Bringing Order to Your Photos: Event-Driven Classification of Flickr Images Based on Social Knowledge

Claudiu S. Firan, Mihai Georgescu, Wolfgang Nejdl, Raluca Paiu
L3S Research Center, University of Hanover
Appelstr. 9a
30167 Hanover, Germany
{firan,georgescu,nejdl,paiu}@l3s.de

ABSTRACT

With the rapidly increasing popularity of Social Media sites, a lot of user generated content has been injected in the Web, thus resulting in a large amount of both multimedia items (music – *Last.fm*, *MySpace.com*, pictures – *Flickr*, *Picasa*, videos – *YouTube*) and textual data (tags and other text-based documents). As a consequence, especially for multimedia content it has become more and more difficult to find exactly the objects that best match the users' information needs. The methods we propose in this paper try to alleviate this problem and we focus on the domain of pictures, in particular on a subset of *Flickr* data. Many of the photos posted by users on *Flickr* have been shot during events and our methods aim to allow browsing and organization of picture collections in a natural way, by events. The algorithms we introduce in this paper exploit the social information produced by users in form of tags, titles and photo descriptions, for classifying pictures into different event categories. The extensive automated experiments demonstrate that our approach is very effective and opens new possibilities for multimedia retrieval, in particular image search. Moreover, the direct comparison with previous event detection algorithms confirm once more the quality of our methods.

Categories and Subject Descriptors

H.3.2 [**Information Storage and Retrieval**]: Metadata; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation, Human Factors, Measurement, Reliability

Keywords

Event Detection, Event Classification, Collaborative Tagging, Machine Learning, Metadata Enrichment

CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.

1. INTRODUCTION

Over the past years we have faced a rapid shift from analog to digital photography, due both to the increase in performance and to the drop of prices for digital photo cameras. With the advent of Social Media sites dedicated to photography (*Flickr*, *Picasa*, etc.), a lot of these users' digital personal photos become available to the public at large. However, users often do not invest much effort in organizing their own pictures and prefer instead to create quite broad sets including hundreds of pictures. Because of this, a huge amount of digital pictures remains untouched unless powerful techniques for image indexing and retrieval become available. Image retrieval is particularly difficult, given the fact that *Flickr* data is noisy and, besides, it is not easy to capture the content of photos.

The key idea of this paper is to use events as the primary means to organize media and in a more concrete scenario, pictures. Our lives are a constellation of events, which one after another, pace our everyday activities and build up our memories. Many of the *Flickr* pictures have been shot during specific events, therefore enabling users to organize or browse this type of media by events is very natural and can potentially facilitate retrieval.

In our paper we define an event like in [1], as "a specific thing happening at a specific time and place". Moreover, we consider events having both a local and a global dimension. Events such as birthdays, a marriage, a summer vacation or a car accident are the lens through which we see and memorize our own personal experiences and are therefore events of local type. In turn, global events, such as world sport championships or global natural disasters (e.g. 2010 Haiti earthquake, 2004 Thailand tsunami, climate change, world recession, etc.) or, on a smaller scale, a local festival or a soccer match, build collective experiences. These types of events allow users share personal experiences as a part of a more social phenomenon – "collective events".

Our approach for classifying pictures into events relies entirely on user generated information, which we gather from the *Flickr* Web site. We experiment both with simple types of features, such as tags, time information, photo titles and descriptions, as well as with different combinations among them. Based on this information we construct classifiers, which automatically assign the pictures to their associated event categories. The method we introduce is very intuitive and the extensive automatic evaluations we perform demonstrate the high accuracy of our algorithms. Compared to similar work (e.g. [5]) trying to solve the same problem, our approach is much simpler and achieves better results.

Moreover, the applicability of the methods we introduce in this paper is not restricted to pictures in *Flickr*, but can be employed for any types of pictures having tags associated with them, as well as to other types of multimedia data, e.g. videos, music, etc. Additionally, the methods can be applied not only for event-based browsing or organization, but also for enabling users to discover other users interested in the same types of events and thus easing social connectivity.

The contributions of this paper are manifold:

- We propose to solve some of the users' knowledge management problems by providing the means to organize and browse content in a very natural way, based on events.
- We introduce some very intuitive classification based methods, which serve the purpose of multimedia organization.
- We address a new aspect of the event detection task, namely the local and the global dimensionality and discuss the potential of our methods for classifying items corresponding to both event categories.
- Last but not least, we evaluate our algorithms on a large *Flickr* data set, and compare our results with those of other existing methods.

The rest of the paper is organized as follows: we start in Section 2 with a review of the relevant literature; next, in Section 3 we present the data set on which our experiments have been performed, focusing in particular on the collection of events and *Flickr* images. In Section 4 we introduce our algorithm, together with details on the data set clustering (Section 4.1) and preprocessing (Section 4.3) steps. The evaluation of our methods and the experimental results are described in Section 5. Finally, in Section 6 we conclude and discuss possible extensions of this work.

2. RELATED WORK

The topic of *event detection* is not new; first papers addressing this domain appeared already in 1998, as part of the *Topic Detection and Tracking (TDT)* initiative [2]. In [23] the authors introduce two different types of event detection methods: retrospective and online detection. The former refers to discovery of previously unidentified events inside a collection, while the latter strives to identify in real time new events from live news feeds. The experiments show that hierarchical clustering methods are highly informative for retrospective detection of previously unidentified events, while temporal distribution patterns of document clusters provide useful information for improvement in both retrospective detection and online detection of novel events. With the algorithms we propose in the present paper we also target the detection of retrospective events.

Arguing that most of the existing research focusing on retrospective news event detection (RED) make use of only the contents of the news articles, the authors of [17] propose to do explorations on both content and time information and introduce a probabilistic model to incorporate these both sources of information in a unified framework. Similarly, the authors of [12] also utilize both time and content information. However, in contrast to TDT, which attempts to cluster documents as events using clustering techniques, in [12] the focus is on detecting a set of bursty features for a bursty event. The main technique employed in the paper is a free probabilistic approach which fully utilizes the time information to determine a set of bursty features which may occur in different time windows. Both [17, 12] differ from our approach, as we do not perform our event-analysis with respect to time windows. Instead we make use of time information in order to help us decide whether items belong to a specific event class or not.

