Estimating the Semantic Type of Events using Location Features from Flickr

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

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

Bart Dhoedt
Ghent University - iMinds
Gaston Crommenlaan 8
Ghent, Belgium
bart.dhoedt@ugent.be

ABSTRACT

Various methods for automatically detecting events from social media have been developed in recent years. However, little progress has been made towards extracting structured representations of such events, which severely limits the way in which the resulting event databases can be queried. As a first step to address this issue, we focus on the problem of discovering the semantic type of events. While current methods are almost exclusively based on bag-of-words methods, we show that additionally using location features can substantially improve the results. In particular, we use the tags associated with Flickr photos and the types of the known events near the venue of the event as context information.

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

Semi-structured Data, Geographic Information Retrieval, Ensemble Learning, Events, Social Media, Flickr

1. INTRODUCTION

Databases of events such as EventFul[1], Upcoming[2] and Facebook Events[3] have become increasingly popular in last few years. These databases are constructed in different ways: EventFul, for instance, combines data from several existing sources such as websites of music venues and ticketing sites. Another method, which is used by Upcoming and Facebook Events, is to let users freely add and modify event information. Finally, social media can also be used to automatically extend such databases [1, 2, 3, 9]. To effectively query or browse through large collections of events, it is important that events have an appropriate associated semantic type. Someone who is interested in cultural events can for instance search for musicals and theatre shows, whereas families visiting a city may be more interested in circus and other family events. Furthermore, the semantic type of events can be used as a first step towards structured representation of events. In particular, relevant properties of an event are often based on the semantic type of the event, e.g. the magnitude for earthquakes and the final score for football matches. These structured representations are for instance needed to answer queries such as 'Who are the artists who played at the most popular festival in London?'. However, more than 10% of the Upcoming events we collected have an unknown semantic type. In addition, events which are automatically detected using social media are often represented as a bag of words, or a set of social media documents, and therefore have no associated type.

Evidence about the semantic type of an event can be obtained by analyzing social media documents, such as Flickr photos taken at the event, which we consider in this paper, or tweets that have been sent about the event. In particular, similar as in e.g. [1, 2, 9], we represent an event as a set of social media documents related to that event, together with an associated event type. Most initial work about discovering the semantic types of events only used the textual information of the social media documents associated with the event [5, 6], which may lead to poor performance when the text is noisy (e.g. in some Twitter posts) or absent (e.g. in some Flickr photos). However, some social media documents are also annotated with geographic coordinates. This can be used to estimate the location of the event, to discover other events organized nearby and to detect social media documents created in its vicinity (not necessarily associated with the event). Our main objective in this paper is to investige to what extent this geographical information can be exploited to discover event types more accurately.

The remainder of this paper is structured as follows. We start with a review of related work in Section 2. Next, in Section 3, we describe our methodology for classifying events based on the meta-data of their associated Flickr photos. Details on the considered training data, test data and data preprocessing steps are provided in Section 4. Subsequently, Section 5 presents the experimental results. Finally, we conclude our work in Section 6.

[1]http://eventful.com/ [2]http://upcoming.org/
[3]https://facebook.com/events/

SIGSPATIAL '14, November 04-07 2014, Dallas/Fort Worth, TX, USA

ACM 978-1-4503-3135-7/14/11$15.00
http://dx.doi.org/10.1145/2675354.2675700

2. RELATED WORK

There has been a lot of interest in detecting events and their associated documents using location, temporal and textual features in social media. In [3], for example, the authors analyzed the locational and temporal distributions of Flickr tag usage to detect bursty tags in a given time window, employing a wavelet transform to suppress noise. Afterwards, the tags were clustered into events such that each cluster consists of tags with similar geographical distribution patterns and with mostly the same associated photos. Finally, photos corresponding to each detected event were extracted by considering their related location, time and tags. Becker et al. [1] represent an event as a cluster of social media documents related to that event. To detect events, they clustered social media documents based on their textual, time and location similarity features. They used a classifier with these similarity scores as features to predict whether a pair of documents belongs to the same cluster. To train the classifier, known clusters of social media documents were used, which were constructed manually and by using the Upcoming database. When the probability that a document belongs to an existing cluster is smaller than a threshold, a new cluster is generated for this document. Becker et al. [2] introduced an additional step which classifies the clusters corresponding to candidate events as 'event' or 'non-event' based on e.g. the burstiness of the most important words in the clusters and the coherence of the content of the social media documents in the cluster. Using the methodology described in [1, 2], the authors were able to detect events using Flickr and Twitter data. Their methodology was evaluated in [1] by comparing the detected photo clusters and the photo clusters collected from the Upcoming dataset. The approach from [9] improved the approach of Becker et al. by only using the k nearest clusters and by using a classifier to determine if a document belongs to an existing cluster or a new cluster.

