Using Social Media to Find Places of Interest: A Case Study

Steven Van Canneyt
Ghent University
Gaston Crommenlaan 8
9050 Gent, Belgium
steven.vancanneyt@ugent.be

Steven Schockaert
Cardiff University
The Parade, 5
Cardiff, United Kingdom
s.schockaert@cs.cardiff.ac.uk

Olivier Van Laere
Ghent University
Gaston Crommenlaan 8
9050 Gent, Belgium
olivier.vanlaere@ugent.be

Bart Dhoedt
Ghent University
Gaston Crommenlaan 8
9050 Gent, Belgium
bart.dhoedt@ugent.be

ABSTRACT

In this paper, we show how the large amount of geographically annotated data in social media can be used to complement existing place databases. After explaining our method, we illustrate how this approach can be used to discover new instances of a given semantic type, using London as a case study. In particular, for several place types, our method finds places in London that are not yet contained in the databases used by Foursquare, Google, LinkedGeoData and Geonames. Encouraged by these results, we briefly sketch how similar techniques could potentially be used to identify likely errors in existing databases, to estimate the spatial extent of places, to discover semantic relationships between place types, and to recommend tags to users who are uploading photos.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining, Spatial databases and GIS*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Experimentation

Keywords

Social Media, Geographic Information Retrieval, Detecting Places Of Interest

1. INTRODUCTION

We are becoming increasingly dependent on databases of places such as Foursquare, Google Places, LinkedGeoData and Geonames to find interesting places. However, due to manual effort to compile and update such databases, they are typically incomplete and partially outdated.

An important source to improve existing databases is geographically annotated data in social media. For example, about 1.5% of all Twitter posts (i.e. tweets) are annotated with geographical coordinates [14]. In addition, there are currently more than 190 million geotagged Flickr photos[1]. This data has been used to e.g. automatically detect events [12, 18, 19, 20], to find popular places [5, 7, 8, 24] and tourist routes [6, 9].

In this paper we discuss how social media can be used to improve existing databases of places. We start from our previous research [23], which presents a method that is able to detect places of interest using geographically annotated information obtained from social media. The method is based on the assumption that the type of a place is indicated by the tags of the Flickr photos and the terms of the Twitter posts associated with locations in the vicinity of the place. For example, if photos around a particular location contain tags such as 'food', 'dinner' and 'eating', this strongly suggests that there is a restaurant at that location. In particular, we first train an SVM classifier for a given place type t based on Flickr tags and Twitter terms which are associated with locations nearby known places of type t. Subsequently, we use this classifier to rank locations which potentially contain a place of interest based on the probability that they contain a place of the given type t. We applied our method to 14 different place types on a training set of 1 292 782 places with known place types and a test set of 323 195 locations, which led to rankings with a mean precision value at 50 (mean P@50) of 85%, mean P@100 of 82% and a mean P@500 of 66%.

In the evaluation of [23], we used a quantitative evaluation to demonstrate that our method is able to detect places which are already included in our dataset. However, a more useful application of our approach is to detect places which are not yet included in existing databases of places. Therefore, we perform a a qualitative evaluation discussing in detail which of the places detected by our method for London are not yet

[1]http://www.flickr.com/map/, accessed on July 3, 2012

ACM SIGSPATIAL GEOCROWD '12, November 6, 2012. Redondo Beach, CA, USA.

Table 1: **The place types which are considered in this paper, together with their corresponding category names in LinkedGeoData (LGD) and Geonames.**

place type	LGD categories	Geonames categories
Place of Worship	PlaceOfWorship	S.CH S.MSQE
School	School University	S.SCH
Shop	Shop	S.RET
Restaurant	Restaurant FastFood	S.REST
Graveyard	GraveYard	S.CMTY S.GRVE
Hotel	TourismHotel Motel Hostel	S.HTL
Pub	Pub Bar Cafe	S.PUB S.CAFE
Station	RailwayStation TramStop	S.RSTN S.RSTP S.RSTN S.MTRO
Hospital	Hospital	S.HSP S.HSPC S.HSPD S.HSPL
Monument	Monument Memorial	S.MNMT
Airport	Airport	S.AIRP
Library	Library	S.LIBR
Museum	TourismMuseum	S.MUS
Castle	Castle	S.CSTL

known by existing databases of places. In particular, we are able to find hotels, monuments and castles which are not yet included in LinkedGeoData, Geonames, Google Places nor Foursquare. Furthermore, we discuss how social media can be further used to improve existing databases.