Similar to [12], in [15] the authors aim to identify feature bursts and their associated bursty periods, by introducing a simple but effective mixture density-based approach. Word trajectories are analyzed in both time and frequency domains, with the specific goal of identifying important and less-reported, periodic and aperiodic events. A set of words with identical trends can be then grouped together to reconstruct an event in a completely unsupervised manner. [11] is also aiming to produce a characterization of the most interesting tags associated with a sliding interval of time. Here however, the focus is not on textual documents, but rather on user generated tags attached to *Flickr* photos. The most important part of the article addresses the problem of visualizing the evolution of tags within the *Flickr* online image sharing community. The authors make a short observation regarding the identified categories of interesting tags, i.e. events, personalities and social media, but no statistics on this are reported.

A different approach for detecting events is presented in [9], where the authors propose to use Web click-through data for this purpose. The click-through data is first transformed to a 2D polar space by considering the semantic and temporal dimensions of the queries. Further, robust subspace estimation techniques are applied in order to detect subspaces consisting of only queries with similar semantics and the uninteresting subspaces containing queries not related to real events are pruned. Finally, events are detected from interesting subspaces using a non-parametric clustering technique.

Most of the existing work on event detection focused on identifying events from news corpus collections, and only recently new methods targeting other types of data have been proposed. For example, in [10] the authors propose an approach for detecting *Flickr* photos depicting events. Given a set of *Flickr* photos with both user tags and other metadata including time and location (latitude and longitude), the algorithm aims to discover a set of photo groups, where each group corresponds to an event. The method consists of three steps: (1) based on temporal and spatial distributions, tags are identified as related to events or not; (2) after detecting event-related tags, they are further classified into periodic- or aperiodic-event tags; (3) finally, for each tag cluster representing an event, the set of photos corresponding to the event are retrieved. This approach differs from ours in that it relies on geographical information – still inexistent for many pictures. Our method uses solely social information and has thus a broader applicability.

Another dimension of investigations refers to the work presented in [19]. Focusing also on the domain of pictures, the paper tries to extract event and place semantics from tags assigned to photos in *Flickr*. The proposed approach relies on bursts analysis: tags referring to event names are expected to exhibit high usage patterns over short time periods (maybe also periodical), while location-related tags show these kinds of patterns in the spatial dimension. However

like [10], [19] also relies on GPS information and has thus a more restricted applicability than our approach.

In [20], the authors also identify event-related tags by using WordNet. However, event detection is not the main focus of the paper, but rather on providing tag recommendations.

The approach presented in [4] is targeting a broader range of data types, namely it tries to identify events and their associated user-contributed social media documents. It thus not only focuses on pictures, but also on music, videos, news and Facebook data. The validation of the accuracy of the introduced algorithms is performed on *Flickr* data, having tags corresponding to entries in the Yahoo!'s Upcoming event database[1]. An extended version of [4] is presented in [5]. [5] discusses in detail the different distinctive representations of social media documents for analyzing their similarity and for identifying which documents correspond to the same events. The authors define similarity methods for each document representation and also explore various techniques for combining them into a single measure of social similarity. Both ensemble and classification-based similarity learning techniques are described and these are used in conjunction with an incremental clustering algorithm to generate a clustering solution where each cluster corresponds to an event and includes the social media documents associated with the event. Compared to [4], in [5] the experiments are much more extensive and classification-based similarity learning methods are discussed in addition to the previously introduced ensemble-based techniques. We also compare our results with the measures reported in [5]. Nevertheless, our methods are purely classification-based, whereas in [5] also clustering techniques are employed. Moreover, in our case the event classes are known beforehand, while in [5] they are determined based on clustering techniques. [7, 8] are similar to the approach we introduce in the present paper, but in this case the focus is on music resources and the authors aim to provide tag recommendations in terms of music themes, moods, genres or styles. Our focus instead is on photo classification and not on tag recommendations.

3. DATA SET DESCRIPTIONS

For the purpose of our experiments we collected an extensive set of *Flickr* pictures. Identifying a good ground truth set for events (in particular for pictures) turned out to be quite difficult. The main reason for this is the lack of a verified and largely accepted event taxonomy (ontology). One of the most known events categorization is the Yahoo! Upcoming events catalog, however it's thirteen categories are not extensive enough. Using WordNet[2] was also not an option in our case, as *Flickr* tags can be written in many different ways (intentionally or unintentionally) by the users and are thus not matchable with the WordNet database. Wikipedia[3] has also its' own event taxonomy[4], but here the categorization is not easily understandable. Besides, several possibilities are listed regarding how to organize events (e.g. by location, topic, year, etc.), and these different categorizations introduce ambiguities, as they are not mutually exclusive and do not cover all possible facets of a good event classification.

To cope with this problem, we decided to make use of the YAGO ontology, which brings together WordNet and Wikipedia knowledge. In the following we present its' most important features and characteristics.

3.1 The YAGO Ontology

YAGO[22][5] is a large and extensible ontology that builds on entities and relations from Wikipedia. Facts in YAGO have been automatically extracted from Wikipedia and unified with semantics from WordNet, achieving an accuracy of around 95%. All objects (e.g., cities, people, even URLs) are represented as entities in the YAGO model. The hierarchy of classes starts with the Wikipedia categories containing a page and relies on WordNet's well-defined taxonomy of homonyms to establish further *subClassOf* relations. We make use of these *subClassOf* relations in YAGO, which provide us with semantic concepts describing Wikipedia entities. We also rely on the *type* relation, deducting what higher level concept a page is about.

3.2 Events Collection

For collecting event names, we made use of the YAGO ontology and selected only those entities having a *type* (YAGO relation) *wordnet_event*. With this method we retrieved a list of 138,794 Wikipedia event page titles, like "Reichstag fire", "Battle of the Nile", "CeBIT", or "Iranian presidential election, 2009". We will refer to this list of events as [*events*] later in our experiments. Starting from [*events*] we retrieved the Wikipedia categories assigned to the Wikipedia pages, i.e. to the events, using the *subClassOf* relation in YAGO. With this method we retrieved a total of 25,223 distinct Wikipedia categories assigned to [*events*] – later called [*categories*]. Furthermore, we also retrieved the WordNet concepts of [*categories*] using again *subClassOf* relations starting from [*categories*]. This set of 1,521 distinct WordNet concepts will be referred as [*concepts*]. Thus we have an extensive list of [*events*], along with the corresponding [*categories*] and the super-[*concepts*], forming a three-level hierarchical taxonomy.

