# Detecting Places Of Interest using Social Media

Steven Van Canneyt*, Steven Schockaert†, Olivier Van Laere* and Bart Dhoedt*

*Department of Information Technology
Ghent University, IBBT, Belgium
Email: {Steven.VanCanneyt,Olivier.VanLaere,Bart.Dhoedt}@intec.ugent.be
†School of Computer Science & Informatics
Cardiff University, United Kingdom
Email: S.Schockaert@cs.cardiff.ac.uk

***Abstract*—Place recommender systems are increasingly being used to find places of a given type that are close to a user-specified location. As it is important for these systems to use an up-to-date database with a wide coverage, there is a need for techniques that are capable of expanding place databases in an automated way. On the other hand, social media are a rich source of geographically distributed information. In this paper, we therefore propose an approach to discover new instances of a given place type by exploiting correlations between terms and locations in geotagged social media. For a variety of place types, our approach is able to find places which are not yet included in popular place databases such as Foursquare or Google Places.**

***Keywords*-Social Media; Geographic Information Retrieval; Detecting Places Of Interest**

## I. Introduction

Databases of places such as Foursquare, Google Places, LinkedGeoData and Geonames have become increasingly popular in the last few years. These databases are constructed in different ways: A first method — which is used by Geonames — is to combine data from several existing sources such as the National Geospatial-Intelligence Agency's and hotels.com. Second, LinkedGeoData [1] uses the data of OpenStreetMaps, which is generated mainly by user generated GPS track logs and by users who explicitly submit information about places. A similar approach is used in Foursquare, where users can freely add places to the database. Finally, some sources, such as Google Places, do not clearly specify their sources, but users can add places after approval of moderators. Regardless of which of these methods is being used, databases may be outdated and incomplete due to the manual effort which is required to keep them up to date, especially outside major cities.

On the other hand, social media contain a lot of geographically distributed information. For example, there are currently more than 185 million[1] photos on Flickr that have been annotated with geographical coordinates. Additionally, an estimated 1.5% of all the Twitter posts (i.e. tweets) are geotagged [2]. Social media can be used to e.g. automatically detect events [3], to find popular places [4], [5] and tourist routes [6].

[1]http://www.flickr.com/map/, accessed on May 22, 2012

The goal of our research is to discover new places of a given type such as 'restaurant' or 'library'. To this end, we use descriptions of locations and places based on data collected from social media. Murdock [2] states that Flickr and Twitter are used for different objectives and by different kinds of users. Flickr photos are mainly generated by visitors, which leads to a bias for tourist attractions. On the other hand, Twitter is mostly used by residents and may therefore lead to better descriptions of places such as libraries, graveyards and schools. To assess the impact of these differences in usage, we evaluate our approach both on data collected from Flickr and Twitter. We also use the density of known place types near a given, unknown, place to determine to which type it belongs. For example, the likelihood that an unknown place is a pub may increase if there are other pubs in the surroundings.

## II. Related work

Automatic detection of the location and characteristics of places using social media has attracted attention by researchers in recent years. A summary of this research was given by Murdock [2] and Naaman [7]. In this section, we focus on the work that is most related to our research.

Initial work on determining points of interest (POIs) from social media has been exclusively based on analyzing the coordinates of geotagged data. For instance, Crandall et al. [4] used Mean Shift to cluster the locations of geotagged Flickr photos to detect POIs. This method has among others been applied in [5] to detect and recommend popular tourist places in cities. A second line of research relevant to our work analyzes text originating from social media, in order to detect places as well as to retrieve characteristics of these places. Rattenbury et al. [3] used multiscale burst analysis to detect place-related Flickr tags. Our work is most closely related to Gazetiki [8]. They detected places by extracting Wikipedia articles and Panoramio titles which contain a given geographical concept. The detected places were georeferenced, categorized and ranked using Flickr and Alltheweb.

However, so far no effort has been devoted to detect places of a particular type using social media, given only some examples of places of that type.

