Categorization of Linear Objects for Map Generalization Using Geocoded Articles of a Knowledge Repository

Tobias Dahinden and Monika Sester

Institut für Kartographie und Geoinformatik, Appelstrae 9a,
Leibniz Universität Hannover, 30167 Hannover, Germany
{tobias.dahinden | monika.sester}@ikg.uni-hannover.de

Abstract. One of the major tasks when generalizing a map is the classification of map objects according to their importance. The target of this task is to retain important objects and eliminate unimportant objects. In order to determine the importance of spatial phenomena we link the notion of importance with the fact and the frequency of their use. This approach has been proposed in vernacular geography to delineate objects with vague boundaries. To classify objects we therefore propose to use the article and link structure of a knowledge repository with geocoded articles. This is done by collecting the coordinates of the linked articles of a class of objects and drawing them in a map as points. Then we consider the importance of an object the higher the denser the points are. To illustrate this method we study an alpine river system. As knowledge repository with a link structure and geocoded articles the German Wikipedia is used.
The result of this classification depends particularly on the kernel density estimation function and its parameter, such as the search distance. Furthermore the result may vary radical if multi-linked articles are weighted more than single-linked and if we take into account that the total of geocoded articles are not distributed equally. We describe the variation of the point density when the parameters are changed according to our example. In addition the results of our calculation are also compared to generalizations derived by traditional techniques.

1 Introduction

In cartographic generalization objects have to be presented according to their relevance and importance. In topographic mapping, there are rules which indicate the (relative) importance of topographic objects: e.g., roads are more important than boundaries of vegetation; therefore, they are emphasized at the cost of other less important objects. A new research direction in mobile cartography is to personalize cartographic visualizations; this also involves that the notion of relevance has to be introduced in a flexible and personalized way. The question is, how to define relevance.

In this paper we use crowd-sourced information as an indication for relevance. The idea is that data in public repositories like Wikipedia are introduced, if

they are of certain relevance—at least for the one who writes the article. A higher number of references to the same object and a more detailed description indicate a higher relevance. In this way, the counts or the distribution of spatial locations of an object are used as indication for the relative importance. We use the Epanechnikov kernel estimation to transform the discrete location data in a continuous form.

The paper is organized as follows. After a review of related work our approach is described in detail. Then we present examples and finish with a summary and outlook on open problems on future work.

2 Related work

The automation of cartographic generalization has been researched since more than 40 years (for a comprehensive overview see [5]). Methods for automatic realizations of most of the operators have been proposed and implemented (e.g. simplification, selection, enhancement, typification, displacement). All operations, however, need prior knowledge to determine the importance of features. Some measures can directly be determined from geometric measures like minimal visible sizes or distances, or pre-given rankings of features. However in the case of selection or typification, there are similar objects which have to be reduced in number, e.g. a set of buildings, roads or rivers of similar category. For solving this problem, density based approaches have been proposed, e.g. by [7] or [9].

Collaborative data and knowledge acquisition is becoming more and more frequent. Knowledge repositories like the German Wikipedia contain 950,000 articles, the English 3,000,000 articles with totally 650,000 coordinates associated (August 2009). Also, people are collaboratively editing spatial objects in project like Open Street Map. To exploit this crowd-sourced information has been proposed in several approaches. Jones et al. [4] have proposed to use the footprints of Websites in order to delineate the boundary of regions with uncertain or vague boundaries like the "Black Forest" or the "British Midlands". They used websites which were scanned for geographical names, which in turn were georeferenced using gazetteers. The underlying idea is that the boundary can be found by inspecting how the location names are used by the people.

Dahinden [1] used a similar approach to delineate areas. He, however, used a repository which already includes spatial references, namely Wikipedia as a basis; in this way he was able to determine the outline of Swiss cantons or the location of linear objects like a motorway. Hecht and Raubal [3] locate non-geographic expressions. They also 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.

Piatti et al. [6] locate activity zones 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.

The exploitation of crowd-sourced data for digitization of spatial objects has been investigated by Sayda [8]: from a set of uploaded GPS-tracks of hikers, he determined the most probably and at the same time reliable track.

TomTom and Vodaphone use the temporal distribution of mobile phones on highways for the prediction of traffic jams [2].