The remainder of this paper is structured as follows. Section 2 summarizes our methodology for obtaining training data. Next, in Section 3, we recall our method from [23] which describes how we model places. Subsequently, we will focus on London to demonstrate how social media can be used to improve existing databases of places in Section 4 and 5. In Section 4, we demonstrate how social media can be used to detect places which are not yet included in existing databases. In addition, Section 5 discusses a number of additional challenges that can be addressed making use of social media data. Finally, we present our conclusions in Section 6.

2. DATA ACQUISITION

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

2.1 Collecting Places of Interest

We have used two open source databases to obtain training data: 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 1.

In LinkedGeoData and Geonames, some places occur multiple times. However, both the name and location of duplicate entries may be slightly different. Therefore, we have used a heuristic based on the approach from [15] to detect and remove duplicates: first, places are indicated as duplicates when they are located closer than 5 meters to 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 distinct places for which locations are plotted in Figure 1. An overview of the number of places per type and source can be found in Table 2.

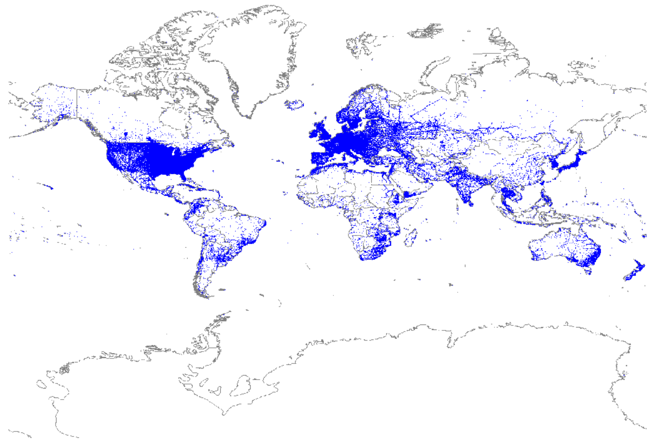

Figure 1: Plot of the places in our dataset.

Table 2: Statistics of the used datasets of places.

place type	LGD	Geonames	combined
Place of Worship	315 532	241 745	356 329
School	284 141	241 041	349 157
Shop	326 388	38	316 773
Restaurant	217 145	1 315	215 613
Graveyard	136 655	125 481	139 096
Hotel	67 563	83 210	136 174
Pub	133 761	0	132 123
Station	80 849	58 484	125 556
Hospital	54 363	24 281	59 599
Monument	35 110	746	32 322
Airport	1 138	24 547	25 591
Library	22 730	11 549	22 946
Museum	18 060	5 000	19 421
Castle	5 043	3 666	8 474
total	1 698 478	821 103	1 939 174

2.2 Collecting Social Media Data

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 [21]. In addition, photos with invalid coordinates or without tags were removed. The dataset thus obtained contains 23 324 644 geotagged photos of which 726 940 are located in London.

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

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

Table 3: Used σ-values in the Gaussian distributions of Equation 1.

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

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 June 23, 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 end up with a total number of 30 095 000 tweets of which 203 885 are located in London.

3. DESCRIBING LOCATIONS

We associate a feature vector V_l to each location l of interest based on the tags of the Flickr photos and the terms from the Twitter posts that are associated with locations nearby l.

Using Flickr and Twitter, we describe a location l by a feature vector $V_l^{F,T}$. Each component of this vector is associated with a word from the dictionary $D^{F,T}$. This dictionary $D^{F,T}$ is the set of all the tags of the Flickr photos and all the terms of the Twitter posts associated with the places in the training set. Formally, for feature vector $V_l^{F,T}$ of location l, the component c_w associated with word $w \in D^{F,T}$ is given by a Gaussian-weighted count of the number of nearby photos and tweets that have been tagged with w. For efficiency, photos and tweets for which distance to l is more than 2σ are not considered:

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

with r a Flickr photo or Twitter post, F_w the set of Flickr photos which contain tag w, T_w the set of Twitter posts which contain term w, and $d(l, r)$ the distance between location l and the coordinates of r.