Initial research has been performed to discover the semantic type of an event using social media. The methodology introduced in [6], for instance, consists in classifying Flickr photos into different event types using their tags, description and title. For this purpose, a Naive Bayes classifier was trained on photos associated with events of known types. Yao et al. [16] detected events using the tagging history of the social bookmarking webservice Del.icio.us. The authors organized the detected events by mapping them to a hierarchy of semantic types, i.e. an automatically generated taxonomy extracted from the same tag space from which bursty events were detected [5]. The detected events were then mapped to an appropriate type at a suitable level based on the coverage of tags of the event in the subtree of the type. The task of classifying photos was also considered in the social event detection challenge at MediaEval 2013 [10]. In particular, photos had to be classified into 'event' and 'non-event' and into event types. Training data was collected using the Instagram API and retrieved photos were manually labeled into event types such as conferences, protests and sport events. Participants mainly used textual features such as the tags, title and description of the photos. Some participants enriched these textual features using e.g. a mapping to Wordnet or by extracting latent topics. In addition, the visual information of the photos were sometimes considered as additional features. However, no participants considered location features to improve the classification performance.

3. METHODOLOGY

The objective of this paper is to analyze the impact of including location features for learning the semantic type of events. In the following, we assume that a training set K is available, containing events with a known semantic type, together with a list of associated Flickr photos. In this paper, the Upcoming event database is used to construct K as described in Section 4.1. Additionally, we consider set U containing events whose semantic type our method will try to estimate. This set contains for instance known events with an unknown semantic type, or events which are automatically extracted from social media and therefore have no associated type.

As mentioned, an event is represented as a set of Flickr photos related to that event and an associated semantic type, wich is similar to the representation used in e.g. [1, 2, 9]. The set of all events is called $E = K \cup U$, the set of photos that are associated with event $e \in E$ is denoted by D_e, and the set of event types associated with e is denoted by $T_e \subseteq T$ where T is the set of all considered event types. Note that an event may have more than one type. For example, an event where a person gives a lecture about art can be classified as both 'education' and 'art'.

In Section 3.1, we describe a number of different ways in which the geographic location of an event can be estimated based on its associated Flickr photos. How these locations can be used to describe the events is explained in Section 3.2. Each of these descriptions is then used to classify the events in U, using an ensemble of classifiers. The output of these classifiers is finally combined to estimate the types of the events in U. More details about the classification framework we used can be found in Section 3.3.

3.1 Locations of Events

To estimate the locations of an event $e \in E$, we use the geographic coordinates of the photos in D_e, where available; we denote this set of coordinates by O_e. When O_e is empty, we consider the location of the event as unknown. We consider three approaches to estimate the location of a given event e. The first approach considers the geometric median of the coordinates in O_e as the location of the event e, denoted by $L_e = \{l\}$. The set of all detected locations associated with events in the training set K is denoted by L. In this approach, we assume that an event has only one location. Therefore, the weight $w(l)$ of the location $l \in L_e$ is set to 1. This approach is called 'median location'. However, photos which are associated with an event may be taken at different locations. For instance, a major sport event like the Olympics takes place across several venues. To estimate different locations for the same event, we use meanshift clustering [4] on the coordinates in O_e. The mean shift $m_b(o)$ of coordinate $o \in O_e$ is given by the difference of coordinate o and the weighted mean of the coordinates in O_e nearby o :

$$m_b(o) = \frac{\sum_{dist(o,o') \leq 2 \cdot b} G_b(o, o') \cdot o'}{\sum_{dist(o,o') \leq 2 \cdot b} G_b(o, o')} - o \qquad (1)$$

with b the bandwidth parameter which is set to 2.5, $dist(o, o')$ the geodesic distance in kilometers between coordinate o and o', and $G_b(o, o')$ the kernel function which determines the weight associated with coordinate o' depending on its dis-

tance to o. We use a Gaussian kernel for a smooth density estimation:

$$G_b(o, o') = e^{-\frac{dist(o,o')^2}{2 \cdot b^2}} \quad (2)$$

The mean shift procedure then computes a sequence starting from all initial coordinates $o_1 \in O_e$ where

$$o_{i+1} = o_i + m_b(o_i) \quad (3)$$

which converges to a location that corresponds to a local maximum of the underlying distribution as $m_b(o_i)$ approaches zero. In this second approach, the coordinates of the center of all the clusters are considered as the locations of event e. We denote the set of these locations by $L_e = \{l_1, l_2 \dots l_k\}$. The weight $w(l_i)$ of location $l_i \in L_e$ is taken as the percentage of coordinates from O_e that are clustered to location l_i. This approach is called 'meanshift all'. In the third approach, called 'meanshift top', we assume that an event only takes place at one location and that photos which were taken far from this location are noise. Therefore, in this approach, the coordinates of the center of the cluster containing most coordinates from O_e is considered as the location of the event e. This location is denoted by $L_e = \{l_1\}$ and the weight $w(l_1)$ of $l_1 \in L_e$ is set to 1.

3.2 Descriptions of Events

In this section, we present different kinds of feature vectors to describe events.

