Noname manuscript No.
(will be inserted by the editor)

Predicting Celebrity Attendees at Public Events using Stock Photo Metadata

**Xin Shuai · Neil O'Hare · Luca Aiello ·
Alejandro Jaimes**

Received: date / Accepted: date

Abstract The large collections of news images available from stock photo agencies provide interesting insights into how different celebrities are related to each other, in terms of the events they attend together and also in terms of how often they are photographed together. In this paper, we leverage such collections to predict which celebrities will attend future events. The main motivation for this is in the event-based indexing of online collections of multimedia content, an area that has attracted much attention in recent years. Based on the metadata associated with a corpus of stock photos, we propose a language model for predicting celebrities attending future events. A temporal hierarchical version of the language model exploits fresh data while still making use of all historical data. We extract a social network from co-appearance of public figures in the events depicted in the photographs, and combine this latent social information with the language model to further improve prediction accuracy. The experimental results show that combining textual, network and temporal information gives the best prediction performance. Our analysis also shows that the prediction models, when trained by the most recent data, are most accurate for political and sports events.

Keywords Language model · Photo metadata · Events attendees prediction · Stock photos

Xin Shuai
Indiana University Bloomington, IN, USA
Tel.: +1-8126068969
E-mail: xshuai@indiana.edu

Neil O'Hare, Luca Aiello, Alejandro Jaimes
Yahoo Research Barcelona, Spain
E-mail: nohare,alucca,ajaimes@yahoo-inc.com

1 Introduction

Many news articles are accompanied by photographs that depict events, and very often such photographs include public figures (e.g., politicians, celebrities, athletes, etc.). The photos, mostly taken by journalists, are usually manually enriched with textual metadata, including a short descriptive caption and additional information such as the location, and possibly people and events, related to the photo. They are distributed via wire services (e.g, AFP), and included in stock photo collections for future use[1].

In aggregate, the large collections of news stock images can provide interesting insights into how different celebrities and public figures are related to each other in terms of the events they attend and also in terms how often they are photographed together. In this work we leverage such collections, in particular the textual descriptions of photos and events, and the latent network of co-appearances, to predict which celebrities will attend future events. The ability to use automatically extracted information about event attendees for the indexing and retrieval of public events will add much value for search applications in particular, although other applications are sure to benefit from such metadata. We also emphasize that, although we evaluate our approaches in the setting of a prediction task, the approaches are also applicable for past events for which a definitive list of attendees is not available. It is likely that the most useful applications of these approaches would be in search and retrieval over such past events, although the ability to predict event attendees will also enable new and exciting applications.

Other data resources (news articles, Wikipedia[2], IMDB[3]) also reflect the relationships between celebrities and events, which could also be used to predict celebrity event attendance. We choose to focus on news photo metadata for a number of reasons. Firstly, news photo metadata provide real-time updates, in that these news photos are often published shortly after the events occur. Secondly, news photo metadata provides richer and clearer information about co-occurrence than news articles, which may only mention a few key attending celebrities at an event: in fact, an application of the current work would be to predict the 'missing' celebrities from news article's description of an event. The third reason we focus on stock photos is that co-depiction in photographs captures social relationships, unlike the connections in Wikipedia and IMDB, and repeated co-depiction gives us an opportunity to measure relationship strength in a way that Wikipedia, especially, cannot. Finally, stock photo data has clearly defined event information, which is not always available elsewhere. Although we use photo metadata for prediction in this paper, the proposed prediction framework can be easily extended to other types of textual data.

[1] The largest agencies are Getty Images, Corbis, and Sipa Press.

[2] http://www.wikipedia.com

[3] http://www.imdb.com

To the best of our knowledge, ours is the first work that has attempted to predict attendees of future public events using stock photograph metadata. The main contributions of this work are:

- The extraction and analysis of a social network of public figures from a large image corpus.
- Using this network, in combination with textual photo metadata, to predict attendees at public events.

While this paper is a direct extension to our previous paper [1], we extend and improve that work here in a number of ways. First, we consider here a much larger dataset of 30 million photos and over 16 thousand events. We implement temporal smoothing models that take advantage of fresh, recent data while still exploiting all the historical data in the corpus. We conduct a deeper analysis of the results of the prediction experiments, in particular showing that we can predict the participants of smaller events more accurately, as we can for Sports and Political events. We also show that model freshness has a large impact on performance. Finally, we perform an extensive parameter tuning to optimise the model parameters, facilitating a more reliable comparison between methods.

The rest of the paper is organized as follows. In the next section we review related work, followed by a description of our dataset in Section 3. Section 4 outlines the prediction methodology. Section 5 describes the evaluation of the models, including an analysis of the results. Finally, in Section 6, we conclude the paper.