Places are represented by one coordinate in existing databases, however places can have a varying spatial extent, according to their type. For example, airports are in general larger than restaurants. Therefore, we determined an optimal σ for each place type by using development set (see Table 3). Finally, we normalize these vectors w.r.t. the Euclidean norm, and denote the resulting vectors by $\mathit{normalized}(V_l^{F,T})$.

4. DETECTING PLACES OF INTEREST

Existing databases of places such as LinkedGeoData, Geonames, Foursquare and Google Places are constructed in different ways: first, LinkedGeoData [4] uses the data of OpenStreetMaps, which is derived from user generated GPS track logs and by users who explicitly submit information about places. A similar approach is used in Foursquare [13], where users can freely add places to the database. A second method — which is used by Geonames — is to combine data from several existing sources such as the National Geospatial-Intelligence Agency and hotels.com. 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. Therefore, the goal of this section is to discover new places of a given type such as 'restaurant' or 'library'.

Related work Initial work on determining points of interests (POIs) from social media has been mainly based on analyzing the coordinates of geotagged data. For instance, Crandall et al. [8] used Mean Shift to cluster the locations of geotagged Flickr photos to detect POIs. This method has among others been applied in [7, 24, 5] to detect and recommend popular tourist places in cities. A second line of research analyzes text originating from social media, in order to detect places and their names. Rattenbury et al. [19] used multiscale burst analysis to detect place-related Flickr tags. This technique was applied in [1] to detect names for arbitrary areas in the world. They first cluster the locations where Flickr photos were taken using k-means. For each cluster, representative tags were searched using an extended version of TF-IDF. The most extensive work to detect places using social media was done by Popescu et al. [16, 17]. They detected places by extracting Wikipedia articles, Panoramio titles and Flickr tags 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. In addition, none of the described work analyzed whether their approach was able to detect places which were not yet included in existing databases.

Ranking locations In this paper, we follow our approach from [23] to assess whether a given place is of a particular type. We start with collecting a training set containing locations of places with known place types as described in Section 2.1. Then, using the descriptions $\mathit{normalized}(V_{l_{tr}}^{F,T})$ of the locations l_{tr} in the training set, we train a classifier for each considered place type. To this end, we use the Support Vector Machine (SVM) implementation of LibLinear [11] with the standard configuration. Using this classifier, the likelihood that a given location l contains a place of a particular type can be estimated based on his description ($\mathit{normalized}(V_l^{F,T})$).

In [23], we evaluated this approach by dividing the dataset of places in a test set and training set. As indicated in the introduction, applying this approach for 14 different types led to rankings with a mean precision at 50 (mean P@50) of 84.9%, mean P@100 of 82.2% and a mean P@500 of 66.3%.

Detecting places in London To further evaluate our approach, it is important to determine if our method is able to find places which are not yet included in existing databases. To this end, we use a grid overlay which divides London in cells of 30m by 30m and consider the centers of the obtained cells as locations which potentionally contain a place of interest. To ensure a fair evaluation, places of London in the training data were removed and we manually evaluated the correctness of the newly discovered places.

Table 4: Top 10 of the detected places which are not yet included in our dataset. Given a particular type, places of that type which can not been found in Google Place and Foursquare are indicated with [Go] and [Fo], respectively[4]. Finally, errors are indicated in italic.

type	1st place	2nd place	3rd place	4th place	5th place
Place of Worship	Imam Khoei Islamic Centre London[Fo]	Baitul Aziz Islamic Cultural Place	Vietnamese Chaplaincy	Westhill Baptist Church[Fo]	London Sri Murugan Temple
School	Ivydale Primary School	Brunswick Park Primary School[Fo]	Lyndhurst Primary School	St Paul's School	Hornsby House School[Fo]
Shop	Tesco Brent Cross	Asda Leyton	*lidl*	Surrey Quays Shopping Centre	Tesco Morning Lane
Restaurant	McDonalds	Pizza Express	Ganapati	*mexican dish*	Beluga
Graveyard	City Of London Cemetery	Camberwell New Cemetery	Nunhead Cemetery	Tower Hamlets Cemetery Park	Abney Park Cemetery
Hotel	Cranbrook Hotel	Novotel Paddington[Go]	Serviced Apartments	Gallions Hotel[Go,Fo]	Premier London Hyde Park
Pub	The Canal Cafe	The Telegraph	The Boathouse	The Crabtree	Bar Bastille
Station	King's Cross	Euston	Clapham Junction	Willesden Junction	Hackney Downs[Go]
Hospital	Whittington Hospital	Dulwich Community Hospital	The Lister Hospital	King's College Hospital	Lewisham Hospital
Monument	Hackney Wick Great War casualties[Go,Fo]	Henry Grey blueplaque[Go,Fo]	New West End Synagogue War Memorial[Go,Fo]	Sir Nigel Playfair blueplaque[Go,Fo]	Sir Stafford Cripps blueplaque[Go,Fo]
Library	The British Library (Euston Road)	Idea Store Chrisp Street	British Library book store (Armstrong Road)	Nunhead Library	British Library book store (Micawber Street)
Museum	Natural History Museum	*geek science museum*	Science Museum	Victoria and Albert Museum	Imperial War Museum
Castle	Eltham Palace	Severndroog Castle[Go]	Vanbrugh Castle[Go,Fo]	*elephant castle*	*flower foxglove*

