MediaEval Benchmark: Social Event Detection in Collaborative Photo Collections

Markus Brenner, Ebroul Izquierdo
School of Electronic Engineering and Computer Science
Queen Mary University of London, London E14NS, UK
{markus.brenner, ebroul.izquierdo}@eecs.qmul.ac.uk

ABSTRACT
In this paper, we present an approach to detect social events in collaboratively annotated photo collections as part of the MediaEval Benchmark. We combine various information from tagged photos with external data sources to train a classification model. Experiments based on the MediaEval Social Event Detection Dataset demonstrate the effectiveness of our approach.

Categories and Subject Descriptors
H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Systems**]: Information Storage and Retrieval

General Terms
Design, Experimentation, Performance

Keywords
Benchmark, Photo Collections, Classification, Event Detection

1. INTRODUCTION
The Internet enables people to host, access and share their photos online, e.g. through websites like Flickr. Collaborative annotations and tags are commonplace on such services. The information people assign vary greatly, but often seem to include some kind of references to *what* happened *where* and *who* was involved. In other words, such references describe observed experiences or occurrences that are simply referred to as *events* [7]. In order to enable users to explore such events in their photo collections, effective approaches to detect events and group corresponding photos are needed. The MediaEval Social Event Detection (SED) Benchmark [4] provides a platform to compare different such approaches.

1.1 Background
There is increasing research in the area of event detection in web resources in general. The subdomain we focus on is photo websites, where users can collaboratively annotate their photos. Recent research like [1] put emphasis on detecting events from Flickr photos by primarily exploiting user-supplied tags. [6] and [5] extend this to place semantics, the latter incorporating the visual similarity among photos as well. Our aim, however, is to also use information from external sources to find photos corresponding to the same events. [2] is an example that goes further in our direction by exploiting Wikipedia classes.

1.2 Objective
In this work we present an approach where we utilize external sources to detect social events and group applicable photos in collaborative photo collections such as Flickr. The approach is tailored to the two challenges laid out by the MediaEval SED Benchmark: The goal of Challenge I relates to soccer events taking place in two given cities, and that of Challenge II to events at two given (music) venues during a given month.

The remainder of this paper is structured as follows: In the next section we set forth how we gather relevant external information and describe the feature extraction from the photos. Then, we explain the design of our classifier-based approach. Using experiments, we test and discuss the overall framework and present our conclusions.

2. GATHERING EXTERNAL DATA

2.1 Challenge I: Soccer Matches
Our strategy for detecting soccer events (or matches) is to first find all soccer clubs and associated stadiums for the given cities in the challenge query. We automatically retrieve this information from DBpedia by means of the SPARQL interface. For each soccer club, we also gather its club- and nickname. Similarly, we request alternative names for the stadiums as well as any location information available. For simplicity, we limit ourselves to bigger soccer events by considering only those clubs whose home stadiums have a capacity of at least 20000 people.

To our knowledge, there is no public dataset or web service that provides all-encompassing statistics related to the world of sports. As for soccer only, there are a few dedicated websites, one of which is *playerhistory.com*. The website does not provide an API, and thus, we manually navigate and parse through the webpages to retrieve the date and opposing team of all matches against any of the home teams found earlier on.

2.2 Challenge II: Music Performances
We define a venue as a place (usually with a physical location) at which events can occur. There are web services like Foursquare that compile and maintain venue directories. We use Last.fm, which specializes in music-related venues and events, to retrieve data such as venue location and performances (date and time, title, artists, etc.) associated with the venues given in Challenge II.

2.3 Generic Terms and Location
For each challenge, we compile a list of generic words relating to the challenge. Examples are *goal* or *stadium* for Challenge I, or *music* and *concert* for Challenge II. We utilize both DBpedia and WordNet for the task. Depending on the country the venue is located in, we additionally get corresponding translations via the Google Translate API.

For each venue, we also gather location-centric information like suburb, region and the geographic coordinates. We employ the Google MAP API to query the mentioned information based on initial evidence from DBpedia (Challenge I) and the venue location available through Last.fm (Challenge II).

3. DETECTING IN- AND OUTLIERS
As geo-tagged photos become more and more popular, we can identify photos as belonging and *not* belonging to a venue (and thus an event when also considering the time). Prior to discarding

Copyright is held by the author/owner(s).
MediaEval 2011 Workshop, September 1-2, 2011, Pisa, Italy

all photo outliers from the dataset at this stage, we extract features of them as well as of the inliers. We later incorporate both in a classification process to train appropriate classes.

In general, the date and time a photo was captured is an effective cue to bound the search and classification space. The MediaEval Benchmark defines an event as a distinct combination of location and date (but not time). As such, we can limit our approach to at most one event per day at the same location. Note that we also try to retrieve the event time so that we can further tighten the bound to within a certain margin.

We do not further classify photos which match both venue and time of an event. If we find multiple photos (at least five) that match only a venue's location but do not fall into any of that venue's events (e.g. gathered through external sources), we consider them as part of another *new* event.