2 Related Work

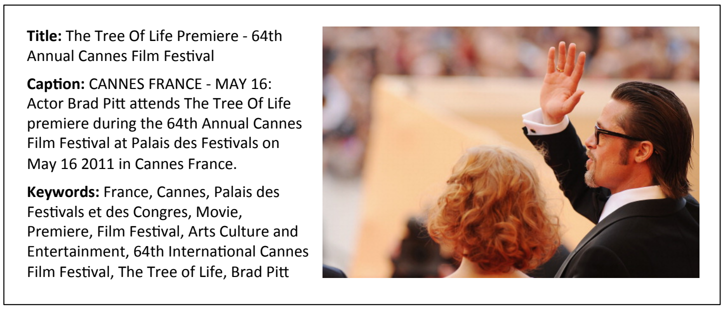
The problem of event recognition from multimedia has been studied extensively in recent years [2]. Approaches for detecting events in personal photo collections based on metadata such as tags, geo-location, and time [3] as well as content-based techniques [4] have been explored. While these works have focused on detecting small, personal, events in personal photo collections, other work has attempted to detect larger events in public photo collections, based on analysis of image content and metadata [5–7].

An important initiative in the large-scale detection of event in large-scale social multimedia has come from the *Mediaeval Social Event Detection (SED)* task [8]. The *SED* task at *Mediaeval* has focused on *detecting* events in a set photographs uploaded to photo sharing platforms; subtasks include classifying media into event types. The organizers of this task have provided large-scale dataset for these tasks, with the 2013 edition providing a dataset of over 400 thousand photos. Example approaches to social event detection within this task include multi-modal feature selection and clustering techniques [9], and combining contextual information into a constraint-based clustering and classification model to classify photos according to event types [10]. Our work differs significantly from this in terms of both the task and the dataset. Our task is, given a set of photos and the events they belong to, predict the people who attend these events (i.e. the people depicted in the event photos). This task is supported by a dataset of over 30 million stock photos dataset that contains reliable annotations for both people and events depicted in photographs, which is very different from the Flickr photos used in the SED task, for which reliable event and people information available.

Identification of people from videos and images has also been investigated in depth [11]. With the advent of social media, a lot of work has focused on using image content. Some approaches exploit the fact that people wear the same clothes during an event and use torso analysis to enhance face recognition [12–14], while other work makes use of spatial and temporal context, or social network information, to improve recognition accuracy [15–17]. All of these approaches make use of image content analysis in the form of face recognition; differently from these, Naaman et al [18] propose an approach to identifying people in images based on spatial, temporal and social context alone, assuming a partially annotated personal photo collection. Our work differs from all of this work on person identification in that we are interested in associating people with events, as opposed to individual photos. We focus on large scale, public photos where the candidate people are typically celebrities; also, we do not use image content analysis, and unlike the prior work, our approach is mostly based on analysis of the text descriptions of the photos people appear in.

In addition to some of the above work on person identification, which constructs social networks from images [15, 18, 17, 19] other efforts have extracted social connections between individuals in images and videos. Recognition of social clusters [20], prediction of social relationship types between individuals in photos [21], and friend recommendation [22] are examples of possible tasks in this context. Similarly to our work, Devezas et al [23] build a coreference network based on entities from photo descriptions, where nodes represent personalities and edges connect people mentioned in the same photo description. They focus on a much smaller collection, and limit their study to an exploration of the structural clusters in the social graph, and to not propose any prediction task or consider events. Other work constructs celebrity social networks from IMDB for the purposes of visualization [24], and automatically detecting influential historical characters [25].

Mantrach & Renders [26] focus on the task of person-finding by creating textual representation of people based on the text they write. This is similar in spirit to the current work, although our work differs in that we rely on the text describing a person, as opposed to the text they create, and we also exploit social network information.



Title: The Tree Of Life Premiere - 64th Annual Cannes Film Festival

Caption: CANNES FRANCE - MAY 16: Actor Brad Pitt attends The Tree Of Life premiere during the 64th Annual Cannes Film Festival at Palais des Festivals on May 16 2011 in Cannes France.

Keywords: France, Cannes, Palais des Festivals et des Congres, Movie, Premiere, Film Festival, Arts Culture and Entertainment, 64th International Cannes Film Festival, The Tree of Life, Brad Pitt

Fig. 1 An example stock photo and its title, caption and keywords.

3 Data Description

3.1 Photo Metadata

We collect the publicly available metadata from approximately 30 million images, covering a time period from 1970 to 2012, from a well-known stock photo agency. The metadata of each image includes the following items of relevance to our study:

