Discovering and Characterizing Places of Interest using Flickr and Twitter

Steven Van Canneyt
Ghent University, iMinds, Belgium
Steven Schockaert
Cardiff University, United Kingdom
Bart Dhoedt
Ghent University, iMinds, Belgium

ABSTRACT

Databases of places have become increasingly popular to identify 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 broad coverage, there is a need for techniques that are capable of expanding place databases in an automated way. In this paper we discuss how geographically annotated information obtained from social media can be used to discover new places. In particular, we first determine potential places of interest by clustering the locations where Flickr photos have been taken. The tags from the Flickr photos and the terms of the Twitter messages posted in the vicinity of the obtained candidate places of interest are then used to rank them based on the likelihood that they belong to a given type. For several place types, our methodology finds places that are not yet contained in the databases used by Foursquare, Google, LinkedGeoData and Geonames. Furthermore, our experimental results show that the proposed method can successfully identify errors in existing place databases such as Foursquare.

Keywords: Social Media, Geographic Information Retrieval, Discovering Places Of Interest

INTRODUCTION

Berners-Lee's vision of the Semantic Web (Berners-Lee, 2001) has become increasingly popular in the last few years. The World Wide Web would evolve to a highly interconnected network of data that could be easily accessed and understood by machines. Applications could for instance use the Semantic Web to construct customized answers to a particular question. In such applications the user is no longer required to search for information or pore through results. A question can e.g. be 'What are locations of the restaurants in London?'. To answer this question, a structured dataset has to be available containing places located in London (entities), associated with their location and semantic type (properties).

However, a lot of information on the Web is still unstructured or only semi-structured. Therefore, there is a need for automated methods to extend structured datasets using existing Web data. Several methods of this form have been proposed, e.g. YAGO2 (Hoffart, 2013) and BabelNet (Navigli, 2010) are knowledge bases that are constructed using Wikipedia and Wordnet. Other research focuses on establishing structured datasets containing information of a specific type. For instance, LinkedGeoData (Auer, 2009) is a dataset of places constructed using OpenStreetMap, an application in which users can submit geographical data such as place semantics.

In this paper, we will focus on improving existing databases of places. More precisely, we will add new places and discover likely errors using data from the Web. Social media data is particularly promising in this respect, due to the large amounts of geographically annotated data produced by these media. For example, about 1.5% of all Twitter posts (i.e. tweets) are annotated with geographical coordinates (Murdock, 2011). In addition, there are currently more than 190 million geotagged Flickr photos (Flickr, 2013). This data has been used to e.g. automatically detect events (Rattenbury, 2007; Sakaki, 2010; Lee, 2011), to find popular places (Crandall, 2009; Van Canneyt, 2011) and tourist routes (Choudhury, 2010; Jain, 2010).

The main focus of this paper is on how geographically annotated information obtained from social media can be used to discover new places of a given type such as 'hotel' or 'school' to extend semantic databases of places. Our hypothesis is that the type of a place can be derived from 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 our previous work (Van Canneyt, 2012a), we have provided evidence for the validity of this hypothesis: given the location of various places of interest (POIs), we addressed the task of identifying those POIs that are most likely to be of a particular type. Our main conclusion was that Flickr tags are a rich source of information for deciding on the type of a place. Using Twitter terms further improved the results although this improvement was more limited. We also considered the correlation between the type of the POIs and the types of the places in the vicinity to categories the POIs. However, this additional information led to a minimal improvement of the performance of our methodology, and in this paper we are mainly interested in the use of social media by itself to improve databases of places. Therefore, we do not consider such correlations here. In (Van Canneyt, 2012b) we considered the more challenging problem of finding locations where places of particular types can be found, without providing a list of candidate locations. Instead, we used a simple grid overlay to find candidate locations and compared the results against existing databases of places. This qualitative analysis demonstrated the potential of the proposed method to find POIs in London that are not yet contained in

Foursquare, Google Places, Geonames and LinkedGeoData. Encouraged by these initial results, we improve the proposed methodology in this paper and present a more detailed experimental evaluation. First, the Support Vector Machine classifier used in (Van Canneyt, 2012a; Van Canneyt, 2012b) is replaced by a language modeling approach, which improves the results significantly. Second, we analyze the behavior of different feature selection techniques. We conclude that for the Flickr data correlation coefficient feature selection (Ng, 1997) performs significantly better than χ^2 feature selection. The performance of the proposed methodology can be further improved when names of cities and countries are removed from the considered features. Finally, we perform a large-scale evaluation on 88 different cities, where we examine the results for London in more detail. Based on this evaluation, we can conclude that our approach is able to extend and validate data sets of places. In particular, our method is able to detect new places of a particular type, even when the locations of places of interest are not given. Furthermore, our experimental results show that the proposed method can also be used to successfully identify errors in existing place databases such as Foursquare.

The remainder of this paper is structured as follows. We start with a review of related work. The subsequent section explains how training and test data have been collected. Thereafter, we describe our methodology of discovering places of a given type. This section is followed by an experimental evaluation. Finally, we conclude our work in the last section.

RELATED WORK

To fill in the gap between the unstructured and semi-structured data from the Web and the structured data in the Semantic Web, a number of methodologies have been proposed. Kwok (2001) and Etzioni (2005), for instance, extracted named entities from unstructured web pages using natural language processing. Other research (Hoffart, 2013; Navigli, 2010) used semi-structured data available in Wikipedia and Wordnet to construct a structured dataset in an automatic way. The used semi-structured data available in Wikipedia have been improved by applying information extraction techniques on the main text of the corresponding Wikipedia article (Wu, 2007; Wu, 2008). To further construct structured data, social media have been used due to the large amount of data available. Social media are, on the one hand, used to derive ontologies which describe relations between words (Schmitz, 2006; Markinez, 2009). This information can for instance be used to improve search results: Given a word as query, similar words can be used to extend the results. Schmitz (2006) detected subsumption relations between Flickr tags using the co-occurrence of the tags. The methodology applied in (Markinez, 2009) measures the similarity of tags used in the social bookmarking system BibSonomy using several statistical measures such as cosine similarity, Jaccard similarity and the mutual information metric. On the other hand, social media can be used to extract entities and their semantics. Sakaki (2010) for example constructed a probabilistic model to detect the location and time of earthquake and typhoon occurrences using Twitter. The researchers in (Lee, 2011) described a method which discovered events by detecting unusual regional activities in Twitter. Other research (Choudhury, 2010; Jain, 2010) developed methodologies which automatically constructs travel itineraries using Flickr.

In this paper, we are focusing on automatic detection of place locations and semantics using social media. This data can e.g. be used for personalized place recommendations. Ozdikis (2011) for instance developed an application which recommends places similar to a user defined place. Initial work on extending semantic datasets of places by discovering points of interest (POIs) from social media has been exclusively based on analyzing the coordinates of geotagged data.

For instance, Crandall et al. (2009) used the mean shift method to cluster the locations of geotagged Flickr photos to detect POIs. This method has among others been applied in (Cao, 2010; Clements, 2010; Van Canneyt, 2011) to detect and recommend popular tourist places in cities. In this paper, mean shift clustering is used as the first step of the proposed methodology to detect candidate locations of places of a given type. The Antourage system (Jain, 2010) on the other hand uses a hexagonal grid overlay over a city map, and associates with each hexagon a weight based on the number of Flickr photos that have been taken within the boundaries of that cell. Given such a weighted grid, the max-min ant system meta-heuristic (Stutzle, 2000) is used to find distance constrained trips in a city covering as much as possible popular POIs. These contributions focus on using locations of geotagged photos to detect POIs. In particular, in the aforementioned works, no attempt is made to associate semantic information with places. In contrast, in this paper we aim to discover places of a given semantic type, and we do not restrict ourselves to tourist places, by also considering e.g. schools, graveyards and libraries.

A second line of research relevant for our work analyzes text originating from social media, in order to discover places and to retrieve semantic information on these places. Rattenbury et al. (2007) used multiscale burst analysis to detect place-related Flickr tags. This technique was applied in (Ahern, 2007) to detect names for arbitrary areas in the world. They clustered the locations where Flickr photos were taken using k-means clustering. For each cluster, representative tags were searched using TF-IDF and the percentage of users in the cluster who have used a given tag. To find landmarks, their names and their most representative photos, Abbasi et al. (2009) proposed a further extension of this approach. To detect landmarks in a city, they first select photos containing the city name. Second, using support vector machines and the tags that have been assigned to the photos, these photos were classified either as being or not being taken of a landmark. As training data, photos were obtained from manually selected photo groups such as ‘landmarks around the world’. Finally, tags and photos which describe landmarks in a given city were extracted based on the obtained photos. This research demonstrates that text from social media can be used to detect POIs and their associated name. However, the semantic type of the obtained places was not determined.