4. COMPOSING FEATURES

We compose text features of each photo's title, description, keywords and username (perhaps linking a user's collection). In our training step, we also include the generic terms we compiled previously as well as the event information.

Then, we apply a Roman preprocessor that converts text into lower case, strips punctuation as well as whitespaces and removes accents from Unicode characters. It also eliminates common (stop) words like *and, cannot, you* etc. Moreover, we discard all words that are less than three characters in length. We also ignore numbers and terms commonly associated with photography. Examples are *Canon, Nikon, 80mm* and *IMG_8152*. Finally, photos with less than two words overall are filtered out.

In the next step, we split the words into tokens. The text assigned to photos by users on online services such as Flickr is often not *clean*: Words have spelling errors and different suffixes and prefixes. Furthermore, traditional natural language processing steps, e.g. word-stemming, are often tailored to the English language. To accommodate other languages, we do not apply a word-based tokenizer but a language-agnostic character-based tokenizer (minimum three, maximum seven characters). However, we exclude the username from this step (it is an ID and has no alternative word forms). We also take all preprocessed words in their full and non-tokenized form into account.

We then use a vectorizer to convert the tokens into a matrix of occurrences. To make up for photos with a large amount of textual annotations, we also consider the total number of tokens. This approach is commonly referred to as Term Frequencies (TF).

5. CLASSIFICATION

After composing the features, we train a Linear Support Vector Classifier [3]. Based on brief internal tests, we use a value of 100 for parameter *C* and otherwise recommended default parameters.

For each event, we train a separate classifier. As mentioned earlier, we only consider testing samples falling on the same day according to each event in the prediction step. Basically, we perform binary classification: Photos which are either related or not related to an event. However, introducing a third class reflecting events from the same challenge seems to perform better.

Given the assumption that both challenges are exclusive, we include the features of each other's challenge in the appropriate class label. We aggregate the features of the location in- and outliers into single samples (starting as a set of distinct terms), as it seems to perform better than considering multiple samples (with the same class label).

6. EXPERIMENTS AND RESULTS

We perform experiments on the MediaEval SED Dataset that consists of 73645 Flickr photos with accompanying metadata.

For Challenge I, we identify two soccer clubs (we discard several smaller ones) for each given city. We find and detect a total of twelve events (two if not considering external event sources as outlined in Section 2) at their according venues (stadiums). For Challenge II, we compile a total of 37 events (six without external event sources).

We find about 14300 geographic outliers not associated with any venue (of both challenges), thus substantially reducing the testing candidates while providing a large amount of training samples for the non-relating class.

Certain samples in our experiments suggest that the number of false positives could potentially be reduced by considering terms reflecting geographic places like *Paris* or *London* that do not correspond to an event's venue location. We also notice the special case where the exemplarily term *London* is part of a particular event's title (with its venue being in Amsterdam), and thus, actually leads to numerous incorrect classifications.

In the following table we present our test results (as evaluated by the organizers of the MediaEval Benchmark).

Table 1: Results depending on configuration

Configuration	Challenge I		Challenge II	
	F-Score	NMI	F-Score	NMI
Complete configuration	45.5	0.28	25.9	0.36
Without generic terms	68.7	0.41	33.0	0.50
Without other challenge	60.3	0.38	25.6	0.20
Without outlier features	43.1	0.19	19.0	0.28

As expected, we see a notable performance gain when using geographic outlier features. This is also true for externally sourced events (omitted above). Surprisingly, generic terms have a negative impact (less precision).

7. CONCLUSION

We present an approach to find and detect social events in tagged photo collections. We combine external information with (mostly textual) data extracted from photos to train a classifier. Based on our experiments, we conclude that external information and identified outliers can aid classification, but challenges such as finding and linking structured external data remain. For future experiments, we intend to additionally detect events from the photos' textual annotations as well as include visual features to further improve results.

8. REFERENCES

[1] Chen, L. and Roy, A. 2009. Event detection from Flickr data through wavelet-based spatial analysis. *ACM CIKM* (2009), 523–532.

[2] Firan, C.S. et al. 2010. Bringing order to your photos: Event-driven classification of Flickr images based on social knowledge. *ACM CIKM* (2010), 189–198.

[3] Keerthi, S.S. et al. 2008. A sequential dual method for large scale multi-class linear SVMs. *ACM KDD* (2008), 408–416.

[4] Papadopoulos, S. et al. 2011. Social Event Detection at MediaEval 2011: Challenges, Dataset and Evaluation. *MediaEval 2011 Workshop* (Pisa, Italy, Sep. 2011).

[5] Papadopoulos, S. et al. 2010. Cluster-based landmark and event detection on tagged photo collections. *Multimedia, IEEE*. 99 (2010), 1–1.

[6] Rattenbury, T. et al. 2007. Towards automatic extraction of event and place semantics from Flickr tags. *ACM SIGIR* (2007), 103–110.

[7] Troncy, R. et al. 2010. Linking events with media. *I-Semantics* (2010), 1–4.