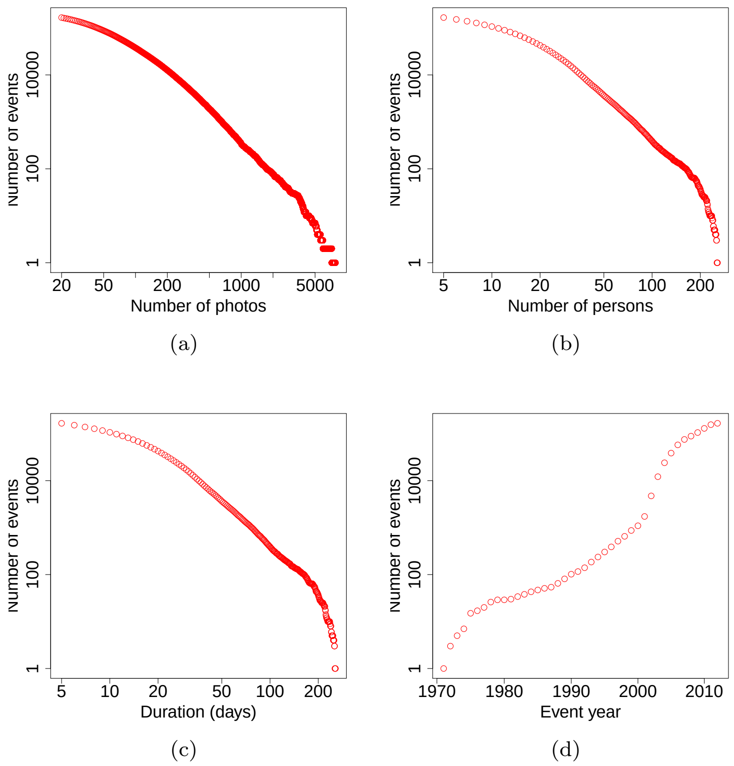
ImageId: A unique identifier for the image
DateCreated: The time stamp of when the image was taken
Title: The name of the image
Caption: A brief description of the image, including, e.g. the location, time and people in the image
EventIds: A list of events that are depicted in the image. Events can be a difficult concept to define precisely, and much of the related work avoids doing so. For our purposes, the events are pre-defined by the stock photos annotations and correspond to a group of photos taken at the same time and place. Typical events in the dataset are sports games (e.g. “Super Bowl 2012”), film premieres and other celebrity events (e.g. “83rd Annual Academy Awards 2011”), fashion events (e.g. “Marithe & Francois Girbaud Fashion Show, Sep 2011”) and political events such (e.g. “President Barack Obama Signs Payroll Tax Bill 2012”).
Keywords: A list of keywords related to the image. They can describe the subject of the image, aspects of the photographic technique, the number of people, the location, etc. The keywords, which can be provided by the stock photo agency or the professional photographer who captured the photo, can also have a *type* attribute, of particular interest to us is the keyword type describing *specific people* depicted in the photo.

Figure 1 shows an example photo from our corpus, along with its title, caption and keywords.

The photos in the corpus are related to more than $500,000$ events of different size and duration, with the average event having 40 photos and 6 attendees, as shown in Table 1. Figure 2 shows the distributions of size and timespan of

Table 1 Statistics of all events

	max	mean	median	std.
# images	65,333	40.291	19	137.7
# attendees	2,948	6.384	3	13.73
duration (days)	2,175	1.398	1	7.629



(a) (b) (c) (d)

Fig. 2 The distribution of (a) number of photos (b) number of persons (c) duration and (d) year of event, for the events captured by the photos in the corpus. (a), (b), (c) show that most of events contain a small number of images and attendees, and last for less than two days; (d) exhibits the increasing trend of stock images by year.

the events. The distributions of size, in terms of number of photos and attendees, are broad. Most of events contain a small number of images and people, and last for only one or two days. In addition, the number of recorded events gradually increases in time, with a rapid increase after year 2000, possibly due to the emergence of digital media technology.

Table 2 Statistics of photo and event co-appearance networks

	co-photo	co-event
nodes	54,152	50,911
edges	1,121,823	9,955,931
avg. edge weight	4.850	2.768
density	$7.651 \cdot 10^{-4}$	$7.682 \cdot 10^{-3}$
degree assortativity	0.035	0.163
avg. degree	41.43	391.1
avg. clustering coefficient	0.288	0.502
size of GCC	53,636	50,906

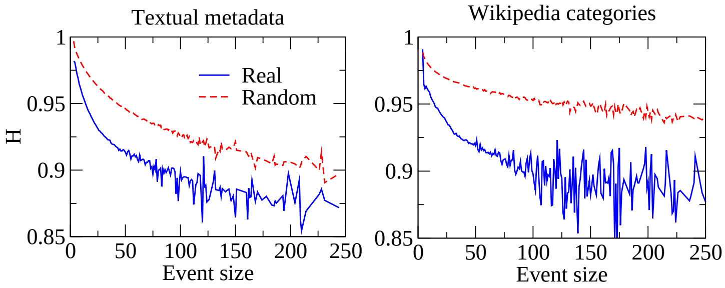


Fig. 3 Average entropy H of keywords in photo metadata and Wikipedia categories for people at the same event, at fixed event sizes (number of attendees).