3.3 Flickr Images

Having now an extensive set of event categories, the next step corresponded to gathering the actual ground truth data, consisting of *Flickr* images. For this purpose we made use of the *Flickr* API[6]. We started from the 138,794 Wikipedia event pages and crawled the corresponding *Flickr* groups. Being explicitly created by users and containing pictures contributed by a multitude of users, these *Flickr* groups quite accurately represent social groups interested in the specific events and represent thus a good ground truth. For gathering the *Flickr* groups we made use of the $'flickr.groups.search'$ method and kept the first hit in the results list. With this method we could gather 29,796 event-related *Flickr* groups.

In the next step, for all retrieved groups, we collected their corresponding group pools, i.e. all pictures contributed by all users to the corresponding groups. In total the number of pictures gathered was 3,600,520 and among them 2,639,254 unique pictures.

Finally, for all collected photos, we needed the tag information, i.e. raw and normalized forms of the attached tags,

[1] `http://www.upcoming.yahoo.com`
[2] `http://wordnet.princeton.edu/`
[3] `http://en.wikipedia.org/wiki/Main_Page`
[4] `http://en.wikipedia.org/wiki/Category:Events`
[5] Available for download at `http://www.mpi-inf.mpg.de/yago-naga/yago/downloads.html`
[6] `http://www.flickr.com/services/api/`

title, descriptions, as well as the Id and name of the user assigning the tag[7]. 22,985,996 tags have been gathered with this method (165,009 unique ones). For 2,837,182 pictures we could find non empty titles and only 1,662,075 had descriptions attached to them (1,623,667 had both title and description). We also wanted to investigate the influence of the location information for our event detection algorithm. However, we could find latitude and longitude coordinates for only cca. 25% of photos in our collection, so in the experiments we finally ommitted this type of information. On average, the event-related groups had 810.56 pictures in their corresponding group pools and 185.65 contributing users (taggers).

3.4 The Upcoming Data Set

For comparing the performance of our event detection methods with previous work [5] trying to solve the same problem, we made use of the *Upcoming* data set, which was kindly provided by the authors. This data set consists of *Flickr* pictures that were manually tagged by users with an event id corresponding to an event from the *Upcoming* event database[8]. For this collection, the Upcoming tags represent the "ground truth", both in the case of the algorithms presented in [5] and in the case of our classification algorithms.

Each photo corresponds to a single event and each event is self-contained and independent of other events in the data set. In total, there are 9,616 unique events and 275,672 unique pictures, with an average of 28.67 photos per event. All pictures have been taken between January 1st, 2006 and August 9th, 2008.

4. EVENT DETECTION FROM TAGS

For classifying images into the different categories of events we base our solution on collaboratively created social knowledge, i.e. tags associated with *Flickr* pictures. Based on already provided user tags, we build classifiers which try to assign the pictures to the corresponding event categories. More specifically, our approach relies on the hypothesis, that the existing tags provided by users for a particular photo carry information which can be used to infer the event category this picture belongs to. We perform a preprocessing step on the data collected from *Flickr* (described in Section 3.3) and we also experiment with different types of clustering methods on the original data sets. Below we describe the details of our algorithm, together with the preprocessing and photo clustering steps.

4.1 Clustering Flickr Pictures

In the approach we propose, we experiment with different ways of organizing the pictures crawled from *Flickr*. Besides the original, unclustered data set, two additional types of hierarchical clustering are considered, based on Wikipedia and WordNet classes, respectively. Below we present the details:

Original data set (Unclustered). As described in Sections 3.1, 3.2 and 3.3, we start collecting the pictures by considering the YAGO entities having as type *wordnet_event* and their corresponding Wikipedia event pages. The *Flickr* groups we can identify as first hits in response to queries consisting of the Wikipedia event page's name, correspond then to the list of [*events*]. For the example in Table 1, the list [*events*] consists of all entries in column $'Event'$. The pictures collected for the corresponding *Flickr* groups identified for these events remain unclustered.

Clustering based on Wikipedia classes. The first method of clustering we applied relies on the Wikipedia classes. In this case, all pictures belonging to the *Flickr* groups having the same Wikipedia class are merged into one cluster. The list of [*categories*] thus contains all unique entries from the column $'Wiki\_class'$, i.e. *united_nations_day* and *auto_shows* (for the example in Table 1). The pictures corresponding to the group Ids in rows 1 and 2 will be merged and the resulting cluster corresponds to the *united_nations_day* 'category'. Similarly, pictures corresponding to the groups from rows 3 and 4 will be put together.

Clustering based on WordNet classes. The second type of clustering we employed made use of the WordNet classes. For this case, similar to clustering based on Wikipedia classes, all pictures belonging to the *Flickr* groups having the same WordNet class are merged into one cluster. Rows 1 and 2 from Table 1 will be merged and will correspond to the *day* WordNet class. Rows 3 and 4 will remain untouched, as their WordNet classes are different. For this particular example, the list of [*concepts*] will be composed of *day*, *show* and *attraction*.

The photo event detection algorithm is then run on all these three variations of our data set, i.e. clustered or unclustered. Moreover, we also perform experiments on the Upcoming data set (see description in Section 3.4) and compare our results with the ones reported in [5]. However, for this last set of experiments, no clustering step is applied.

4.2 Event Detection Algorithm

The core of our photo event detection algorithm is a probabilistic classifier trained on the *Flickr* ground truth using tags as input features. Separate classifiers correspond to the different types of event classes that we extracted from YAGO. For building the classifiers, we use the open source machine learning library Weka[9]. In the experiments presented in this paper, we use the Naïve Bayes Multinomial implementation available in Weka. We also experimented with other classifiers (e.g. Support Vector Machines, Decision Trees), which resulted in similar classification performances, but were much more computationally intensive. Moreover, we also experimented with feature selection based on automatic methods (e.g. Information Gain) but the results showed that the full set is better suitable for learning, even though it contains some noise.