3.2.1 Baseline: Bag-of-Words

As a baseline approach to describe the events we use the textual content of the associated Flickr photos. The textual content associated with a photo consists of a set of tags, a title and a description. In previous work, the textual content of social media documents has already been used to classify events [6, 16]. In this 'bag-of-words' approach, a vector describing an event $e \in E$ is constructed, whose components are associated with a word that appears in dictionary W. This dictionary W is the set of all terms from the textual meta-data associated with the photos (i.e. their tags, title, and description) of the events in the training set K. For feature vector V_e^b of event e, the component $comp_w^b$ associated with word $w \in W$ is given by its number of occurrences in D_e:

$$comp_w^b = \sum_{d \in D_e} |d_w| \quad (4)$$

with $|d_w|$ the number of times photo $d \in D_e$ contains word w. Finally, we use the Euclidean norm to normalize these feature vectors. The set of all the bag-of-words feature vectors corresponding to the events in K is denoted by $V^b(K)$.

3.2.2 Nearest Events

If we know the type of some events which have taken place near the location of the considered event e, then this might be used as further evidence about the type of e. For example, when a lot of music events were organized nearby e, it is more likely that e is also a music event. For the nearest event based feature vector V_e^n of event e, the component $comp_t^n$ associated with event type $t \in T$ is given by the Gaussian-weighted number of nearby events of type t:

$$comp_t^n = \sum_{l \in L_e} \sum_{\substack{l' \in L_t \setminus L_e \\ dist(l,l') \leq 2 \cdot \sigma}} w(l) \cdot w(l') \cdot e^{-\frac{dist(l,l')^2}{2 \cdot \sigma^2}} \quad (5)$$

with σ the standard deviation, $dist(l, l')$ the geodesic distance in kilometers between location l and l', and L_t the locations from L which are associated to an event of type t. Set L_e contains the locations of event e and are obtained using the 'median location', 'meanshift top' or 'meanshift all' approach described in Section 3.1. Instead of using a Gaussian weighting, we also consider the following alternative, in which the k nearest events are considered for a fixed k, each being weighted based on their distance to the event:

$$comp_t^n = \sum_{l \in L_e} \sum_{l' \in N_{k,l,t}} w(l) \cdot w(l') \cdot \frac{1}{1 + dist(l, l')} \quad (6)$$

with set $N_{k,l}$ containing the k locations from $L \setminus L_e$ which are closest to l, and $N_{k,l,t}$ contains the locations from $N_{k,l}$ which are associated to an event of type t. Finally, we use the Euclidean norm to normalize these feature vectors. The set of all the nearest event based feature vectors corresponding to the events in K is denoted by $V^n(K)$.

3.2.3 Nearest Documents

This type of feature vector is inspired by the approach described in [13], which uses the tags of Flickr photos taken nearby a place to discover its semantic type. Our assumption is that the textual content of all Flickr photos taken in the vicinity of an event may provide evidence about its type. In contrast to the photos in D_e, there is no guarantee that these nearby photos are associated to the event itself. For instance, the photos may have been created years before the event took place. However, these nearby photos may contain words which relate to the place type of the venue of the event, the types of the events organized in the past at that place, etc. This information can then be used to discover the semantic type of the event.

We consider F as a large set of Flickr photos. Using the textual content of the photos in F which have been created nearby the location of the events, we describe an event e as a feature vector V_e^f. Similar as in Section 3.2.2, we consider the 'median location', 'meanshift top' and 'meanshift all' approaches described in Section 3.1 to estimate the location of the event. Each component of this vector is associated with a term from the dictionary W^f. This dictionary W^f is the set of all the tags of the photos which have been taken nearby events in the training set. In the first representation, component $comp_w^f$ associated with term $w \in W^f$ is given by the Gaussian-weighted number of times a nearby photo contains w:

$$comp_w^f = \sum_{l \in L_e} \sum_{\substack{d \in F \\ dist(l,d) \leq 2 \cdot \sigma'}} w(l) \cdot |d_w| \cdot e^{-\frac{dist(l,d)^2}{2 \cdot \sigma'^2}} \quad (7)$$

with $dist(l, d)$ the geodesic distance in kilometers between location l and the coordinates of the photo $d \in F$, and $|d_w|$ the number of times photo $d \in F$ contains term w. For the second representation, the component $comp_w^f$ is given by:

$$comp_w^f = \sum_{l \in L_e} \sum_{d \in N'_{k',l}} w(l) \cdot |d_w| \cdot \frac{1}{1 + dist(l, d)} \quad (8)$$

with set $N'_{k',l}$ containing the k' photos from F which are closest to l. Finally, we use the Euclidean norm to normalize these feature vectors. The set of all the nearest documents based vectors of to the events in K is denoted by $V^f(K)$.

3.3 Classification Framework

For each type of feature vector from Section 3.2, we learn a separate classifier. Each type of feature vector described in Section 3.2 is used to classify the events in U. The output of these classifiers is then combined to estimate the semantic types of the events in U. To achieve this, we use a method which is based on the stacking framework introduced by Wolpert [15] and Ting and Witten [12].

In the first phase, a set of learning algorithms L^b, L^n, L^f is selected, one for each feature vector described in Section 3.2. A learning algorithm is a function which maps a set of training items (i.e. feature vectors) to a classifier. The optimal learning algorithm for each vector is selected using 5-fold cross-validation on the training set K (see Section 5.2). For each type of feature vector $x \in \{b, n, f\}$, a base-level classifier C^x is trained on $V^x(K)$ using learning algorithm L^x, i.e. $C^x = L^x(V^x(K))$. Using this classifier, we can classify each event e from set U using its associated feature vector V_e^x. We denote the resulting classification for event e by $pred^x(e)$, and the confidence that e belongs to type $t \in T$ is denoted by $conf^x(t|e)$.

