Categorical Prominence and the Characteristic Description of Regions

Martin Tomko[1] and Ross S. Purves[1]

Department of Geography, University of Zurich - Irchel, Winterthurerstrasse 190,
CH-8057 Zurich, Switzerland
{martin.tomko, ross.purves}@geo.uzh.ch

Abstract. The annotation of georeferenced information objects is related to the annotation of the content of the containing spatial regions. Not all spatial features contained in the regions, however, present characteristic attributes of the region described. In this paper, we present a method designed to select prominent spatial features in a region in order to improve the annotation of the region. The method is demonstrated on an artificial dataset and preliminary results show that the method results in a reduction of the number of terms typically describing a region which is statistically significant.

1 Introduction

A key challenge for the Semantic Web is the identification of not only terms to describe information objects (such as images, audio or video recordings), but also the attachment of semantics to those terms which are relevant to a particular context of use [1]. From a geospatial perspective, a key task for the Semantic Web is therefore to describe the geographic context of such information objects. Much of the information found on the Web relates in some way to specific geographic locations [2, 3].

Recently, increased attention has been paid to the automated description of information objects which are georeferenced, and in particular to the production of so-called *tags* which characterise such information objects. One area where these techniques show particular promise is in the description of georeferenced images. Given a geographic footprint related to an information object, it is possible to query spatial data to identify, for example, toponyms related to that location and thus likely to prove useful in describing an information object related to the location. A number of researchers have proposed a variety of techniques to automatically produce image annotations. The NameSet application [4] does just this to identify toponyms describing sets of images. However, simply selecting the toponym which lies nearest to the region captured in pictures will be unlikely to provide relevant descriptions for many users, and NameSet thus attempts to identify the most appropriate and useful toponym for a given set of locations based on the assessment of the importance of the possible toponyms.

The European project TRIPOD (`http://www.projecttripod.org/`) focuses on building multi-faceted text descriptions of the landscape and permanent man-made features pictured in georeferenced photographs by exploiting spatial data related to the photographs' locations. One important task is determining which spatial data provide the most concise and useful description of an image location.

We have argued in previous work [5] that in producing image annotations we should consider the *specific of*, the *generic of* and the *about* as defined in the Panofsky-Shatford matrix [6]. The *specific of*, in a geographic context, generally refers to the use of individually named geographic locations to describe an information object, whilst the *generic of* is concerned with kinds of location. Finally, the *about* refers to more abstract notions conveyed by a location.

We hypothesise that descriptions of images can be derived by querying spatial data within some footprint related to the image, for example the footprint of the region containing the image, and that by applying the ideas proposed by Shatford [6] we can annotate different aspects of the image. In this paper, we focus on the characterisation of regions in terms of the *generic of*. People typically characterise geographic regions generically by keywords relating to prominent *kinds* of objects found within a region (e.g. *castles* and *vineyards* in the Loire Valley). While references to numerous spatial objects are possible, prominent objects succinctly characterise the region in question.

The research presented therefore addresses the hypothesis that the descriptors relevant for the characterisation of a geographic region can be identified as a function of their relative distribution in space. We present a method to identify references for characteristic geographic descriptions, based on the assessment of the distribution of spatial objects in a hierarchical partitioning of space, where the descriptive relevance of a category is assessed with respect to its frequency. The method is demonstrated in an introductory study based on an artificially generated dataset and the results are discussed with regard to future application to real-world datasets.

2 Characterising Regions

Typical descriptions of named regions appear to focus on naming specific objects (Grossmnster), examples of prominent kinds of objects (churches, banks), and emotions or feelings invoked by a region (rich, tidy) [7]. In this work, we define a region as a bounded geographic footprint, though we are aware that many regions have no agreed boundaries [8].

The prominence of a spatial object or a group of spatial objects is a function of the contrast between the properties of those spatial objects with the properties of the remaining spatial objects in the region. For example, visual salience, assessed based on the properties of façades, can be used to infer prominence and has been used to retrieve landmarks for use in wayfinding [9]. In the method proposed we explore the identification of prominent categories of objects for inclusion in region descriptions. In this respect, the method presented belongs to generalization selection operators executed on the model schema [10].