3.2 Stock Photo Co-occurrence Network

The rich metadata in this image corpus allows us to infer social connections between the people depicted in the photos. These social connections can be inferred based on co-occurrence in photos, where two people appear in the same photo, or by co-occurrence in events, where two people appear in the same event but not necessarily in the same photo. These connections can be modeled as ties in an undirected, weighted graph, where nodes are people and edges represent co-occurrence. We define this graph as $G = (V, E, W)$, where V is the set of people appearing in the photo corpus, $E = V \times V$ is the set of co-occurrence relations in an image or an event, and $W = \{w(v_i, v_j)\}, v_i, v_j \in V$ is the set of weights indicating the co-occurrence frequency of v_i and v_j.

Table 2 shows statistics of the event-based and photo-based co-occurrence networks, based on the training set of photos used in the experiments described in Section 5. Since the co-event social graph expresses weaker but broader interpersonal relationship than co-photo, many more edges are involved in the co-event network.

Apart from the graph properties, it is interesting to investigate the relation between the structure of the network and the metadata that defines events and their participants. Verifying local alignment of node-level properties could unveil the factors that correlate with social linking [27], thus helping understanding features that could be predictive of future aggregations of groups of people. First, we observe that events tend to aggregate people with similar profiles. This can be quantified by describing each person with a vector of

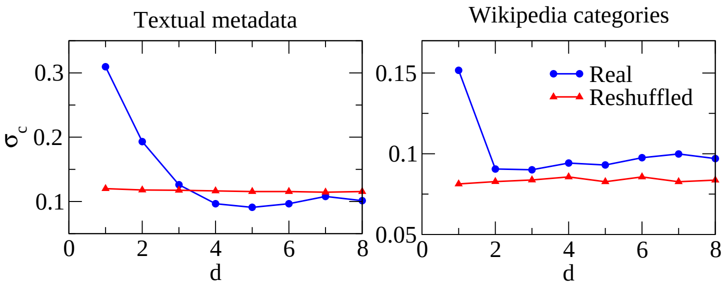


Fig. 4 Cosine similarity σ_c between keyword/category vectors of pairs of people at distance d in the social graph.

terms and measuring the entropy of the overall term distribution inside the event, compared to the entropy of an "artificial" event of the same size composed of random people. In a scenario in which people's profiles condition the participation in events, events will end up aggregating people that are more similar than the random case. The overall variety of the attendee's profiles (measured as entropy) will be lower than a case in which people participate in events regardless of their profile (e.g., their job). We can build such person profiles from the aggregation of all the textual metadata associated to the images they appear in. In this section, to investigate whether the effects we study are limited to our dataset, we also build profiles from the categories in people's Wikipedia pages. Figure 3 shows that, in both cases, the entropy in random events is appreciably higher than in real events, meaning that, to some extent, events tend to aggregate similar people.

Profile alignment can be also tested at network level [28], rather than at event level. From a network perspective, we expect social proximity to be related to person profile similarity. Figure 4 shows that people tend to be more similar to people who are closer in the network. The similarity decay with the distance is evident for user profiles created from photo descriptions and those taken from Wikipedia profiles. The difference of the similarity decay applied on a reshuffled version of the network, where links are rewired at random, shows that profile alignment of close people is not determined by chance or by pure assortativity. Edge strength is also correlated with profile similarity, meaning that people often co-appearing in a photo share many profile features (Figure 5). This also holds for the Wikipedia categories, showing that it is not simply an artifact of people "sharing" the annotations of photos they co-appear in.

4 Predicting Event Attendees

Our task is to predict attendees of future events. Assuming that a textual description of the event and, optionally (for network based methods) that at least one event attendee is known in advance, it is possible to leverage the metadata of the photos taken during past events to infer the most likely

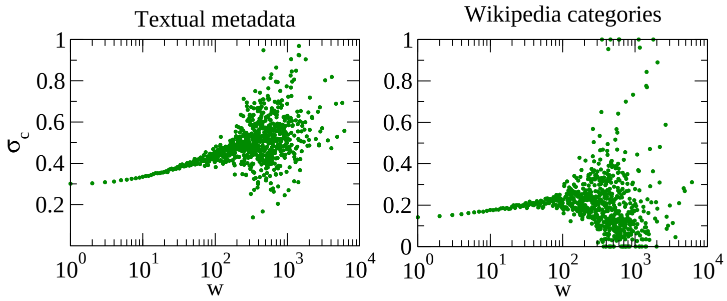


Fig. 5 Average cosine similarity of profiles of people connected with a edge of weight w. Positive slopes show positive correlation.

attendees. In this work we use the textual metadata associated with photos as the textual event description; in practice, any textual description, possibly written in advance of the event, could serve this purpose. In this section we describe an approach that uses the textual description of photos of past events, and an approach based on the past co-occurrence network of people, followed by a combination of the two approaches.

4.1 Data Pre-processing

Event Description Generation For each *eventId* in the corpus, we generate a textual description of the event by extracting the metadata from all photos belonging to the event (i.e. containing that *eventId*), concatenating all *Title*, *Caption* and *Keywords* metadata that co-occur with the *eventId*. After tokenizing the remaining string using a standard tokeniser, we remove duplicate terms, so that an event is represented by the unique set of terms contained in title, caption and keywords of the photos depicting that event. Additionally, all person names are removed from the event description.

