Measuring Similarity of Geographic Regions for Geographic Information Retrieval

Andreas Henrich and Volker Lüdecke

University of Bamberg, D-96045 Bamberg, Germany
{andreas.henrich,volker.luedecke}@uni-bamberg.de

Abstract. Representations of geographic regions play a decisive role in geographic information retrieval, where the query is specified by a conceptual part and a geographic part. One aspect is to use them as query footprint which is then applied for the geographic ranking of documents. Users often specify textual descriptions of geographic regions that are not contained in the underlying gazetteer or geographic database. Approaches that automatically determine a geographic footprint for those locations have a strong need for measuring the quality of this footprint, for evaluation as well as for automatical parameter learning. This quality is determined by the 'similarity' between the footprint and a correct representation of that region.

In this paper we introduce three domain-specific points of view for measuring the similarity between representations of geographic regions for geographic information retrieval. For each point of view (strict similarity, visual similarity and similarity in ranking) we introduce a dedicated measure, two of which are novel measures that we propose in this paper.

1 Measuring Similarity between Region Representations

1.1 Points of View

There are simple measures for measuring the quality of region approximations or more generally the similarity between two polygons. Overlap-percent of the areas or percent-inside versus percent-outside are typically used for that purpose. A comparison of several of those measures can be found in [3]. These are binary measures that only consider a point to be part of a polygon or not, and they do not take into account the degree of variation. A small part just a bit outside the correct region is just as wrong as the same small part being hundreds of miles away, which does not seem appropriate for our purposes. The similarity of the shapes itself (trying to match them by rotating or scaling, like in [4]) is obviously also not very useful in our scenario.

The above considerations show that the intention behind the computation of approximated regions has to be considered carefully when talking about the quality of an approximation. In this paper we distinguish three points of view:

– **Strict similarity.** In some scenarios it is very important that an approximation of a geographic region is strictly within the borders of the original

region. For example, a query might refer to a law that is only valid inside a region delimited by a border (like a US state). Even a small crossing of that border makes the geographic position worthless. This is also a common baseline assumption for measures that consider regions of overlap between approximated and correct areas and regions of non-overlap between them.

- **Visual similarity.** You might want to create geographic footprints for visualizing the geographic correlation of certain concepts or to answer where-is-like queries like *where were the Olympic games 1972*. We can assume soft borders here: a part of the query region R_q only slightly outside the correct region counts as "quite good", while it loses value the further it is away from the actual correct region.
- **Ranking similarity.** This is the perspective of a search engine. Since the geographic footprints can be used for creating a geographic relevance ranking in order to fulfil the geographic constraint of a query, the footprint itself does not matter much. Instead, it is more important how two different footprints influence the ranking decision of the search engine. Two geographic footprints for a query region are equally well suited if they lead to the same geographic ranking. A footprint is considered better than another one if the resulting geographic ranking leads to better results.

Each of those points of view has different implications for a suitable similarity measure. Therefore, we propose to use a different measure for each semantics.

1.2 Strict Similarity

Measures that consider overlap of regions typically follow this semantics, like the one mentioned in [2]. $A(R)$ is the surface area of a region R, R_q is the approximation of the query region, and R_c is the correct region:

$$sim_{strict} = \frac{2 * A\left(R_q \cap R_c\right)}{A\left(R_q\right) + A\left(R_c\right)}$$

As we pointed out earlier, the semantics behind sim_{strict} is a very strict, though, and one that implies some limitations, as we will show in section 2.

1.3 Visual Similarity

Figure 1 shows several examples of approximations R_i for a correct region R_c. Looking at R_1 to R_3, it is obvious that R_1 is a better approximation than R_2,

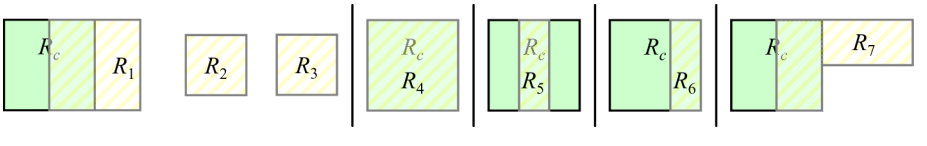

Fig. 1. Approximated regions R_i relative to correct region R_c

which is not good overall, but better than R_3. R_4 is identical to R_c and should therefore be the best approximation possible. While R_5 and R_6 both lie within the bounds of R_c and are of equal size, R_5 is a better "visual" approximation of R_c. Comparing R_1 and R_7, which both have 50% of their surface overlapping R_c and 50% outside R_c, however R_7 is "visually" worse than R_1.