A full description of a region would contain references to all contained spatial objects. References to such spatial objects could then be extracted and used as keywords or a vector of annotation terms suitable for use in text-based information retrieval tasks [11]. However, it is clear that categories of spatial objects may not be equally prominent within a region or relevant for the description of the region with respect to the purpose of description.

We can then define a *characteristic description* of a region as being one which only contains references to the most prominent objects contained within a region. Such a description need not unambiguously identify a region, but should provide the best, and thus most relevant, characterisation on the basis of the available data.

3 Deriving Prominence of Points of Interest

In this paper we propose and demonstrate a method to describe regions, and thus information objects related to such regions, based on the notion of occurrence of spatial objects, specifically point-like objects. While many spatial datasets are available with which we may attempt to characterise these regions, we demonstrate the method using simulated Points of Interest (POI) data. We discuss the reason for not using currently available real-world datasets in the Conclusions.

POI data are typically divided into a range of categories and sub-categories. We assume here that the category itself has no importance in defining relevance– we are only interested, in this first stage, in the concept of contrast from background occurrence. Thus, prominent categories of POIs, which may be relevant for the description of a region, are those who's instances occur in the given region relatively more frequently than the instances of the majority of other represented categories. As descriptions of regions rarely refer to the lack of a certain quality, categories who's instances appear in the region less frequently than in the whole area of interest are considered of low descriptive relevance.

The prominence of a category of objects c is derived from the variation of occurrence in the region described and its neighbourhood e.g., containing (higher order) regions. The categories can then be classified as *relatively over* or *under-represented* in a given region. The difference in occurrence grants contrast.

For example, one could describe dominant properties of a Saharan desert landscape through references to the types of objects most frequently found, such as sand dunes. A specific region of the Saharan landscape may, however, contain an oasis with a spring, a rare feature in a Saharan natural environment. In the context of the specific region, the oasis is unique. In the region of the oasis, the spring is overrepresented, compared to the occurrence of springs in the whole of Sahara.

3.1 Occurrence Cases

We define relative over-representation and under-representation as follows:

Relative over-representation of a category: a category c is relatively over-represented in a region A^x if and only if the frequency of occurrence of objects of category c in the region A^x is higher than the frequency of occurrence of an object of category c in a containing region A^{x-1}.

Relative under-representation of a category: a category c is relatively under-represented in region A^x if and only if the frequency of occurrence of an object of category c in the region A^x is lower than the frequency of occurrence of an object of the category c in a containing region A^{x-1}.

Let A^1, A^2, A^3 be hierarchically nested geographical regions, where A^1 represents the whole area of interest and A^3 is the smallest region fully containing the footprint captured in the image. The regions at the level A^2 should represent an intermediate level between A^1 and A^3. We assess the descriptive relevance of an object to the characterisation of a region by comparing the frequency of occurrence of a category in these hierarchically nested regions, where over or under-representation of a category at each granularity can be typified into eight cases (Table 1). Over or under-representation at the global level is determined by comparison with the occurrence of other categories of object.

Table 1. Occurrence cases. + stands for over-representation, − for under-representation.

Region	A^3		A^2		A^1	
	(small)		*(medium)*		*(large)*	
Case	+	-	+	-	+	-
1	X		X		X	
2		X		X		X
3		X	X		X	
4	X			X		X
5	X			X	X	
6		X	X			X
7		X		X	X	
8	X		X			X

The eight cases presented can be verbally characterized. For example, Case 4 can be described in the following manner:

Case 4: The category c is a globally underrepresented category, abundant in the limited space defefined by region A^3. A reference to category c should therefore be included in the characteristic description of A^3.

Example: fruit trees (category c) in a remote oasis (region A^3) in the middle of a large desert (region A^2).

