The Wikipedia Location Network – Overcoming Borders and Oceans

Johanna Geiß[1], Andreas Spitz[1], Jannik Strötgen[1,2], Michael Gertz[1]
[1] Institute of Computer Science, Heidelberg University, Germany
[2] Max-Planck-Institute for Informatics, Saarbrücken, Germany
{johanna.geiss, spitz, stroetgen, gertz}@informatik.uni-heidelberg.de

ABSTRACT

In social network analysis and information retrieval, research has recently been devoted to the extraction of implicit relationships between persons from unstructured textual sources. In this paper, we adapt such a person-centric approach to the extraction of locations and build the *Wikipedia Location Network* based on co-occurrences of place names in the English Wikipedia. We summarize the network's characteristics and demonstrate its value for future location relationship analysis tasks.

CCS Concepts

•**Computing methodologies** → **Natural language processing;** •**Information systems** → *Information retrieval;*

Keywords

Location similarity, location network, Wikipedia

1. INTRODUCTION AND RELATED WORK

Wikipedia is a free and convenient source of data that is increasingly being used to construct knowledge bases for natural language processing tasks. While some approaches focus on the extraction of structured information, most data on Wikipedia is contained in unstructured text. Here, a natural assumption is that entities have something in common if they are mentioned in the same context (i.e., on the same Wikipedia page). As a result, these co-occurrences of entity mentions can be exploited to create a network of relations between locations that is not based on geographic information but rather on some semantic connection. For example, such a network might contain a relation between *Wacken, Germany*, and *Springfield, Massachusetts*, because they are both known for music festivals.

Recently, Geiß et al. showed that such an approach is feasible for the construction of social networks, which can be built not only from data that contains explicit relationships, but also from unstructured text based on co-occurrences [1].

GIR '15, November 26-27, 2015, Paris, France

DOI: http://dx.doi.org/10.1145/2837689.2837694

In this paper, we adapt their approach and extract the *Wikipedia Location Network*[1]. We focus on relations between locations instead of persons to obtain a network that can be used in NLP tasks such as disambiguation, or even coreference resolution. To this end, we identify significant co-occurrences of location mentions on Wikipedia by weighting relations based on toponym distances within the text.

While social network analysis is focussed on relations between persons, there are also some works that deal with geographical data in a similar fashion. In particular, Liu et al. present an approach to study the relatedness of toponyms [2]. By using provinces of China as a fixed set of locations, they search for province toponym co-occurrences on news Web pages with an IR system, analyze the co-occurrence matrix and show that neighbouring locations have similar co-occurrence patterns. In contrast to this approach, we do not limit the locations to a small fixed set but use Wikipedia's wikilinks and Wikidata information to extract location mentions from the Wikipedia content pages.

The probably most similar work to our approach is by Quercini and Samet [3]. They study the spatial relatedness in Wikipedia, explore graph-based similarity measures to determine related concepts for each location and build so-called local lexicons for all occurring locations, which are a valuable resource for toponym disambiguation. While they also use wikilinks to extract similarities of place mentions, our approach differs in a significant way, namely the textual distance of distinct mentions within each of the Wikipedia documents. Since it has been shown that co-occurrence by itself is not sufficient to distinguish between spurious and meaningful co-mentions of persons in Wikipedia [1], we adapt this inclusion of a text-based distance between mentions to the extraction of a network of locations that is more finely nuanced than previously extracted local lexicons.

2. WIKIPEDIA LOCATION NETWORK

We build the Wikipedia Location Network by using wikilinks (WL) in the English Wikipedia and information on locations and places extracted from Wikidata. WLs are links within a Wikipedia page to another Wikipedia page. To determine which WLs refer to a location, we use the link information in Wikidata: if the target of the link is equal to the link of a Wikidata item that refers to a place or location, we assign the Wikidata entity id to the WL. A Wikidata item is considered to refer to a place if it has coordinate location. As of June 2, 2015, the English Wikipedia contains

[1]The network is available for download at our website: `http://dbs.ifi.uni-heidelberg.de/index.php?id=data`

4.8M content pages and 78.8M WLs, of which 20.6M refer to 724,989 different locations that we use to construct the network.

2.1 Wikidata

Wikidata is a free, collaboratively edited, multilingual database launched in October 2012 [5]. As of June 1, 2015, Wikidata includes more than 17.7M items representing real life topics, concepts, and subjects, 957,412 of which refer to a location. Where available, we extract additional information for these locations (e.g. , the population). Based on this information, we create a conceptual hierarchy for involved toponyms, namely *city* < *country* < *continent*.