People Description Generation For all person names, we generate a ‘person document’ (i.e. a textual representation of the person) by extracting all *Title*, *Caption* and *Keywords* metadata from all photos containing that person (in the *Specific People* keyword type). For large events containing many photos, many of those photos may contain duplicate or near-duplicate textual metadata. To prevent these sets of near duplicates dominating a person’s textual representation, when multiple images depicting a person belong to the same event, we only consider one (randomly chosen) representative image from the event as part of the person’s textual description.

4.2 Text-based Prediction

We construct a document to represent each person, as described above, based on which we build a model for event attendance. Below we describe the basic language model for event attendee prediction, a temporal hierarchical smoothing extension, a network based approach to prediction and, finally, a combination of textual and network information.

4.2.1 Language Model

This classical IR *query likelihood* language model creates a generative language model for each document in the collection and, given a query, ranks documents against that query according to the probability that their language model 'generated' the query [29]. In the current scenario, our language modeling approach is based on the assumption that the textual description of an event is *generated* by the people who participate in the event, i.e. the text used to describe the event is, to some extent, determined by the people in the event.

Representing each person by a multinomial probability distribution over the vocabulary of words in the collection (i.e. a language model), we can think of the textual description of an event as being drawn from the language models of all of the people in the event. For any given event, any description of that event (e.g. a news article, press release, etc) could for used for this purpose.

Continuing the analogy with classical information retrieval, in the current framework we can think of 'people' as documents and 'events' as queries. That is, using an event description as a query, we rank people by the probability that their language model generated that event:

$$P(p|e) = \frac{P(e|M_p)P(p)}{P(e)} \tag{1}$$

where $P(p)$ is the prior probability of person p, $P(e|M_p)$ is the probability of event e, given person p's language model M_p, and $P(e)$ is the prior probability of event e. Since $P(e)$ can be ignored for ranking as it is constant for all people being ranked, $P(p|e)$ is determined by $P(p)$ and $P(e|M_p)$. Although $P(p)$ is often treated as a constant, some previous work has shown that when reliable prior estimates, when available, can improve rankings for some applications [30,31]. We propose two alternative methods of estimating this prior. The first method simply calculates the *global prior* of the person, $\hat{P}_{global}(p)$, as their relative frequency in the collection. The second method is a network-based estimate, or *network prior*, which exploits a person's social distance from known event participants, and is described in Section 4.3.

Assuming independence between terms, we can calculate $P(e|M_p)$ as:

$$P(e|M_p) = \prod_{t \in e} P(t|M_p) \tag{2}$$

The simplest way to estimate the parameters of the language model (i.e. the probabilities of the individual terms), is to calculate a maximum likelihood estimate (MLE), based on the relative frequency of the term in the model (i.e. the frequency of the term divided by the total number of terms). MLE estimates will result in zero probabilities for any people (documents) that do not contain all of the query terms from the event description. Thus, to give some probability mass to unseen words, we smooth probabilities using the standard Jelinek Mercer smoothing approach [32]:

$$\hat{P}_{jel}(t|M_p) = \lambda \hat{P}_{mle}(t|M_p) + (1-\lambda)\hat{P}_{mle}(t|M_c) \tag{3}$$

where λ is the *collection smoothing* parameter between 0 and 1, and M_c is the collection language model, estimated based on the entire collection. Although there are a number of other smoothing methods available, and the choice of smoothing method can be important, an in-depth study of smoothing is not the focus of this work. Also, a previous study has found that, for similar models to ours, albeit on a slightly different task, there is no significant difference between Jelinek Mercer smoothing and Dirichlet smoothing, another widely used smoothing method [30].

4.2.2 Temporal Smoothing for Term Estimates

The interactions between public events and their participant are highly dynamic, and can be expected to change over time. For example, in the domain of Sports, if a player signs for a new team, he will not have the same strong relationship with his old team, or with his former teammates. For this reason, we expect that the most recent photos, and their event-participant relationships, to be more predictive of future events attendees than older photos are.

One way to capitalize on this would be to only use the most recent data to build our models, although this approach is susceptible to the problem of creating less robust models, by ignoring most of the data. For this reason, we propose a hierarchical temporal smoothing method for estimating the term weights for a given person's language model. Hierarchical smoothing methods have previously been used in video retrieval [33], where they can represent the hierarchy of shots, scenes and videos, in context-based person recognition in personal media collections, where they represent temporal or spatial hierarchies [16], and in other applications such as estimating the location of photos [34,30].

In Equation 3, we smooth the specific model for a person, M_p, with the more general model of the entire collection, M_c. To incorporate the temporal factor in the language model, we define another language model M_{tp}, based on all metadata about person p within the most recent time period t. The period t is a time window with its' end point fixed on the latest date in the collection. The size of this time window is a parameter of the model: we explore time windows of one month, six month, one year, two years and five years and ten years.