Frequently occurring object categories are found for Cases 1, 3, 5 and 7, where Cases 1 and 5 are relevant for the descriptions of regions at the level A^1 and relate to objects which typify a region. Unique or rare objects belong to the object categories that fall in the long tail of the distribution, classified under Cases 2, 4, 6 and 8. Here, we assume that Case 4 has the highest descriptive power, since it occurs in small regions as rare category.

3.2 Example of Characteristic Region Description

Based on the above analysis, we can construct reduced term vectors containing only references to object categories that are considered relevant. Consider the example of the region A^3, containing spatial objects of categories $c1$, $c5$ and $c7$. The original term vector for content-based characterization of region A^3 would thus be: $v(A^3) = \{c1, c5, c7\}$. Casing results in the following: $cases(A^3) = \{Case1, Case6, Case8\}$. Case 6 has a low descriptive relevance since it is both rare in region A^3 and globally (region A^1), but common in region A^2 - objects of category $c5$ do not characterise region A^3. Objects $c1$ and $c7$ are however, in the case of $c1$ typical, and in the case of $c7$ rare globally, but typify regions A^2 and A^3. Thus, the reduced vector $v(A^3) = c1, c7$ could be used to describe region A^3.

4 Initial Method Testing and Results

4.1 Experimental Dataset

As the definition of over-representation or under-representation relates to the relative frequency of occurrence of a category within a given region, the datasets used must satisfy several conditions. The comparison of the relative occurence frequencies of objects belonging to a given category asumes that the distribution of the spatial objects in the dataset is determined by the spatial variation of the phenomenon, and not distorted as a consequence of, for example, cartographic generalization. Many currently available POI datasets are generated from cartographic data and hence do not meet this requirement. However, as the volume of volunteered geographic information increases [12] we suppose that these biases will decrease and datasets meeting our assumptions will become commonly available. In this paper, the method is demonstrated on an artificially generated dataset not influenced by generalization or other biases.

Altogether, 22677 points were generated and classified into 11 categories. The frequency of occurrences of the individual categories follows a long-tailed distribution, as shown in Figure 1. To simulate spatial correlation of occurrence of objects within categories, objects from each category were generated around seed points following a two-dimensional normal distribution.

The descriptive characterisation of regions was tested on a set of 150 randomly located A^2 regions (10 by 15 units), located in a rectangular region A^1 (100 by 150 units). Each region A^2 fully contained ten randomly located smaller regions A^3 (2 by 3 units). The content of the regions A^3 was then characterised based on the contained points.