Our work is most closely related to Gazetiki (Popescu, 2008), a gazetteer which was automatically derived from Wikipedia, Panoramio and web search using a four step approach. The method proposed in (Popescu, 2008) first collected Wikipedia articles which contain associated geographic coordinates, and a geographical concept (i.e. a place type from Geonames) in their first sentence. From these articles, links to other Wikipedia articles with a geographical concept in their first sentence were extracted. For each obtained Wikipedia article, a candidate geographical entity was constructed with the name equal to the title of the article and the type extracted from the first sentence of the article. If the article also contains a coordinate, this coordinate was used as coordinate of the geographical entity. Second, named entity recognition was used to find additional geographical names in the titles of Panoramio photos. In addition, a candidate geographical entity was constructed for each obtained geographical name. For the candidate geographical entity obtained using Panoramio, the place type was determined by taking into account the number of search results of queries such as ‘<geographical name> is a <place type>’ on the Alltheweb search engine. Third, the obtained candidate geographical entities without associated coordinates were geotagged by the center of gravity of the coordinates of the Panoramio photos that have been tagged with the geographical name. Finally, the obtained places were ranked using the number of results by searching for the geographical name in Alltheweb and the number of times the place was photographed.

However, to the best of our knowledge, no effort has so far been devoted to discover 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 or is able to detect incorrect data in existing databases of places.

DATA ACQUISITION

Our goal is to determine the locations in a city *C* which correspond to places of a given type *t* (e.g. schools, hospitals, train stations and restaurants), based on the tags of the Flickr photos taken in the city and the terms of the tweets posted in the city. To obtain training and test data, we have collected a set of places with known location and type for several cities. We have subsequently mined Flickr and Twitter to find metadata about these places. We now explain these steps in more detail.

Collecting Bounding Boxes of Cities

The considered cities have been selected by first selecting the names of all cities with a population of more than 15,000 inhabitants using Geonames. For each of the obtained cities, its bounding box has been determined using Yahoo! PlaceFinder. When the two bounding boxes of two different cities overlap, only the city with the largest bounding box is kept to ensure that there is no overlap in the bounding boxes of the cities in the training, test and development set. After identifying the bounding boxes of the cities, only cities where more than 1000 Flickr photos were taken and more than 1000 tweets were posted were retained. As a result of this process, we collected bounding boxes of 530 cities, whose locations are plotted in Figure 1. A more detailed plot of the locations of the obtained cities in Europe and the USA is shown in Figure 2 and 3, respectively. In these figures, the radius of the circles is proportional to the number of Flickr photos and tweets posted in the city.

Finally, the dataset has been split in three parts: two thirds of the cities were used as training data (called the training set, $S_{training}$), while one sixth of the cities were used to find optimal values of the parameters in our method (called the development set, S_{dev}). The remaining sixth was used for evaluation (called the test set, S_{test}). This was done by ranking the cities in descending order based on the number of tweets and Flickr photos taken in the city. The cities ranked 1^{st}, 7^{th}, 13^{th}… were considered as development set, the cities ranked 2^{nd}, 8^{th}, 14^{th}… as test set, and the remaining cities as training set. In this way we obtain three sets that contain cities with all varieties of number of Flickr photos and tweets.

Figure 1. Plot of the cities in our dataset.

Figure 2. Plot of the considered cities in Europe, with the radius of the circles proportional to the number of Flickr photos and tweets posted in the city.

Figure 3. Plot of the considered cities in the USA, with the radius of the circles proportional to the number of Flickr photos and tweets posted in the city.

Collecting Places of Interest

To obtain locations of known places of different types, we have used two open source databases: LinkedGeoData (LGD) and Geonames. We have in particular collected all places in these databases of the types shown in the first column of Table 1. These are the types with the highest number of instances in the union of the LinkedGeoData and Geonames database.

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

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 (Ozdikis, 2011) 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 of the lengths of both names. As a result of this process, we obtained 1,939,174 distinct places of which 312,478 are located in the considered cities. We define *K* as the set of known places located in the cities of $S_{training}$, which are used to train our model. The places located in S_{dev} and S_{test} are used as ground truth to respectively optimize and evaluate our methodology. An overview of the number of places per type and source can be found in Table 2.

Table 2. Statistics of the used datasets of places.

place type	LGD	Geonames	combined	in considered cities
Shop	326 388	38	316 773	64 124
Restaurant	217 145	1 315	215 613	51 647
School	284 141	241 041	349 157	46 473
Place of Worship	315 532	241 745	356 329	45 227
Pub	133 761	0	132 123	32 829
Hotel	67 563	83 210	136 174	28 567

Station	80 849	58 484	125 556	18 225
Hospital	54 363	24 281	59 599	8 400
Monument	35 110	746	32 322	4 598
Library	22 730	11 549	22 946	4 373
Graveyard	136 655	125 481	139 096	3 524
Museum	18 060	5 000	19 421	3 328
Airport	1 138	24 547	25 591	753
Castle	5 043	3 666	8 474	410
total	1 698 478	821 103	1 939 174	312 478

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 photosharing 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 (Serdyukov, 2009). In addition, photos with invalid coordinates or without tags were removed. The dataset thus obtained contains 23,324,644 geotagged photos of which 9,516,714 are located in the considered cities.

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 ended up with a total number of 30,095,000 tweets of which 8,138,974 are located in the considered cities.

METHODOLOGY

In this section we describe various aspects of our proposed approach to discover places of interest. More precisely, we want to determine the locations in a city *C* which correspond with places of a given type *t*. Therefore, we first cluster the locations where Flickr photos have been taken to obtain the locations which potentially correspond to places of interest (POIs) in city *C*. We then associate with each candidate place of interest a feature vector based on the tags of the Flickr photos that are associated with locations nearby. Afterwards, a query associated with place type *t* is constructed based on the descriptions of known places of type *t*. This query is then used to rank the obtained POIs in *C*, based on the likelihood that they belong to type *t*. Finally, we discuss how Twitter can be used to further improve the results.

In the remainder of this section we describe the steps of our proposed approach in more detail. To further clarify our methodology, we explain in a running example how each step could be applied to a fictional city *E*. This city contains three places of interest: a museum, a church and a monument located inside the church. These places are considered as the ground truth of *E*.

Detecting Places of Interest

In the first step, we determine the locations in city $C \in S_{test}$ that potentially correspond to a place of given type *t*. To this end, we cluster the locations where Flickr photos have been taken using mean shift clustering (Cheng, 1995). Mean shift clustering is a straightfoward iterative procedure that shifts each coordinate to the mean of the coordinates in its neighborhood. This algorithm is particularly suitable for this task, as it is scalable, does not require a predefined number of clusters, and allows us to adapt the scale at which clusters should be identified. Moreover, mean shift clustering has already been successfully applied to detect POIs from Flickr photos (Crandall, 2009).

Let L be the set of coordinates where Flickr photos have been taken in city *C*. The mean shift $m_b(l)$ of coordinate $l \in L$ is then given by the difference of *l* and the weighted mean of the coordinates nearby l:

$$m_b(l) = \frac{\sum_{l' \in L \wedge d(l,l') \leq 2b} G_b(l,l') \cdot l'}{\sum_{l' \in L \wedge d(l,l') \leq 2b} G_b(l,l')} - l \qquad (1)$$

with *b* the bandwidth parameter, $d(l,l')$ the geodesic distance in meters between coordinate l and l', and $G_l(l,l')$ the kernel function which determines the weight associated with coordinate l' depending on its distance to *l*. We use a Gaussian kernel for a smooth density estimation:

$$G_b(l,l') = e^{-\frac{d(l,l')^2}{2b^2}} \qquad (2)$$

The mean shift procedure then computes a sequence starting from all initial coordinates $l_1 \in L$ where

$$l_{i+1} = l_i + m_b(l_i) \qquad (3)$$

which converges to a location that corresponds to a local maximum of the underlying distribution as $m_b(l_i)$ approaches zero. Based on the obtained clusters, we consider the center of each cluster as a candidate points of interests, called set U^F.

In our running example, two candidate points of interests corresponding to the center of the museum (poi_1) and the center of the church (poi_2) may be detected. This is formally noted by $U_E^F = \{poi_1, poi_2\}$, where poi_i is represented by a latitude and a longitude value.

Describing Places of Interest

We associate a feature vector V_{poi} to each candidate point of interest $poi \in U^F$ based on the tags of the Flickr photos that are associated with locations nearby *poi*. Let D be the dictionary containing all the tags of the Flickr photos in our dataset, the vector contains a component associated with each word $w \in D$. Formally, for feature vector V_{poi} of candidate point of interest $poi \in U^F$, the component $c_{poi,w}$ associated with word $w \in D$ 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 *poi* is more than $2\sigma_U$ are not considered:

$$c_{poi,w} = \sum_{f \in F_w \wedge d(poi,f) \leq 2\sigma_U} e^{-\frac{d(poi,f)^2}{2\sigma_U^2}} \quad (4)$$

with *f* a Flickr photo, F_w the set of Flickr photos that contain tag *w*, σ_U the deviation value used for the description of the candidate POIs in set U^F and *d*(*poi*, *f*) the geodesic distance in meters between *poi* and the coordinates of the photo *f*.

