Building Geospatial Data Collections with Location-based Games

Sebastian Matyas, Peter Wullinger, and Christian Matyas

Chair of Computing in the Cultural Sciences
Laboratory for Semantic Information Processing
Otto-Friedrich-University Bamberg
{firstname.lastname@uni-bamberg.de}

Abstract. The traditional, expert-based process of knowledge acquisition is known to be both slow and costly. With the advent of the Web 2.0, community-based approaches have appeared. These promise a similar or even higher level of information quantity by using the collaborative work of voluntary contributors. Yet, the community-driven approach yields new problems on its own, most prominently contributor motivation and data quality. Our former work [1] has shown, that the issue of contributor motivation can be solved by embedding the data collection activity into a gaming scenario. Additionally, good games are designed to be replayable and thus well suited to generate redundant datasets. In this paper we propose semantic view area clustering as a novel approach to aggregate semantically tagged objects to achieve a higher overall data quality. We also introduce the concept of semantic barriers as a method to account for interaction betwen spatial and semantic data. We also successfully evaluate our algorithm against a traditional clustering method.

Key words: Games and interactive entertainment, Knowledge acquisition

1 Introduction

The traditional method of collecting geospatial data (geographic information combined with semantic annotations) is to send surveyors out into the field and subsequently postprocess the collected data, a process which is both expensive and time consuming. Goodchild [2] suggests replacing the experts by a community of voluntary contributors and thus to reduce survey costs significantly.

However, there are good arguments in favor of the traditional approach: Keeping the motivation of the voluntary contributors high is not an easy task. Additionally, ensuring acceptable data quality is also a problem as voluntary contributors are often inexperienced and community collected data is often of low quality.

Finding solutions to these problems is not trivial, as both are closely related to the well-known knowledge acquisition bottleneck (see e.g. [3]). Ahn and Dabbisch [4] argue that contributor motivation may be upheld by embedding the data collection process into an enjoyable game. Nonetheless, data quality assurance remains a tedious task and it is a challenge to integrate quality control into a game without spoiling the overall game experience.

One potential solution to the data quality problem is game replayability: Games, and in particular location based games are usually designed to be replayable, i.e. let the player pass through similar situations repeatedly in different game instances. If it is possible to adapt the game rules, such that each repeated situation produces a data point, this creates a large, inaccurate, but redundant dataset. To successfully exploit the gathered data, two new challenges have to be tackled: Designing replayable games and aggregating the redundant data points. In this paper we describe a method to aggregate geospatial data points.

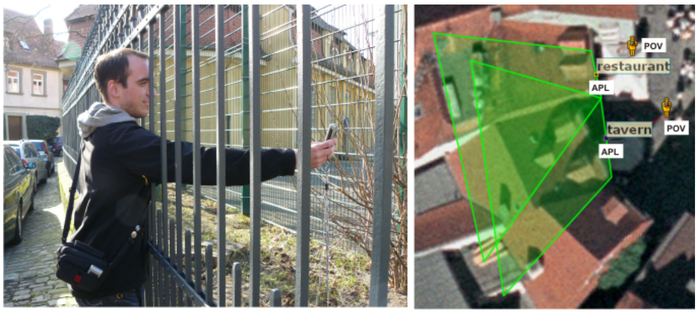

Fig. 1. Inaccessibility of an object and reconstructed view areas for two markers

The rest of the paper is structured as follows: After a review of related work (section 2), we proceed to the main contributions of our paper: (1) We illustrate how specific game rules can promote the collection of redundant data (section 3). (2) We use the data points (section 4) obtained in CityExplorer game rounds as input for a novel clustering algorithm that makes use of the unique structure of the collected data points (section 5). (3) We evaluate our algorithm against real game data to show the improvements in comparison to a traditional clustering algorithm (section 5.1). (4) We conclude the paper with the introduction of the concept of "*semantic barriers*" (section 5.2) as a method to account for semantic dissimilarity in spatial clustering and give an outlook onto further improvements (section 6).

2 Related Work

It has long been discussed in the games research community to gather geospatial data in a game [5]. CityExplorer [1], however, is the first working location-based game that specifically collects geospatial and semantic data of real world objects.

Earlier work like [6] either focuses on only one kind of data (semantic or geospatial) and/or does not feature a quality control element. Lately, Bell et al. [7] presented the Eyespy game to identify digital photos suitable for navigation tasks, but their approach does not make proper use of data semantics, too.