3 Approach

3.1 Knowledge repository and gazetteer used

In principle our approach works with any knowledge repository with an associated gazetteer. For our research we used the German Wikipedia as knowledge repository and the collection of coordinates of Wikipedia-World [11] as gazetteer.

The German Wikipedia contains more than 900,000, the English more than 3,000,000 articles. They contain texts about geographical features from all over the world. Thus it should be useful for all kind of maps. Yet some places are missing. For example the place Negenborn exists three times in Germany, but there are only two articles in Wikipedia by now (August 2009).

Analyses of Hecht and Raubal [3, p. 102] show a relation between the topics of the articles and the language, i.e., a domination of German topics in the German Wikipedia.

The names of the places with its coordinates 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 in a 1:1 relation. For this reason we do not have to use named entity recognition to match ambiguousness.

Yet the coordinates of several language versions may differ. For example the Turkish town 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. There are also articles with missing coordinates (e.g., Schloss_Neuenhinzenhausen) and the gazetteer could be out of date.

In the gazetteer about 72,000 entries have information about the dimension of the objects, where approx. 17,000 correspond to 2,500 m in diameter or smaller, 10,000 to 5,000 m, 27,000 to 10,000 m, 6,000 to 25,000 m and 1,000 to 50,000 m or larger. The mean value of this granularity of the object is approx. 13,000 m.

In addition there is also some information about the type of the object (e.g. city, monument, river) and the ISO-3316 Country Code of the area the coordinate belongs to.

3.2 Processing of the data

The investigation of an object has to be based on one article or on a list of certain articles that describe the object. Both article and list can be found in Wikipedia, e.g. for the river system Reuss you may use its category [12].

The link-list of a certain article can be requested through the Mediawiki API. This list has to be compared with the entries in the gazetteer. As a result a list with coordinates associated to the object is derived.

The list with coordinates can be seen as random variables of an unknown probability distribution describing the relevance of the object. To estimate the probability distribution we use the Epanechnikov kernel density estimation [10]. Unfortunately the parameters of the density estimation have to be selected according to the distribution of the coordinates. As a hint we may use the dimension of the objects.

The relevance of a linear system is determined by integrating the density along a line segment. This leads to a value that depends on the density and the length of the line segment. As a consequence the result is different if a long line is divided in segments. To avoid the influence of the length of the line segments, it is possible to divide the value by the length of the line segment.

4 Examples and results

The approach is tested with the linklist "Kategorie: Flusssystem Reuss" [12] of German Wikipedia. A river system (Flusssystem) is a collection of rivers that constitute a major river. For its cartographic representation, the different river sections have to be evaluated with respect to their importance. A classical approach is to calculate the so-called Horton order [5] to determine the relative importance. Here, we extract the importance value from the analysis of Wikipedia links.

A major problem is the occurrence of the same link in several articles. When using a single article as origin each link is usually unique. But when using a list, this is certainly not the case. In each article there is usually an entry about the country the object belongs to. Thus you get the centroid of the country from each article of the list. The same problem arises with objects that are superior such as the main river.

There are three possibilities to solve this problem. Either all links are used, or the links are weighted according to their dimension, or each link is used only once. If all links are used, it is assumed that superior objects are more important than inferior. But the superior object can be of a more abstract type than the investigated, e.g., in our example the centroid of the Switzerland lies in the investigated area. If the border of Switzerland would be changed also the

centroid would and thus the result of our calculation. Yet the border of a county seems irrelevant to the categorization of a river system.

Unfortunately the dimension of river objects is often missing in the gazetteer. So most of the weight can only be guessed. This method may be investigated in the future.

The third possibility is to use each coordinate only once. Figure 1 shows the distribution of the footprints of the articles.