In the second phase, a meta-level classifier is learned that combines the outputs of the base-level classifiers. To generate a training set for learning the meta-level classifier, a 5-fold cross-validation procedure on the training set K was used. We train each of the base-level classifiers using 80% of the training set K. We then use the learned classifiers to classify the remaining 20% of the training data. Repeating this process five times results in predictions $pred^x(e)$ and $conf^x(t|e)$ for each event e in K, each type of vector x and each event type $t \in T$. Similar as proposed in [12], the meta-level feature vector V_e^m is then constructed by combining the $conf^x(t|e)$ values for each $x \in \{b, n, f\}$ and $t \in T$. We can also use the $pred^x(e)$ values as described in [15] or both the $pred^x(e)$ and $conf^x(t|e)$ values. Initial experiments have shown that these alternatives yield worse results, which is why we do not consider them in the remainder of the paper. Finally, a classifier C^m is trained on vector set $V^m(K)$ using a learning algorithm L^m, i.e. $C^m = L^m(V^m(K))$. For each event $e \in U$, this classifier is then used to estimate its type $pred(e)$ and the confidence that it belongs to semantic type $t \in T$, denoted by $conf(t|e)$.

4. DATA ACQUISITION AND PREPROCESSING

To obtain training and test data, we have collected event information from the Upcoming event database (set E), together with the Flickr photos associated with each of these events. In addition, we crawled a large set of Flickr photos which is used to calculate the nearest-documents features of the events (set F). All data which is needed to replicate the experiments has been made publicly available.[2] We now explain these steps in more detail.

4.1 Events and associated Documents

Similar to [1], we based our ground truth on the Upcoming event database. The Upcoming database contains information about a large set of events. For each event, it stores an ID, an event type and references to a set of Flickr photos associated with the event. In addition, these Flickr photos contain the ID of their associated Upcoming event as one of their tags. Using the Flickr API, we first collected all photos which are tagged with an event ID from the Upcoming database. In this way we obtained 373 494 Flickr photos which were taken between January 1, 2000 and April 30, 2013 and which are associated with 22 290 events. Note that one photo may be associated with more than one event, e.g. a photo may be associated with an event such as a conference and one of its subevents such as the social dinner. Second, we retrieved the semantic types of the collected events from the Upcoming database. The 2 670 events (12%) with an unknown semantic type were removed. Finally, events with the same set of associated documents were considered as duplicates and only one of these events was retained in our dataset. As a result of this process, we obtained 16 469 events with a known type (set E) and 347 320 Flickr photos which are associated with at least one of these events. We collected the tags, title, description, user, creation date and geographic coordinates of the photos, where available. In particular, for 40% of the photos in our dataset, geographic coordinates were available. This means that 35% of these events have at least one associated photo which contains geo-coordinates. The considered types and the number of examples of each type in our dataset can be found in Table 1. Note that the sum of the number of events per type (20 647) is larger than the total number of obtained events (16 469) because one event may have more than one type. Finally, the dataset of events has been split in two parts: 5/6th of the dataset was used as training data (called the training set, K) and 1/6th was used for testing (called the test set, U). For a fair evaluation, we ensured that no Flickr photos were associated with both an event in the training set and an event in the test set.

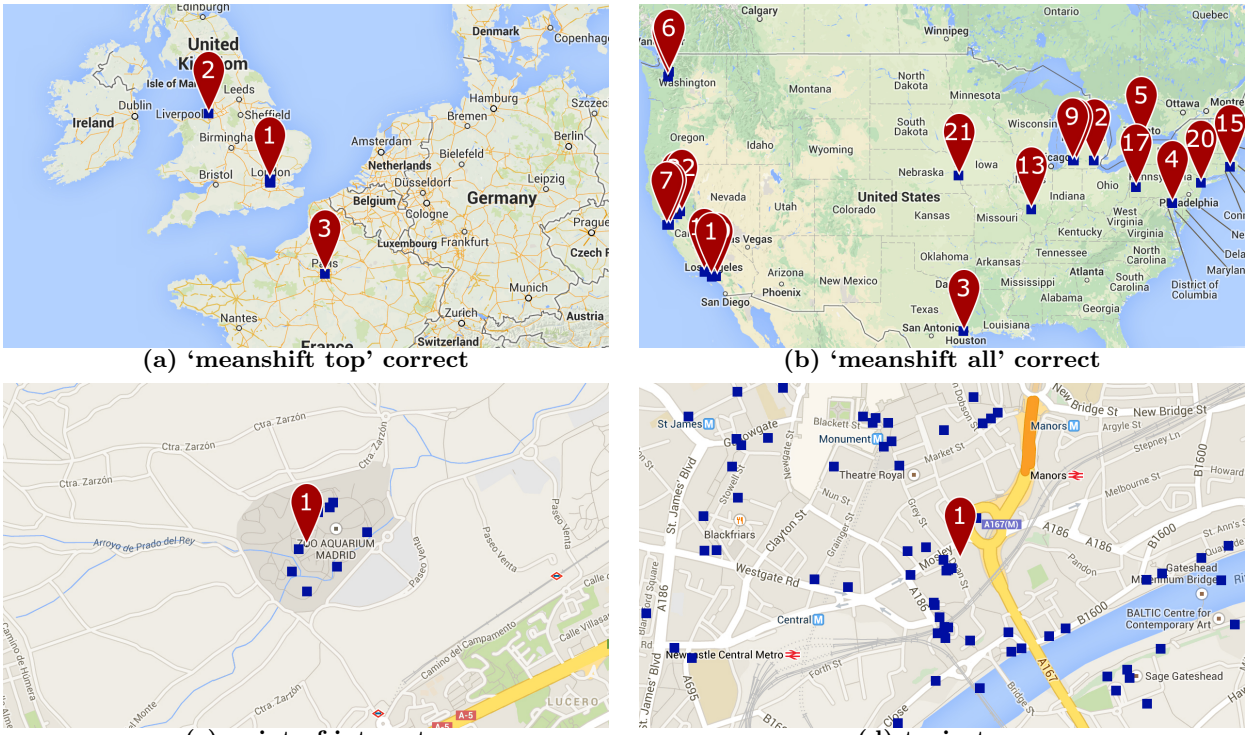
Table 1: Number of events per type.