The candidate points of interests poi_1 and poi_2 in the fictional city *E* have associated feature vectors V_{poi_1} and V_{poi_2}, respectively. Assume for instance that $V_{poi_1} = (5.99, 3.81, 0.76, 0, 0, 0)$ and $V_{poi_2} = (0, 0, 0, 7.87, 6.74, 6.63)$ where the six components of these vectors respectively refer to 'museum', 'art', 'bike', 'church', 'statue' and 'monument'.

Constructing a Query

To rank the candidate POIs based on the likelihood that they are associated with the given type *t*, we first construct an associated query q_t. Let K_t be the set of all known places of type *t* located in the cities of the training set $S_{training}$ and D_t the dictionary of all words which are indicative for place type *t*. A query q_t of type *t* is represented as a vector with one component $q_{t,w}$ associated with each word $w \in D_t$ given by

$$q_{t,w} = \sum_{p' \in K_t} c_{p',w} \quad (5)$$

where $c_{p,w}$ similar defined as in (4):

$$c_{p,w} = \sum_{f \in F_w \wedge d(poi,f) \leq 2\sigma_K} e^{-\frac{d(poi,f)^2}{2\sigma_K^2}} \quad (6)$$

with *f* a Flickr photo, F_w the set of Flickr photos that contain tag *w*, σ_K the deviation value used for the descriptions of the places in the training set *K* and *d*(*poi*, *f*) the geodesic distance in meters between *poi* and the coordinates of the photo *f*.

Starting from dictionary *D* containing all the tags of the Flickr photos in our dataset, dictionary D_t is defined as a subset of all the words that are likely to be indicative for type *t*. To identify such words, feature selection techniques can be used. We discuss in this paper chi-square (χ^2) and correlation coefficient (*CC*) based feature selection. The dictionary D_t is then obtained by taking the *m* tags with the highest χ^2, respectively CC, value.
Chi-square based feature selection has been successfully applied in other research (e.g. Van Laere, 2011) and is defined as

$$\chi^2 = \frac{N \times (O_{w,t} \cdot O_{\neg w,\neg t} - O_{\neg w,t} \cdot O_{w,\neg t})^2}{(O_{w,t} + O_{\neg w,t}) \times (O_{w,\neg t} + O_{\neg w,\neg t}) \times (O_{w,t} + O_{w,\neg t}) \times (O_{\neg w,t} + O_{\neg w,\neg t})} \quad (7)$$

where the values are defined as

$$O_{w,t} = \sum_{p' \in K_t} c_{p',w} \quad (8)$$

with $c_{p',w}$ as defined in (4),

$$O_{w,\neq t} = \sum_{p' \in K \setminus K_t} c_{p',w} \quad (9)$$

in which *K* is the set of all known places located in cities of $S_{training}$,

$$O_{\neq w,t} = \sum_{w' \in D \setminus \{w\}} \sum_{p' \in K_t} c_{p',w'} \quad (10)$$

with *D* the dictionary of all the tags of the Flickr photos in our dataset,

$$O_{\neq w,\neq t} = \sum_{w' \in D \setminus \{w\}} \sum_{p' \in K \setminus K_t} c_{p',w'} \quad (11)$$

and

$$N = \sum_{w' \in D} \sum_{p' \in K} c_{p',w'} \quad (12)$$

Value $O_{w,t}$ is the number of occurrences of word *w* in the descriptions of places of type *t*, $O_{w,\neq t}$ the number of occurrences of *w* in the descriptions of places of another type than *t*, $O_{\neq w,t}$ the number of occurrences of all words $w' \in D \setminus \{w\}$ in the descriptions of places of type *t*, $O_{\neq w,\neq t}$ the number of occurrences of all words $w' \in D \setminus \{w\}$ in the descriptions of places of a different type than *t*, and *N* the total number of occurrences of all words $w' \in D$ in the descriptions of all places in K. The correlation coefficient *CC*, introduced in (Ng, 1997), is a variant of the more popular χ^2 feature selection metric, where $CC^2 = \chi^2$:

$$CC(w,t) = \frac{\sqrt{N} \times (O_{w,t} \cdot O_{\neq w,\neq t} - O_{\neq w,t} \cdot O_{w,\neq t})}{\sqrt{(O_{w,t} + O_{\neq w,t}) \times (O_{w,\neq t} + O_{\neq w,\neq t}) \times (O_{w,t} + O_{w,\neq t}) \times (O_{\neq w,t} + O_{\neq w,\neq t})}} \quad (13)$$

CC can be viewed as a "one sided" χ^2 metric. The correlation coefficient *CC* selects the words that are highly indicative of membership in a category, whereas the χ^2 metric will also pick out words that are indicative of non-membership in the category. In the evaluation section, we compare the results of our methodology using the *CC* and χ^2 metric in more detail.

As a final optimization, we exclude from D_t the names of the cities in the training set and the names of the countries in which these cities are located. Lists of alternative names of the cities and their corresponding countries were obtained using Geonames. The rationale behind filtering the names of the cities and countries is as follows: A lot of names of cities from the training set and their corresponding countries have high *CC* and χ^2 values because some cities have a disproportional number of places of particular types. For example, 5% of the stations in the training set are located in Tokyo leading to a high *CC* and χ^2 value for the word 'tokyo' when type *t* is equal to 'station'. This may result in a false positive observation of a station when the word 'tokyo' is used in other cities. The impact of introducing this additional filter step is described in more detail in the evaluation section. In future work, word sense disambiguation and relatedness measures will be considered to cluster tags by meaning (Gracia, 2009).

In our running example, we consider three types of places, i.e. museums, places of worship and monuments. The query $q_{\text{'museum'}}$ associated with type 'museum' contains weighted components $q_{\text{'museum','museum'}} = 103.92$ and $q_{\text{'museum','art'}} = 78.75$. We note that $D_{\text{'museum'}}$ for instance could initially contain the word 'paris' which is eliminated in the final optimization step in the query constructing phase. In addition, $q_{\text{'placeofworship'}}$ contains components $q_{\text{'placeofworship','cathedral'}} = 50.81$ and $q_{\text{'placeofworship','church'}} = 46.80$; and $q_{\text{'monument'}}$ the components $q_{\text{'monument','monument'}} = 80.97$ and $q_{\text{'monument','statue'}} = 78.94$. For clarity of the example, only non-zero $q_{t,w}$ values are mentioned.

Ranking Places of Interest

Using the locations and descriptions of the candidate points of interest U^F in city *C* and a query q_t associated with place type *t*, we rank the points of interests based on the likelihood that they belong to type t using a language modeling approach. Other classification methods may be used, e.g. methods based on k-nearest neighbors or decision trees. However, preliminary experiments have shown that the use of language models outperforms the other methods. The probability $\mathrm{P}[poi \mid q_t]$ that $poi \in U^F$ belongs to type *t* is estimated as

$$\mathrm{P}[poi \mid q_t] \propto \prod_{w \in D_t} \mathrm{P}[w \mid poi]^{q_{t,w}} \qquad (14)$$

where $q_{t,w}$ is the weighted number of occurrences of word *w* in query q_t as defined in (5). We estimate $\mathrm{P}[w \mid poi]$ using Jelinek-Mercer smoothing as

$$\mathrm{P}[w \mid poi] = \lambda \cdot \frac{c_{poi,w}}{\sum_{w' \in D} c_{poi,w'}} + (1-\lambda) \cdot \mathrm{P}[w \mid K] \qquad (15)$$

with $\lambda \in [0,1]$ and the background model $\mathrm{P}[w \mid K]$ is estimated using maximum likelihood:

$$\mathrm{P}[w \mid K] = \frac{\sum_{poi' \in K} c_{poi',w}}{\sum_{poi' \in K} \sum_{w' \in D} c_{poi',w'}} \qquad (16)$$

As the value of $\mathrm{P}[poi \mid q_t]$ may be very small, the values are calculated in log-space to avoid significant loss of precision and underflow:

$$\mathrm{P}[poi \mid q_t] \propto \log \prod_{w \in D_t} \mathrm{P}[w \mid poi]^{q_{t,w}} = \sum_{w \in D_t} q_{t,w} \cdot \log \mathrm{P}[w \mid poi] \qquad (17)$$

We denote the right-hand side of (17) as $score(poi \mid t)$:

$$score(poi \mid t) = \sum_{w \in D_t} q_{t,w} \cdot \log \mathrm{P}[w \mid poi] \qquad (18)$$

Finally, the candidate points of interest from set U^F are ranked based on their $score(poi \mid t)$ value, in descending order.

For each considered place type in the running example (i.e. museum, place of worship and monument) we rank the candidate POIs in U_E^F according to the likelihood that they belong to the

given type. For museum, we get a $score(poi_1 | \text{'museum'})$ of -139 and a $score(poi_2 | \text{'museum'})$ of $-\infty$ when we set λ equal to 1. This leads to a list where poi_1 is ranked above poi_2. Note that a score of $-\infty$ indicates a likelihood of 0. In a similar way, for both the 'place of worship' and the 'monument' place type, poi_2 is ranked above poi_1. Note that when a candidate point of interest corresponds to several places of different types, it can be ranked first for different types.