We have one classifier for each event category that we aim to learn to classify. The positive examples are represented by the pictures gathered from the event's corresponding *Flickr* group / resulted cluster, while the negative ones are randomly selected from the pictures corresponding to the rest of the event classes. The number of positive and negative examples is almost equally balanced.

Algorithm 1 presents the main steps of our approach. This corresponds to the general approach, when solely tags are used as features of the classifiers. Additionally, we also experiment with classifiers relying on features consisting of the

[7]We assume that the user tagging a photo belonging to a group pool corresponds in general to the author of the picture.

[8]Upcoming. `http://www.upcoming.org`

[9]`http://www.cs.waikato.ac.nz/~ml/weka`

Nr.	*Flickr Group Id*	*Event*	*Wiki_class*	*WordNet_class*
1	24165441@N00	*intl_day_of_peace*	*unit_nat_day*	*day*
2	602639@N25	*intl_holocaust_rembr_day*	*unit_nat_day*	*day*
3	1172355@N23	*motorama*	*auto_shows*	*show*
4	84783197@N00	*australian_intl_motor_show*	*auto_shows*	*attraction*

Table 1: Example for clustering Flickr pictures

words from the photos' titles, descriptions, time information and combinations of them, i.e. textual – tags, title, description, and textual+time – tags, title, description, time. For the case of features based on time information we substitute the Naïve Bayes with an SVM classifier, as this one is more suitable for the type of representation used for time. Moreover, for the combination of textual and time features, we use a linear combination of Naïve Bayes classifiers for the pure textual features with the SVM classifier, which we use for time. The two types of classifiers are equally weighted. For the combination of textual features the vectors are constructed similarly with the case of tag-based feature vectors.

Algorithm 1. Event detection

1: **Cluster *Flickr* pictures** (*optional*, see Section 4.3)
2: **Data preprocessing** step (see Section 4.3)
3: **For each** *event / category / concept*, E_x
4: **Split picture collection,** P_{total} into
5: P_{train} = pictures for training
6: P_{test} = pictures for testing
7: **Select tag features** for training the classifier
8: **For each** photo $p_i \in P_{train}$
9: **Create feature vector** $F(p_i) = \{t_j | t_j \in T\}$,
10: T = set of tags from all photos
11: $t_j = \begin{cases} 1, P_i \text{ has tag } t_j; \\ 0, \text{otherwise.} \end{cases}$
12: **Train Naïve Bayes** classifier on P_{train} using $\{F(p_i); p_i \in P_{train}\}$
13: **For each** photo $p_i \in P_{test}$
14: **Compute** classifier **output,** $NB(p_i)$
15: If $(NB(p_i) >= 0.5)$ classify p_i in E_x

Step 1 of the algorithm above aims at reducing the number of event classes to be predicted for the photos. This step is optional (described in detail in Section 4.1), as we experiment with all classes of events extracted from YAGO, as well as with a subset resulted from applying one of the clustering methods on the original set. If two or more classes are clustered based on one the methods described in Section 4.1, the resulted class will contain all pictures which have been originally assigned to the composing classes. As we need a certain amount of input data in order to be able to consistently train the classifiers, we discard those classes containing less than 100 photos (step 2) and the details of this pruning step are described in Section 4.3.

After selecting separate sets of pictures for training and testing (steps 4 - 6), we build the feature vectors corresponding to each picture in the training set (lines 7 - 11). The vectors have as many elements as the total number of distinct tags assigned to the images belonging to the event / category / concept classes. The elements of a vector will have values of either 1 or 0, depending on whether the tag has been assigned to the particular photo, or not. Once the feature vectors are constructed, they are fed into the classifier and used for training (step 12). A model is learnt and afterwards is applied to any new, unseen data. For any unseen picture, if the output of the classifier is greater or equal 0.5, the picture will be assigned to the current class.

4.3 Data Preprocessing

Since for training the event classifiers we need enough data at our disposal, we need to perform a preprocessing step and remove the groups / group clusters not having sufficient photo instances. The preprocessing actions' flow looks as follows:

Algorithm 2. Data preprocessing

1: **For each** *event / category / concept* class
2: **Repeat until** nothing to discard anymore
3: **Discard tags** corresponding to:
4: - names of Wikipedia event pages
5: - words composing event names
6: - synonyms of words composing event names
7: - combinations of words / synonyms
8: **Discard tags,** t_i, where $freq(t_i) < 10$ over all *event/category/concept* classes
9: **Discard photos** p_i, where $nrTags(p_i) \notin [2, 75]$
10: **Discard class** c_i of *event/category/concept* if $nrPhotos(c_i) < 100$

As we can see in lines 3 through 7, all tags appearing in the name of the Wikipedia event page, together with their combinations and synonyms are removed from the pictures in the collection corresponding to the specific *Flickr* group (or resulted cluster, as described in Section 4.1). With this step we avoid the potential bias of the classifiers towards words which might indicate the appartenance of the photos to the corresponding classes of events / categories / concepts. This step is necessary due to the crawling methodology. Likewise, for features in the form of titles and descriptions, the matching words are discarded. In case of the Upcoming data set, just like the authors of [5], we did not apply any discarding procedure. Here the crawling method was based on the constraint that pictures have as tag an Upcoming identifier, and in this case the ids corresponding to the Upcoming events do not reveal any information about the topic/scope of the events.

Tags which do not appear together with at least 10 photos throughout the collection are discarded (line 8), as they might represent too obscure annotations, or even misspellings – and thus do not have any positive influence on the classification, or might even introduce noise. Similarly, photos with less than 2 tags are removed from the collection, as well as photos having more than 75 tags, since they might

contain spam-tags [16] (line 9). Finally, the classes of events / categories / concepts with less than 100 photos are also removed (line 10), since we need sufficient instances in order to be able to train the classifiers. The whole process is repeated until no more pruning can be made. Except of lines 3 through 7, all the other steps of Algorithm 2 are applied also to the Upcoming data set.

