Assigning Textual Names to Sets of Geographic Coordinates

[Extended Abstract]

Mor Naaman, Yee Jiun Song, Andreas Paepcke, Hector Garcia-Molina
Stanford University

1. INTRODUCTION

In many situations, it is necessary for a set of geographic coordinates to be described with textual place names that are familiar to humans. One reason to do so is to convert to text a list of coordinates, that may appear on a web page. For example, a list of some geo-referenced observations, or a set of geo-referenced digital photographs that appear on a page. This textual name can later serve a number of uses, including text-based retrieval and textual representation of the coordinate set. We focus and report here on the latter. However, the techniques can be applied towards text-based retrieval as well.

The problem can be simply stated as follows: *Given a set of geographic coordinates, find a textual name that describes them best.* However, it is assumed that the set is somewhat coherent — it is not the case that some coordinates are in Switzerland while other coordinates in the same set are in England.

In [3] we explore a "flip" version of this problem: given a set of coordinates, each associated with some free-text caption, propose a good geographically-meaningful name for the set, or for a new un-labeled coordinate that occurred in the same area.

We first describe a sample application of the naming technique (Section 2). Section 3 describes the algorithm itself.

2. SAMPLE APPLICATION

In [4] we used location information to automatically organize collections of geo-referenced digital photographs. Our system, PhotoCompas, groups the photos into sets that represent different events and locations where photos were taken. Once this step is completed, PhotoCompas needs a way to present the results in a user interface, without the benefit of a map. The second processing step is therefore to assign textual names to the nodes in the location and event hierarchies. The names for the nodes are based on the coordinates of the photos belonging to these nodes.

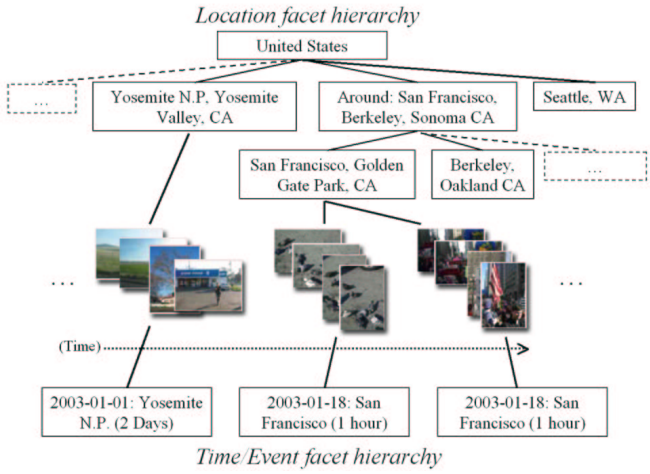

Figure 1: Sample PhotoCompas structure. Parts of the location and time/event hierarchies for an actual collection of photos, including names as generated by our algorithm.

Figure 1 shows a subset of a sample location and event grouping created by this algorithm. The nodes are annotated using the textual captions generated by the naming algorithm described below.

The coordinate sets are guaranteed to be relatively coherent due to the specifics of the algorithm that generates them, thus matching the problem we wish to solve in this paper.

3. THE NAMING ALGORITHM

Our naming process has three steps. First, for each latitude/longitude pair, we find the state, city and/or park that contain it. This is done using an off-the-shelf geographic dataset of administrative regions.[1] For example, a particular coordinate may be inside of California (state), San Francisco (city), and Golden Gate National Recreational Area (park). Another coordinate may be in Washington (state)

[1]Regretfully, we only have access to a database of US cities and parks. Thus, we have only tested our naming procedure on US coordinates.

and Seattle (city), but not in any park.

We compute the frequency at which each city and park occur in the set of coordinates, building a term-frequency table. We weigh each type differently, with national parks weighed more heavily than cities, and cities weighed more heavily than other parks such as state parks or national forests. The different weights allow us to give more importance to names that are more likely to be recognizable to users. At the end of this process, we have a *containment table* with terms and their score.

In the second step, we look for *neighboring* cities. By locating cities that are close to the coordinates in this set, and computing the distance from the center of the set to the city, we are able to produce textual names for these clusters such as "40 KMs south of San Francisco". We pick neighboring cities based on their "gravity": a combination of population size, the city's "Google count", and (inversely) the city's distance from the center of the set of photos. The "Google count" of a city is the number of results that are returned by Google [1] when the name of the city (together with the state) is used as a search term. We use this as a measure of how well known a city is, and thus, how useful it would be as a reference point. For example, a set of coordinates on the Stanford campus may be captioned "40 KMs South of San Francisco", or "30 KMs North of San Jose". The population of these cities is comparable, but since the Google count for San Francisco is much higher than San Jose's, the former is chosen despite being further away. This step creates a *nearby-cities table*, again with terms and their scores.

The final step involves picking 1–3 terms from the tables to appear in the text caption of each set of coordinates. For example, a possible caption can include the two top terms from the containment table, and the top nearby city: "Stanford, Butano State Park, 40 KMs South of San Francisco, CA". Our method of picking the final terms varies according to the semantics of the set of photos we are trying to name, but we do not expand on it here for lack of space.

We have also experimented with the Alexandria Digital Library's gazetteer [2]. While it seems like there are a number of ways in which Alexandria would have been useful for this task, the limited on geographic representation used by Alexandria for features (a bounding box) prevented us from using it. We omit this discussion here for lack of space.

4. EVALUATION

As the context of our application was collections of geo-referenced digital photographs, we evaluated the results using three real-life collections. The number is low due to the current scarcity of large geo-referenced photo collections. We expect more collections to be available in the future, but today, we could only find three subjects with a large enough collection of such photos (each spanning thousands of photos, and at least one year of photo-taking). Our results are then, by necessity, case studies rather than statistically significant analysis. However, these collections supplied dozens of distinct sets of coordinates for the algorithm to name; in that sense, the evaluation was quite broad.

Specifically, we evaluated the naming algorithm through interviews with the owners of the test collections. Our evaluation goals were to verify that the produced textual names are:

- Useful to the subjects, in that a) the name includes terms that are familiar to the subjects and help them understand which geographic area is covered by the cluster, b) the subject is able to tell the cluster apart from other clusters, based on the name and c) the subject can tell which pictures belong to this cluster based on the name.
- Similar to the names that the subjects would have generated themselves.

For each collection, and each cluster, we performed several tests. The results of our tests are not reported here in detail due to lack of space. In brief, the results showed that our algorithms performed very well. In most cases, our algorithm picked at least one name in common with the human subject. In cases where our algorithm picked different names, the subjects found the names picked by our algorithm to be useful.

5. CONCLUSIONS

We have shown that our system, PhotoCompas, can automatically generate a meaningful and useful textual name to sets of coordinates in the context of a geo-referenced personal photo collection. In the future we plan to expand this work to other domains, using different semantics.

6. REFERENCES

[1] Google inc. http://www.google.com.

[2] L. L. Hill, J. Frew, and Q. Zheng. Geographic names - the implementation of a gazetteer in a georeferenced digital library. *CNRI D-Lib Magazine*, January 1999.

[3] M. Naaman, A. Paepcke, and H. Garcia-Molina. From where to what: Metadata sharing for digital photographs with geographic coordinates. In *10th International Conference on Cooperative Information Systems (CoopIS)*, 2003.

[4] M. Naaman, Y. J. Song, A. Paepcke, and H. G. Molina. Automatic organization for digital photographs with geographic coordinates. In *Proceedings of the Fourth ACM/IEEE-CS Joint Conference on Digital Libraries*, 2004.