Improving Results using Twitter

In the same way as for the Flickr data, we can obtain a ranked list of POIs only using the Twitter data. First, the locations where the tweets have been posted are clustered to find locations of candidate POIs. We will refer to this clustering as U^T. Second, these candidate POIs are ranked based on the terms of the Twitter posts that are associated with locations nearby. This is performed in a similar way as described in the previous sections, where the Flickr data is replaced by the Twitter data.

We can also use the Flickr and Twitter data together to improve the results. To this end, we again use the clustering U^F, which is only based on the Flickr data. We have also tested other clustering approaches to detect locations of candidate POIs. In one approach, the candidate POIs set obtained using Twitter (U^T) was used. In a second approach, we clustered both the locations where Flickr photos have been taken and tweets have been posted, called set $U^{F \cup T}$. Finally, the sets U^F and U^T have been combined to $U^F \cup U^T$ in the last approach. Experiments have shown that these alternatives to yield worse results, which is why we do not consider them in the remainder of this paper.

After the clustering step, we use the Flickr data and Twitter data separately to get two estimates which indicates the likelihood that a $poi \in U^F$ belongs to a given type *t*. More precisely, we first use the Flickr data to describe the POIs in U^F, to construct the queries associated with the place types, and to estimate for each $poi \in U^F$ the likelihood that *poi* belongs to a given type *t*. The log of this likelihood is indicated by $score^F(poi | t)$ as defined in (18). In a similar way, the Twitter data is used to describe the POIs in U^F, to construct the queries, and to estimate for each $poi \in U^F$ the log of the likelihood that *poi* belongs to type *t*, given by $score^T(poi | t)$.

Afterwards, the $score^F(poi | t)$ and $score^T(poi | t)$ are combined to obtain a $score^{F,T}(poi | t)$ value which indicates the log of the likelihood that *poi* belongs to type *t*:

$$score^{F,T}(poi | t) = \eta \cdot score^F(poi | t) + (1-\eta) \cdot score^T(poi | t) \qquad (19)$$

with $\eta \in [1, 0]$.

Finally, the candidate points of interest from set U^F are ranked based on their $score^{F,T}(poi | t)$ value, in descending order.

In the running example, we obtained a $score^F(poi_1 | \text{'museum'})$ value of -139. Recall that this value corresponds to the log of the likelihood that poi_1 belongs to place type 'museum', based on the Flickr data. Using the Twitter data, an additional feature vector $V^T_{poi_1}$ describes poi_1 using the tweets in the vicinity of this POI. The components of this vector are for instance equal to $c^T_{poi_1,\text{'exposition'}} = 9.03$ and $c^T_{poi_1,\text{'flower'}} = 2.68$. The query $q^T_{\text{'museum'}}$ associated with type 'museum' is

constructed using the tweets in the vicinity of known museums and contains weighted components $q^T_{\text{'museum','museum'}} = 149.29$ and $q^T_{\text{'museum','exposition'}} = 132.10$. Using $V^T_{poi_1}$ and $q^T_{\text{'museum'}}$ we get $score^T(poi_1 \mid \text{'museum'}) = -34$. When we set η equal to 0.75, a $score^{F,T}(poi_1 \mid \text{'museum'})$ value of -113 is obtained.

EVALUATION

In this section, we describe how we optimized the parameters using the development set. Subsequently, we use the test set to examine to what extent our methodology is able to discover places which are not yet known by existing databases and to identify errors in existing databases of places.

Parameter Optimization

The task we consider is to discover the locations of possible POIs in a city *C* and to rank them according to the likelihood that they belong to a given type *t*. In this section, we use the development set to optimize the quality of these ranked POIs by determining the impact of different parameter values and feature selection techniques. When optimizing the parameter settings, it is useful to consider only one metric to measure the performance of our methodology. Additionally, the used metric has to summarize the performance of our methodology in one value (e.g. between 0 and 100). In particular, this metric must have an optimal value when the distances between the discovered POIs and the places of type *t* in our ground truth are minimal, and when the POIs which are located very close to a place of type *t* in our ground truth are ranked at the top. To this end, the quality of a ranked list of POIs associated with city *C* and type *t* is measured using the Normalized Discounted Cumulative Gain metric:

$$NDCG(C,t) = \frac{DCG(C,t)}{IDCG(C,t)} \times 100 \qquad (20)$$

with $DCG(C,t)$ the Discounted Cumulative Gain of the ranking

$$DCG(C,t) = rel(C,t)@1 + \sum_{i=2}^{|U^F|} \frac{rel(C,t)@i}{\log_2(i)} \qquad (21)$$

where U^F is the set of all candidate POIs in city *C*, and $rel(C,t)@i$ the relevance of the POI at position *i* in the ranked list, defined as

$$rel(C,t)@i = e^{-\frac{d(poi_i, nn_{i,t})^2}{2h^2}} \qquad (22)$$

with poi_i the POI at position *i* of the ranked lists of POIs for city *C* and type *t*, $nn_{i,t}$ the place of type *t* in the ground truth which is nearest to poi_i, $d(poi_i, nn_{i,t})$ the geodesic distance in meters between poi_i and $nn_{i,t}$, and *h* the deviation value which is set to 40. Furthermore, the Ideal Discounted Cumulative Gain, $IDCG(C,t)$, is defined as the *DCG* value of the optimal ranking, i.e. when the POIs located in *C* are ranked by relevance. Finally, we calculate the mean *NDCG* of all cities in the development set, which is given by

$$MNDCG(t) = \frac{\sum_{C' \in S_{dev}} NDCG(C',t)}{|S_{dev}|} \qquad (23)$$

with S_{dev} the set of all cities in the development set.

Table 3. Optimal parameter values.

place type	b	σ_K	σ_U	η	m *(Flickr)*	m *(Twitter)*	λ *(Flickr)*	λ *(Twitter)*
Shop	5	25	80	0.44	1000	7000	0.75	0.70
Restaurant	5	5	60	0.43	20	300	0.95	0.95
School	5	5	55	0.82	1600	2900	0.85	0.10
Place of Worship	5	10	35	0.10	400	50	0.55	0.95
Pub	5	5	45	0.87	100	5000	0.95	0.95
Hotel	5	5	50	0.23	1400	1200	0.95	0.95
Station	5	15	50	0.49	50	1500	0.85	0.35
Hospital	5	15	100	0.24	4500	100	0.80	0.45
Monument	5	5	40	0.65	400	6000	0.80	0.90
Library	5	45	45	0.99	50	50	0.95	0.80
Graveyard	25	10	75	0.32	1800	20	0.70	0.90
Museum	5	15	45	0.88	1100	6000	0.75	0.75
Airport	25	60	60	0.76	700	20	0.10	0.50
Castle	25	15	70	0.35	3000	100	0.95	0.90

We first optimize for each considered place type the deviation value σ_K from (6) and σ_U from (4) which is used to describe the known places from the training set *K* and the obtained candidate POIs U^F, respectively. The optimal value for each considered place type can be found in the second and third column of Table 3. The most informative words associated with the given place type *t* can be found in the tags of the Flickr photos taken close nearby the places of type *t* in the training set. To detect new POIs on the other hand, also tags of Flickr photos taken further away from the POI may be useful to determine its place type. For example, Flickr photos may have been taken at some distance of the actual POI. This observation leads to a smaller deviation value σ_K for the descriptions of the places in the training set than the deviation value σ_U for the description of the obtained candidate POIs.

Using these σ values, we compare the feature techniques described above, i.e. χ^2, the correlation coefficient (*CC*) and *CC* after filtering city and country names (*CC+filter*). For these experiments, we use for each place type their optimal σ values, a bandwidth value *b* (see Equation 2) of 5, and a λ value (see Equation 15) of 0.9. Tables 4 and 5 show for each described feature selection technique the optimal number of features *m* and their corresponding *MNDCG* value. We indicate that for some place types such as libraries, restaurants and places of worship the informative terms are located at the very top of the ranked features leading to an optimal number of features around 150. For other types, e.g. schools and hotel, the informative terms are more distributed leading to larger *m* values. Note that the optimal number of features may also vary depending on when Flickr data or Twitter data is used.
When only the Flickr data is used (Table 4), we find that using *CC* results in a significant improvement over χ^2 (Wilcoxon signed ranks test, $p < 0.01$). As an example, the *MNDCG* values of the χ^2 and *CC* feature selection for place type 'monument' are plotted in Figure 4. The top

300 ranked words according to χ^2 do not contain words that strongly characterize places of other types than monuments resulting in behaviour which is similar to *CC* when *m* is smaller than 300. However, χ^2 ranks for place type 'monument' the words 'nationalcemetery' and 'food' at respectively position 359 and 384. While such terms are potentially useful to exclude particular other places types, they appear to be less effective than the terms that are directly indicative of monuments that are preferred by *CC*. For the other considered place types, a similar observation can be made.

Figure 4. MNDCG values for different number of tags when χ^2 and CC feature selection is used on the place descriptions from Flickr, for place type 'monument'.