event type	#events
Music	6401
Social	4571
Performing Arts	2412
Education	1726
Festivals	1149
Community	886
Sports	767
Family	600
Comedy	544
Commercial	543
Media	540
Conferences	209
Technology	171
Politics	128

4.2 Flickr Photos

We crawled an additional set of Flickr photos, called set F. In particular, we crawled the tags, user, creation date and geographic coordinates of around 70% of the georeferenced photos from the photo-sharing site Flickr that were taken before April 2014 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 [11]. In addition, photos with invalid coordinates or without tags were removed. The dataset thus obtained contains 56 660 850 geotagged photos. It is used to calculate the nearest-documents features of the events.

[2] https://github.com/semantictype/data/



(a) 'meanshift top' correct (b) 'meanshift all' correct

(c) point of interest (d) trajectory

Figure 1: Estimated event locations using meanshift clustering.

Table 2: Number of events per number of locations found by the meanshift clustering approach.

#locations	#events	#locations	#events
0	10776	7	3
1	5441	9	2
2	190	10	1
3	30	11	1
4	9	14	1
5	6	17	1
6	7	23	1

5. EVALUATION

We first give some statistics and examples of the introduced location estimation approaches which are performed on the collected Upcoming dataset. Subsequently, we describe how we determine the optimal learning algorithms and event representations using 5-fold cross-volidation on training set K. Finally, we use test set U to examine to what extent the classification performance increases when location features are used in addition to textual features.

5.1 Locations of Events

We described different approaches to estimate the location of an event in Section 3.1. In this section, we illustrate the result of applying these methods on the set E of collected Upcoming events. Table 2 shows a histogram of how many locations were found for the events. As mentioned in Section 4.1, 5693 out of the 16469 events in E (35%) have at least one associated photo which contains geographic coordinates. In other words, an event location and thus its nearest events and documents features can be estimated for only 35% of the considered events. Therefore, we also experimented with automated methods for estimating the coordinates of Flickr photos in D_e for each considered event e based on their tags [14]. However, initial experiments did not yield better results. As can be concluded from Table 2, only for 1.5% of the events in E more than one cluster is found by the meanshift method. Thus, changing the location estimation approach will only have effect on the nearest events and nearest documents features of a small number of the events (see Section 5.3).

Plots of the proposed meanshift clustering for four events are shown in Figure 1. For an event e, the dots indicate the geographic coordinates of its associated photos in D_e. The markers indicate the center of the clusters found by the meanshift clustering approach, ordered based on the number of associated photos. Figure 1(a) shows the photos and estimated locations of the Yahoo! BBC Hackday 2007 event at London (Upcoming id 173371). 33 out of the 35 associated photos are indeed taken at the venue of the event (number 1). However, some participants took photos of event items at their home location. Thus, for this event, the estimated location with most associated photos is the real venue of the event. This location is extracted by the 'meanshift top' approach. On the other hand, all the detected locations for Upcoming event with id 472136 indicate true locations of the event (Figure 1(b)). In particular, 23 locations are detected using the 'meanshift all' approach, of which 20 are located in the USA. This event is called 'the day of the donut' and was held on April 16, 2008. At this day, people came together at different restaurants, pubs, bakeries and shops to share and eat donuts.