M_{tp}, M_p and M_c form a hierarchical series of language models, from specific (most recent data) to general (all data), which are combined with Jelinek Mercer Smoothing:

$$\hat{P}_{hier}(t|M_h) = \lambda_1 \hat{P}_{mle}(t|M_{tp}) + \lambda_2 \hat{P}_{mle}(t|M_p) + \lambda_3 \hat{P}_{mle}(t|M_c) \tag{4}$$

λ_1, λ_2 and λ_3 are the smoothing parameters, and are subject to the constraint that they must sum to 1.

4.3 Network-based Prediction

For network-based approaches to event participant prediction, we consider the event and photo co-occurrence networks described in Section 3.2. We assume that one attendee is known for each target event, and is used as a *seed* to discover related people in its ego-network. We compute the probability of co-occurrence of a candidate attendee p with the seed person s by dividing the weight of the edge that connects them by the weighted degree of the seed:

$$\hat{P}_{net}(p) = \frac{w(p,s)}{\sum_{n \in \Gamma(s)} w(n,s)} \tag{5}$$

Time may also be a factor when determining the value of P_{net}, since the latest co-occurrences might be more predictive than earliest ones. To study this, we compare the co-occurrence probabilities in the network that embeds the co-occurrence information based on all data with that based only on more recent time windows.

Since the network based estimate can be affected by the sparseness of the data, we also smooth the network estimate for each candidate person using Jelinek Mercer smoothing, to interpolate it with the global prior of the person:

$$\hat{P}_{net\_jel}(p) = \alpha \hat{P}_{net}(p) + (1-\alpha)\hat{P}_{global}(p) \tag{6}$$

4.4 Combining Text and Network Information

We explore two alternative approaches to combining textual and network based prediction. The first is to expand the textual event description with the name of the seed attendee, and use the standard text-based language model approach. Since people who are depicted in the same photo will have their names in each other's textual representation, this approach will give a higher rank to candidate people who co-occur with the seed person. Strictly speaking this method is a pure text-based approach, but it is does represent a simple way of making use of the co-occurrence information.

The second approach is to use the *network prior* (Equation 5) or the *smoothed network prior* (Equation 6) as the estimate of the prior probability of a person, $P(p)$, in Equation 1.

5 Evaluation

We evaluate our approaches to person prediction using the public photo corpus described in Section 3. We remove extremely small events containing less than

Table 3 Evaluation results of language model and network-based event attendance prediction algorithms.

	MAP	P@1	P@3	P@5	P@10	P@20
Language Model (LM)	0.1552	0.3086	0.2577	0.2273	0.1825	0.1365
LM (frequency prior)	0.1665	0.3351	0.2783	0.2454	0.1963	0.1466
LM (with seed name)	0.1808	0.3582	0.2979	0.2618	0.2079	0.1546
Network (co-photo)	0.0615	0.1522	0.1184	0.1007	0.0778	0.0596
Network (co-event)	0.0785	0.1661	0.1346	0.1152	0.0900	0.0702
LM (network prior)	0.1645	0.3336	0.2772	0.2443	0.1949	0.1449
LM (smoothed network prior)	0.1835	0.3587	0.3021	0.2683	0.2145	0.1602
LM + Seed Name (smoothed network prior)	**0.1868**	**0.3588**	**0.3034**	**0.2701**	**0.2162**	**0.1614**

5 persons or 20 images, as they are likely to be less important, and extremely large events containing more than 300 persons or lasting more than a week, as they are likely to be a combination of multiple events. This leaves around $166,000$ events as the final dataset for experimentation. For each *eventId*, we extract the names of all people in the event (from the *Specific People* keyword type), to be used as a ground truth for evaluation.

Sorting the events by the time of the first photo in the event, the photos from the first 80% of the events (i.e. 133,173) are used as a training corpus to generate the textual representations of people and the social network among them, which gives a very large training period that spans over 30 years, from February 1st, 1970 to February 19th, 2011. The remaining 20% of events are randomly split in two: 10% (i.e., 16,647) of the events are used as a tuning set, to optimize the various model parameters, and the other 10% are used as a test set to evaluate the accuracy of the prediction. This gives a large testing period that spans almost 16 months, from February 19th, 2011 to June 12th, 2012. The test corpus contains total of 51,983 people, 98% of whom (50911) can be found in training corpus.

Person and event descriptions are created as described in Section 4.1. To avoid training the model with information that is part of the expected prediction, all person names are removed from the event description. For the textual approaches, we rank candidate people by the likelihood that their language model created this event. For the social network approaches we select, as the 'seed person', the person within the event having the highest degree. For all approaches, this seed name is removed from the ground truth and the result predictions.

We evaluate our models using two common Information Retrieval evaluation measures: *Precision at K* (P@K) and *mean average precision*(MAP). For all experiments, we use a brute force search to find the parameters of the model the optimize the MAP measure on the tuning set. We then use these optimal parameters on the test events.