5. EXPERIMENTS AND RESULTS

5.1 Evaluation Methodology

As already described in Section 4, we experiment with four different data sets:

- We aim to classify the events themselves and we create separate classifiers for all [*events*]. The data is coming from the *Flickr* groups we can identify for the YAGO events and their corresponding Wikipedia event pages. Below we will refer to the results for this data set as *Event Groups*.

- We also experiment with classifications of Wikipedia categories, containing the events. As described in Section 4.1, *Flickr* groups get clustered based on their common Wikipedia categories. In this case, we build classifiers for all [*categories*] and we will refer to this set of results as *Wikipedia Categories*.

- Moreover we aim to classify the WordNet concepts describing the Wikipedia categories, i.e. some higher level event-centered concepts. For this, we build classifiers for all [*concepts*], which we feed with data resulted from aggregating the *Flickr* group pools having the same WordNet class. In the results section below, this will be referred to as *WordNet Concepts*.

- Last but not least, we perform experiments on the Upcoming data set (see description in Section 3.4). We can thus also compare the performance of our algorithms against the one of the methods introduced in [5]. We will refer to this set of results as *Upcoming*.

For these four data sets we experiment with classifiers relying on different types of features: tags, titles, descriptions, time information, as well as combinations of them, such as tags+titles+descriptions (later referred as "textual") and tags+titles+descriptions+time (later referred as "textual+time"). As already mentioned, for time the Naïve Bayes is substituted with an SVM classifier, and in the case of "textual+time"-related set of experiments we use a linear combination of Naïve Bayes and SVM classifiers. The NB classifiers are constructed for the textual features (i.e. tags, titles and descriptions), whereas the SVMs correspond to time information. Equal weight is assigned to the two types of classifiers.

With this evaluation we focus on automatically measuring the quality of the photo event classification algorithm. For the first three sets of experiments, the ground truth data is represented by the information collected from the *Flickr* photo groups, or the resulted clustered sets. Being manually created by humans, the assignments of photos to the different classes of events can be considered correct and thus accepted as ground truth. Besides, through the collaborative participation of more users to the groups, i.e. by both joining the emerging networks and by contributing content in terms of pictures, comments, tags, etc., we can ensure that the spam-groups will be filtered out[10]. For the case of the *Upcoming* data set, the ground-truth consists of *Flickr* pictures tagged with event ids corresponding to events from the Upcoming database.

For the different types of experiments we present in Table 2 statistics regarding the number of classifiers built (column '# Classif.'), average number of instances for each classifier (column 'Avg. Inst.') and the number of features, respectively ('# Feat.').

Tag_features	**# Classif.**	**Avg.Inst.**	**# Feat.**
Event Groups	4,289	1,622	158,385
Wikipedia Categories	725	2,126	62,339
WordNet Concepts	161	26,456	129,621
Upcoming	559	476	8,803
Title_features	**# Classif.**	**Avg.Inst.**	**# Feat.**
Event Groups	3,651	1,509	44,043
Wikipedia Categories	668	1,838	17,722
WordNet Concepts	154	22,689	36,179
Upcoming	386	435	3,386
Description_features	**# Classif.**	**Avg.Inst.**	**# Feat.**
Event Groups	2,917	1,291	76,031
Wikipedia Categories	552	1,580	34,854
WordNet Concepts	132	18,315	65,887
Upcoming	601	511	3,277
Time_features	**# Classif.**	**Avg.Inst.**	**# Feat.**
Event Groups	4,091	1,594	1
Wikipedia Categories	728	2,068	1
WordNet Concepts	161	25,681	1
Upcoming	607	516	1
Textual_features	**# Classif.**	**Avg.Inst.**	**# Feat.**
Event Groups	4,146	1,656	278,459
Wikipedia Categories	729	2,165	114,915
WordNet Concepts	162	26,744	231,687
Upcoming	607	516	15,562

Table 2: Statistics for the experiment sets

The numbers in Table 2 are all computed after performing the pruning step, as described in Algorithm 1 (see Section 4.2). As we can observe, many groups have been discarded because of not containing enough data.

For evaluating the performance of our algorithms, we inspect the classification accuracy (*Acc*), precision (*P*) and recall (*R*) measures, when performing 10-fold cross-validation on the data sets. Moreover, for allowing the comparison with the results reported in [5], we also inspect the Normalized Mutual Information (NMI) [18, 21] and B-Cubed [3] values.

NMI measures how much information is shared between the actual ground-truth events (each with an associated set of pictures) and the clustering / classification assignment. For a set of classes $C = \{c_1, ..., c_j\}$ and events $E = \{e_1, ..., e_k\}$, where each c_i and e_i is a set of documents and n is the total number of documents

$$NMI(C, E) = \frac{I(C, E)}{(H(C) + H(E))/2} \quad (1)$$

[10]Like in the case of tagging, correct and suitable tags will get more and more employed, while obscure / misspelled tags will be pushed to the tail of the power law frequency distribution [13, 14]

where

$$I(C,E) = \sum_k \sum_j \frac{|e_k \cap c_j|}{n} log \frac{n \cdot |e_k \cap c_j|}{|e_k| \cdot |c_j|} \quad (2)$$

$$H(C) = -\sum_j \frac{|c_j|}{n} log \frac{|c_j|}{n} \quad (3)$$

$$H(E) = -\sum_k \frac{|e_k|}{n} log \frac{|e_k|}{n} \quad (4)$$

B-Cubed estimates the precision and recall associated with each document in the data set individually, and then uses the average precision P_b and average recall R_b values for the data set to compute B-Cubed as:

$$B - Cubed = \frac{2 \cdot P_b \cdot R_b}{P_b + R_b} \quad (5)$$

For each document, precision is defined as the proportion of items in the document's cluster that correspond to the same event, and recall is defined as the proportion of documents that correspond to the same event, which are also in the document's cluster. Both NMI and B-Cubed can take values between 0 and 1.