In our approach, we always assume that events are held at one or more points of interest. For instance, the Madrid Flickr meet was held on July 10, 2008 at the zoo of Madrid (Upcoming id 865742). The locations of the photos taken at

Table 3: Optimal learning algorithms for each type of feature vector.

feature vector	learning algorithm
Bag-of-Words	L2-regularized L2-loss SVM (dual) [8]
Nearest Events (5)	L2-regularized logistic regression (primal) [8]
Nearest Events (6)	L2-regularized logistic regression (primal) [8]
Nearest Documents (7)	L2-regularized L1-loss support vector classification (dual) [8]
Nearest Documents (8)	L2-regularized L1-loss support vector classification (dual) [8]
Meta-level Classifier	L2-regularized logistic regression (primal) [8]

Table 4: Optimal parameters (par) and related average classification accuracy (ACA) for different nearest-events and nearest-documents representations using cross-validation on the training set.

	median location		meanshift top		meanshift all	
	par	ACA	par	ACA	par	ACA
Nearest Events (5)	$\sigma = 0.440$	44.07	$\sigma = 0.440$	44.23	$\sigma = 0.370$	44.36
Nearest Events (6)	$k = 5$	45.85	$k = 6$	**45.99**	$k = 6$	**45.99**
Nearest Documents (7)	$\sigma' = 0.038$	47.09	$\sigma' = 0.039$	47.29	$\sigma' = 0.047$	47.38
Nearest Documents (8)	$k' = 16$	47.80	$k' = 16$	47.97	$k' = 16$	**48.05**

this event are plotted in Figure 1(c). However, some events in our Upcoming dataset are not held at a fixed point of interest. Figure 1(d), for instance, shows the locations where the photos of the UK Flickr Meet were taken (Upcoming id 1827864). A group of photography enthusiasts took a walk in the Tyne and Wear county of England, and took photos at different locations during that walk. In this case, the location of the event takes the form of a trajectory, rather than a point of a fixed set of (disjoint) points.

5.2 Optimal Learning Algorithms

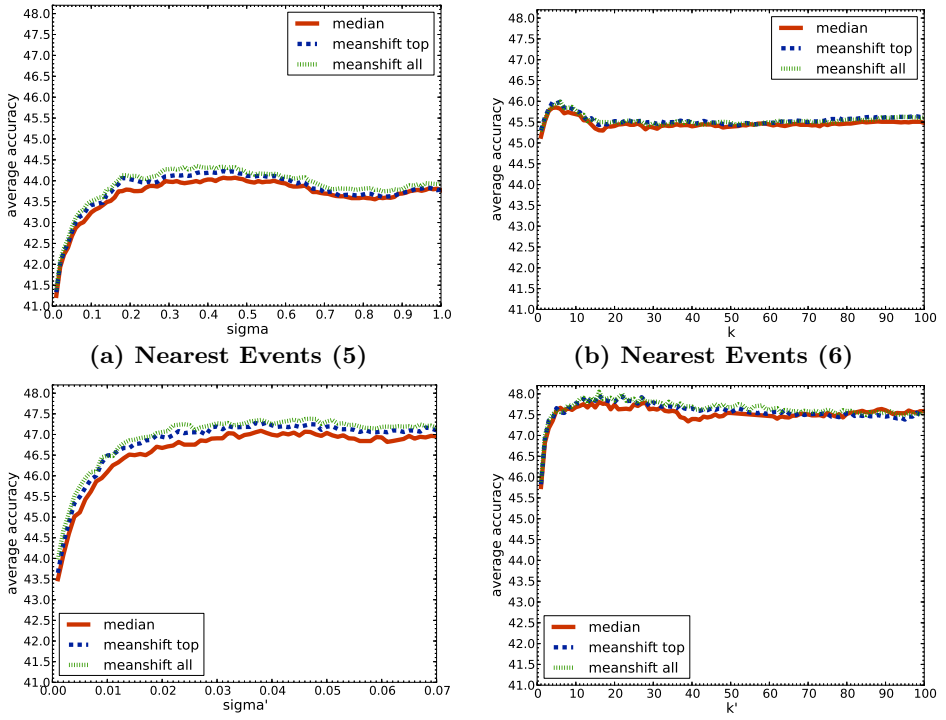
We used 5-fold cross-validation on the training set to find the learning algorithms which optimize the classification accuracy. In particular, the training dataset K was randomly partitioned in five equally sized subsets. The following process was repeated 5 times. Each time, one of the five subsets was used for validation (set K_v) and the remaining four sets were used to construct training set K_t. We trained a classifier using set K_t, which was then used to classify the events of K_v and to calculate its classification accuracy. The settings which optimized the average accuracy of the five folds were found by repeating this cross-validation approach for several learning algorithms. As candidate learning algorithms, we considered all methods implemented in WEKA [7] as well as the Support Vector Machine (SVM) implementations of LibLinear [8]. We used the standard configurations of the learning algorithms, both for WEKA and LibLinear. The learning algorithms which were obtained from the training set can be found in Table 3.

5.3 Optimal Event Representations

We also used the 5-fold cross-validation process to determine the optimal nearest-events and nearest-documents representations, as described in respectively Section 3.2.2 and 3.2.3. As mentioned, we consider three approaches to estimate the location of an event, called 'median location', 'meanshift top' and 'meanshift all'. Additionally, two types of feature vector representations has been considered, one based on a Gaussian distribution (5) (7) and another based on the k nearest neighbours of the event (6) (8). The average accuracies for different parameter values of the six considered nearest-events representations can be found in Figure 2(a,b), and for the nearest-documents representations in Figure 2(c,d). The optimal parameter values and their associated average accuracy values can be found in Table 4.