5.1 Results

We show our main evaluation results in Table 3. The baseline language model achieves a MAP score of 0.155 and a P@1 score of 0.3086. Adding the frequency-based prior to the model increases MAP to 0.1665 (a relative improvement of 7%), and improves P@K by a similar amount, for all values of k. Adding the seed people names to the query gives a further 8% improvement in MAP, showing that this naive method of using social network information can be effective.

Unsurprisingly, the network-based approaches do not perform as well as the text-based approaches. It is notable, however, that the event-based co-occurrence network performs much better than the photo based co-occurrence network. Although this seems somewhat surprising, as photo co-occurrence should be a stronger indicator of relatedness, it suggests that the event-based network is more robust, possibly because it is less affected by data sparseness, and because this network is more relevant to our prediction task.

Using the raw network prior in combination with the language model improves over the basic language model, although it does not perform as well as the frequency-based prior. Smoothing the network-based prior with the frequency-based prior, however, gives a 12% improvement, and adding the seed name to the query gives further relative improvement of 1.8%. Overall, using priors and exploiting the latent social network (via network-based and text-based methods) gives an overall 20% improvement in MAP over the baseline Jelinek Mercer language model (increasing MAP from 0.155 to 0.187).

In Table 4 we show the results for the hierarchical temporal language model. As a baseline, we show the best time-agnostic results, which uses the smoothed network prior, and includes the seed name in the event description. All of the temporal models also use the smoothed network prior and the seed name. The first column shows the results from using temporal windows with the standard language model, meaning that only those events falling in the temporal window under consideration are used for training, and standard collection smoothing is used. The 1-month and 6-month standard language models are outperformed by the baseline, clearly showing that relying solely on the most recent information leads to poor performance. The 1-, 2- and 5-year standard language models, however, all outperform the baseline. For the 2 year model, the MAP is 3% better than the model built from all of the data, suggesting that older, ‘stale’, data can have a negative impact by introducing noise.

Looking at the hierarchical smoothing approaches, we see that they almost always improve over the time-agnostic baseline. The only exception is for the 1 month model with the temporal network prior, which calculates the network prior based on 1 month of data: the poor performance of this model suggests that a reliable social network cannot be constructed from 1 month of stock photo data. Also, when a ten year time window is used, the performance is almost the same as the baseline, showing that data older than 10 years has already lost its prediction power. In general, these results also show that, even for the temporal models, it is better to use the entire data to construct the social network, although the difference is relatively minor. The best overall result is achieved with 6 month hierarchical smoothing, using the entire social network. This method shows an 8% improvement over the best time agnostic

Table 4 MAP of temporal models. All methods use the smoothed network prior, and include the seed name in the query. *Temporal network prior* - creates the social network only using data from the time window being considered. *Full network prior* - creates the social network from all data.

	Standard LM	Hierarchical Smoothing (temporal network prior)	Hierarchical Smoothing (full network prior)
LM + Seed Name (smoothed network prior)	0.1868	-	-
1 month	0.0965	0.1801	0.1933
6 months	0.1689	0.1982	0.**2019**
1 year	0.1893	0.2002	0.2002
2 years	0.1935	0.1964	0.1965
5 years	0.1882	0.1904	0.1889
10 years	0.1868	0.1872	0.1875

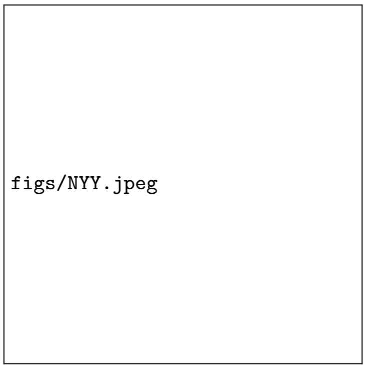


Fig. 6 A baseball match between New York Yankee and Boston Red Sox. Two players in the photo are *Alex Rodriguez* and *Bret Gardner* from New York Yankee.

method, and a 30% improvement over the standard Jelinek Mercer model. Note that Jelinek Mercer smoothing is a strong baseline, often used in information retrieval applications. In particular, when combined with a reliable estimate of the document prior it performs similarly to Dirichlet Smoothing, with the advantage that it can benefit from improved estimates of the prior[30,31]. All of this shows that we achieve the best performance when textual, network and temporal information are integrated into the language model.

5.1.1 Case Study: New York Yankees vs Boston Red Sox

To better illustrate how network and temporal information improve people prediction, in this section we have a deeper look at a concrete example of an event from our data. Figure 6 photo from US Major League Baseball between the New York Yankees and the Boston Red Sox in April 10, 2011, depicting the players *Alex Rodriguez* and *Bret Gardner* from New York Yankees. In our approach, *Alex Rodriguez* is selected as the seed person for this event and a subset of his ego-network, with four most frequently co-occurring teammates, is shown in Figure 7. In addition, the top five predicted persons using three algorithms, i.e. LM, LM with smoothed network prior as well as LM with

Table 5 Top five predicted attendees of a baseball match using three algorithms. Adding network and temporal information improves the prediction results.