type	6th place	7th place	8th place	9th place	10th place
Place of Worship	St Dunstan and All Saints	St James's Church	The Temple Church	St Marys Roman Catholic Church[Fo]	St Stephen Walbrook Church
School	Lauriston Primary School[Fo]	Henwick Primary School	Evelyn Grace Academy[Fo]	Frank Barnes School (Harley Road)[Go]	Donnington Primary School
Shop	Asda Bugsby's Way	Aldi	gasstation Morrisons	Superblooms	Sainsbury's
Restaurant	Yo Sushi	Dinner By Heston	McDonalds	McDonalds	McDonalds
Graveyard	St Mary's Catholic Cemetery	Brompton Cemetery	Brockley Cemetery	Highgate Cemetery	Wandsworth Cemetery
Hotel	Novotel London West	Crowne Plaza Docklands	*exhibition hotel*	Georgian House	B&B Belgravia
Pub	Boogaloo	*thewoodsman*	The Greyhound	The Railway (Wells Terrace)[Go]	The Hospital Tavern[Fo]
Station	Poplar	Gospel Oak	Paddington	New Cross	*railways signalbox*
Hospital	*mrfox hospital*	*hbm hospital*	*surgery happyparents*	*appointment hospital*	*greenwichdistricthospital*
Monument	George Goodwin blueplaque[Go,Fo]	Vladimir Lenin blueplaque[Go,Fo]	War memorial City and Midland Bank[Go,Fo]	George Moore blueplaque[Go,Fo]	Memorial Bermondsey and Rotherhithe[Go,Fo]
Library	Battersea Library	Barons Court Library[Fo]	Barking Library[Go]	Durning Library	*lewishamarthouse library*
Museum	Geffrye Museum	*signs museum*	British Museum	Royal Artillery Museum	Blythe House museum store
Castle	*scene castle*	*house castle*	*castle trees*	*langleycastle*	*big castle*

To obtain a first indication of the performance of our method, we determined whether our approach is able to detect familiar places in London. In particular, we determined the most likely locations in London to contain a place of a given type, according to our method. In this way, we were able to find the most popular places such as the Stratford station, the Whittington hospital, the British library and the Natural History Museum.

Closer examination of the detected places revealed a number of misleading tweets and Flickr tags. For example, the Flickr photo taken of a furniture shop with a misleading description such as 'we can now make and install complete libraries' may hint towards a library instead of a shop. In addition, tweets such as 'Science museum today #geek' are not always related to a place nearby the user. Furthermore, old photos may indicate the previous type of a place, for example for places which are converted to an other type (e.g. from pub to restaurant) recently.

Next, we determine if our method can be used to detect places which are not yet included in existing databases. In order to find such places, when places of type t are detected we filter out locations which have already a place of type t in LinkedGeoData or Geonames. In particular, locations that have a place of type t located closer than 4σ in the dataset described in Section 2.1 are removed.

Table 4 shows the top 10 of the resulting rankings, and indicates which places can not been found in Google Places (Go) and Foursquare (Fo) when a user searches for places of a particular type[4]. The place names mentioned in the table are manually determined, as detecting place names is out of the scope of this paper. For each of the discovered places, we manually assessed whether they were of the correct type. The erroneously detected places are indicated with tags in italic. To eliminate duplicates, locations are filtered out if a higher ranked location is located closer than 4σ (see Table 3).