Spatial clustering has so far relied mainly on geometric similarity measures between points of interest (POIs), consisting of a single spatial coordinate pair (e.g. [8] or [9]) to find assortative objects. Semantic information is most often used only for preliminary filtering if at all [10]. Snavely et al. [11] present an new approach to cluster images based on viewpoints and SIFT keypoints. Their similarity measurement is based solely on the keypoints and uses GPS data only for visualization purposes. They are able to propagate the semantic data associated to parts of the image to other images that feature the same part, but do not use this kind of data in their similarity computation process.

3 Game Rules Promoting Multiple Measurements

The goal of CityExplorer is to seize the majority of segments in an initially unexplored city-wide game area by placing virtual "followers" (*markers*) within them (see [1] for details). CityExplorer consists of an outdoor component where players use mobile phones to place markers and an online component in which they upload their markers and the game status is visualized [1]. Placing markers is only admissible at real-world locations of predefined categories of non-movable objects. Two basic rules make CityExplorer replayable:

Player-generated Content: In CityExplorer players can choose object categories freely. This makes it possible to reuse the same game area several times.

Freely Eligible Game-relevant Locations: Markers may be set anywhere in the game area. The only restriction is the list of available categories. The fact that players can transform any non-moveable object of interest (OI) into a game-relevant location greatly improves replayability.

Together, both rules make up an interesting property of CityExplorer: OIs are usually tagged not only once but multiple times, sometimes even in the same game round.

With only these basic rules, however, each marker would be represented only by a single spatial coordinate associated with a category. To improve clustering, we have requested the players to follow a certain procedure when *marking* an OI: Players have to take a photo of the object and subsequently move as near as possible towards the object (see figure 1) before selecting the object's category. We record the position (as GPS coordinates) of the player in both steps.

[1] see `http://www.kinf.wiai.uni-bamberg.de/cityexplorer`

4 Aggregating Semantically Tagged Spatial Data Points

Community collected spatial data points are usually not accurate. Errors may occur in both the spatial coordinates as well as the semantic data of the dataset: The measuring errors of consumer-grade GPS devices are sometimes in the range of several meters and proper categorization of objects is often non-trivial even to experts.

In general, we can view the collected data points as approximations of an abstract, errorless structure. Formally, we call the errorless structure an **object of interest** (OI). An OI is a pair $OI \equiv_{\text{def}} \left\langle \overrightarrow{poi}, cat \right\rangle$, consisting of

poi The **point of interest**, a 2D-dimensional spatial coordinate that yields the geographic location of the POI.

cat A piece of semantic information. For simplicity, we assume that *cat* is one from a predefined set of categories. Our algorithm is extensible to more complex information sets.

As described above, players of CityExplorer place markers on the game fields to identify the approximate locations of OIs. These markers only represent vague information. Formally, a marker is a 3-tuple $mrk \equiv_{\text{def}} \left\langle \overrightarrow{apl}, va, cat \right\rangle$ consisting of

APL The **assumed POI location**, a 2D-dimensional geographic coordinate vector, that gives an estimate of the actual POI.

VA **view area** (VA). While the APL is an estimate of the POI, the view area represents the geographic area, where the POI is assumed to be found. Shape and size of the view area are depend on the measuring error.

cat A **category identifier** (*cat*) that uniquely associates a marker with a user defined category.

Semantically tagged data points may contain errors in two dimensions: The location and the tag. Since all game-relevant locations can be freely chosen by the players not all of them may actually be accessible (see e.g. figure 1). In this case, the APL is really only an approximation of the actual POI. Given the above game rules, we make the following assumptions: (1) The viewing direction is from the POV in the direction of the APL. (2) The actual object is “behind” the APL (as viewed from the POV) (3) The cameras’ aperture angle further limits the location of the object. We have therefore decided to restrict the potential locations of objects to viewing trapezoids. A schematic overview is given in figure 1. Miscategorization errors occur, when the tag assigned by the user does not match with the desired categorization. In particular, user assigned categories rarely extend to the desired level of intensional information about an object. We propose to use semantic similarity to tackle this problem.

5 Semantic View Area Clustering

Different players of the same or other game sessions can place a marker on the same object. It is even possible for a single player to set different markers on the same object using different categories if applicable. For example, a player could place a marker at *objectA* using "*building*" as a category and a second marker at *objectA* using "*restaurant*", if both categories are available in the game session. To expedite data quality, it is desirable to re-integrate those duplicate markers, a process commonly known as "*clustering*".