Table I
INFORMATION OF THE THE PLACE TYPES WHICH ARE CONSIDERED IN THIS PAPER.

| place type | LGD categories | Geonames categories | #places |
|---|---|---|---|
| Place of Worship | PlaceOfWorship | S.CH S.MSQE | 356 329 |
| School | School University | S.SCH | 349 157 |
| Shop | Shop | S.RET | 316 773 |
| Restaurant | Restaurant FastFood | S.REST | 215 613 |
| Graveyard | GraveYard | S.CMTY S.GRVE | 139 096 |
| Hotel | TourismHotel Motel Hostel | S.HTL | 136 174 |
| Pub | Pub Bar Cafe | S.PUB S.CAFE | 132 123 |
| Station | RailwayStation TramStop | S.RSTN S.RSTP S.RSTN S.MTRO | 125 556 |
| Hospital | Hospital | S.HSP S.HSPC S.HSPD S.HSPL | 59 599 |
| Monument | Monument Memorial | S.MNMT | 32 322 |
| Airport | Airport | S.AIRP | 25 591 |
| Library | Library | S.LIBR | 22 946 |
| Museum | TourismMuseum | S.MUS | 19 421 |
| Castle | Castle | S.CSTL | 8 474 |

## III. DATA ACQUISITION

To obtain training and test data, we have collected a set of places with known location and type, based on existing databases. For each of these places, we have subsequently mined Flickr and Twitter to find metadata associated with their locations. We now explain these two steps in more detail.

### A. Collecting Places of Interest

To obtain training and test data, we have used two open source databases: LinkedGeoData[2] (LGD) and Geonames[3]. We have in particular collected all places in these databases of the types with the highest number of places: place of worship, school, shop, restaurant, hotel, graveyard, pub, station, hospital, monument, airport, library, museum and castle. The corresponding categories of LGD and Geonames are specified in Table I. As a result, we obtained 1 698 478 places from LGD and 821 103 from Geonames.

In LinkedGeoData and Geonames, some places occur multiple times, and to ensure a fair evaluation, it is important to detect such duplicates. However, both the name and location of duplicate entries may be slightly different. Therefore, we have used a heuristic based on the approach from [9] to detect and remove duplicates: first, places are indicated as duplicates when they are located closer than 5 meters of each other. Second, to detect additional duplicates of a given place $p$ all neighboring places of the same type in a range of 100 meter were selected as candidate duplicates. Each of the names of these candidates have been converted to lower case, and have been stripped of category words such as 'restaurant', 'bar', 'tavern', etc. A place from the candidate set is assumed to be a duplicate of $p$ if its Damerau-Levenstein distance to $p$ is sufficiently small. For our experiments, we have used a threshold of $x/3$, with $x$ the maximum length of the two names. As a result of this process, we obtained 1 939 174 places. An overview of the number of places per type can be found in the last column of Table I.

The obtained dataset was globally split in three parts: two thirds of the places were used as training data (called the training set, 1 292 782 places) while one sixth of the places were used to find optimal values of the parameters in our methods (called the development set, 323 197 places). The remaining sixth was used for evaluation (called the test set, 323 195 places).

[2]http://www.linkedgeodata.com, release of April 6, 2011

[3]http://www.geonames.com, accessed on March 13, 2012

### B. Collecting Social Media Data

We have collected data from Flickr and Twitter to obtain textual descriptions of places, which will be used to estimate their semantic type.

**Collecting Flickr data.** We crawled the metadata of around 70% of the georeferenced photos from the photo-sharing site Flickr that were taken before May 2011 and which contain a geotag with street level precision (geotag accuracy of at least 15). Once retrieved, we ensured that at most one photo was retained in the collection with a given tag set and user id, in order to reduce the impact of bulk uploads [10]. In addition, photos with invalid coordinates or without tags were removed. The dataset thus obtained contains 23 324 644 geotagged photos.