Our method is able to find places of worship, schools, shops, restaurants, graveyards, hotels, pubs, stations, libraries, museums and monuments that are not yet included in our dataset which was constructed by combining LinkedGeoData and Geonames. Our method was not able to find new airports because the observed region in London contains only one airport which was already included in our dataset. Table 4 shows that our method is also able to find places that are not in Google Places and Foursquare. As shown in Table 4, places not present in Google Places are e.g. schools (e.g. Frank Barnes School on the Harley Road), hotels (e.g. Novotel Paddington), pubs (e.g. The Railway at Wells Terrace), libraries (e.g. Barking Library) and castles (e.g. Severndroog Castle). Additionally, our method is able to extend Foursquare with places of worship (Imam Khoei Islamic Centre, West Hill Baptist Church and the St. Dunstan and All Saints Church), schools (Brunswick Park Primary School, Hornsby House School, Lauriston Primary School and Evelyn Grace Academy), pubs (The Hospital Tavern) and libraries (e.g. Barons Court Library). Finally, some places such as the Gallions Hotel and the Vanbrugh Castle were retrieved which are neither included in Foursquare nor Google Places. It is remarkable that among the top 10 discovered monuments which are not yet included in LinkedGeoData and Geonames, there are none which were already included included in Foursquare and Google Places. We note that one war memorial was found in the New West End Synagogue. This indicates that one place of interest can contain another place of interest, for example, a monument in a place of worship, a restaurant in an airport, a shop in a hospital. Most gazetteers only collect the main places of interest. However, finer granularity can be useful and social media may be a good source to enrich existing place databases in this way.

Conclusions We can conclude that social media can be used to find places of several types which are not yet included in LinkedGeoData, Geonames, Google Places and Foursquare, and that some places (e.g. synagogues) may contain other places of interest (e.g. monuments). However, there are some challenges with using social media for detecting places. For example, tweets and Flickr tags may not be related with the place nearby the location of the user (e.g. a picture taken from the tower bridge at the tower of london), misleading (e.g. a photo of a drink with friends at a user's home) or out-of-date (e.g. a picture of a shop which has been replaced by a pub).

[4]Databases accessed on July 4, 2012

Table 5: Most informative Flickr features for each place type.

type	1st feature	2nd feature	3rd feature	4th feature	5th feature
Place of Worship	church	cathedral	mosque	catedral	stainedglass
School	school	university	campus	highschool	college
Shop	market	shopping	christmas	shop	mall
Restaurant	restaurant	food	bean	dinner	pizza
Graveyard	cemetery	graveyard	grave	headstone	cemetary
Hotel	hotel	casino	beach	ponte	hotels
Pub	pub	pubs	bar	beer	publichouse
Station	train	station	railway	subway	metro
Hospital	hospital	newborn	birth	baby	medical
Monument	monument	memorial	statue	parliament	obelisk
Airport	airport	cessna	airplane	aviation	aircraft
Library	library	libraries	publiclibrary	librariesandlibrarians	biblioteca
Museum	museum	museo	dinosaur	aquarium	museums
Castle	castle	castello	castillo	burg	schloss

Table 6: Most informative Twitter features for each place type.

type	1st feature	2nd feature	3rd feature	4th feature	5th feature
Place of Worship	church	jobs	wanna	get	misa
School	school	class	nigga	ass	teacher
Shop	shopping	comprando	store	condes	apple
Restaurant	dinner	lunch	food	burger	breakfast
Graveyard	cemetery	erhaltende	decompsable	exhaustedd	schwester
Hotel	hotel	beach	room	pool	conference
Pub	pub	pint	bar	beer	coffee
Station	sta	train	station	tspot	metro
Hospital	hospital	surgery	hospitals	costanera	clénica
Monument	monument	monumen	monumento	obelisco	fadas
Airport	pirep	filed	airport	aeroporto	flight
Library	library	biblioteca	providence	trekanten	liberry
Museum	museum	exhibit	museo	art	monumenta
Castle	castle	estresas	prinsenzaal	artera	poslovnoj

5. DISCUSSION

In the previous section, we demonstrated how social media can be used to detect places which are not yet included in existing databases. In this section, we discuss other ways to use social media to improve such databases.