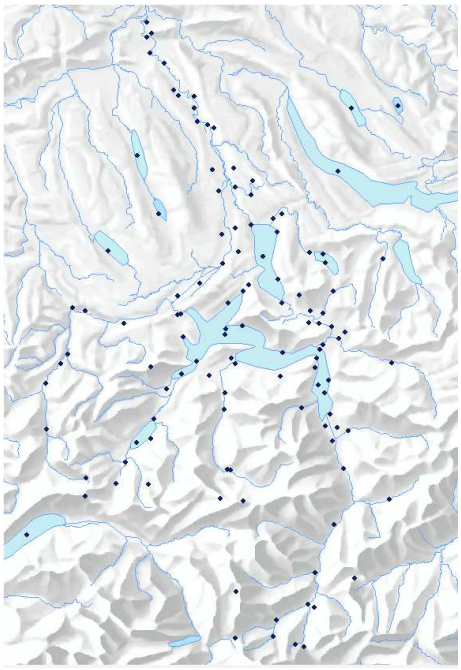

Fig. 1. Distribution of the footprints of the articles about the river system Reuss in German Wikipedia. Geometry of waterbodies: geodata @ swisstopo.

We estimated the kernel density of these points with a search distance of 6 km. In Figure 2 the kernel density estimation for the river system is depicted. The result seems to reflect the river system quite nicely.

On the basis of a vector dataset with river for each part of the river system Reuss the integral over the estimated density was calculated. This leads to a product of the relevance and the length. The relevance value can be calculated by dividing value of the integral with the length of the waterbodies. In Table 1 the ten waterbodies with the largest integral are named. A map with nine of these ten waterbodies is shown in Figure 3.

With this method it is possible to compare different kinds of categorization of vector data. For the river system under investigation we have the possibility

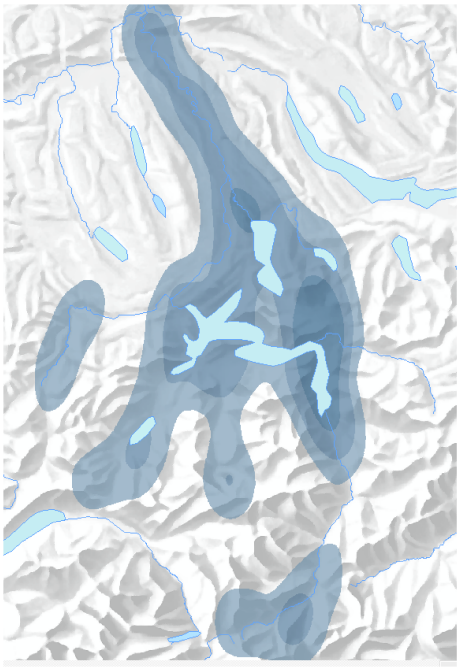

Fig. 2. Kernel density estimation of river system Reuss. In the dark blue area the density is high, in the light blue low. The blue lines correspond to all rivers shown in the national map of Switzerland 1:25.000. Geometry of waterbodies: geodata @ swisstopo.

Table 1. The 10 most relevant waterbodies according to the integral of the kernel density estimation along the waterlines.

Name	*Integral*	*Length*	*Relevance*
Vierwaldstaettersee	6972047	130061	53.6
Reuss	6359911	152205	41.8
Zuger See	1944531	38226	50.9
Lorze	1534526	30633	50.1
Sarner Aa	1200075	30450	39.4
Rigiaa	1110433	22338	49.7
Lauerzer See	1076590	12190	88.3
Engelberger Aa	1045192	39526	26.4
Muota	951644	30107	31.6
Kleine Emme	846280	36100	23.4

to compare the categories of the river system Reuss of swisstopo's Vector25 data for watercourses with the Atlas of Switzerland data and our representation. The topographic data set for rivers of swisstopo has 7 categories, yet there is only one category for lakes (Figure 4). Atlas of Switzerland has 3 categories for describing the rivers and lakes (Figure 5). In both datasets the first category contains the most important objects, the second some less important and so on.

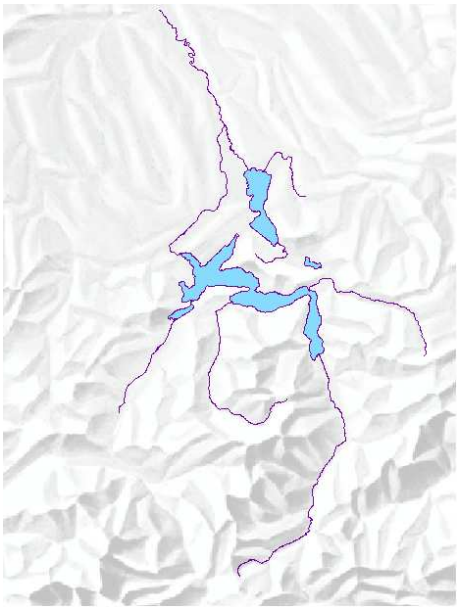

Fig. 3. The 9 most important waterbodies of the river system Reuss according to the density of articles in German Wikipedia. Geometry of waterbodies: geodata @ swisstopo.