**Collecting Twitter data.** We used the Twitter Streaming API to collect tweets. Using the 'Gardenhose' access level, we collected about 10% of the public geotagged tweets posted between March 13, 2012 and May 12, 2012. Because we were specifically interested in the added value of using Twitter, we have removed content which was automatically created by other services. More precisely, automatic generated content from Foursquare, Instagram, Path and Yahoo! Koprol has been removed. Finally, the tweets were converted to lower case, and urls and special characters such as #, & and punctuations were removed. After filtering, we ended up with a total number of 10 042 000 tweets.

## IV. DESCRIPTION OF LOCATIONS

A variety of features can be used to estimate whether a given location contains a place of a given type. In particular, in this paper we use two different kinds of features to describe a location. First, a location $l$ can be described using the tags of the Flickr photos and the terms from the Twitter posts that are associated with nearby locations. Second, the density of places of different types in the neighborhood of the location can be used as additional evidence. The optimal settings for each method were determined by optimizing the Mean Average Precision (MAP) of the rankings of the places in the development set.

### A. Textual Descriptions

Social media provide a rich source of geographical information. As described in [11], [10], [12], [13], the tags of Flickr photos and the terms in Twitter posts often describe the geographic location that is associated with these items. This suggests that social media can offer relevant information to derive the type of a place.

**Flickr.** Using Flickr data, we describe a location $l$ as a feature vector $V_l^F$. Each component of this vector is associated with a word from the dictionary $D^F$. This dictionary $D^F$ is the set of all the tags of the Flickr photos associated with the places in the training set.

Table II
USED $\sigma$-VALUES IN THE GAUSSIAN DISTRIBUTIONS OF EQUATION 1, 2 AND 3.

| Place of Worship | School | Shop | Restaurant | Graveyard | Hotel | Pub |
|---|---|---|---|---|---|---|
| 15 | 50 | 25 | 15 | 30 | 20 | 15 |
| **Station** | **Hospital** | **Monument** | **Airport** | **Library** | **Museum** | **Castle** |
| 30 | 40 | 10 | 50 | 10 | 30 | 35 |

Formally, for feature vector $V_l^F$ of location $l$, the component $c_w$ associated with word $w \in D^F$ is given by a Gaussian-weighted count of the number of nearby photos that have been tagged with $w$. For efficiency, photos whose distance to $l$ is more than $2\sigma$ are not considered:

$$c_w = \sum_{f \in F_w \ \wedge \ d(l,f) \leq 2 \cdot \sigma} e^{-\frac{1}{2 \cdot \sigma^2} \cdot d(l,f)^2} \quad (1)$$

with $f$ a Flickr photo with its tags, $F_w$ the set of Flickr photos which contain tag $w$ and $d(l, f)$ the distance between location $l$ and the coordinates of the photo $f$. We determined an optimal $\sigma$ for each place type using the development set (see Table II). Finally, we use the Euclidean norm to normalize these feature vectors, formulated as *norm*$(V_l^F)$.

**Twitter.** From Twitter, we derive a vector $V_l^T$ for each location $l$, with one component $c_w$ for every term $w \in D^T$:

$$c_w = \sum_{t \in T_w \ \wedge \ d(l,t) \leq 2 \cdot \sigma} e^{-\frac{1}{2 \cdot \sigma^2} \cdot d(l,t)^2} \quad (2)$$

with $w \in D^T$, $D^T$ the set of all the terms of the Twitter posts associated with the places in the training set, $T_w$ the set of Twitter posts which contain term $w$ and $\sigma$ the deviation value (see Table II). Similar to Flickr, we introduce the normalized feature vector *norm*$(V_l^T)$.

**Flickr and Twitter combined.** We combine the Flickr tags and Twitter terms to determine a vector $V_l^{F,T}$ for each location $l$, with one component $c_w$ for every term $w \in D^F \cup D^T$:

$$c_w = \sum_{r \in T_w \cup F_w \ \wedge \ d(l,r) \leq 2 \cdot \sigma} e^{-\frac{1}{2 \cdot \sigma^2} \cdot d(l,r)^2} \quad (3)$$

Similar to the vectors of Flickr and Twitter, we also introduce *norm*$(V_l^{F,T})$.

