LOCALIZATION OF UNCERTAIN AND FUZZY-BORDERED AREAS BY GEOCODED ARTICELS OF A KNOWLEDGE REPOSITORY

Tobias Dahinden

Institut für Kartographie und Geoinformatik, Leibniz Universität Hannover, Appelstraße 9a, 30167 Hannover, Germany, tobias.dahinden@ikg.uni-hannover.de

Keywords: Information retrieval, localization, labelling, uncertainty, geovisualization

Abstract

In this article we present a method to determine geographic line and area features. The idea is to investigate an article about a geographic feature in a knowledge repository with linked articles including geocoordinates. We collect the coordinates of every linked article and estimate the density of the coordinates in space. It is possible to compare the density of several geographic features and to derive a boundary between these features. As prove of concept we use Wikipedia as such a repository.

Introduction

It is possible to determine a precise position for *point objects*, surveyed *lines* (such as railways, streets, rivulets and rivers) and *areas* (such as lakes and administrative units like parcels, hunting grounds and official borders). However there are geographical features that can be located only rather imprecise. As an example a *landscape's name* belong to a certain range, but this range has no certain border.

Tobler (1970, p. 236) defined "the first law of geography" as following: "Everything is related to everything else, but near things are more related than distant things." Based on this law we assume, that the expressions used in a text about a geographic feature are not only related to this feature. The majority of them are about places inside the geographic feature or on its border. This means also, places outside the geographic feature are referred seldom.

In this paper we present a method to determine such imprecise bordered geographic features. This method consists of comparing links about a geographic feature in a knowledge repository with a corresponding gazetteer. So we need a text about a geographic feature, and a list of coordinates that fits some expressions of the text precise. The *German Wikipedia* is an example of this. Its articles contain a hyperlink structure and some of the linked articles are *geocoded*, that means that linked articles have coordinates assigned. So we do not need any named entity recognition.

For each feature we get a list with geographical places related to. We interpolate these data and get a probability space for the feature. To determine a boundary of a feature we compare

its probability space with the corresponding spaces of its neighbours. The comparison is just to look which space is more possible.

Related work

Label placement in printed maps

In common printed maps landscapes do not use a well defined border – it is not necessary. The labelling gives only a hint, where a certain landscape is located. An example of label-placement for larger areas can be found e.g. in the *Altante mondiale svizzera* (Spiess, 2002). In figure 1 we show an extract containing mountain areas.

Figure 1: Example of label placement for mountain areas (after *Spiess*, 2002, p. 60).

Maps with landscape-borders according to topographic and biologic conditions

The extension of a landscape or a macrochore depends on the concept used for its definition. The landscape is defined according to its use, its soil conditions, and its vegetation. There are several maps made by experts with the borders of such landscapes. Liedtke (2002) made a map with the boundaries of landscapes of the Federal Republic of Germany. We reproduce an extract in figure 2. A similar map was made by Meynen and Schmithüsen (1953-1962). There are also services that provide such information like "LANIS-Bund" with 24 types of landscapes and 858 landscapes.

Figure 2: Lipper-, Weser-, and Leinebergland. Extract form the map *Landschaften* Bundesamt für Kartographie (2002).

Determining districts according to political or technical conditions

Boundaries are assigned by planning or redistricting political districts, sales force deployments, social organization districts, school districts, and territories of distributed service networks (salt spreading operation, winter gritting, waste-collection, health care). Typically technical and political constraints have to be fulfilled. The constraints can even be interdependent.

The approaches for the solution contain location-allocation, set partitioning and some heuristic methods. Usually it is assumed that a region consists of a definite number of smallest units, each one with exact one centre. So this technique cannot be used to define landscapes. An in-depth discussion on area assignment using operation research can be found in a thesis of Schröder (2001).

Organizations for determining names

The topographic names are defined in special commissions, e.g. *UNGEGN - United Nations Group of Experts on Geographical Names.* The names usually are saved in toponymic information systems, also known as gazetteers like *GNIS* (USA), *geoXwalk* (UK), *Geo-Info* (Poland), *GN-DE* (Germany) or *SwissNames* (Switzerland). For German-speaking areas the *permanent committee on geographical name* authors recommendations, e.g. for the boundaries of a certain area (Sievers et al., 2000). An example is the determination of the major regions of Europe (Jordan, 2005).