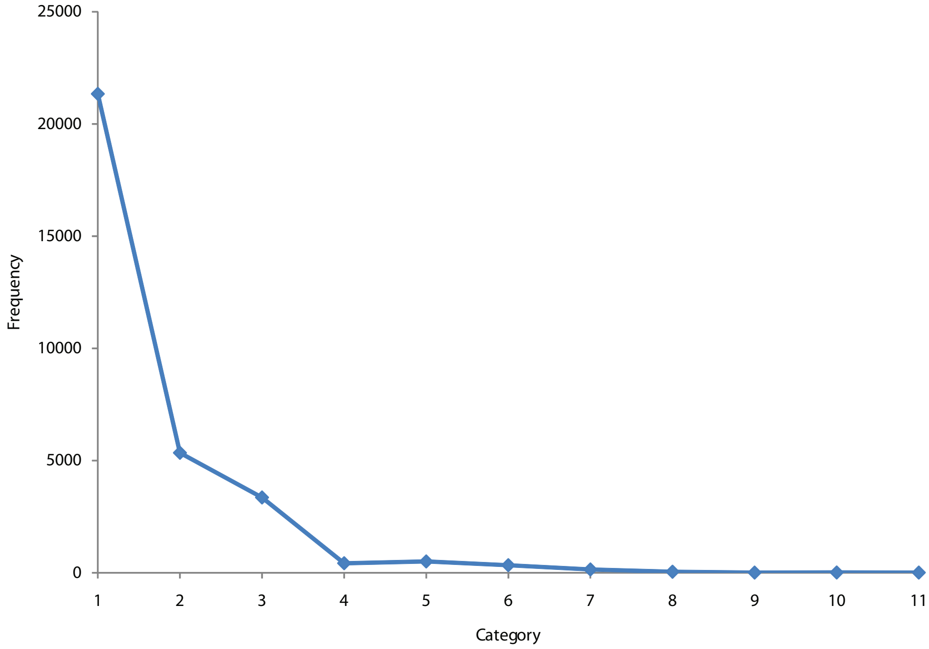

Fig. 1. Frequency of categories

4.2 Results

The 1500 regions A^3 contained a total of 2463 points, with a mean of 1.64 (s=0.973) points per region. The classification of the object categories into cases is shown in Table 2. The rare categories $c11$, $c13$, $c14$, $c15$ did not occur in any of the A^3 regions.

Since we suppose that only cases where an object category is prominent in region A^3 have descriptive relevance—that is to say Cases 1, 4, 5 and 8, where Case 1 (typical of the region at all scales) and Case 4 (typical only in the region A^3) are the most relevant for characteristic descriptions of regions. Out of the 2463 points, 67.68% fall into one of the descriptively relevant categories, with a mean of 1.11 (s=0.649) descriptively relevant terms per region. Such a reduction of descriptive terms is significant ($p < .001$) and should therefore, we hypothesise, improve the precision of the annotation of information objects.

The ratio of references that have descriptive power is visualized in Figure 2 for each category. The trend shows how globally frequently occurring categories have lower relative characteristic descriptive power then rarely occurring categories of objects. Thus, a rare historical railway would provide a better characterization of the region in its immediate neighbourhood than a post-office.

Table 2. Frequency of Cases per category (null values omitted)

Category	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10	c11
Case 1	499	249	354								
Case 2				8							
Case 3	305	191	199								
Case 4				17	11	33	5	7	2	4	
Case 5	27	18	55								
Case 6				38	25	8	4				
Case 7	2	9	7								
Case 8				97	133	95	46	9	2	1	3
Total	833	467	615	160	169	136	55	16	4	5	3

5 Conclusions and Further Work

The method proposed allows the identification of categories of spatial objects which are prominent in, and we assume therefore relevant to, characteristic descriptions of regions. The method can be used to characterise regions and information objects attached to them, such as photographs, and potentially for the identification of relevant terms for the construction of thematic descriptions in geographic information retrieval [13, 3].

The categories identified are considered relevant, based on the assessment of their prominence in a hierarchical system of nested regions. The assessment of prominence is based on the variation of occurrence frequency of a category within the nested regions. However, we have not considered the semantics of the different categories in our classification of relevance and in future work we will consider how this can be incorporated in our method.

The results produced by the model will be influenced by generalization and cartographic biases of the datasets used. We assume that the spatial variation in occurrence of a category of POI in the dataset is exclusively due to the spatial variation of the occurrence of the POI category in reality. Further work will explore the use of the methodology with real datasets with and without supposed cartographic biases for building annotations for georeferenced images. The technique will be evaluated by comparing identified categories of prominent spatial objects for a given footprint with human-generated textual characterisations sourced from the Web, as well as in a subject testing experiment.

Acknowledgments

The research reported in this paper is part of the project TRIPOD supported by the European commission under contract No. 045335.