### B. Neighborhood Description

We add context information of a location by using the density of places of different types in its vicinity. When, for example, a location is surrounded by a large number of shops, this may increase the probability that there is a shop located at this location as well, or even a restaurant or bar, while decreasing the probability that it contains a graveyard or school.

Each location $l$ is represented as a feature vector with one component per place type, in which the features correspond to the weighted number of places in the training set that belong to the corresponding type. The component $c_t$ associated with type $t \in D^N$ in the feature vector $V_l^N$ of location $l$ is given by

$$c_t = \sum_{l' \in Loc_t \ \wedge \ d(l,l') \leq 2 \cdot \sigma} e^{-\frac{1}{2 \cdot \sigma^2} \cdot d(l,l')^2} \quad (4)$$

Based on the results on the development set, we get an optimal performance when $\sigma$ is around 350 meter.

Table III
RESULTS OF THE SVM-M ALGORITHM USING FLICKR TAGS TO DESCRIBE LOCATIONS.

| | P50 | P100 | P500 | P2500 | AP | Prandom |
|---|---|---|---|---|---|---|
| **Place of Worship** | 98.00 | 94.00 | 91.80 | 76.16 | 21.51 | 18.37 |
| **School** | 94.00 | 92.00 | 82.40 | 63.64 | 17.45 | 18.06 |
| **Shop** | 72.00 | 68.00 | 68.60 | 59.48 | 31.40 | 16.27 |
| **Restaurant** | 78.00 | 73.00 | 69.00 | 54.88 | 41.42 | 11.15 |
| **Graveyard** | 86.00 | 88.00 | 72.00 | 19.36 | 6.33 | 7.17 |
| **Hotel** | 86.00 | 87.00 | 83.40 | 66.12 | 20.10 | 7.02 |
| **Pub** | 90.00 | 90.00 | 75.00 | 60.56 | 27.43 | 6.78 |
| **Station** | 94.00 | 90.00 | 88.00 | 67.32 | 18.01 | 6.54 |
| **Hospital** | 82.00 | 81.00 | 62.60 | 19.92 | 5.51 | 3.01 |
| **Monument** | 88.00 | 84.00 | 55.00 | 22.92 | 7.99 | 1.67 |
| **Airport** | 84.00 | 68.00 | 25.00 | 5.68 | 3.08 | 1.33 |
| **Library** | 90.00 | 88.00 | 33.80 | 8.80 | 4.64 | 1.18 |
| **Museum** | 78.00 | 73.00 | 61.40 | 25.24 | 13.92 | 1.02 |
| **Castle** | 66.00 | 59.00 | 34.20 | 10.28 | 8.31 | 0.44 |
| **Mean** | 84.71 | 81.07 | 64.44 | 40.03 | 16.22 | 7.14 |

## V. RANKING PLACES

The task we consider in this paper is to rank locations based on the likelihood that they contain a place of a given type. In particular, given a feature vector $V_l$ describing the location $l$ and a place type $t$, we want to determine a confidence value *conf*$(V_l, t)$.

We use the Support Vector Machine (SVM) implementation of LibLinear [14] to this end. We examine both the one-vs-rest SVM and the multiclass SVM. As one-vs-rest SVM we use the *L2-regularized L2-loss support vector classification (primal)* option of LibLinear (SVM-o). For multiclass SVM we use the SVM classification by Crammer-Singer [15], which is optimized to efficiently classify large datasets (SVM-m). We use the standard configuration of LibLinear both for the one-vs-rest and the multiclass SVM. We have also tested other classifiers such as k-Nearest Neighbors, Naive Bayes and Decision Trees, but as the results were worse, we will not consider them here.