Localization of non-geographic features

Hecht and Raubal (2008) locate non-geographic expressions. They use the *Wikipedia Article Graph* (WAG). In this graph all articles are nodes and the links are the edges of the graph. The edges of the graph are weighted by the semantic relatedness of the articles. This is a measure based on the number of links in article A and B and the number of links that point from A to B and from B to A. They describe why the WAG is easier to use than the Wikipedia-text-structure.

The major technique is to follow the links of the page. If they find a geocoded article they add its weighted coordinate to the non-geographic feature. The weight is calculated according to the semantic relatedness.

Localization of vague places with knowledge from the web

Jones et al. (2008) tried to retrieve locations by extracting place names by searching the web. They used the first 100 results from search engines like Google. From these pages they extracted the place names by a *named entity recognition* (NER) method. With this method they were able to localize regions and concepts (like *hotel*). In a test for the correct assignment a value of 57% was reached.

Localization of scenes of literature

Piatti et al. (2008) locate the activity zone of literature. The works of fiction can be seen as knowledge repository. However the assignment of scenes to geographic places is not unique. This leads to some problems:

- First, there are several places with the same name (e.g. *Santiago*).

- Second, there are names of people that sound like places (e.g. *Hilton, Paris*).
- Third, some names are alienated or fictitious (e.g. *Gotham City* or Gottfried Keller's *Seldwyla*).

A problem is also to show uncertain areas. They are using fuzzy shapes and animations (ibidem, p. 15 ff).

Visualization of point densities

It is very common to show point distributions with the number of points per unit. Wolff and Asche (2009) show an example. They use the *kernel density estimation* (KDE), also known as *Parzen*-window. An explanation of KDE can be found in de Smith et al. (2006, chapter 4.2.1).

Non-topographical representation of landscapes

Hermann and Leuthold (2003) show in the *Atlas der politischen Landschaften* the distribution of the population of Swiss Cantons in the space of ideology. In this space it is possible to retrieve real landscapes like the *Napf*-region (ibidem, p. 40 ff), although the representation is rather different to the topographic representation.

One map shows whole Switzerland in the space of ideology (ibidem, p. 59). This map is available on http://www.vdf.ethz.ch/service/Atlas/Politlandkarte.jpg.

Wikipedia as knowledge repository and gazetteer

Advantages

The German Wikipedia contains more than 800 000, the English more than 2 Mio. articles. They contain texts about geographical features from all over the world. Thus it should be useful for all kind of maps.

Analyses of Hecht and Raubal (2008, p. 102) support our assumption that the geographical terms used in an article lay inside or on the boundary of that region.

The links of a certain article can be requested through the *Mediawiki* API. The names of the places are provided in a separate *MySQL*-Database. This Database is an extract of Wikipedia and thus the names in the database correspond to the links in the articles. Because of this a request is unique. For example the link *Altdorf* which points to the article *Altdorf _(disambiguation)* has no coordinate but the link *Altdorf_(JU)* has one, because it points to a location. So we do not need NER.

Disadvantages

Although Wikipedia has more than 800 000 articles, some places are missing. For example the place *Negenborn* exists three times in Germany, but there is only one article in Wikipedia by now (Feb. 09). There are also articles with missing coordinates in Wikipedia (e.g. *Schloss_Neuenhinzenhausen*) or in the gazetteer (e.g. *Sempach*, *Sursee*).

Further there are several discrepancies between the different language versions of Wikipedia.

- This concerns the topics: Hecht and Raubal (2008, p. 112) found a domination of German topics in the German Wikipedia.
- It also concerns the content and the number of links: The article *Livinental* (Valle Leventina) has 50 links in the German, 11 in the Romanian, 16 in the Dutch, and 34 in the French. The area is located in the Italian language area but it is not discussed in the Italian Wikipedia.
- As a third it concerns the coordinates that may differ. For example *Patara* is located in the German Wikipedia with 36° 16′ N, 29° 19′ E, and in the English with 36° 15′ 37″ N, 29° 18′ 51″ E.

All these articles were visited in February 2009.