5.2 Local vs. Global Events

Our experiments do not resume just to the simpler task of event detection for photos. What we also aim to investigate is the suitability of our method for correctly classifying pictures into event classes corresponding to "local" and "global" categories. Personal experiences, such as birthdays, marriages or vacations are just a few examples of local events. On the other hand, sport competitions, concerts or natural disasters build collective experiences and are thus corresponding to the global-type of events. The most distinctive characteristic differentiating the local and global types of events is represented by the number of users contributing information to the corresponding event-data clusters. For local events, a very limited set of users contributes most of the event data, whereas in general for global events a numerous user participation is to be expected.

As a social network for sharing pictures, on *Flickr* we expect that most of the groups we collected correspond to global events. Sharing the personal experiences from a global type of event is the most common motivation for users to upload their photos to the platform [6]. Moreover also within the Wikipedia pages, from which our crawling mechanism was initiated, the number of global events exceeds the number of local or personal events. This fact directly influences our data collection.

We base our decision for separating the *Flickr* event groups into local or global events on the assumption that local event groups contain pictures contributed by at most 5 users, whereas global event groups have at least 50 contributors. These values have been selected based on the insights we got by manually assessing some randomly selected event classes. Below we present statistics corresponding to the number of distinct classes, average number of instances and number of features for the two local and global dimensions. All statistics are presented for the more general case when using just tag information for the event classification.

This type of experiments is performed on the initial set of *Flickr* event groups only. In the case of Wikipedia categories and of WordNet concepts, the initial set of *Flickr* groups is first clustered, such that we lose the local and global aspects of the groups, as created by the platforms' users.

	# Classif.	Avg.Inst.	# Feat.
Local Event Groups	244	611	4,269
Global Event Groups	2,713	2,284	146,152

Table 3: Statistics for Local / Global experimental setups

5.3 Results

The results of the evaluation runs are presented in Table 4. We observe that the average classification accuracy for all four data sets is very good when using tags as input features, almost in all cases being above 70%. Using this type of information as input features for the classification is thus very convenient, as tags can be easily collected along with the resources they are attached to. These results confirm once more the quality of user provided tags – a result also observed in [6] – as well as the hypothesis on which our approach relies (see Section 4). Using solely title or description information for the classification translates into poorer classification accuracy and strongly depends on the characteristics of the undelying dataset. On the other hand, combining all types of textual information (i.e. tags, titles and descriptions – "textual") results in similar classification accuracy as in the case of simple tag-based classifiers. As the results indicate, time information is not sufficient for correctly distinguishing between the different types of events. However, the linear combination of Naïve Bayes and SVM classifiers, corresponding to textual and time features, respectively, boosts the performance above the one of tag based classifiers.

For the *Upcoming* dataset we obtain very good values for the NMI and B-Cubed scores. For comparison, the authors of [5] report in the case of the best performing methods values of 0.94 and 0.82 for NMI and B-Cubed, respectively. For the other datasets (i.e. *Event Groups*, *Wikipedia Categories* and *WordNet Concepts*) these measures are not directly comparable, and the classes are not mutually exclusive, as for the *Upcoming* dataset. Moreover, our approach is complementary to the methods described in [5] – in our case the classes of events are known beforehand, while in [5] they are not. One should thus decide on the type of algorithm to be used depending whether this type of information is known or not.

The experimental results for the picture classification into local and global events (Table 5) indicate also good classification accuracy. Here we present the results for the general case, when using solely tags as input features. The performance is better for local than for global events. In the case of local events the information on which we base the classification is more heterogeneous, since it is created by a smaller set of users. For global events, the performance of the classification is still very good, with recall values above 80%. Recall is more important for a setting where users' pictures automatically get classified into the corresponding event clusters. This way the event classes will present a complete view of the different events and will be populated with enough items.

In Figure 1, for a better comparison among the four experimental runs, we depict the averaged values of accuracy, precision and recall. Even if in the case of *Event Groups*, i.e.

Features	Classes	Acc[%]	P	R	NMI	B-Cubed
Tags	***Event Groups***	72.46	0.68	0.87	0.85	0.69
	Wikipedia Categories	83.40	0.78	0.94	0.88	0.80
	WordNet Concepts	78.91	0.75	0.89	0.60	0.66
	Upcoming Events	93.39	0.89	0.99	0.99	0.99
Title	***Event Groups***	56.55	0.55	0.68	0.68	0.46
	Wikipedia Categories	66.66	0.64	0.78	0.68	0.53
	WordNet Concepts	62.70	0.61	0.72	0.41	0.51
	Upcoming Events	81.83	0.77	0.94	0.90	0.82
Description	***Event Groups***	60.96	0.60	0.70	0.72	0.50
	Wikipedia Categories	72.45	0.70	0.80	0.73	0.60
	WordNet Concepts	66.00	0.65	0.72	0.44	0.53
	Upcoming Events	66.62	0.69	0.82	0.66	0.48
Time (SVM)	***Event Groups***	51.69	0.51	0.57	0.52	0.28
	Wikipedia Categories	52.99	0.52	0.59	0.48	0.32
	WordNet Concepts	52.36	0.53	0.58	0.46	0.30
	Upcoming Events	52.74	0.52	0.65	0.57	0.36
Textual	***Event Groups***	72.20	0.67	0.84	0.84	0.68
	Wikipedia Categories	83.99	0.80	0.93	0.88	0.80
	WordNet Concepts	78.62	0.75	0.87	0.60	0.65
	Upcoming Events	92.35	0.88	0.99	0.99	0.98
Textual+Time (linear combination NB+SVM)	***Event Groups***	75.46	0.72	0.88	N/A	N/A
	Wikipedia Categories	83.66	0.81	0.93	N/A	N/A
	WordNet Concepts	78.84	0.75	0.88	N/A	N/A
	Upcoming Events	92.62	0.88	0.99	N/A	N/A

Table 4: Averaged classification results showing Accuracy, Precision, Recall, NMI, and B-Cubed (NMI and B-Cubed values are not available for the linear combination of the two classifiers)

	Acc[%]	P	R	NMI	B-Cubed
Local Ev.	89.32	0.84	0.99	0.99	0.99
Global Ev.	69.67	0.66	0.83	0.82	0.67

Table 5: Averaged classification results for local and global event groups using only tags as features

the original, unclustered data set, where we have 4,289 event classes, the average accuracy is 72.46%. For *Wikipedia Categories* the performance is best among the first three sets of results. In this case we cluster the initial *Flickr* groups' pools of pictures. However, even if the resulted clusters gather also all pictures from the composing *Flickr* groups, the resulting sets are still homogeneous. This is due to the fact that the Wikipedia categories are more abstract than the event classes, yet the abstraction is not introducing any noise in the classification. The clustering in this case is even reducing some of the ambiguities of the *Event Groups* sets of photos, these initial sets being perhaps a bit too fine grained.