When we compare *CC* without and with the filtering step we also get a significant improvement (Wilcoxon signed ranks test, $p < 0.01$). For example, for place type 'hotel' the word 'italy' receives the largest *CC* value because a lot of hotels in the training set are located in Italy (more precisely in Venice), but this word may also refer to Italian design, cars or restaurants. By filtering such terms, the effectiveness of the method is improved. In particular, we get significant improvement of the optimal *MNDCG* value. Finally, by comparing the *MNDCG* values of the χ^2 and *CC+filter* we conclude that latter technique performs significantly better (Wilcoxon signed ranks test, $p < 0.01$).

Surprisingly, in contrast to Flickr, there is no clear difference in the use of the different feature selection techniques for the Twitter data (Table 5; Wilcoxon signed ranks test, $p > 0.2$). This relates to the fact that Twitter data contains a lot of non-informative terms such as opinions, statements and personal status updates (Naaman, 2010). Moreover, tweets contain less geographic information such as city and country names than Flickr (Van Laere, 2013), which reduces the impact of our filtering step. We note that filtering out names of cities and countries may even decrease the performance. For schools, for instance, the word 'lacrosse' is excluded because it may refer to the city La Crosse, located in Wisconsin, United States. However, 'lacrosse' may also refer to a team sport which is played in many US colleges and the occurrence of the word 'lacrosse' may therefore indicate the presence of a school. For several place types, the result is not very sensitive to the actual number of features which is used. In such a case, the optimum number of features may change quite drastically between *CC* and *CC+filter*. This is most pronounced in the case of 'station', as shown in Figure 5. In the rest of this paper, we will use *CC+filter* to obtain a fair comparison between Flickr and Twitter.

Figure 5. MNDCG values for different number of tags when CC and CC+filter feature selection is used on the place descriptions from Twitter, for place type 'station'.

Using the optimal settings of our methodology, we finally optimize the remaining parameters *b* (see Equation 2), λ (see Equation 15) and η (see Equation 19). The optimal values of all parameters can be found in Table 3.

Table 4. Optimal number of features (m) and corresponding MNDCG values for χ^2, CC and CC+filter on the place descriptions from Flickr.

place type	**Optimal number of features (*m*)**			***MNDCG***		
	χ^2	***CC***	***CC+filter***	χ^2	***CC***	***CC+filter***
Shop	2000	1000	1000	50.13	50.13	**50.33**
Restaurant	20	20	20	56.61	56.61	**56.62**
School	1600	1600	1600	35.36	35.36	**36.98**
Place of Worship	300	400	400	58.38	58.41	**58.46**
Pub	100	100	100	64.14	64.77	**67.79**
Hotel	1400	1400	1400	59.61	59.61	**60.38**
Station	50	50	50	74.08	**74.09**	**74.09**
Hospital	4500	4500	4500	42.12	42.14	**42.36**
Monument	500	500	400	63.30	63.64	**64.33**
Library	50	50	50	52.49	52.61	**54.44**
Graveyard	1700	1800	1800	74.34	74.34	**74.55**
Museum	100	1100	1100	60.96	61.24	**61.36**
Airport	700	700	700	67.55	67.56	**67.57**
Castle	2800	2800	3000	85.58	85.60	**85.63**

Table 5. Optimal number of features (m) and corresponding MNDCG values for χ^2, CC and CC+filter on the place descriptions from Twitter.

place type	**Optimal number of features (*m*)**			***MNDCG***		
	χ^2	***CC***	***CC+filter***	χ^2	***CC***	***CC+filter***
Shop	300	7500	7000	**46.00**	45.48	45.58
Restaurant	300	300	300	52.14	52.14	**52.15**
School	4500	4500	2900	32.38	**32.41**	32.40
Place of Worship	50	50	50	36.14	**36.15**	36.14
Pub	6000	5000	5000	54.25	**54.30**	54.28
Hotel	1200	1200	1200	**48.25**	47.70	47.71
Station	100	100	1500	52.17	**52.63**	52.43
Hospital	100	100	100	**33.96**	**33.96**	**33.96**
Monument	4500	6000	6000	47.01	**47.18**	47.17
Library	50	50	50	**36.68**	36.45	36.61
Graveyard	20	20	20	**59.00**	**59.00**	**59.00**
Museum	6000	5500	6000	39.89	**39.90**	**39.90**
Airport	20	20	20	65.38	65.38	**65.39**
Castle	100	100	100	**81.51**	**81.51**	**81.51**

Quantitative Evaluation

In this and the following section, we apply our proposed methodology to the cities from the test set, using the optimal parameters that were obtained from the cities in the development set. In this section we perform a quantitative evaluation, where we compare the difference in performance when Flickr or Twitter data is used, and demonstrate how Flickr and Twitter can be combined to optimize the results. Given a city $C \in S_{test}$ and a place type *t*, we evaluate the rankings of POIs using the A*verage Precision* $AP(C,t,y)$ metric in addition to the *NDCG* metric defined in (20). The average precision metric is added because it can be used for a more detailed analysis than the *NDCG* metric by using different distance thresholds (indicated by parameter *y*). On the other hand, the use of *NDCG* is more useful for parameter optimization because it summarizes the performance of our methodology in one metric. As for *NDCG*, the *AP* value lies between 0 and 100, in which a higher value means a better performance.

To define the average precision metric, we first define *Precision at position n*, $P(C,t,y)@n$, which is the fraction of the top *n* ranked POIs that are relevant to the user's information need. $P(C,t,y)@n$ is formally given by

$$P(C,t,y)@n = \frac{\sum_{i=1}^{n} relevant(C,t,y)@i}{n} \times 100 \qquad (24)$$

For the calculation of the precision, a point of interest $poi \in U^F$ is considered as relevant if the ground truth of city *C* contains a place of type *t* within *y* meters of *poi*, where we will consider different values of *y*:

$$relevant(C,t,y)@i = \begin{cases} 1 & \text{if } d(poi_i, nn_{i,t}) \le y \\ 0 & \text{otherwise} \end{cases} \qquad (25)$$

The *Mean Precision at position n*, $MP(C,t,y)@n$, is defined as the mean of the *Precision at position n* values of all cities in the test set:

$$MP(t,y)@n = \frac{\sum_{C' \in S_{test}} P(C',t,y)@n}{|S_{test}|} \qquad (26)$$

with S_{test} the set of all cities in the test set.

In addition, the *Recall at position n* metric, $R(C,t,y)@n$, corresponds to the fraction of POIs that are relevant which are successfully ranked in the top *n* POIs:

$$R(C,t,y)@n = \frac{\sum_{i=1}^{n} relevant(C,t,y)@i}{\sum_{i=1}^{|U^F|} relevant(C,t,y)@i} \times 100 \qquad (27)$$

By computing a precision and recall for each $n \in [1, |U^F|]$ one can plot a precision-recall curve and consider the area under the curve as the relevance of the ranked list. This value can be approximated using the *Average Precision* metric $AP(C,t,y)$ (Zhu, 2004), which is defined as

$$AP(C,t,y) = \frac{\sum_{n=1}^{|U^F|} relevant(C,t,y)@n \cdot P(C,t,y)@n}{\sum_{n=1}^{|U^F|} relevant(C,t,y)@n} \quad (28)$$

We finally define the *Mean Average Precision* $MAP(t,y)$ as the mean of the *Average Precision* values of all cities in the test set:

$$MAP(t,y) = \frac{\sum_{C' \in S_{test}} AP(C',t,y)}{|S_{test}|} \quad (29)$$

with S_{test} the set of all cities in the test set.

The *MAP* and *MNDCG* values of each considered approach are listed in Tables 7 and 8. In the remainder of this section, we discuss each approach in more detail.

Flickr results. We start our experiments by only using the Flickr data to discover POIs in a city *C* and to rank them according to the likelihood that they are associated with a given place type *t*. First, we use the methodology from our previous work as a baseline (Van Canneyt, 2012a; Van Canneyt, 2012b). To this end, we cluster the locations where Flickr photos has been taken to obtain the locations of candidate POIs and associate a description to them as described above. We then train a multi-class support vector machine (SVM) classifier (Crammer & Singer, 2012) for a given place type *t* based on the descriptions of the places in the training set. Subsequently, we use this classifier to rank the locations which potentially contain a place of interest based on the probability that they contain a place of the given type *t*. The performance of this approach can be found in the columns labeled with 'Flickr (SVM)' in Tables 7 and 8.

Second, we used the approach described in this paper on the Flickr data to discover places of a given type. The main difference with our previous work is that we replaced the SVM classifier by a language model approach and introduced a feature selection technique. The performance of this approach is shown in the 'Flickr' columns of Tables 7 and 8. By comparing the performances of the two approaches, we found that our new approach significantly outperforms the SVM based baseline (Wilcoxon signed ranks test, $p < 0.01$).