Determine landscapes

Modelling of scatter plots as continuous areas by kernel density estimation

Each coordinate found in a linked article can be seen as a measure for the localization of the geographic feature (random variable). All links have the same accuracy, if not using concepts like the semantic relatedness. That is why we only use direct links as a first approach. Further we assume that every link has the same variance in X and Y.

With the a KDE we can estimate the probability density function of the geographic feature. As a first approach we used the *Epanechnikov*-kernel *(de Smith et. al.* 2006, chapter 4.2.1).

Unfortunately in every kernel it is necessary to fix a bandwidth where the KDE is working on.

- If the bandwidth is too large, the output is the mean of the linked coordinates weighted by the number of links. As an effect small areas disappear.
- If the bandwidth is too small, each linked coordinate can be seen as circles.

A useful bandwidth is depending on the number and distribution of linked coordinates and the map scale.

Boundaries of regions

We estimate the location of regions by comparing the KDE of the linked coordinates of German Wikipedia. An area is allocated to a certain region, if the KDE is higher than the KDE of every other region in consideration.

When we do this, we assume that the Wikipedia knows which sub regions are in a certain region. That this assumption does not hold can be seen on the discussion-page of the article about *Weserbergland* (http://de.wikipedia.org/wiki/Diskussion:Weserbergland).

As a test we determined the *Swiss Cantons* with our method with a bandwidth of 0.25°. The result is displayed in figure 3. Counting the topologic errors, the solution is superior to the ones with bandwidth of 0.2° and 0.3°.

- Positive: The topology for *Geneva*, *Thurgau* and *Glarus* is correct. The borders of *Geneva* and *Basel-City* are very close to the real borders.
- Negative: *Appenzell* (IR and AR) is missing, furthermore the exclaves of *Fribourg*. Instead the map shows some exclaves that do not exist, e.g. there is a part of *Solothurn* and *Zurich* in the northern part of *Ticino*.

As a quantitative result we calculated that 77.8% of the area was allocated correct while 22.2% is wrong.

Figure 3: Swiss Cantons determined using the geocoded links in Wikipedia. The bandwidth for KDE is 0.25°.

Line structures

The determine line structures also the Epanechnikov-KDE is used. We used again the bandwidth of 0.25°. As a test we determined the line where the autobahn A2 of Switzerland would take place. In figure 4 the real autobahn is shown in green, the determined with Wikipedia links in orange-red. It is possible to recognize that the autobahn seems to be important in north and central Switzerland. Unfortunately there is a gap in our result. This gap can be leaded to missing coordinates in the gazetteer.

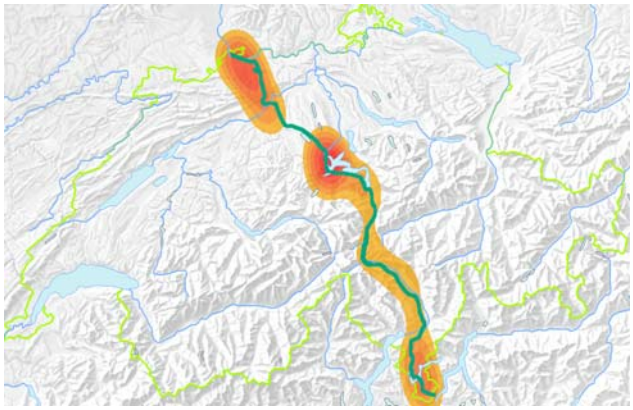

Figure 4: Autobahn A2 of Switzerland. Orange-Red: sphere of influence according to the Wikipedia-Links. Dark-green: real route.

Examples of Applications

- Localization and determination of toponymes of areas: With our technique it could be possible to determine regions (such as *Central-Europe*), landscapes (*Weserbergland*), mountain areas (*Urner Alps*) or parts of a sea (*Aegean Sea*). As a successful example we show in figure 5 where the *Weserbergland* is located according to our technique. Figure 6 shows an example where the major regions of Europe where determined. Unfortunately there were not enough links in the German Wikipedia article to get a reasonable result.