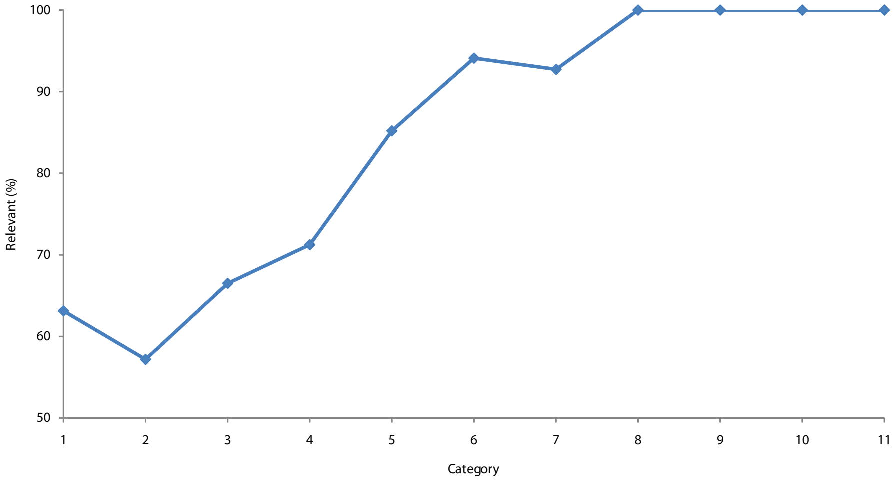

Fig. 2. Ratio of references with descriptive power, per category

References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American **184** (2001) 34–43
2. Himmelstein, M.: Local Search: The Internet Is the Yellow Pages. Computer **38** (2005) 26–34
3. Purves, R., Clough, P., Jones, C.B., Arampatzis, A., Bucher, B., Finch, D., Fu, G., Joho, H., Syed, A.K., Vaid, S., Yang, B.: The Design and Implementation of SPIRIT: a Spatially Aware Search Engine for Information Retrieval on the Internet. International Journal for Geographical Information Science **21** (2007) 717–745
4. Naaman, M., Songa, Y.J., Paepckea, A., Garcia-Molina, H.: Assigning Textual Names to Sets of Geographic Coordinates. Computers, Environment and Urban Systems **30** (2006) 418–435
5. Purves, R., Edwardes, A.J., Sanderson, M.: Describing the Where - Improving Image Annotation and Search Through Geography. In: 1st Intl. Workshop on Metadata Mining for Image Understanding (MMIU 2008), Funchal, Madeira - Portugal (2008)
6. Shatford, S.: Analyzing the Subject of a Picture: A Theoretical Approach. Cataloging and Classification Quarterly **6** (1986) 39–62
7. Edwardes, A.J., Purves, R.: A Theoretical Grounding for Semantic Descriptions of Place. In Ware, J.M., Taylor, G.E., eds.: Web and Wireless Geographical Information Systems. 7th International Symposium, W2GIS 2007, Cardiff, UK, 28-29th Nov. 2007, Proceedings. Volume 4857. Springer-Verlag, Berlin, Heidelberg (2007) 106–120
8. Montello, D.R., Goodchild, M., Gottsegen, J., Fohl, P.: Where's Downtown?: Behavioral Methods for Determining Referents of Vague Spatial Queries. Spatial Cognition and Computation **3** (2003) 185–204
9. Nothegger, C., Winter, S., Raubal, M.: Computation of the Salience of Features. Spatial Cognition and Computation **4** (2004) 113–136

10. Regnauld, N., McMaster, R.B.: A Synoptic View of Generalisation Operators. In Mackaness, W.A., Ruas, A., Sarjakoski, L.T., eds.: Generalisation of Geographic Information: Cartographic Modelling and Applications. ICA/Elsevier, Amsterdam (2007) 37–66
11. Salton, G., Buckley, C.: Term-Weighting Approaches in Automatic Text Retrieval. Information Processing & Management **24** (1988) 513–523
12. Goodchild, M.: Citizens as Sensors: the World of Volunteered Geography. GeoJournal **69** (2007) 211–221
13. Cai, G.: GeoVSM: An Integrated Retrieval Model for Geographic Information. In Egenhofer, M.J., Mark, D., eds.: Geographic Information Science: Second International Conference, GIScience 2002, Boulder, CO, USA, September 25-28, 2002. Proceedings. Volume 2478 of Lecture Notes in Computer Science. Springer-Verlag, Berlin, Heidelberg (2002) 65–79