We first discuss the performance of the different nearest-events representations described in Section 3.2.2. Figure 2(a) shows the average accuracies for different σ values when the Gaussian-weighted features are used (5). We can observe that the average accuracy increases when σ increases from 0.010 to 0.200, and stagnates when a larger σ value is used. As (5) only considers nearby events located at a maximum of $2 \cdot \sigma$ kilometers of the given event e, this means that events up to 400 meters of e tend to be relevant for determining its semantic type. The average accuracies when using different k values for the k nearest neighbours of an event used in (6) are shown in Figure 2(b). The average accuracy increases between $k = 1$ and $k = 6$, then decreases until $k = 12$ and afterwards stagnates. To further compare the performance of each considered nearest-events representation, we look at their average classification accuracy when their optimal parameters are used (Table 4). For each location estimation approach, the representation from (6) significantly outperforms the representation from (5) (sign test, $p < 0.001$). The average accuracy for 'meanshift top' and 'meanshift all' is significantly higher than for 'median location' (sign test, $p < 0.001$) when (5) is used, but the difference in accuracy is not significant in case of (6). Note that one reason why the use of a clustering method instead of the median location is limited is because only for 1.5% of the events in the training data more than one cluster is found by the meanshift method. We get no significant difference when 'meanshift top' or 'meanshift all' are used for both (5) and (6) (sign test, $p > 0.05$). As the 'meanshift all' location estimation in combination with (6) gives the best average accuracy, we will use this representation in the rest of the paper.

The average accuracies for the considered nearest documents representations are shown in Figure 2(c,d). Similar as for the nearest-events vectors, the average classification accuracy first increases when σ' of the Gaussian-weighted features (7) increases, after which it stagnates. However, the turning point for the nearest-documents representations ($\sigma' = 0.025$) is much smaller than for the nearest-events representations ($\sigma = 0.200$). One of the reasons is that the set of potential nearest documents (set F, $|F| = 56.7$ million) is much larger than the set of potential nearest events (set K, $|K| = 13\,725$). Therefore, a lot of photos are taken near most events, which means that even with a small σ' value enough information can usually be obtained. Figure 2(d) shows the average accuracies when (8) is used. The



(a) Nearest Events (5)

(b) Nearest Events (6)

(c) Nearest Documents (7)

(d) Nearest Documents (8)

Figure 2: The average accuracy for different nearest-events and nearest-documents representations.

average accuracy increases substantial between $k' = 1$ and $k' = 5$, and is optimal for $k' = 16$. The average classification accuracy for each considered nearest-documents representation is shown in the last two columns of Table 4. In each case, we assume that the optimal parameters are used. Similar to the nearest-event representations, the representation from (8) significantly outperforms the representation from (7) (sign test, $p < 0.001$). For both (8) and (7), 'meanshift all' and 'meanshift top' performs significantly better than 'median location'. However, there is no significant difference between 'median location' and 'meanshift top' (sign test, $p > 0.05$). As the 'meanshift all' location estimation in combination with (8) gives the best average classification accuracy, we will use this representation in the rest of the paper.

5.4 Experimental Results

The task we consider in this section is to estimate the semantic type of the events in test set U. Similar as in the previous section, the accuracy metric is considered to determine if the difference in quality of the classifications are statistically significant when different approaches are used. However, the differences in accuracy are sometimes limited because the test set is imbalanced. Even a naive classifier returning the most occuring category ('Music') achieves an accuracy of 38%, for instance. Therefore, the average precisions of the events from U which are ranked based on the confidence $conf(t|e)$ that they belong to type t are also considered. Tables 5 and 6 summarize the result of our evaluation on test set U.

We observe that the average precision when using all the proposed feature vectors is always higher than for the baseline. The mean average precision increases from 30% to 44%

Table 5: Classification accuracy and mean average precision (MAP) per feature vector type.

feature vector type	accuracy (%)	MAP (%)
Bag-of-Words (baseline)	65.38	30.21
Nearest Events	46.68	15.80
Nearest Documents	48.47	16.11
Bag-of-Words + Nearest Events	66.95	43.67
Bag-of-Words + Nearest Documents	65.93	40.51
All Features	**67.46**	**43.81**

and the accuracy from 65.4% to 67.5% (sign test, $p < 0.001$). Furthermore, the average precision improves substantially for relatively rare event types (e.g. 'family', 'comedy', 'commercial', 'conferences' and 'technology'). For instance, the average precision for conferences increases from 4% to 30% when locations features are used. For instance, the baseline approach was unable to discover that a skateboard race event in our test set (Upcoming id 318498) was of type 'sport' because its associated tags were not sufficiently informative. However, the photos taken close to the event contain words such as 'ferrari' and 'race' which may indicate that the event was held on a race track. Together with the information that all known nearby events are of type 'sport', the ensemble learner was able to discover the correct type.