Classifying places and data cleaning Databases of places are often used to search for nearby places of a given type. In particular, databases such as Google Places and Foursquare are constructed for this purpose. However, existing categorization of the places can be too broad ('building' instead of 'museum'), outdated (a shop converted to a pub), wrong or even completely absent. This may lead to sub-optimal performance of the applications that use these databases. For example, about 53% of the places from London[5] in Google Places are not classified.

To automatically estimate the semantic type of places, some authors have suggested to classify places based on their name [15], the content of webpages which contain the place name [3] or both [16]. However, only a few broad place types were considered in these works and the proposed methods have only been tested on small test sets ranging from 59 to 1160 places. Furthermore, the names of the places may be misspelled or may be absent in which case the aforementioned approaches fail or need further fine-tuning. In such cases, social media can be used to improve the performance of existing classification methods. In particular, places can be classified based on the Flickr tags and Twitter terms in their vicinity using a similar methodology as described in Section 4. We note that for our method, only the coordinates of the place are needed. In this way, we detected for example that the unclassified place Evelyn Grace from the Foursquare database has 'school' as semantic type.

[5]collected in November, 2011

In addition to absent, too broad or outdated categorization of places, the place types may also be wrong. This is especially a problem for datasets which are constructed by volunteers with little moderation (e.g. Foursquare). Our method can be used to rank the places of a given type in the dataset based on the probability they really belong to that type. This should make it easier to detect errors in existing databases.

Boundary estimation Most databases of places only contain one coordinate to describe the location of places. However, in many contexts, it would be useful to have some knowledge about the shape and size of a place, especially for spatially extended place types such as parks, graveyards, and schools. Furthermore, a better estimation of the spatial extent of places can be used to more accurately predict if a user is at a given place.

To this end, a similar approach as described in Section 4 can be used to rank locations of a city based on the likelihood that they are associated with a place of a given type. In this way, nearby locations classified as the same place type can be clustered to determine the size of the places. Based on this idea, the boundaries of the Camberwell New Cemetery and the Nunhead Cemetery can roughly be estimated. However, it is important to remark that this method only works for very popular places with a lot of geotagged social media, distributed all over the surface of the place of interest. For example, this method is not able to determine the boundaries of an airport because only the terminal of an airport is publicly available.

Semantic characterization of place types It is important to add semantics to place types for better interaction with gazetteers [10]. For example, when a user wants to use public transport, an application can recommend bus, train and metro stations, because they are all subtypes of place type 'station'. Furthermore, when a user really likes to go to pubs, also clubs, restaurants and even cinemas can be recommended because these types are semantically related. To determine the most informative terms for each place type we used χ^2 feature selection based on the description of the places in our dataset. The most informative Flickr tags and Twitter terms for each place type are shown in Table 5 and Table 6, where we filtered out words which do not correspond to nouns, verbs or adjectives. We can observe more informative features for Flickr than for Twitter, especially for graveyards. This corresponds to the results in [23], in which we determined that Flickr on its own is a better source to detect places than Twitter on its own.

These features of Table 5 and Table 6 can be used to discover synonyms and translations, for instance 'castle', 'castello', 'castillo', 'burg' and 'schloss'. In addition, subtypes can be found such as 'church', 'cathedral' and 'mosque' for type 'place of worship'. This information may be used to enrich existing ontologies of place types. Furthermore, affordances associated with place types can be estimated to improve results of affordance based queries such as 'i want to have lunch in London' [2, 25]. For this approach, it is important to establish methodologies which can automatically determine if a tag indicates a subtype, affordance, synonyms or something else.

Finally, given the tags associated to each place type, similarities between place types can be estimated to enrich place type ontologies [15]. For example, using the Jensen-Shannon divergence (JSD) between the tag probabilities of two place types, we can determine that restaurants are semantically most related to pubs (JSD of 255 748), and that restaurants are more related to shops (JSD of 801 100) than to museums (JSD of 941 575).

Tag recommendations When a user visits a place, she may want to publish a tweet or a Flickr photo with tags about that place. In such a context, tag recommendation can help users to find meaningful tags [22]. For example, the most informative of the type of the place a user is visiting can be recommended, see Table 5 and 6.