The measure for "visual similarity" should take that into account. The basic idea of the measure we propose is to take into account the absolute error of each point P_q of the approximation R_q not within the bounds of the correct region R_c ($P_q \in R_q \setminus R_c$), while also considering each point P_c of R_c that is not covered by R_q ($P_c \in R_c \setminus R_q$). Since this is a domain specific measure, we can specify a "maximum error". In practice this means to define a distance M where the value of the result becomes zero. On a global scale in an extreme situation this "maximum error" might be about 20000 kilometers meaning that the considered point is lying on the opposite side of the globe. On the other hand, we could set M to 50 kilometers meaning that every result with a distance to R_c of more than 50 kilometers is worthless.

Let the distance measure $dist_{\min}(P, R)$ calculate the minimum distance of point P to region R. For an easier computation, we discretise both region R_q and region R_c to n discrete and evenly-distributed points within their boundaries. The visual similarity can then be calculated by:

$$sim_{visu} = 1 - \frac{\sum_{i=1}^{n} \min(M, dist_{\min}(P_{q,i}, R_c)) + \sum_{j=1}^{n} \min(M, dist_{\min}(P_{c,j}, R_q))}{2 * n * M}$$

For simplicity we use the liner distance (as the crow flies) to calculate $dist_{\min}(P, R)$ in the following, even though we are aware that there are other possible measures. Furthermore, we use $M = 20000$ kilometers making this a global measure that is able to create a ranking between any two geographic regions in the world. As mentioned above, other settings for M are also possible: Whenever we assume that regions R_q at a certain distance to R_c are equally worthless and are to be ranked with a score of 0, we can set M to that distance.

1.4 Ranking Similarity

Since the original idea of the approaches presented in this paper is to provide geographic footprints as approximations for locations or regions a user specifies in a query, and these footprints are to be used by a geographic search engine to provide a geographic ranking of documents, it is sensible to consider the perspective of the search engine when evaluating the quality of the approximations. If a footprint R_q has a completely different shape than the correct region R_c, but the ranking done by the search engine is the same both for R_q and R_c, R_q is in that sense equal to R_c.

We try to measure the effect an approximation R_q has on a potential ranking. Therefore, we assume that geographic footprints of documents to be ranked are evenly distributed in the data space and can be represented as points P. We then iterate over all points P and calculate the difference in distance from P to R_q, or R_c respectively. To make it an absolute measure, we normalize it to the interval

[0;1] by assuming a theoretical maximum error of M, as described for the visual similarity measure. n is the number of points P (representing documents in the data space).

$$sim_{rank} = 1 - \frac{\sum_{i=1}^{n} \min\left(M, |dist_{\min}(P_i, R_q) - dist_{\min}(P_i, R_c)|\right)}{n * M}$$

For our experiments, where geographic regions are restricted to Germany, we found that the effect n had on the results was not that big, as long as n was reasonably high. We used $n = 10000$ for our experiments, but $n = 1000$ works just as fine.

2 Evaluation

We earlier introduced several exemplary regions R_i sketched in figure 1. For comparing the three measures, we calculated the similarity scores for R_i with each measure (see table 1). For this experiment we have positioned the exemplary regions in the south-western part of Germany. The strict measure sim_{strict}, which is based solely on overlapping and non-overlapping regions, cannot differentiate between areas R_2 and R_3, R_1 and R_7 or R_5 and R_6. The visual measure sim_{visu} ranks those regions according to the postulations from section 1.3, while the ranking measure sim_{rank} leads to similar but slightly different results. As an example for the differences, we can consider R_5 and R_6. These two approximations are equally good with respect to sim_{strict}. With respect to sim_{visu} R_5 is a better approximation than R_6 in accordance with our impression. With respect to sim_{rank} we have to consider that the regions were positioned in the south-western part of Germany. Therefore, for most points P_i within Germany, the distances to R_6 and R_c are the same because they have the same eastern border. A ranking with respect to R_6 therefore yields a better approximation for a ranking with respect to R_c than a ranking with respect to R_5 does.