By comparing Figure 3, 4 and 5 it becomes obvious that there are some differences in the categorization of the waterbodies. Concerning rivers the categories of swisstopo and Atlas of Switzerland tend to be distributed uniformly. In our method the rivers in the center of the river system tend to be accentuated. Concerning lakes the swisstopo and the Atlas of Switzerland data categorize the objects by its size. In contrast in our method some small objects like lake Lauerz (Lauerzersee) are categorized as very relevant.

We can make a quantitative comparison of the datasets. This is done by adding up the integral value of each element that is shown in the generalized data set and divide it by the length of all elements. For Vector25 we use Category 1 and 2, for Atlas of Switzerland Category 1-3. Table 2 shows the name of the datasets, the number of elements selected and the relevance value according to our method.

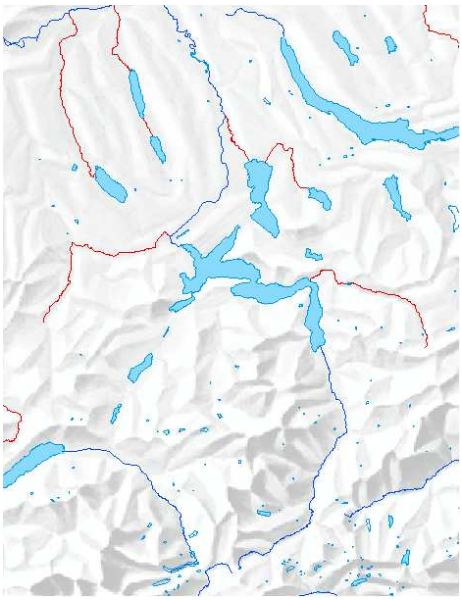

Fig. 4. Lakes and category 1 (blue) and 2 (red) of rivers in the dataset Vector25 of swisstopo (geodata @ swisstopo).

Table 2. Comparing the categorization of several datasets. The relevance is calculated for the "number of elements" most important object.

Dataset	*Number of elements*	*Relevance value*
Swisstopo Vector25 (only rivers)	4 elements	38.9
Our approach only rivers	4 elements	43.3
Atlas of Switzerland only rivers	5 elements	38.0
Our approach only rivers	5 elements	40.8
Atlas of Switzerland with rivers and lakes	9 elements	43.2
Our approach with rivers and lakes	9 elements	45.6

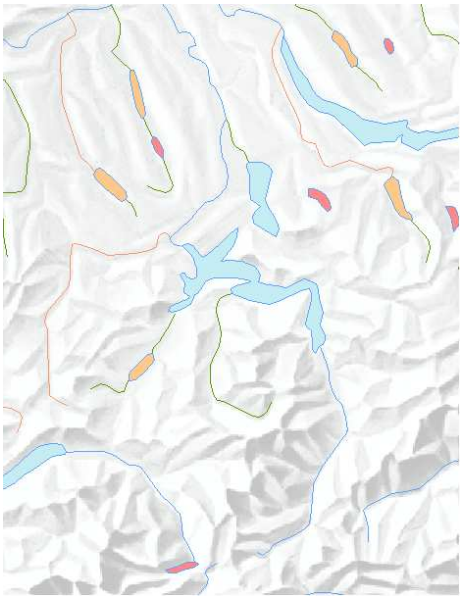

Fig. 5. Lakes and rivers according to the Atlas of Switzerland. Category 1: blue, category 2: orange, category 3: red.

In Table 2 the difference of the representation of the rivers are quantified. Of course, the objects selected according to our method have the higher relevance value than the other datasets under investigation. That's obvious because we selected the elements in a manner to maximize this value. Anyhow it is not possible to conclude, that these datasets are categorized inappropriate.

The more elements are selected according to our method the more the relevance value should decrease. Looking at Table 2 this seems not to hold. But actually the relevance value for four and for five elements is calculated for the river system only. Per contra the relevance value for nine elements is calculated for rivers and lakes.