- Linguistics: There are also some articles about literature in Wikipedia. Theoretically it could be possible to analyse the literature like *Piatti et. al.* (2008) did. But practically it will often not lead to a useful result because there are too few geocoded links in the according article. For example the article *Wilhelm Tell (Schiller)* has only four geocoded links: *Altdorf (UR)*, *Rütli*, *Vierwaldstättersee* and *Küssnacht am Rigi*.

- Another Linguistic feature could be the determination of the linguistic variety (map with dialects).

- Determination of the importance of parts of a structure: In figure 7 we show the location of the *Reuss*-valley. It is possible to distinguish between to northern part and the southern part. The estimated density in the northern part is larger than in the southern part. Thus it is possible to assume that the northern part is more important.

- Classification of a network structure and areas: In figure 8 we show the rivers and rivulets of Lower-Saxony. The estimated density of the waterbodies in this area is visualized as blue areas. By comparing these two things it is possible to make a selection of waterbodies for generalization that is based on the importance. The importance can be controlled by the KDE and by merging Wikipedia articles.

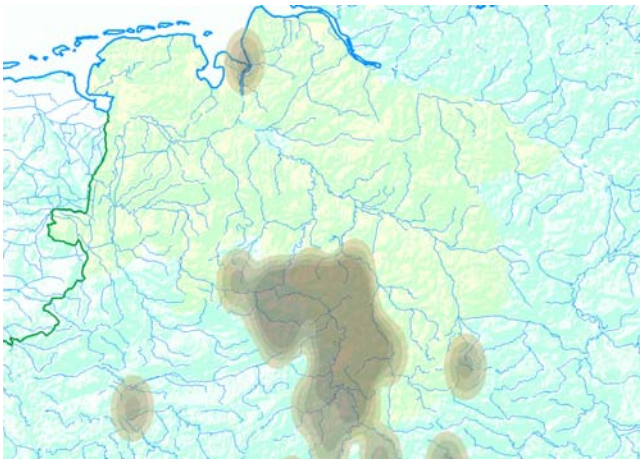

Figure 5: Localization of Weserbergland.

Figure 6: An example with few links in the article: The major regions of Europe.

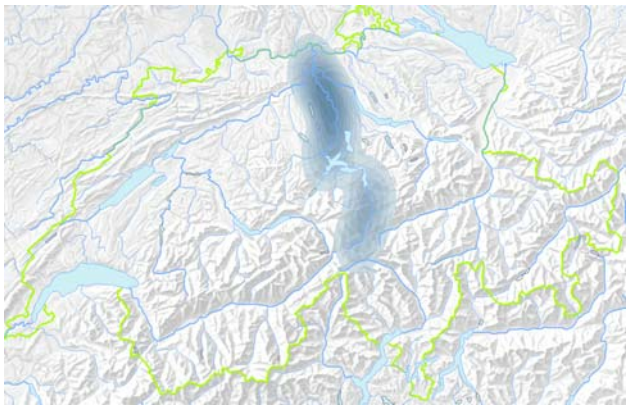

Figure 7: Importance of Reuss-valley. The KDE bandwidth is 0.25°.

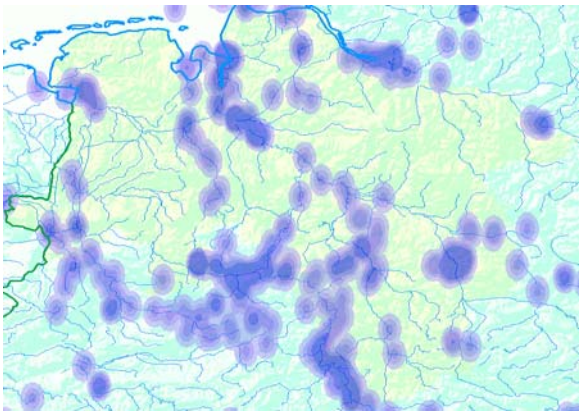

Figure 8: Waterbodies and importance of them in Lower-Saxony.

Conclusion and further work

The presented technique offers a possibility to determine well known but not well-defined areas. But the result depends on the completeness of the knowledge repository and its database with coordinates. A possible solution could also be using the links of the links and the back-links to get more coordinates. But in this case the coordinates have to be weighted according to their semantic relatedness.