The absolute numbers resulting from the visual similarity measure are of course very close to 1, the optimal value. This is because we used relatively small regions in a global scope ($M = 20000$) here for comparison, which leads to the following consideration: if you were looking for an area like Northern Ireland,

Table 1. Ranking of regions R_1–R_7 from figure 1

	sim_{strict}		$sim_{visu,global}$		$sim_{rank,global}$		$sim_{visu,local}$	
Region	value	rank	value	rank	value	rank	value	rank
R_1	0.50000	4	0.99861	4	0.99585	4	0.950	4
R_2	0.00000	6	0.98286	6	0.97398	6	0.383	6
R_3	0.00000	6	0.97255	7	0.95872	7	0.012	7
R_4	1.00000	1	1.00000	1	1.00000	1	1.000	1
R_5	0.66667	2	0.99961	2	0.99718	5	0.986	2
R_6	0.66667	2	0.99929	3	0.99878	3	0.974	3
R_7	0.50000	4	0.99792	5	0.99879	2	0.925	5

but find the Republic of Ireland instead, this is not bad compared to areas in Africa or Australia! For ranking purposes, the absolute numbers of the results do not matter, since they can easily be ranked. For other scenarios, you could lower the parameter M, as described above. We did the same calculation with $M = 1100$ reflecting a more local scenario for sim_{visu} (last column of table 1).

We also tried to get a first impression of the impact of using our measures in geographic ranking. For this purpose we calculated rankings for approximations of 115 geographic regions for each measure. We used our system presented in [1] for that. Assuming that the strict similarity sim_{strict} can be seen as baseline measure and the resulting ranking $ranking_{strict}$ as baseline ranking, we compared the other two rankings $ranking_{visu}$ and $ranking_{rank}$ to $ranking_{strict}$ using Spearman's rank correlation coefficient ρ. This resulted in $\rho_{strict,visu} = 0.97$ and $\rho_{strict,rank} = 0.94$, which is at least an indication that the measures indeed lead to different rankings. The differences are relatively small, because the approximations itself are quite accurate and have a high degree of overlap with the correct region. As an example, the approximation for the region *Chiemsee* is the best according to sim_{visu} and sim_{rank}, whereas it is only on rank 8 by sim_{strict}.

3 Conclusion

In this paper we proposed different points of view for measuring the similarity between geographic regions. For the points of view "visual similarity" and "ranking similarity" we introduced novel similarity measures, while we chose a common measure for "strict similarity". The results show that the measures work as intended and reflect the perceived quality of approximated regions in the given scenarios.

References

1. Henrich, A., Lüdecke, V.: Determining geographic representations for arbitrary concepts at query time. In: LOCWEB 2008: Proc. of the First Intl. Workshop on Location and the Web, pp. 17–24. ACM, New York (2008)
2. Hill, L.L.: Access to Geographic Concepts in Online Bibliographic Files: Effectiveness of Current Practices and the Potential of a Graphic Interface. Ph.D thesis, University of Pittsburgh (1990)
3. Larson, R.R., Frontiera, P.: Spatial ranking methods for geographic information retrieval (gir) in digital libraries. In: Heery, R., Lyon, L. (eds.) ECDL 2004. LNCS, vol. 3232, pp. 45–57. Springer, Heidelberg (2004)
4. Veltkamp, R.C.: Shape matching: Similarity measures and algorithms. In: SMI 2001: Proceedings of the International Conference on Shape Modeling & Applications, Washington, DC, USA, p. 188. IEEE Computer Society, Los Alamitos (2001)