5 Summary and outlook on future work

The paper presented an approach to determine the importance of geographic features in order to use it for cartographic visualization and generalization. The underlying idea is the fact that geographic features that are mentioned in public knowledge repositories give an indication to their importance; the relative frequency determines the relative importance. In this way it was possible to identify different grades of importance which can be used for generalization and visualization.

Issues to consider are the following:

- The frequency measures are not necessarily objective, as they will be higher in more populated areas; there might even be no web-articles for some spatial feature, although it is of importance in its local environment. In this way, the proposed measure reflects the usage of these features in the public awareness.
- As described above, there are problems with the completeness of the knowledge repository. Non-existing links may indicate non-relevance. Here general quality measures should be derived that give hints to the completeness of a dataset. E.g., there are measures in OpenStreetMap that calculate an expected road density depending on the number of inhabitants in a city. If the actual number of roads is below this average value, it is an indication that some information is missing.
- There is a dependency on the interpolation scheme, especially on the parameters of the density kernel. This leads to the problem that the combination of objects to object classes in combination with the density estimation is not necessarily distributive.
- It may be difficult to compare the relative measures for different kinds of topographic features. E.g., a city area will probably be mentioned more often in web-repositories than a river system. So relative weights between different feature classes have to be investigated.

These issues will be addressed in future work.

References

1. Dahinden, T.: Localization of uncertain and fuzzy-bordered areas by geocoded articles of a knowledge repository. To appear ICC, Santiago de Chile (2009)
2. Friedrich, M., Jehlicka, P., Otterstätter, T., Schlaich, J.: Mobile Phone Data for Telematic Applications. Proceedings of International Multi-Conference on Engineering and Technological Innovation: IMETI 2008. International Institute of Informatics and Systemics (IIIS), Orlando, Florida, USA (2008)
3. Hecht, B., Raubal, M.: GeoSR: Geographically Explore Semantic Relations in World Knowledge. In L. Bernard, A. Friis-Christensen and H. Pundt (eds.): The European Information Society. Lecture Notes in Geoinformation and Cartography, Springer, Berlin (2008)
4. Jones, C., Purves, R., Clough, P., Joho, H.: Modelling Vague Places with Knowledge from the Web. International Journal of Geographical Information Science, **22**(10) (2008) 1045-1065
5. Mackaness, W., Ruas, A., Sarjakoski, L. (Eds.): Generalisation of Geographic Information: Cartographic Modelling and Applications. Elsevier, published on behalf of the International Cartographic Association (2007)
6. Piatti, B., Bär, H., Reuschel, A., Hurni, L., Cartwright, W.: Mapping Literature: Towards a Geography of Fiction. Proceedings of the Cartography and Art—Art and Cartography Conference. International Cartographic Association, Vienna, Austria, http://www.literaturatlas.eu/downloads/vienna_piatti-mapping_literature.pdf (2008)
7. Regnauld, N.: Recognition of Building Clusters for Generalization. In Kraak, M., Molenaar, M. (Eds.): Advances in GIS Research. Vol. 1, Faculty of Geodetic Engineering, Delft, The Netherlands, (1996) 4B.1-4B.14

8. Sayda, F.: Zur Aktualisierung von Geodaten eines LBS unter Einbeziehung der Nutzer. Dissertation, Universität der Bundeswehr München (2006)
9. Sester, M.: Self-Organizing Maps for Density-Preserving Reduction of Objects in Cartographic Generalization. In Agarwal, P., Skupin, A. (Eds.): Self-Organising Maps. Wiley (2008)
10. de Smith, M.J., Goodchild, M.F., Longley, P.A.: Geospatial Analysis. Second Edition, Troubador Publishing Ltd. (2007)
11. Wikipedia: WikiProjekt Georeferenzierung / Wikipedia-World. `http://de.wikipedia.org/wiki/Wikipedia:WikiProjekt_Georeferenzierung/Wikipedia-World`. Visited: 31.8.2009.
12. Wikipedia: Kategorie: Flusssystem Reuss. `http://de.wikipedia.org/wiki/Kategorie:Flusssystem_Reuss`. Visited: 31.8.2009.