We propose *semantic view area clustering* (svac) as an algorithm that makes use of marker structure without increasing complexity. The algorithm consists of a preprocessing step followed by the actual clustering operation.

```
CropViewAreas($Markers$, $threshold$)
for $\langle m1, m2 \rangle \in Markers \times Markers$
    do if $m1 \neq m2$ and If $sim_C(m1, m2) < threshold$
        ▷ Crop marker view area, unless semantically similar
          then $m1.VA \leftarrow m1.VA - m2.VA$
```

Fig. 2. View Area Cropping

.

```
SVAC($Markers$)
$O \leftarrow Markers$                                  ▷ Open/unprocessed marker list
$C \leftarrow \emptyset$                                                     ▷ Cluster list
while $O \neq \emptyset$
    ▷ Create a new cluster with a marker from the open list
    ▷ Combine $currentArea$ with the view area of the new marker
      do $m1 \leftarrow$ pickAny($O$); $O \leftarrow O \setminus \{m1\}$
         $currentArea \leftarrow$ getViewArea($m1$)
         $c \leftarrow$ newCluster($m1$)
         $m2 \leftarrow$ findOverlapping($currentArea$, $O$, $m1.cat$)
             ▷ Find all markers with overlap in the same category.
         while $m2 \neq$ nil
             do $O \leftarrow O \setminus \{m2\}$
                $currentArea \leftarrow$ combineViewAreas(
                    getViewArea($currentArea$, $m2$))
                $c \leftarrow c \cup \{m2\}$
                $m2 \leftarrow$ findOverlapping($currentArea$, $O$, $m1.cat$)
         $C \leftarrow C \cup \{c\}$
return C
```

Fig. 3. Simple View Area Clustering

In the preprocessing step, we modify the view areas of the individual markers taking into account the semantic data associated with the markers. We measure the similarity between two marker categories as $sim_C(m1, m2)$. For each pair of markers, where the category-similarity is below a certain threshold, we set $m1.VA \leftarrow m1.VA - m2.VA$, where $-$ is the topological difference operation (see figure 2, line 4). This cropping step removes any overlap between markers, whose categories are too different with regard to $sim_C(m1, m2)$.

In the main clustering step, we make use of the cropped view areas. We first pick a marker *m1* from the list of unclustered markers and create a new cluster containing only *m1*. We store the view area of *m1* as the *currentArea*. The cluster is extended by searching for an additional unclustered marker *m2* whose view area overlaps with *currentArea*. If such a marker is found, it is added to the current cluster and its view area is combined with *currentArea* (see line 10 and 11 in figure 3). If no more overlapping markers are found, we finalize the current cluster and repeat the above steps until all markers are clustered. Pseudo-code for the algorithm is given in figure 3.

5.1 Evaluation

For evaluation, we compared svac against the $k - means$ algorithm found in [9]. We computed Ashbrooks $k - means$ algorithm with a cluster radius of 80m. We test two versions of svac that differ in the implementation of the COMBINEVIEWAREAS() function: (1) view area union and (2) view area intersection.

Table 1. Clustering Results

Algorithm	Precision	Recall	F_β
k-means	0.85	0.82	0.80
svac_catsim (intersect)	0.90	0.72	0.85
svac_catsim (union)	0.89	0.74	0.81

To compute the similarity of two markers our *svac* algorithm needs an external similarity function $sim_C(A, B) : Cat \times Cat \mapsto \mathcal{R}$. We use Resnik's terminological similarity [12] measure with a WordNET backend, using the existing tags as instances.

Precision and recall values of the algorithms were compared against a manual reference clusterinq of the around 700 data points from four different CityExplorer game rounds played in Bamberg (see [1] for details). Additionally, we computed the F_β score (see [13]), using the weighted (with $\beta = 3$) harmonic mean of the precision and recall values for a class of data points in the reference clustering.

The results in table 1 show an increase in precision with regard to k-means. The difference between the intersection and the union variant of svac algorithm is only marginal for our test data set. Nonetheless, we note, that the intersection variant is less prone to the *chaining effect*, known from single-linkage algorithms [14]. The lower recall is due to the fact that although more clusters with only one member were correctly found the percentage of correctly identified larger clusters went slightly down.

5.2 Semantic Barriers

Despite its simplicity, the algorithm exhibits some interesting properties, which we illustrate by the following example: Consider the scenario in figure 4. $mrk1$ and $mrk2$ are of the same, $mrk3$ of a different category. If we do not perform view area cropping, $mrk1$ and $mrk2$ would be incorrectly combined into a single cluster, albeit both are clearly separated by the view area for $mrk3$. We call this separative marker, respectively its corresponding object, a *semantic barrier* as it forms a spatial barrier between semantically similar markers $mrk1$ and $mrk2$. Note that *semantic barriers* are not only found in this particular scenario, but play an important part in every aggregation scenario [15].