Finally, the confidence values *conf*$(norm(V_l^{F,T}), t)$ and *conf*$(V_l^N, t)$ are combined as follows:

$$conf(norm(V_l^{F,T}), V_l^N, t) = \lambda \cdot conf(norm(V_l^{F,T}), t) + (1 - \lambda) \cdot conf(V_l^N, t) \quad (5)$$

with $\lambda \in [0, 1]$. Optimal results on the development set were obtained for $\lambda = 0.9$.

## VI. EVALUATION

In this section, we will rank a set of a priori chosen locations, which were obtained by extracting the locations of all places in our test set, i.e. the places of a different type are used as negative examples of the considered type. We evaluate the rankings using Average Precision (AP), Precision at 50 (P50), Precision at 100 (P100), Precision at 500 (P500) and Precision at 2500 (P2500). When we compare rankings, we calculate a $p$-value using the Wilcoxon signed-rank test to determine if the differences in the performances are significant. We state that a difference is statistical significant if $p < 0.01$.

**Flickr results.** We start our evaluation by ranking the locations using their *norm*$(V_l^F)$ feature vector. By comparing the results of the classifiers, we could conclude that SVM-m classifier performs significantly better than SVM-o classifier. In the remainder of this section, we

Table IV
RESULTS OF THE SVM-M ALGORITHM USING TWITTER TERMS TO DESCRIBE LOCATIONS.

| | P50 | P100 | P500 | P2500 | AP | Prandom |
|---|---|---|---|---|---|---|
| **Place of Worship** | 74.00 | 57.00 | 24.80 | 14.60 | 16.38 | 18.37 |
| **School** | 86.00 | 73.00 | 63.60 | 52.36 | 16.67 | 18.06 |
| **Shop** | 62.00 | 66.00 | 60.40 | 49.40 | 29.88 | 16.27 |
| **Restaurant** | 64.00 | 64.00 | 61.20 | 47.92 | 53.29 | 11.15 |
| **Graveyard** | 4.00 | 3.00 | 1.40 | 0.84 | 4.36 | 7.17 |
| **Hotel** | 74.00 | 82.00 | 62.80 | 36.40 | 12.79 | 7.02 |
| **Pub** | 54.00 | 47.00 | 42.40 | 32.92 | 21.61 | 6.78 |
| **Station** | 78.00 | 70.00 | 49.40 | 25.28 | 9.46 | 6.54 |
| **Hospital** | 76.00 | 71.00 | 62.40 | 16.96 | 6.44 | 3.01 |
| **Monument** | 20.00 | 13.00 | 5.80 | 2.12 | 1.59 | 1.67 |
| **Airport** | 64.00 | 35.00 | 8.20 | 2.04 | 1.64 | 1.33 |
| **Library** | 20.00 | 12.00 | 4.80 | 2.60 | 1.20 | 1.18 |
| **Museum** | 34.00 | 29.00 | 9.20 | 4.32 | 1.45 | 1.02 |
| **Castle** | 10.00 | 5.00 | 1.20 | 0.32 | 0.44 | 0.44 |
| **Mean** | 51.43 | 44.79 | 32.69 | 20.58 | 12.66 | 7.14 |

will therefore focus on the results of SVM-m classifier. A detailed view of its performance is shown in Table III, with the expected Average Precision of a random ranking specified in the 'Prandom' column.

We observe a remarkable performance regarding the P50, P100 and even P500 values. We get P50 and P100 values of more than 90% for places of worship, schools, pubs and stations. In contrast with what was suspected in [2], there is no clear over-representation of tourist places and an under-representation of ordinary places. For example, for many castles, the associated tags were not informative enough to find out their place type. On the other hand, we obtain good results for schools and graveyards, which are not usually associated with tourism. A closer look at the Flickr data shows that the tags which are indicative for schools are mostly collected from photos of school events, art photos or photos of school buildings in the neighbourhood of a user's residential area. On the other hand, the photos from graveyards were mostly taken by members of the 'Cemetery Central (Geo-tagged Photos)' Flickr group. Their goal is to locate, celebrate, and preserve gravesites by uploading geotagged photos of them.