To further interpret the results we calculated *Mean Precision at 1* $MP(t,500)@1$ (see Table 9). Based on these metric values, we can conclude that for 81% of the cities in the test set the highest ranked POI is located within 500 meter of a known station. This is due the fact that in the most cities there are a lot of pictures taken nearby the main train station, and such pictures typically have highly indicative tags such as 'train', 'station' and 'railway'. However, there are some challenges with using Flickr for detecting places. For instance, for 48% of the cities in the test set, the highest ranked POI is not located within 500 meter of a known hospital. One reason is that the Flickr tags can be misleading (e.g. a photo of an ill person far away from a hospital). Second, our Flickr data may contain no photos with descriptive tags taken nearby a smaller hospital of the city. Third, some hospitals are found by our methodology using Flickr, but the distance between the location of the detected POI and the hospital is larger than 500 meter. This is mainly due the fact that most of the pictures at the hospital are taken in the hospital rooms, whereas the ground truth may refer to another part of the hospital (e.g. the main entrance). Finally, it may be the case that an actual hospital is found, which is not contained in our ground

truth, as LGD and Geonames are inherently incomplete. The effect of missing places in the ground truth will be investigated in more detail in the next section.

Twitter results. The results for the Twitter data are shown in the columns labeled with 'Twitter' in Tables 7 and 8. Although a large number of tweets are not informative (see Murdock (2011)), we have observed that some tweets are very useful to recognize place types. For example tweets such as 'About to have dinner #feelsgood', '@DavidSahadi: Enjoying the first (of a few) micro beers at the outdoor bar at Big River Brewery' and 'waiting for the train...' may indicate the occurrence of restaurants, pubs and stations, respectively. However, our training data hardly contain tweets describing places of types such as places of worship and hospitals, resulting in low *MNDCG* and *MAP* values for these types.

Flickr and Twitter combined. Comparing the results obtained using the Twitter data with the results obtained using the Flickr data, we find significant better *MAP* and *MNDCG* values for Flickr (Wilcoxon signed ranks test, $p < 0.01$). Based on this observation, we may conclude that Flickr tags are more informative for finding places of a given type than Twitter posts. However, when we use both Flickr tags and Twitter terms ('Flickr+Twitter' colums Table 7 and 8), we get a further significant improvement for the average *MNDCG* and *MAP* values over all considered place types (Wilcoxon signed ranks test, $p < 0.01$). However, no clear improvement can be observed for place types with only a few instances in our ground truth dataset such as castles and airports.

We observed the best performance for London. With over 600,000 Flickr photos and over 120,000 tweets this is the city with most social media data in our test set. The *MNDCG* values for London are shown in the column labeled with 'London' in Table 7. These values suggest, somewhat unsurprisingly, that the number of available photos and/or tweets substantially impacts the performance of our method. With our current dataset, the performance for London is sufficiently high to support practical applications, but this may not yet be the case for some smaller cities. As more and more geo-annotated social media becomes available, however, we could expect to see a comparable performance for a wider range of cities. Still, even for London, the performance varies substantially across different place types. As people are less likely to tweet from a library than from a pub, it should perhaps not come as a surprise that the method works better for pubs. To further widen the applicability of the proposed method, a wider range of sources, beyond Flickr and Twitter, may need to be considered. The ability of discovering new places of interest in London will be further investigated in the next section.

Table 7. MNDCG of the ranked points of interest when Flickr and/or Twitter data is used. The last column indicates the MNDCG values for London when both the Flickr and Twitter data is used.

place type	Flickr (SVM)	Flickr	Twitter	Flickr+Twitter	London
Shop	43.14	46.27	42.26	**46.73**	89.20
Restaurant	50.99	54.19	52.36	**56.82**	92.79
School	34.12	34.19	33.20	**35.26**	68.89
Place of Worship	49.37	49.77	33.13	**49.83**	78.76
Pub	61.87	66.17	57.77	**66.97**	95.27
Hotel	50.97	51.69	41.69	**54.08**	88.17
Station	69.70	69.89	49.98	**71.73**	87.86
Hospital	36.24	36.14	33.56	**37.68**	71.97
Monument	66.91	67.71	55.45	**67.76**	83.10
Library	53.65	**54.22**	40.37	54.07	59.08
Graveyard	74.90	**74.97**	60.60	74.42	-
Museum	60.88	64.76	41.73	**65.52**	76.17
Airport	68.39	**68.65**	60.59	68.62	-
Castle	84.79	**86.45**	79.00	86.38	95.02

Table 8. MAP values of the ranked points of interest Flickr and/or Twitter data is used.

place type	Flickr (SVM)			Flickr			Twitter			Flickr + Twitter		
	25m	100m	1km	25m	100m	1km	25m	100m	1km	25m	100m	1km
Shop	6.40	18.70	63.69	8.03	25.32	66.88	5.74	17.59	62.54	**8.14**	**26.13**	**68.18**
Restaurant	14.99	23.84	64.58	16.45	32.31	67.78	15.74	28.10	64.74	**17.31**	**36.23**	**71.09**
School	4.97	8.76	62.00	4.83	10.64	64.98	2.18	5.91	66.89	**5.82**	**11.03**	**68.64**
Place of Worship	12.12	19.86	61.66	12.10	24.22	64.98	5.15	6.57	57.33	**12.13**	**24.37**	**66.61**
Pub	29.07	37.70	69.73	31.31	46.90	75.15	26.01	36.04	70.37	**32.60**	**49.01**	**78.83**
Hotel	7.30	20.67	60.28	7.92	26.77	66.73	3.99	14.03	52.13	**9.25**	**29.82**	**70.30**
Station	35.32	54.71	63.23	35.71	58.95	65.68	21.65	31.39	58.54	**36.34**	**61.36**	**69.72**
Hospital	12.14	19.78	36.54	12.90	21.15	41.64	10.99	16.22	36.23	**13.63**	**22.44**	**42.86**
Monument	44.63	49.87	68.02	44.62	53.96	75.64	40.60	43.31	63.66	**44.67**	**54.04**	**76.92**
Library	28.52	34.02	53.90	**28.97**	**36.63**	55.27	25.31	26.41	49.58	28.96	36.40	**59.56**
Graveyard	57.75	66.20	61.96	**58.50**	66.28	**62.47**	54.55	54.69	59.67	58.10	**66.53**	61.84
Museum	30.30	40.01	62.34	32.18	46.50	71.68	24.36	25.80	50.34	**33.07**	**46.69**	**72.87**
Airport	49.78	54.08	67.96	49.91	54.51	**68.51**	49.38	53.01	54.64	**50.73**	**55.11**	68.17
Castle	76.27	81.13	80.35	**77.34**	82.84	82.10	75.58	76.04	77.65	77.28	**82.88**	**82.67**

Table 9. Mean Precision at 1, $MP(t,500)@1$, of the ranked points of interest when Flickr is used.

Shop	Restaurant	School	Place of Worship	Pub	Hotel	Station
60.23	59.09	54.55	63.64	76.14	68.18	81.48
Hospital	**Monument**	**Library**	**Graveyard**	**Museum**	**Airport**	**Castle**
51.59	69.32	57.95	55.68	64.77	67.05	78.41

Qualitative Evaluation

Discovering New Places of Interest

In this section, we will analyze to what extent our method can discover places of type *t* in a city *C* that are not yet contained in LinkedGeoData, Geonames, Foursquare and Google Places. To find such places, we first remove from the results those places that are within distance $2\sigma_U$ from a place in the ground truth of the same type; see Table 3 for the values of sigma for each place type. We will focus on London to get a deeper insight in the ability of our methodology to detect new places.

Table 14 shows the top 10 of the resulting rankings, and indicates which places can not been found in Google Places or Foursquare when a user searches for places of a particular type (databases accessed on February 27, 2013). This may be because the places are not included in Google Places or Foursquare at all, or because they are included but classified as another type. Entries in Table 14 are shown in bold if they are not included in Google Places or Foursquare. Additionally, they are marked with Go and Fo is they are not included in Google Places and Foursquare, respectively, and with Go and Fo if they are only included with a different type. The place names mentioned in the table have been manually determined, as detecting place names is outside 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 those shown in italic.

In London, our method is able to find places of worship, schools, shops, restaurants, graveyards, castles, hotels, pubs, stations, libraries, museums and monuments that are not yet included in our LinkedGeoData and Geonames. Our method was not able to find new airports because the considered region of London contains no airports. Several of these places are not yet included in the Google Places and Foursquare database. As shown in Table 14, places not present in Google Places are for instance shops, restaurants, hotels, monuments, libraries, graveyards and museums. Additionally, our method is able to extend Foursquare with shops, schools, monuments and graveyards. Finally, some places such as the Savanna shop at Portobello Road, the Women of World War II monument, and All Saints Church Cemetery are neither included in Foursquare nor Google Places. Furthermore, several places were detected which were already present in Foursquare and Google Places, but without the desired type associated. For example the House of Gifts, the Kensington Close hotel and Royal Hospital Chelsea are contained in Foursquare, but without an associated semantic type.