For *WordNet Concepts* on the other hand although there are much less classes to distinguish among, the performance is a bit poorer than in the case of *Wikipedia Categories*. The reason for these results is the fact that the WordNet event categories represent more abstract event-related concepts. By clustering the initial set of *Flickr* social groups based on their common WordNet categories, the resulting sets of pictures corresponding to each of these WordNet event concepts are becoming too heterogeneous. Thus, it becomes more difficult for the classifiers to correctly distinguish among classes.

For the *Upcoming* dataset we achieve the best results. The reason is given by the fact that here the ground truth is manually created by users, who link their pictures to event ids from the Upcoming database. In this case the different classes of events are mutually exclusive, making thus easier for the classifiers to distinguish among them. The performance of the local event classifiers is comparable with the classification quality we achieve for the Upcoming dataset, these two datasets having more similar characteristics. Global event classification on the other hand is more similar w.r.t. performance with the event groups.

The results presented so far (Tables 4, 5 and Figure 1) indicate the performance of our algorithms in classifying photos into the different classes of [*events*], [*categories*] and [*concepts*], i.e. *macro* evaluation results. However, we were also interested in *micro*-evaluating our algorithms. More specifically, we also analyzed the results per specific event / category / concept class to find out which classes offer the best performances and which classes are more difficult to learn. Table 6 shows the *Acc*, P and R values, together with the available number of photo instances for training the classifiers. We selected some of the best and worst performing classifiers and the results are grouped based on the three different experimental runs.

The differences show that while some classes are relatively easy to learn, others may require special attention or some level of disambiguation. Also, classes which are hard to learn are ambiguous and the annotations are mostly subjective. As we can see, a higher number of instances available for training, definitely improves the classification accuracy: five of the nine best performing classifiers had more than 500 instances at their disposal. Nevertheless, for very clear event categories, like *wikicategory_Auto_shows*

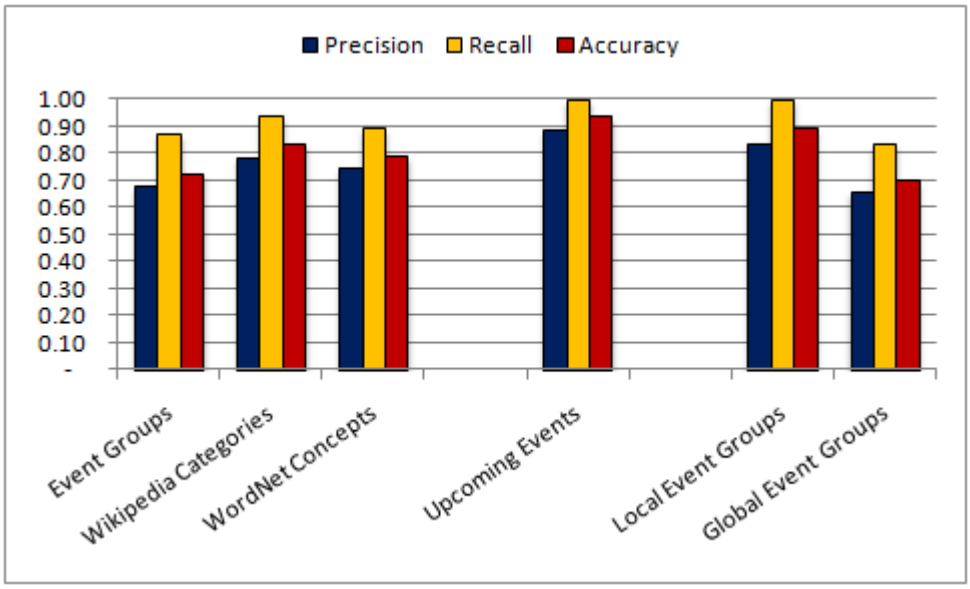

Figure 1: Classification results (Acc, P, R) for the experimental runs using only tags as features

or *wordnet_championship*, the classifiers still can achieve very good performance, since in these cases the corresponding sets of *Flickr* photos and their associated tags are very homogeneous.

At the other end, we have the more ambiguous event classes, such as $Before\_Sunrise$, *wordnet_motion*, or *wordnet_execution*, for which the classification is much more difficult. For these cases (e.g. *wordnet_television_program*) even a high number of training instances does not improve much the classification accuracy. The main reason is that for these types of event classes, the underlying photo sets contain too many different types of pictures, depending on each user's understanding of the corresponding event concept. This of course, translates into a very heterogeneous tag vocabulary, which negatively influences the classification performance.

6. CONCLUSIONS AND FUTURE WORK

With the rapid shift from analog to digital photography we have faced over the past years and with the advent of Social Media sites dedicated to photography, a huge amount of digital photos became available on the Web. However, advanced techniques for easily browsing and intuitively organizing these photo collections are still missing. Image retrieval is still in its infant stage, due to the fact that the available data is noisy and it is not easy to capture the content of photos. On the other hand, collaborative tagging is a valuable source of semantically rich metadata, especially useful for repositories covering multimedia resources, whose non-textual content is not easily indexable / searchable. To tap this potential, in this paper we developed a new algorithm for automatically classifying pictures into classes of events. With the method we proposed, we enable event-based indexing and browsing of photo collections, i.e. a very intuitive way of organizing one's memories.

The algorithms described in this paper rely on user annotations in the form of tags, titles, descriptions, time information and combinations of them and we experiment on two subset of pictures crawled from *Flickr*. The approach is however not restricted to this collection, but being applicable to any other photo set or other types of multimedia content (e.g. videos, music, etc.) containing similar metadata.