**Twitter results.** The results of using the Twitter-based feature vectors $norm(V_l^T)$ in combination with the SVM-m classifier are shown in Table IV. Although a large number of tweets are not informative (see [16]), we have observed that some tweets are very useful to recognize place types. For example 'okay i'm ready for the bell to ring', 'waiting for the train...' and 'sitting in the hospital...not fun', leading to higher P50 and P100 values for schools, stations and hospitals, respectively. Similar to the Flickr results, we can not observe a clear difference between the results of tourist places and non-tourist places.

**Flickr and Twitter combined.** Comparing the results obtained by using the $norm(V^F)$ vectors (Table III) and $norm(V^T)$ vectors (Table IV), we can detect that we get significant better precision values when we use Flickr tags than when we use Twitter terms to describe locations. However, when we use both Flickr tags and Twitter terms in the $norm(V^{F,T})$ feature vectors, we get a significant improvement for precision at 2500 (Table V).

**Neighborhood results.** As a third source to detect locations containing a place of a given type, we use the density of places of different types in the neighborhood around the locations ($V^N$). By comparing the results of the classifiers, we could conclude that the SVM-o classifier performs significantly better than the SVM-m classifier; a summary of its performance can be found in Table VI.

We observe P50 and P100 values of more than 80% and P500 of more than 70% for places of worship, schools, hotels and hospitals, which indicates a strong relation between these types of places and the types of the places in their neighborhood. A more detailed analysis of the results has suggested that there are three classes of place types. The first class is where places of the same type are located near each other. For example, school, airport and hospital contain different building at locations near each other. A second class of places are located in dense regions, such as pubs, shops, hotels and restaurants. Finally, for some types there is hardly any connection between the place type and the place types of their neighbors such as libraries and castles.

Table V
RESULTS OF THE SVM-M ALGORITHM USING FLICKR TAGS AND TWITTER TERMS TO DESCRIBE LOCATIONS.

| | P50 | P100 | P500 | P2500 | AP | Prandom |
|---|---|---|---|---|---|---|
| **Mean** | 84.86 | 82.21 | 66.34 | 42.23 | 16.79 | 7.14 |

Table VI
RESULTS OF THE SVM-O ALGORITHM USING $V_l^N$ FEATURE VECTORS TO DESCRIBE LOCATIONS.

| | P50 | P100 | P500 | P2500 | AP | Prandom |
|---|---|---|---|---|---|---|
| **Place of Worship** | 82.00 | 86.00 | 72.80 | 62.56 | 31.77 | 18.37 |
| **School** | 90.00 | 88.00 | 85.20 | 74.12 | 31.36 | 18.06 |
| **Shop** | 62.00 | 60.00 | 64.20 | 59.56 | 29.02 | 16.27 |
| **Restaurant** | 50.00 | 48.00 | 46.00 | 47.96 | 19.39 | 11.15 |
| **Graveyard** | 60.00 | 52.00 | 29.00 | 20.00 | 13.12 | 7.17 |
| **Hotel** | 86.00 | 81.00 | 72.00 | 64.4 | 27.93 | 7.02 |
| **Pub** | 66.00 | 59.00 | 46.20 | 39.84 | 11.20 | 6.78 |
| **Station** | 52.00 | 51.00 | 44.60 | 54.44 | 22.78 | 6.54 |
| **Hospital** | 90.00 | 84.00 | 72.00 | 49.72 | 13.10 | 3.01 |
| **Monument** | 60.00 | 68.00 | 43.60 | 21.08 | 6.49 | 1.67 |
| **Airport** | 80.00 | 76.00 | 42.60 | 8.52 | 5.53 | 1.33 |
| **Library** | 10.00 | 8.00 | 4.00 | 1.64 | 0.97 | 1.18 |
| **Museum** | 88.00 | 78.00 | 29.80 | 10.44 | 5.07 | 1.02 |
| **Castle** | 22.00 | 23.00 | 16.20 | 4.40 | 2.01 | 0.44 |
| **Mean** | 64.14 | 61.57 | 47.73 | 37.05 | 15.70 | 7.14 |