Closer examination of the detected places revealed some challenges with using social media. The first challenge is that Flickr photos may be taken at a far distance from the place of interest (e.g. a photo taken from the St. Thomas' Hospital taken at Leathermarket Gardens more than 500 meter away). Second, the used Flickr and Twitter data may be out-of-date (e.g. a photo of the Hops bar which is closed). Third, the Twitter term and Flickr tags corresponding to a name of a place or region may incorrectly suggest the presence of a place of a particular type (e.g. the tag 'Elephant and Castle', corresponding to a major Junction in London, incorrectly suggest the presence of castles). Finally, the Flickr tags and Twitter terms may not be related with the place nearby the location of the user (e.g. the tweet 'waiting for a taxi to go to the hospital').

Table 14. Top 10 of the discovered places in London which are not yet included LGD and Geonames. Places are shown in bold if they are not included in Google Places or Foursquare. Additionally, they are marked with Go and Fo if they are not included in Google Places and Foursquare, respectively, and with Go and Fo if they are only included with a different type. Finally, errors are indicated in italic.

place type	1st place	2nd place	3rd place	4th place	5th place
Shop	Nippon and Korea Centre	New Look Oxford Street	Marks and Spencer	Selfridges and Co	**Savana (Portobello Road)Go,Fo**
Restaurant	Zizzi	Otto	Pain QuotidienGo,Fo	Carluccio's	Tay Do
School	University College London	Imperial College	City of Westminster College	University of the Arts	UCL Cruciform Building
Place of Worship	Westminster Cathedral	Southwark Cathedral	St Stephen Walbrook	St Sophia Greek Cathedral	St Mary Aldermary Church
Pub	Off Broadway	The Anchor	The White Horse	The Roxy	Green Man and French HornGo,Fo
Hotel	Bayswater Inn Hotel	Premier Inn Hotel	The Dorchester	Kensington CloseFo	Vicarage Private
Station	Queens Grove	*train*	Waterloo	Pimlico	Fenchurchstreet
Hospital	Royal HospitalFo	*Photo of the St Thomas' Hospital*	St Mary's Hospital	*Temperance Hospital*	*abandoned children's hospital*
Monument	Victoria Memorial	Albert MemorialGo	Tower Hill MemorialGo,Fo	**Webmister Abbey Lions MemorialGo,Fo**	Monument of the Great Fire of London
Library	**Birkbeck LibraryGo**	Science Museum Library	Maughan Library	**SOAS LibraryGo**	Peckham Library
Graveyard	Brompton Cemetery	Bunhill Fields Burial Ground	Nunhead Cemetery	Saint Pancras CemeteryGo,Fo	St. George's GardensGo
Museum	Science Museum	Imperial War Museum	Design Museum	Saatchi Gallery	Clink Prison Museum
Castle	*elephant and castles*	The Pirate CastleGo,Fo	Buckingham Palace	Kensington Palace Fo	*Castle Battersea*

place type	6th place	7th place	8th place	9th place	10th place
Shop	House of Gifts Go,Fo	*shoppers*	**National Portrait Gallery ShopFo**	Rokit	Harrods
Restaurant	**Sen Nin on Islington Park St.Go**	Jamie Oliver's Fifteen	Antariya Sushi Bar	*Hive Bar*	RegencyFo
School	King's College	The Barlett faculty	Royal College of Surgeons	College of Communication	**Spa schoolFo**
Place of Worship	St Olave Church	St Martin in the Fields Church	Christ Church	St James' Church	St George's Cathedral
Pub	Daily Grind	Horse and Groom	*Hops bar*	Draft House Tower Bridge	The Elgin
Hotel	The Sanctuary Fo	St Pancras Renaissance Hotel	*Chelsea Bridge*	**Great Northern Kings CrossGo,Fo**	The Wellesley
Station	*railway track*	*from the train*	*central station*	*rail tracks*	*train portrait*
Hospital	*Science Museum*	The Royal Marsden	*londonroyalhospital*	St Pancras Hospital	*Old Royal Free*
Monument	**Statues on entrance County HallGo,Fo**	**The Women of World War IIGo,Fo**	**Statue of Richard the LionheartFo**	**St George StatueGo,Fo**	Buxton Memorial Fontain
Library	Borough Road Library	Senate House Library	**SSEES LibraryGo**	Idea Store Whitechapel	LSE Library
Graveyard	**All Saints Church CemeteryGo,Fo**	Postman's ParkGo,Fo	Paddington Cemetery	St John's Wood Church GardensGo,Fo	**Royal Hospital Old Burial GroundGo,Fo**
Museum	**Sir John Ritblat GalleryGo**	Foundling Museum	Wellcome collection	*Museum of London sign*	Kirkaldy Testing Museum
Castle	*The Castle*	Victoria Tower	*Dublin Castle*	*elephant and castles*	*elephant and castles*

Validation of Known Places of Interest

In the previous sections, we have described how social media can be used to extend databases of places. Another way of improving databases of places is to identify and remove incorrect information in these databases. The presence of incorrect place information may be due to various reasons: places of interest may have been closed, their type may have been changed (e.g. a shop converted in a pub), or the places may even have been incorrectly added to the database. However, it is very time-consuming to manually check the correctness of the data in existing databases. We describe in this section how our methodology can be used to facilitate this data validation process. In particular, given a type *t* and the locations of the places in the databases which are associated with this type, we indicate which places are most likely incorrect. Places are considered as incorrect when there is no place of type *t* at their location.

For this case study we use the database of Foursquare, a platform on which users can freely add places to the database. The database contains a lot of unverified places, indicating that the owners of the places of interest have not claimed and did not verify the place information. For instance, about 95% of the places we collected from London were not verified. For these unverified places in particular, a method to automatically assess the likelihood that they are accurate would be useful. We first collected 21,436 Foursquare places from London with a type corresponding to one of the considered place types in this paper. The task we consider is to determine which of the collected places are most likely incorrect. In particular, given a type *t* and the locations of the places in the Foursquare database which are associated with this type, we used the Flickr and Twitter data posted nearby the locations to rank them based on the likelihood that there is no place of type *t* located nearby.

The results are shown in Table 15. Places are marked with [1] if the type of the place is incorrect, with [2] if the place is incorrect located and with [3] if the Foursquare place is no place of interest at all. Finally, detected Foursquare places that are correct are indicated in italic. Most of the places which are considered most likely to be incorrect are indeed incorrect, most often because they have an incorrect location. Additionally, some places have a wrong associated type. For instance Epio HQ is a software company which is incorrectly categorized as museum. Finally, some places in the Foursquare databases do not correspond with a general place of interest. Examples are the ‘pub’ labeled ‘Home Of Morris’ which corresponds with someone’s home, and the fictive place ‘Behind you’ which is categorized as cemetery. These results confirm that our method is able to facilitate the detection of incorrect information in databases of places.

Table 15. Top 5 of the Foursquare places in London which are most likely incorrect. Places are marked with [1] if the place type is incorrect, with [2] if the place is incorrect located and with [3] if the Foursquare place is no place of interest at all. Finally, errors are indicated in italic.

place type	1st place	2nd place	3rd place	4th place	5th place
Shop	William Hill [1 2]	*Londis*	The Pantry [2]	Specsavers [2]	nternational Food Centre [2]
Restaurant	BananaTree [2]	Favourite Chicken [2]	Quality Café [1 2]	*5 Skehans Thai*	Sticky Fingers [2]
School	IBAM London [2]	London Studio Centre [2]	*School of Pharmacy*	*SAE Institute*	*London Knowledge Lab*
Place of Worship	Tasmin [3]	St Anne's Church [2]	*MRBC*	Jon Bon Jovi's Dressing Roo [3]	*Church on the Corner*
Pub	V.V. Coffee Bar [2]	*Charlton*	The Clarrence [2]	The Asylum [2]	Home Of Morris [3]
Hotel	*Quality Maitrise Hotel*	5 Doughty Street [3]	Alternative Urban Residence [3]	Jury's Inn [2]	Dylan Apartments [2]
Station	Clapham High [2]	Euston Station [2]	Platform 13 - Gatwick Express [2]	London Field [2]	Brondesbury [2]
Hospital	*London Chest Hospital*	*Tavistock Centre*	*Brondesbury medical center*	*BMI The London Independent Hospital*	Ruskin Wing [2]
Monument	Harley Street and Cavendish Street [3]	New River Walk [2]	The Hellotxtplex [3]	Carlyle's House [1]	Tower Hamlets Labour Party [1]
Library	Clapham Library [2]	Nunhead Library [2]	*Regents Park Library*	CLR James Library [2]	*Paddington Library*
Graveyard	Kensal Green [2]	Behind you [3]	Bunhill fields [2]	*St Paul's Churchyard*	The Stylenoir Lair [3]
Museum	18 Stafford Terrace [3]	*The Jewish Museum*	The Pill Box [1 2]	*Royal Mews*	Epio HQ [1]
Airport	TfL Bus 12 [3]	Admirals Club T3 [1]	Biggin Hill Airport [2]	Heathrow Airport [2]	*The London Heliport*

CONCLUSION

In this paper, we demonstrated how social media can be used to improve existing databases of places. We first used mean shift clustering on the locations of a set of Flickr photos to obtain the locations which potentially correspond to places of interest (POIs) in a given city *C*. We then associated with each candidate POI a feature vector based on the tags of the Flickr photos that are associated with locations nearby. Afterwards, we associated a query with each place type *t* based on the descriptions of known places of that type. The obtained query is used to rank the candidate POIs based on the likelihood that they belong to type *t*. To produce this ranking, we relied on a language modeling approach, which performed significantly better than the Support Vector Machine classifier used in our previous work (Van Canneyt, 2012a; Van Canneyt, 2012b). Finally, we discussed how Twitter can be used to improve the results.