For one set of experiments we rely on the YAGO event ontology, and as ground truth we made use of the *Flickr* group photo organization, taking thus the users' judgments regarding the pictures' assignment to classes of events as golden standard. We experimented with different levels of abstraction of the YAGO event ontology and implicitly clustering of the original picture collection and observed that while some classes are relatively easy to learn, others require more attention or some level of disambiguation. However, which is the right level of abstraction for events, that is still understandable and accepted by users is an interesting question for further investigations. Another set of experiments was carried out on a *Flickr* subset where the ground truth is manually created by users, who link their pictures to event ids from the Yahoo! Upcoming database. Overall, the results of our evaluations show that photo event-based classification is feasible and confirm once more the quality of the user provided tags. Basing the classification decision on this type of data is not only much simpler than approaches combining several other sources of information (e.g. titles, descriptions, time, etc.), but also achieves at least comparable performance. As the results indicate, our approach is also suitable for correctly classifying both local and global types of events. Users' picture collections can thus be correctly automatically categorized into classes of events of either type. Moreover, these findings open new possibilities for multimedia retrieval, in particular image search.

For the future we plan to improve this algorithm, as well as the feature selection mechanism by automatic identification of tag types (e.g. Topic, Author, Usage context, etc.). Learning the coefficients for the linear combinations of the Naïve Bayes and SVM classifiers, or training different types of classifiers for all different kinds of features and combining

	Performance	Event/Categ./Concept Class	Acc[%]	P	R	# Inst.
Event Groups	**Best**	*Clare_Minor_Hurling_Championship*	98.82	0.98	1.00	254
		Grand_Prix_de_Pau	98.09	0.97	0.99	680
		Colorado_Castle_Rocks	98.03	0.97	0.99	304
	Worst	*Before_Sunrise*	65.20	0.12	0.60	102
		A_Night_Full_of_Rain	62.61	0.19	0.59	234
		Heavy_Metal_Parking_Lot	60.29	0.01	0.56	136
Wikipedia Categories	**Best**	*wikicategory_Disney_parks_and_attractions*	99.28	0.99	1.0	1,392
		wikicategory_Sports_in_Pittsburgh_Pennsylvania	97.87	0.96	0.99	1,832
		wikicategory_Auto_shows	97.00	0.94	1.00	200
	Worst	*wikicategory_CBC_network_shows*	74.09	0.18	0.71	440
		wikicategory_Beauty_pageants	73.24	0.13	0.68	284
		wikicategory_2008_NASCAR_Sprint_Cup_races	65.31	0.27	0.63	490
WordNet Concepts	**Best**	*wordnet_championship*	93.84	0.93	0.98	326
		wordnet_park	90.30	0.84	1.00	1,392
		wordnet_battle	88.79	0.86	0.93	2,462
	Worst	*wordnet_execution*	70.33	0.65	0.87	364
		wordnet_motion	69.08	0.67	0.76	760
		wordnet_television_program	66.36	0.62	0.84	1,400

Table 6: Examples of best and worst performing (by *Acc*) classifiers for the different experimental runs

their results are just a few other steps worth investigating. Additionally we also plan more work on the types of events that are mostly employed by users when referring to their event-related memories, as well as experiments considering also the location information. Moreover, merging our approach with content-based methods trying to solve the same task is worth examining.

7. ACKNOWLEDGMENTS

We are greatly thankful to Hila Becker, Mor Naaman and Luis Gravano (the authors of [4] and [5]) for providing us with the Upcoming data set. This work was partially supported by the GLOCAL project funded by the European Commission under the 7^{th} Framework Programme (Contract No. 248984).

8. REFERENCES

[1] Topic detection and tracking evaluation. http://www.itl.nist.gov/iad/mig//tests/tdt/.

[2] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In *SIGIR*, 1998.

[3] E. Amigo, J. Gonzalo, J. Artiles, and F. Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 2008.

[4] H. Becker, M. Naaman, and L. Gravano. Event identification in social media. In *WebDB*, 2009.

[5] H. Becker, M. Naaman, and L. Gravano. Learning similarity metrics for event identification in social media. In *WSDM*, 2010.

[6] K. Bischoff, C. S. Firan, W. Nejdl, and R. Paiu. Can all tags be used for search? In *CIKM*, 2008.

[7] K. Bischoff, C. S. Firan, W. Nejdl, and R. Paiu. How do you feel about "dancing queen"?: deriving mood & theme annotations from user tags. In *JCDL*, 2009.

[8] K. Bischoff, C. S. Firan, and R. Paiu. Deriving music theme annotations from user tags. In *WWW*, 2009.

[9] L. Chen, Y. Hu, and W. Nejdl. Deck: Detecting events from web click-through data. In *ICDM*, 2008.

[10] L. Chen and A. Roy. Event detection from flickr data through wavelet-based spatial analysis. In *CIKM*, 2009.

[11] M. Dubinko, R. Kumar, J. Magnani, J. Novak, P. Raghavan, and A. Tomkins. Visualizing tags over time. In *WWW*, 2006.

[12] G. P. C. Fung, J. X. Yu, P. S. Yu, and H. Lu. Parameter free bursty events detection in text streams. In *VLDB*, 2005.

[13] S. A. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *J. Inf. Sci.*, 2006.

[14] H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. In *WWW*, 2007.

[15] Q. He, K. Chang, and E.-P. Lim. Analyzing feature trajectories for event detection. In *SIGIR*, 2007.

[16] G. Koutrika, F. A. Effendi, Z. Gyöngyi, P. Heymann, and H. Garcia-Molina. Combating spam in tagging systems. In *AIRWeb*, 2007.

[17] Z. Li, B. Wang, M. Li, and W.-Y. Ma. A probabilistic model for retrospective news event detection. In *SIGIR*, 2005.

[18] C. D. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[19] T. Rattenbury, N. Good, and M. Naaman. Towards automatic extraction of event and place semantics from flickr tags. In *SIGIR*, 2007.

[20] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *WWW*, 2008.

[21] A. Strehl and J. Ghosh. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 2003.

[22] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In *WWW*, 2007.

[23] Y. Yang, T. Pierce, and J. Carbonell. A study of retrospective and on-line event detection. In *SIGIR*, 1998.