**Combined results.** When we compare the results of the textual description (Table V) with the results of the neighborhood description (Table VI), we can detect that the P50, P100 and P500 values are significant higher for the textual description than for the neighborhood description. The reason is that textual information is more informative than a description based on the places in the neighborhood of the locations. However, when we combine the textual and the neighborhood description as described in Section V we get a significant improvement of the P500 and P2500 values (Table VII).

**Qualitative evaluation.** Finally, we present a qualitative evaluation of our approach. In particular, we want to find new places of a particular type for a given region. For this purpose, we first use a grid overlay over the map of the region of interest and consider the centers of the obtained cells as locations which potentially contain places of a given type. As training set we use the whole dataset because we want to detect new places, i.e. places which were not yet included in our dataset.

Table VII
RESULTS OF USING ALL THE DESCRIPTIONS TO DESCRIBE AREAS.

| | P50 | P100 | P500 | P2500 | AP | Prandom |
|---|---|---|---|---|---|---|
| **Place of Worship** | 96.00 | 96.00 | 91.60 | 77.84 | 31.58 | 18.37 |
| **School** | 94.00 | 91.00 | 84.20 | 70.64 | 29.44 | 18.06 |
| **Shop** | 78.00 | 71.00 | 69.80 | 63.60 | 27.22 | 16.27 |
| **Restaurant** | 78.00 | 78.00 | 70.80 | 58.04 | 34.14 | 11.15 |
| **Graveyard** | 86.00 | 85.00 | 72.60 | 22.64 | 14.98 | 7.17 |
| **Hotel** | 92.00 | 87.00 | 85.20 | 68.28 | 26.44 | 7.02 |
| **Pub** | 88.00 | 86.00 | 84.00 | 61.68 | 21.43 | 6.78 |
| **Station** | 96.00 | 93.00 | 87.60 | 89.08 | 27.48 | 6.54 |
| **Hospital** | 86.00 | 83.00 | 74.20 | 40.48 | 13.17 | 3.01 |
| **Monument** | 90.00 | 88.00 | 57.20 | 24.24 | 10.24 | 1.67 |
| **Airport** | 90.00 | 82.00 | 29.60 | 11.60 | 6.53 | 1.33 |
| **Library** | 90.00 | 85.00 | 35.80 | 9.04 | 4.23 | 1.18 |
| **Museum** | 80.00 | 76.00 | 64.00 | 25.60 | 14.26 | 1.02 |
| **Castle** | 74.00 | 64.00 | 34.20 | 11.84 | 9.41 | 0.44 |
| **Mean** | 87.00 | 83.21 | 67.20 | 45.33 | 19.33 | 7.14 |

This method has been applied to the city of London using cells of 30 by 30 meter. The resulting ranking contained places of various types which are not included in Geonames or LinkedGeoData. Examples are the Surrey Quays Shopping Centre, Elistano restaurant, London Hostel Association, the Animals in War monument and the Imam Khoei Islamic Centre. Along similar lines, some of these places were not included in the database of Google Places (e.g. the Animals in War monument and the Frank Barnes School) or even in the Foursquare database (e.g. the Imam Khoei Islamic Centre). These observations clearly support our hypothesis that deriving places from social media can provide a useful way to complement existing place databases.

## VII. CONCLUSIONS

In this paper, we described a method to detect places of different types. We first derived for each considered location a feature vector from the tags of the Flickr photos and the terms from the Twitter posts that are associated with coordinates in their vicinity and another feature vector from the number of places of each type that are located in their neighborhood. Next, we used support vector machines to rank these locations based on the likelihood that they contain a place of a given type.