As future work it could be interesting to compare several density estimation kernels. We only gave a quantitative result for one kernel by comparing the real size of Cantons with the determined. It would be useful to investigate how this value changes by changing the numbers of links or the size of the kernel. To improve the number of links for an object, it would be meaningful to follow *interwiki*-links. It would be also useful to define a measure for line structures.

In this paper we completely ignored that the coordinates in Wikipedia have a certain precision. There are also plans to link form Wikipedia to objects in *Openstreetmap* and vice versa. So it could be possible to include this information in the calculation of the probabilities. We also ignored that some articles are more related than others (semantic relatedness). Thus we should improve the method by weighting the coordinates according to these.

As a further step is could be tested, how lines can be defined and how areas could be labelled automatic.

References

Bundesamt für Kartographie und Geodäsie, Ständiger Ausschuss für Geographische Namen 2002, *Landschaften*, 1:1.000.000, Herausgegeben vom Bundesamt für Kartographie und Geodäsie, Ständiger Ausschuss für Geographische Namen, 3., überarbeitete Ausgabe, Frankfurt am Main.

Hecht, B & Raubal, M 2008, 'GeoSR: Geographically Explore Semantic Relations in World Knowledge' in L Berhard, A Friis-Christensen & H Pundt (eds.), *The European Information Society, Lecture Notes in Geoinformation and Cartography*, Spinger, Berlin, Heidelberg.

Jones, C, Purves, R, Clough, P & Joho, H 2008, Modelling Vague Places with Knowledge form the web, *International Journal of Geographical Information Science*, 1365-8816.

Liedtke, H 2002, *Namen und Abgrenzungen von Landschaften in der Bundesrepublik Deutschland*, Band 239 der Reihe "Forschungen zur deutschen Landeskunde" 3. Auflage, Deutsche Akademie für Landeskunde.

Hermann, C & Leuthold, M 2003, *Atlas der politischen Landschaften Ein weltanschauliches Porträt der Schweiz,* VdF Hochschulverlag AG an der ETH Zürich.

Jordan, P 2005, 'Großgliederung Europas nach kulturräumlichen Kriterien', *Europa Regional*, 13. Jahrgang, 2005, Heft 4, pp. 162 – 173, Leibniz-Institut für Länderkunde, Leipzig.

Meynen, E & Schmithüsen, J 1953-1962, *Handbuch der naturräumlichen Gliederung Deutschlands*, Selbstverlag der Bundesanstalt für Landeskunde, Remagen.

Piatti, B, Bär, H, Reuschel, A, Hurni, L & Cartwright, W 2008, 'Mapping Literature: Towards a Geography of Fiction', *Proceedings of the Cartography and Art - Art and Cartography Conference, Vienna*, International Cartographic Association Working group Art and Cartography, http://www.literaturatlas.eu/downloads/vienna_piatti-mapping_literature.pdf

Schröder, M 2001, *Gebiete optimal aufteilen OR-Verfahren für die Gebietsaufteilung als Anwendungsfall gleichmäßiger Baumzerlegung*, Dissertation, Universität Fridericiana zu Karlsruhe, 254 S.

Sievers, J & Schneider, T (eds.) 2001, *Second International Symposium on Geographical Names GeoNames 2000*, Mitteilungen des Bundesamtes für Kartographie und Geodäsie, Band 19, Frankfurt am Main.

de Smith, MJ, Goodchild, MF & Longley, PA 2007, *Geospatial Analysis*, Second Edition, Troubador Publishing Ltd.

Spiess, E 2002, *Atlante mondiale svizzero*. Ed. dalla Conferenza svizzera dei direttori cantonali della pubblica educazione (CDPE), Editione 2002, aggiornata e ampliata: Lehrmittelverlag des Kantons Zürich.

Tobler, W 1970, 'A computer movie simulating urban growth in the Detroit region', *Economic Geography*, 46, pp. 234-240.

Wolff, M & Asche, H 2009, 'Towards Geovisual Analysis of Crime Scenes – A 3D Crime Mapping Approach', in Sester, M, Paelke, V & Berhard L (eds.) *Advanced in GI-Science - The European Information Society*, Lecture Notes in Geoinformation and Cartography, Spinger, Berlin, Heidelberg, pp. 429-448.