6. CONCLUSIONS

In this paper, we showed how social media can be used to improve existing databases of places. Using places from LinkedGeoData and Geonames as training data, our method is able to select locations which are likely to contain places of a given type, where candidate locations are chosen as grid cells of 30m by 30m. In this paper, we have presented a detailed case study on London of our method's performance. In this way, we were able to detect places which were not yet included in LinkedGeoData, Geonames, Google Places and Foursquare. Second, we discussed the idea that similar techniques can be used to identify potentional error in existing databases, to estimate boundaries of places, to determine subtypes, synonyms and affordances of places type, to discover semantic relationships between place types, and to recommend tags to user when they publish a tweet or Flickr photo.

7. REFERENCES

[1] 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*, pages 1–10, 2007.

[2] A. Alazzawi, A. Abdelmoty, and C. Jones. An ontology of place and service types to facilitate place-affordance geographic information retrieval. In *Proceedings of the 6th Workshop on Geographic Information Retrieval*, pages 1–2, 2010.

[3] A. Alves, B. Antunes, F. C. Pereira, and C. Bento. Semantic enrichment of places: Ontology learning from web. *International Journal of Knowledge Based Intelligent Engineering Systems*, 13:19 – 30, 2009.

[4] 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*, pages 731–746, 2009.

[5] L. Cao, J. Luo, A. Gallagher, X. Jin, J. Han, and T. S. Huang. A worldwide tourism recommendation system based on geotagged web photo. In *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing*, pages 2274–2277, 2010.

[6] M. D. Choudhury, M. Feldman, S. Amer-Yahia, N. Golbandi, R. Lempel, and C. Yu. Automatic construction of travel itineraries using social breadcrumbs. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, pages 35–44, 2010.

[7] M. Clements, P. Serdyukov, and A. de Vries. Using Flickr geotags to predict user travel behaviour. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 851–852, 2010.

[8] 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*, page 761, 2009.

[9] 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*, pages 1121–1122, 2010.

[10] K. Janowicz and C. Keß ler. The role of ontology in improving gazetteer interaction. *International Journal of Geographical Information Science*, 22(10):1129–1157, 2008.

[11] 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*, pages 408–416, 2008.

[12] R. Lee, S. Wakamiya, and K. Sumiya. Discovery of unusual regional social activities using geo-tagged microblogs. *World Wide Web*, 14(4):321–349, 2011.

[13] J. Lindqvist, J. Cranshaw, J. Wiese, and J. Hong. I'm the mayor of my house: Examining why people use Foursquare - a social-driven location sharing application. In *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems*, 2011.

[14] V. Murdock. Your mileage may vary: on the limits of social media. *SIGSPATIAL Special*, 3(2):62–66, 2011.

[15] 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.

[16] 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*, pages 85–93, 2008.

[17] A. Popescu, G. Grefenstette, and H. Bouamor. Mining a multilingual geographical gazetteer from the web. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, pages 58–65, 2009.

[18] A.-M. Popescu and M. Pennacchiotti. Detecting controversial events from Twitter. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, page 1873, 2010.

[19] T. Rattenbury, N. Good, and M. Naaman. Towards automatic extraction of event and place semantics from Flickr tags. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 103–110, 2007.

[20] T. Sakaki. Earthquake shakes Twitter users : real-time event detection by social sensors. In *Proceedings of the 19th International Conference on*

World Wide Web, pages 851–860, 2010.

[21] P. Serdyukov, V. Murdock, and R. van Zwol. Placing Flickr photos on a map. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 484–491, 2009.

[22] B. Sigurbjörnsson. Flickr Tag Recommendation based on Collective Knowledge. In *Proceedings of the 17th International Conference on World Wide Web*, pages 327–336, 2008.

[23] S. Van Canneyt, S. Schockaert, O. Van Laere, and B. Dhoedt. Detecting Places Of Interest using Social Media. In *Proceedings of the 2012 IEEE/WIC/ACM International Conference on Web Intelligence, accepted*.

[24] S. Van Canneyt, S. Schockaert, O. Van Laere, and B. Dhoedt. Time-Dependent Recommendation of Tourist Attractions using Flickr. In *Proceedings of the 23rd Benelux Conference on Artificial Intelligence*, pages 255–262, 2011.

[25] V. Zheng, Y. Zheng, and X. Xie. Collaborative location and activity recommendations with gps history data. *Proceedings of the 19th International Conference on World Wide Web*, pages 1029–1038, 2010.