Based on the classification accuracies in Table 5, we can conclude that the bag-of-words representation leads to the best individual classification accuracy. However, the use of the known type of the nearest events improves the MAP score of the baseline with more than 13 percentage points and significantly improves the classification accuracy (sign test, $p < 0.001$). A similar observation can be made when the text of the nearest photos is used. Comparing the performance when nearest events or nearest documents features are used, we observe that the use of nearest documents features leads to the highest individual classification

Table 6: Average precision per event type and feature vector type.

event type	bag-of-words	nearest events	nearest docs	bag-of-words + nearest events	bag-of-words + nearest docs	all features
Music	85.98	49.57	64.30	86.92	85.16	**87.12**
Social	68.14	38.02	46.61	73.75	72.37	**73.81**
Performing Arts	42.69	23.07	25.72	57.01	56.67	**57.12**
Education	40.86	17.79	17.93	52.20	51.50	**52.22**
Festivals	23.31	12.21	8.63	41.48	38.96	**41.61**
Community	18.63	11.94	10.99	28.79	24.55	**28.82**
Sports	55.64	14.34	20.12	71.53	70.06	**71.87**
Family	10.58	13.21	9.17	**23.28**	23.13	23.25
Comedy	23.49	7.49	5.93	**47.98**	39.95	47.96
Commercial	14.26	13.41	5.65	35.89	29.71	**36.44**
Media	23.42	5.33	4.56	**36.37**	31.01	36.28
Conferences	4.36	3.03	3.03	29.77	21.06	**30.46**
Technology	9.20	6.35	1.84	17.70	14.34	**17.76**
Politics	2.42	5.43	1.12	**8.78**	8.72	8.74

accuracy and MAP score (sign test, $p < 0.001$). However, when one of these location-based features are combined with the bag-of-word features, we obtained the best classification performance for the 'bag-of-words + nearest events' combination (sign test, $p < 0.001$). Finally, we note that the classification accuracy is significantly better when the bag-of-words, nearest-documents and nearest-photos vectors are used ('all features') compared to only using the bag-of-words and nearest-events vectors or only using the bag-of-words and nearest-documents vectors (sign test, $p < 0.001$).

6. CONCLUSIONS

In this paper, we have proposed a methodology to discover the semantic type of events, using the meta-data of their associated Flickr photos. In addition to textual features, we also considered context information derived from the location of the event. In particular, we first estimated the location of the event based on the coordinates of its associated Flickr photos. Then, we used this location to find nearby Flickr photos (not necessarily associated with the event) as well as nearby events. Combining this knowledge improved the standard bag-of-words approach significantly. In future work, we will explore the use of additional geographical features such as the nearby points of interests.

7. 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 some of the data we have used in this paper.

8. REFERENCES

[1] H. Becker, M. Naaman, and L. Gravano. Learning similarity metrics for event identification in social media. In *Proc. of the 3rd ACM Int. Conf. on Web Search and Data Mining*, pages 291–300, 2010.

[2] H. Becker, M. Naaman, and L. Gravano. Beyond trending topics: Real-world event identification on Twitter. In *Proc. of the 5th Int. AAAI Conf. on Weblogs and Social Media*, pages 438–441, 2011.

[3] L. Chen and A. Roy. Event detection from Flickr data through wavelet-based spatial analysis. In *Proc. of the 18th ACM Conf. on Information and Knowledge Management*, pages 523–532, 2009.

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

[5] B. Cui, J. Yao, G. Cong, and Y. Huang. Evolutionary taxonomy construction from dynamic tag space. *World Wide Web*, 15(5):581–602, 2012.

[6] C. Firan, M. Georgescu, and W. Nejdl. Bringing order to your photos: Event-driven classification of Flickr images based on social knowledge. In *Proc. of the 19th ACM Int. Conf. on Information and Knowledge Management*, pages 189–198, 2010.

[7] M. Hall, H. National, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software : An update. *SIGKDD Explorations*, 11(1), 2009.

[8] S. Keerthi, S. Sundararajan, and K. Chang. A sequential dual method for large scale multi-class linear SVMs. In *Proc. of the 14th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pages 408–416, 2008.

[9] T. Reuter and P. Cimiano. Event-based classification of social media streams. In *Proc. of the 2nd ACM Int. Conf. on Multimedia Retrieval*, page 22, 2012.

[10] T. Reuter, S. Papadopoulos, G. Petkos, V. Mezaris, Y. Kampatsiaris, P. Cimiano, C. de Vries, and S. Geva. Social event detection at MediaEval 2013: Challenges, datasets, and evaluation. In *Proc. of the MediaEval 2013 Multimedia Benchmark Workshop*, 2013.

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

[12] K. M. Ting and I. H. Witten. Issues in stacked generalization. *Journal of Artificial Intelligence Research*, 10:271–289, 1999.

[13] S. Van Canneyt, S. Schockaert, and B. Dhoedt. Discovering and characterizing places of interest using Flickr and Twitter. *Int. Journal on Semantic Web and Information Systems*, 9(3):77–104, 2013.

[14] O. Van Laere, S. Schockaert, and B. Dhoedt. Georeferencing Flickr resources based on textual meta-data. *Information Sciences*, 238:52–74, 2013.

[15] D. H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–260, 1992.

[16] J. Yao, B. Cui, Y. Huang, and Y. Zhou. Bursty event detection from collaborative tags. *World Wide Web*, 15(2):171–195, 2012.