In the optimization phase of our proposed methodology, we analyzed the behaviour of different feature selection techniques. We concluded that for the Flickr data, correlation coefficient feature selection (Ng, 1997) performs significantly better than χ^2. The performance of the proposed methodology was further significantly improved when names of the cities in the training set and the names of the countries in which these cities are located were removed from the features. Surprisingly, in contrast to Flickr, we did not find a clear difference in performance between the use of the different feature selection techniques for the Twitter data.

We performed a large-scale evaluation on 88 different cities. Using Flickr, our methodology was for instance able to find a location which is within 500 meter of a known station for 81% of the cities in the test set. We concluded that Flickr tags are more informative for finding places of a given type than Twitter posts. However, as we have demonstrated in this paper, using tweets in addition to Flickr photos can still be used to improve the quality of the results. We further

examined the results for London in more detail to analyze to what extent our approach can discover new places of a particular type. Based on this evaluation, we could conclude that our method is able to detect places which were not yet included in LinkedGeoData, Geonames, Google Places and Foursquare. Additionally, we explained how our methodology can be used to identify errors in existing databases of places such as Foursquare.

Acknowledgments: Steven Van Canneyt is funded by a Ph.D. grant of the Agency for Innovation by Science and Technology (IWT). We are grateful to Olivier Van Laere for his help with collecting and processing some of the data we have used in this paper.

REFERENCES

Abbasi, R., Chernov, S., Nejdl, W., & Paiu, R. (2009). Exploiting flickr tags and groups for finding landmark photos. In M. Boughanem, C. Berrut, J. Mothe, & C. Soule-Dupuy (Eds.), *Advances in Information Retrieval* (pp. 654–661). Springer Berlin / Heidelberg.

Ahern, S., Naaman, M., & Nair, R. (2007). World explorer: visualizing aggregate data from unstructured text in geo-referenced collections. *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 1–10). New York, NY, USA: ACM.

Auer, S., Lehmann, J., & Hellmann, S. (2009). LinkedGeoData: Adding a spatial dimension to the web of data. In A. Bernstein, D. Karger, T. Heath, L. Feigenbaum, D. Maynard, E. Motta, & K. Thirunarayan (Eds.), *Proceedings of the 8th International Semantic Web Conference* (Vol. 5823, pp. 731–746). Chantilly, VA, USA: Springer Berlin Heidelberg.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American Magazine*, *284*(5), 34–43.

Cao, L., Luo, J., Gallagher, A., Jin, X., Han, J., & Huang, T. S. (2010). A worldwide tourism recommendation system based on geotagged web photo. *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing* (pp. 2274–2277). Dallas, TX: IEEE.

Cheng, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *17*(8), 790–799.

Choudhury, M. De, Feldman, M., Amer-Yahia, S., Golbandi, N., Lempel, R., & Yu, C. (2010). Automatic construction of travel itineraries using social breadcrumbs. *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia* (pp. 35–44). New York, NY, USA: ACM.

Clements, M., Serdyukov, P., & Vries, A. de. (2010). Using Flickr geotags to predict user travel behaviour. *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 851–852). New York, NY, USA: ACM.

Crammer, K., & Singer, Y. (2002). On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, *2*, 265–292.

Crandall, D. J., Backstrom, L., Huttenlocher, D., & Kleinberg, J. (2009). Mapping the world's photos. *Proceedings of the 18th International Conference on World Wide Web* (pp. 761–770). New York, NY, USA: ACM.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A., Shaked, T., Soderland, S., Weld, D. S., et al. (2005). Unsupervised named-entity extraction from the Web: An experimental study. *Artificial Intelligence Journal*, *165*(1), 1–42.

Flickr (2013). Explore everyone's photos on a Map. Retrieved February 26, 2013, from http://www.flickr.com/map

Gracia, J., & Mena, E. (2009). Multiontology semantic disambiguation in unstructured web contexts. In *Proceedings of the 2009 K-CAP Workshop on Collective Knowledge Capturing and Representation* (pp. 1–9).

Hoffart, J., Suchanek, F. M., Berberich, K., & Weikum, G. (2013). YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence Journal*, *194*(1), 28–61.

Jain, S., Seufert, S., & Srikanta, B. (2010). Antourage: mining distance-constrained trips from Flickr. *Proceedings of the 19th International Conference on World Wide Web* (pp. 1121–1122). New York, NY, USA: ACM.

Kwok, C., Etzioni, O., & Weld, D. S. (2001). Scaling question answering to the web. *ACM Transactions on Information Systems*, *19*(3), 242–262.

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

Markines, B., Cattuto, C., Menczer, F., Benz, D., Hotho, A., & Stumme, G. (2009). Evaluating similarity measures for emergent semantics of social tagging. *Proceedings of the 18th International Conference on World Wide Web* (pp. 641–650). New York, NY, USA: ACM.

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

Naaman, M., Boase, J., Lai, C., & Brunswick, N. (2010). Is it really about me? Message content in social awareness streams. *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work* (pp. 189–192). New York, NY, US: ACM.

Navigli, R., Informatica, D., & Ponzetto, S. P. (2010). BabelNet : Building a Very Large Multilingual Semantic Network. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 216–225). Stroudsburg, PA, USA: Association for Computational Linguistics.

Ng, H. T., Goh, W. B., & Low, K. L. (1997). Feature selection, perceptron learning, and a usability case study for text categorization. In N. J. Belkin, A. D. Narasimhalu, P. Willett, W. Hersh, & F. Can (Eds.), *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 67–73). New York, NY, USA: ACM.

Ozdikis, O., Orhan, F., & Danismaz, F. (2011). Ontology-based recommendation for points of interest retrieved from multiple data sources. *Proceedings of the International Workshop on Semantic Web Information Management*. New York, NY, USA: ACM.

Popescu, A., & Grefenstette, G. (2008). Gazetiki: automatic creation of a geographical gazetteer. *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 85–93). New York, NY, USA: ACM.

Rattenbury, T., Good, N., & Naaman, M. (2007). Towards automatic extraction of event and place semantics from Flickr tags. *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 103–110). New York, NY, USA: ACM.

Sakaki, T. (2010). Earthquake shakes Twitter users : real-time event detection by social sensors. *Proceedings of the 19th International Conference on World Wide Web* (pp. 851–860). New York, NY, USA: ACM.

Schmitz, P. (2006). Inducing ontology from Flickr tags. *Proceeding of the Collaborative Web Tagging Workshop at the World Wide Web Conference* (pp. 3–6). New York, NY, USA: ACM.

Serdyukov, P., Murdock, V., & Van Zwol, R. (2009). Placing Flickr photos on a map. *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 484–491). New York, NY, USA: ACM.

Stützle, T., & Hoos, H. H. (2000). Max-min ant system. *Future Generation Computer Systems*, *16*(8), 889–914.

Van Canneyt, S., Schockaert, S., Van Laere, O., & Dhoedt, B. (2011). Time-dependent recommendation of tourist attractions using Flickr. In P. De Causmaecker, J. Maervoet, T. Messelis, K. Verbeeck, & T. Vermeulen (Eds.), *Proceedings of the 23rd Benelux Conference on Artificial Intelligence* (pp. 255–262). Ghent, Belgium.

Van Canneyt, S., Schockaert, S., Van Laere, O., & Dhoedt, B. (2012a). Detecting places of interest using social media. *Proceedings of the 2012 IEEE/WIC/ACM International Conference on Web Intelligence* (pp. 447–451).

Van Canneyt, S., Schockaert, S., Van Laere, O., & Dhoedt, B. (2012b). Using social media to find places of interest: A case study. *Proceedings of the 2012 ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information* (pp. 2–8).

Van Laere, O., Schockaert, S., & Dhoedt, B. (2011). Finding locations of Flickr resources using language models and similarity search. *Proceedings of the 1st ACM International Conference on Multimedia Retrieval* (pp. 48-55). New York, NY, USA: ACM.

Van Laere, O., Schockaert, S., & Dhoedt, B. (2013). Georeferencing Flickr resources based on textual meta-data. *Information Science*, accepted.

Wu, F., & Weld, D. S. (2007). Autonomously semantifying Wikipedia. *Proceedings of the 16th ACM Conference on Information and Knowledge Management* (pp. 41–50). New York, NY, USA: ACM.

Wu, F., & Weld, D. S. (2008). Automatically refining the wikipedia infobox ontology. *Proceedings of the 17th International Conference on World Wide Web* (pp. 635–644). New York, NY, USA: ACM.

Zhu, M. (2004). *Recall, precision and average precision*. (Tech. Rep. No. 1). University of Waterloo.