By far the most useful of the considered sources is Flickr. Surprisingly, we could not notice a clear difference in performance between tourist and non-tourist places. This stands in contrast with the hypothesis from [2] that there is an under-representation of 'ordinary places' on Flickr. In addition we have used Twitter as a source to detect places. Naaman et al. [16] determined that only about 20% of the tweets are used to share information. However, as we have demonstrated in this paper, tweets can still be used to detect places of different types. For instance, places of worship, schools, hotels and stations were detected with a precision at 50 of more than 74%, although using twitter alone leads to a poor performance for graveyards, libraries, museums, monuments and castles. As a third source of information, we have looked at the distribution of places of different types in the neighborhood of a location of interest. We observed precision at 50 and 100 of more than 80% for places of worship, schools, hotels and hospitals, which indicates a strong relation between these types of places and the type of the places in their neighborhood.

A careful analysis of our results has shown that several places can be discovered in this way, that are not yet contained in well-known places databases, clearly illustrating the potential of social media for populating place databases.

## REFERENCES


[1] S. Auer, J. Lehmann, and S. Hellmann, "LinkedGeoData: Adding a spatial dimension to the web of data," in *Proceedings of the 8th International Semantic Web Conference*, 2009, pp. 731–746.

[2] V. Murdock, "Your mileage may vary: on the limits of social media," *SIGSPATIAL Special*, vol. 3, no. 2, pp. 62–66, 2011.

[3] T. Rattenbury, N. Good, and M. Naaman, "Towards automatic extraction of event and place semantics from Flickr tags," in *Proceedings of SIGIR*, 2007, pp. 103–110.

[4] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, "Mapping the world's photos," in *Proceedings of the 18th International Conference on World Wide Web*, 2009, p. 761.

[5] S. Van Canneyt, S. Schockaert, O. Van Laere, and B. Dhoedt, "Time-Dependent Recommendation of Tourist Attractions using Flickr," in *Proceedings of the Benelux Conference on Artificial Intelligence*, 2011, pp. 255–262.

[6] S. Jain, S. Seufert, and B. Srikanta, "Antourage: mining distance-constrained trips from Flickr," in *Proceedings of the 19th International Conference on World Wide Web*, 2010, pp. 1121–1122.

[7] M. Naaman, "Geographic information from georeferenced social media data," *SIGSPATIAL Special*, vol. 3, no. 2, pp. 54–61, 2011.

[8] A. Popescu and G. Grefenstette, "Gazetiki: automatic creation of a geographical gazetteer," in *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries*, 2008, pp. 85–93.

[9] O. Ozdikis, F. Orhan, and F. Danismaz, "Ontology-based recommendation for points of interest retrieved from multiple data sources," in *Proceedings of the International Workshop on Semantic Web Information Management*, 2011.

[10] P. Serdyukov, V. Murdock, and R. van Zwol, "Placing Flickr photos on a map," in *Proceedings of SIGIR*, 2009, pp. 484–491.

[11] S. Ahern, M. Naaman, and R. Nair, "World explorer: visualizing aggregate data from unstructured text in geo-referenced collections," in *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, 2007, pp. 1–10.

[12] O. Van Laere, S. Schockaert, and B. Dhoedt, "Finding locations of Flickr resources using language models and similarity search," in *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, 2011, p. 48.

[13] Z. Cheng and J. Caverlee, "You are where you tweet: a content-based approach to geo-locating twitter users," in *Proceedings of CIKM*, 2010, pp. 759–768.

[14] S. Keerthi, S. Sundararajan, and K. Chang, "A sequential dual method for large scale multi-class linear SVMs," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008, pp. 408–416.

[15] K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *The Journal of Machine Learning Research*, vol. 2, pp. 265–292, 2002.

[16] M. Naaman, J. Boase, C.-h. Lai, and N. Brunswick, "Is it really about me? Message content in social awareness streams," in *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 2010, pp. 189–192.