in space. In the last decades the World Wide Web has also become a widely used source for this kind of information. People use web search engines to query spatial information, for example to look up the best French restaurant, a still open grocery store, or a scenic hiking route.

Most research around “Neogeography” or “Big Geo-Data” focuses on already geotag data originating on social media. But in comparison to the web, this is only a small portion of all the data, and it can be skewed by a vocal minority [1]. To leverage a big corpus of web data, this corpus needs to be geotagged and the relation between geotag and data needs to be proven.

In previous work a geotag on a website was defined as a human readable postal address. The study was done for a select subset of HTML documents belonging only to the .at top-level domain and the address pool was restricted to addresses from the inner Viennese districts (1.-9. and 20.) [2]. This work resulted in an information distribution landscape that can be seen in **Figure 1**.

The current spatial area of interest is the urban system of Randstad in the Netherlands. The Website dataset will again be sourced from the Common Crawl Project, who provides semimonthly raw crawled websites data. One crawl of which is about 81 terabyte in compressed archives [3].

The apache spark cluster frame work is used to process this dataset. After preprocessing the data an address matching routine, will be employed to find addresses within HTML documents.



Number of Direct Website Matches per Address for the Inner Districts of Vienna Austria

Date: ETRS89 Projection: UTM 33N Sources: openstreetmap.org; open.wien.gv.at By: Alexander Czech / alexander.czech(at)googlemail.com (CC-BY 4.0) 09.06.2015

Figure 1 Distribution of addresses joined to HTML documents [2].

The next step is to show that the data that was geotagged is related to the addresses. Tools from the natural language processing and information retrieval domain are used to achieve this. All documents related to the addresses are transformed into vector representations. The paragraph vector distributed bag of words (PV-DBOW) method is applied for here [4], it builds conceptually upon the skip-gram model to represent words in a vector space [5]. Four models for each language, Dutch, English, French and German are trained on respective Wikipedia corpora. These models are then used, to infer document vectors for documents in the corresponding language, which have been geotagged to an address. This approach was selected because PV-DBOW has been shown to perform robustly in similarity tasks when trained on external corpora [6].

To test the relation between data and address the average similarity between documents geotagged to the same address is calculated through cosine similarity of the document vectors [7] and compared to the cosine similarity of a random sample. The expectation being that the documents matched to the same addresses are more similar to each other than a random sample of documents to each other, thus showing that there is a relation of documents and address.

To further verify this relation, all addresses are grouped into predefined address group (e.g. schools, stores, hotels, and so on). All documents related to all addresses of one group are then compared to each other similar to the process above. The expected results again being that documents in one group are more similar to each other than compared to a random sample.

After being able to proof the relation between data and geotag, the idea is to use this new source of data to support the research on urban spaces and urban systems. As an example the web can be seen as a force of globalization. These have very different and uneven effect on space[8] and **Figure 1** already suggests such uneven effects, because a spatial pattern in the distribution can be observed. Further the concept of relational (Urban-)spaces can be explored through the relation of data and space [9]. Also the research on city networks and urban systems that focuses on the connection between cities [10, 11, 12] could be improved by examining the connection of cities through the webgraph. There is already research around identifying, measuring and quantifying communities through the webgraph structure that could be adapted to such a task [13, 14].

REFERENCES

[1] J. W. Crampton, M. Graham, A. Poorthuis, T. Shelton, M. Stephens, M. W. Wilson, and M. Zook, “Beyond the geotag: situating ‘big data’ and leveraging the potential of the geoweb,” *Cartogr. Geogr. Inf. Sci.*, vol. 40, no. 2, pp. 130–139, 2013.

[2] A. Czech, “Geospatial Information Retrieval for POIs with the use of a Data Mining System,” 2015.

[3] Common Crawl, “Common Crawl.” Mar-2012.

[4] Q. Le and T. Mikolov, “Distributed Representations of Sentences and Documents,” *Int. Conf. Mach. Learn. - ICML 2014*, vol. 32, pp. 1188–1196, 2014.

[5] T. Mikolov, G. Corrado, K. Chen, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” *Proc. Int. Conf. Learn. Represent. (ICLR 2013)*, pp. 1–12, 2013.

[6] J. H. Lau and T. Baldwin, “An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation,” in *ACL*, 2016, pp. 78–86.

[7] C. D. Manning, P. Raghavan, and H. Schütze, *An Introduction to Information Retrieval*. 2009.

[8] S. Sassen, *The global city*. 1991.

[9] P. Weichhart, *Entwicklungslinien der Sozialgeographie*. Steiner, 2006.

[10] J. V. Beaverstock, R. G. Smith, and P. J. Taylor, “A roster of world cities,” *Cities*, vol. 16, no. 6, pp. 445–458, 1999.

[11] J. H. Choi, G. A. Barnett, and B. S. Chon, “Comparing world city networks: A network analysis of Internet backbone and air transport intercity linkages,” *Glob. Networks*, vol. 6, no. 1, pp. 81–99, 2006.

[12] H. Kramar and J. Kadi, “Polycentric city networks in central-eastern Europe: Existing concepts and empirical findings,” *Geogr. Pol.*, vol. 86, no. 3, pp. 183–198, 2013.

[13] J. M. Kleinberg, “Hubs, authorities, and communities,” *ACM Comput. Surv.*, vol. 31, no. 4es, p. 5, 1999.

[14] G. Flake, S. Lawrence, L. Giles, and F. Coetzee, “Self-Organization of the Web and Identification of Communities,” *IEEE Comput.*, vol. 35, no. 3, pp. 66–71, 2002.