2.2 Network Construction

We construct the location network from co-occurrences of individual location mentions on Wikipedia based on the approach for extracting social networks [1]. First, we build a multigraph M in which nodes represent toponyms. Edges between nodes v and w correspond to an instance i in which the toponyms co-occur on a Wikipedia page and are weighted by $\varphi(v,w,i) = \exp\left(-\frac{1}{2}dist(v,w)\right)$, where $dist$ is the distance in sentences between v and w. In order to aggregate parallel edges of M and obtain a simple graph $G = (V,E)$, a new weight for the resulting edges is computed as the cosine similarity of adjacency vectors in the weighted node-edge incidence matrix of M. An equivalent formulation that does not require a matrix representation and is thus easier to compute is based on the neighbourhoods n_v and shared neighbourhoods $n_v \cap n_w$ of nodes, i.e., the distance-cosine weight for edges:

$$dicos(v,w) := \frac{\sum_{e \in n_v \cap n_w} \varphi(e)^2}{\sqrt{\sum_{e \in n_v} \varphi(e)^2}\sqrt{\sum_{e \in n_w} \varphi(e)^2}}$$

3. RESULTS

Based on the Wikipedia dump from June 2, 2015, we use the method described above to construct a network containing 178,890,238 edges between 723,779 locations.

3.1 Network Properties

The network has a number of interesting characteristics that set it apart from the Wikipedia Social Network, despite the similarities in its construction. Most notably, the clustering coefficient of the network is high ($cc = 0.56$) while the density is low ($\delta = 6.8 \cdot 10^{-4}$). If we apply a threshold of $t = 0.0002$ to the edge weights, the clustering coefficient of the resulting network increases to $cc = 0.96$ (i.e., almost perfect clustering), while we still retain 56.6% of edges. This makes it possible to extract cohesive local structures around given place mentions even without further clustering or community detection. For an example, see Figure 1, where we applied a higher threshold to obtain a finer clustering.

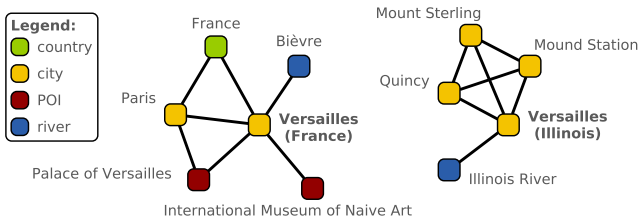

Figure 1: Direct neighbourhood for two instances of *Versailles* in the network with an edge weight threshold of $t = 0.0004$.

3.2 Hierarchy-based Evaluation

Since the Wikipedia Location Network is constructed from co-occurrences within Wikipedia pages, the relations we find between locations represent a semantic relationship that is not limited to spatial proximity. Thus, the network contains similarities between locations around the world that host similar events, such as the Formula One racing street circuit in *Sochi, Russia*. For example, we find the *Circuit de Monaco*, *Hockenheimring* (Germany), *Melbourne Grand Prix Circuit* (Australia), *Circuit of the Americas* (USA), *Circuit Gilles Villeneuve*, (Canada), *Pescara Circuit* (Italy) and the *Red Bull Ring* (Austria) within the top 15 most similar places in the network.

Although we do not take geographic proximity into account in its construction, the network still contains representations of geographic structures. For an evaluation, we use the hierarchical information extracted from Wikidata as ground truth. For each city and country we select the top ranked neighbour of the next hierarchy level from our network and compare this to the hierarchical Wikidata information. We are able to identify the correct country for 75.4% of the cities, whereas for 7.5% of the cities we find no country in the network. This results in a precision of correctly identified countries of 81.6%. For 163 of 203 countries the correct continent is found, i.e., the precision is 80.3%.

4. CONCLUSIONS AND ONGOING WORK

Based on our evaluation of the Wikipedia Location Network, we find that it represents both spatial similarity in the sense of geographical distance as well as a semantic relatedness between objects that are connected by important events. The structural characteristics of the network allow us to extract such similar objects in meaningful local clusters. We also find that the Wikipedia Location Network is just one of a number of possible networks that can be extracted from the co-occurrences of named entities. Based on these, we currently work on the extraction of events based on connections between named entities and temporal expression, akin to the approach by Spitz et al. [4] and consider the disambiguation and resolution of toponyms and other named entities by leveraging their position and neighbourhood in these Wikipedia networks.

5. REFERENCES

[1] J. Geiß, A. Spitz, and M. Gertz. Beyond Friendships and Followers: The Wikipedia Social Network. In *ASONAM'15*, 2015.

[2] Y. Liu, F. Wang, C. Kang, Y. Gao, and Y. Lu. Analyzing Relatedness by Toponym Co-Occurrences on Web Pages. *T. GIS*, 18(1), 2014.

[3] G. Quercini and H. Samet. Uncovering the Spatial Relatedness in Wikipedia. In *SIGSPATIAL '14*, 2014.

[4] A. Spitz, J. Strötgen, T. Bögel, and M. Gertz. Terms in Time and Times in Context: A Graph-based Term-Time Ranking Model. In *TempWeb '15*, 2015.

[5] D. Vrandečić and M. Krötzsch. Wikidata: A Free Collaborative Knowledgebase. *C. ACM*, 57(10), 2014.