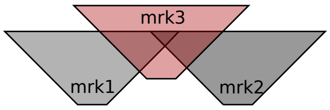

Fig. 4. Inter-category overlap

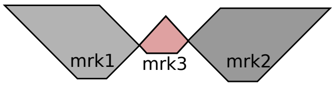

Fig. 5. Cropped overlap
Splinter polygons non-adjacent to the APL have been removed.

To evaluate the usefulness of this modification, the collection of more evaluation data is required, as the current evaluation data set does not contain semantic barrier situations.

6 Conclusion and Outlook

In this paper we illustrated that the addition of two simple game rules are sufficient for a location-based game to produce redundant, semantically tagged geospatial data points. In [1] we already showed that a game like CityExplorer that features these two rules is able to produce large amounts of such data. We presented *svac*, a novel clustering algorithm, that enables us to aggregate multiple, inaccurate measurements of a real-world object to obtain better quality data. We presented semantic barriers as an approach to use semantic information to separate spatially proximate data points, which we expect to generalize well to different scenarios (e.g. [11]).

We believe that the algorithm yields a good balance between complexity and performance and may serve as grounds for future improvements. Manual evaluation of the cropped view areas showed, that the binary, "all-or-nothing" decision to crop view areas is often too radical. A future version of the algorithm should take into account the relative area of overlap and support weighting of overlapping regions using semantic similarity. Additionally, The current algorithm to determine semantic similarity only makes use of the hyponym/hypernom relationship of concepts. We can improve clustering performance, if we take into account the part-of relationship. For example, the categories "*gargoyle*" and "*historical building*" share a high spatial coappearance.

References

1. Matyas, S., Matyas, C., Kiefer, P., Schlieder, C., Mitarai, H., Kamata, M.: Designing location-based mobile games with a purpose - collecting geospatial data with cityexplorer. In: ACM SIGCHI International Conference on Advances in Computer Entertainment Technology: ACE. ACM, New York, NY (2008) 244–247
2. Goodchild, M.F.: Citizens as voluntary sensors: Spatial data infrastructure in the world of web 2.0. International Journal of Spatial Data Infrastructures Research **2** (2007) 24–32
3. Olson, J., Rueter, H.: Extracting expertise from experts: Methods for knowledge acquisition. Expert systems **4**(3) (1987) 152–168
4. von Ahn, L., Dabbish, L.: General techniques for designing games with a purpose. Communications of the ACM **August** (2008) 58–67
5. Capra, M., Radenkovic, M., Benford, S., Oppermann, L., Drozd, A., Flintham, M.: The multimedia challenges raised by pervasive games. In: ACM international conference on Multimedia, New York, NY, USA, ACM (2005) 89–95
6. Casey, S.K.B., D., R.: The gopher game: a social, mobile, locative game with user generated content and peer review. In: International Conference on Advances in Computer Entertainment Technology: ACE. Volume 203. (2007) 9–16
7. Bell, M., Reeves, S., Brown, B., Sherwood, S., MacMillan, D., Ferguson, J., Chalmers, M.: Eyespy: supporting navigation through play. In: CHI: International Conference on Human Factors in Computing Systems, New York, NY, USA, ACM (2009) 123–132
8. Morris, S.M.A., K., B.: Digital trail libraries. In: ACM/IEEE-CS Joint Conference on Digital Libraries. Volume JCDL '04. ACM, New York, NY (2004) 63–71
9. Ashbrook, D., Starner, T.: Learning significant locations and predicting user movement with gps. In: ISWC: IEEE International Symposium on Wearable Computers, Washington, DC, USA, IEEE Computer Society (2002) 101–108
10. Gösseln, G.v., Sester, M.: Integration of geoscientific data sets and the german digital map using a matching approach. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences **35** (2004)
11. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: exploring photo collections in 3d. In: ACM SIGGRAPH 2006 Papers, New York, NY, USA, ACM (2006) 835–846
12. Resnik, P.: Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence **11**(11) (1999) 95–130
13. Zhao, Y., Karypis, G.: Evaluation of hierarchical clustering algorithms for document datasets. In: CIKM '02: International Conference on Information and Knowledge Management, New York, NY, USA, ACM (2002) 515–524
14. Eckes, T. und Rossbach, H.: Clusteranalysen. Kohlhammer (1980)
15. Matyas, S.: Collaborative spatial data acquisition - a spatial and semantic data aggregation approach. In: AGILE International Conference on Geographic Information Science, Aalborg University (2007)