LM	LM with smoothed network prior	LM with temporal smoothing
Joba Chamberlain*	Joba Chamberlain*	Brett Gardner*
Hideki Matsui	Robinson Cano*	Joba Chamberlain*
Brett Gardner*	Hideki Matsui	Robinson Cano*
Derek Jeter*	Derek Jeter*	Nick Swisher*
Bernie Williams	Brett Gardner*	Derek Jeter*

Fig. 7 An ego-network of Alex Rodriguez based on historical co-occurrence relationships.

temporal smoothing are shows in Table 5, with the correct prediction marked with an asterisk.

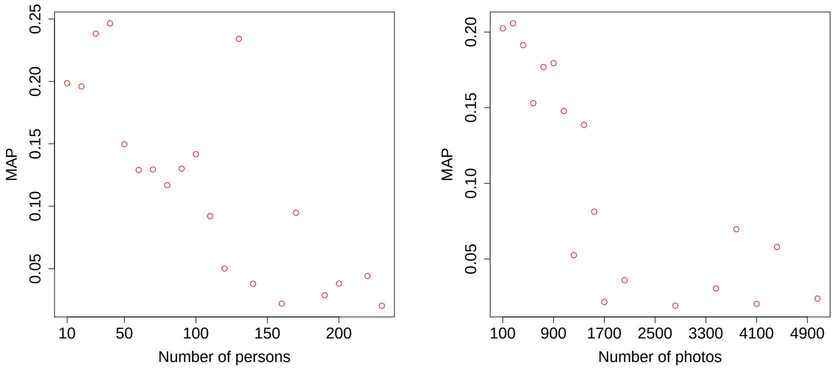
The basic LM algorithm correctly predicted three attendees, namely *Joba Chamberlain*, *Brett Gardner* and *Derek Jeter* who played in that match. The other two people, *Hideki Matsui* and *Bernie Williams*, were retried Yankee players at the time of this match. Adding the network prior ("LM with smoothed network prior") successfully identified another player, *Robinson Cano*, who frequently co-occurred frequently with the seed person, *Alex Rodriguez* (Figure 7). Finally, the LM with temporal smoothing algorithm surfaces another relevant player *Nick Swisher*. Particularly, *Nick Swisher* was a junior Yankee player at that time who joined the team after 2009. Therefore, compared with other senior Yankee players, there are relatively few photos of this player before 2009, but the number of photos increased dramatically after 2010. Such information is captured by the temporal smoothing version of LM by giving certain weight to more recent photos. This example shows that adding network and temporal information can improve prediction results.

5.2 Analysis

In the results above, even the best ranking yields relatively low scores for MAP (0.2019) and P@1 (0.3727), suggesting that predicting events attendees itself is a very difficult task. To put this into perspective, consider that the average event in our data only has 7 attendees, while the collection includes around 50,000 unique people. The dynamic nature of the events and their high diversity further increases prediction uncertainty. Despite this difficulty, we aim to distinguish several properties of events and analyze how they affect the prediction accuracy, and to uncover cases where more accurate prediction can be achieved. To this end, using the best performing predictive model, we investigate the variation of the prediction accuracy based on three event properties and their combination: size, recency, and topic.

5.2.1 Event Size

In our dataset, the size of an event can be measured by the total number of attendees or the total number of images depicting the event. We group all



(a) (b)

Fig. 8 The effects of event size, measured by the (a) the number of persons or (b) number of photos, on the prediction precision of events attendees. Attendees of smaller events are easier to predict.

test events into bins, according to the number of attendees or the number of photos, and for each bin we calculate the MAP for all events in that bin. Figure 8(a) shows the distribution of MAP scores across events grouped by the number of attendees, while Figure 8(b) shows the scores when events are grouped by the number of images. The results show a clear downward trend in the accuracy of predictions as the event size is increased, suggesting that attendees of small events are easier to predict. In particular, when an event has more than 1000 images, there is a sudden drop in MAP. We would suggest that the reason for this is because bigger events are generally more comprehensive, with various participants and broad topics, making the prediction more difficult. For instance, the Oscar Awards Ceremony generally contains hundreds of actors/actress/directers. By contrast, smaller events that have a clearer focus, like the premiere of a movie, will be easier to predict based on historical data.

5.2.2 Recency

The motivation for the hierarchical temporal language model is that by incorporating more recent and fresh data it should have stronger predictive power. Since the test data covers a period of over a year, however, many of the test events take place a long time after the data used to train the models. Therefore, we would expect the prediction accuracy to be higher for testing events that are temporally closer to the training events. To quantify this, and to understand the extent to which having fresh training data is important for this task, in Figure 9(a), and 9(b) we plot the prediction results, in MAP, grouped by the number of weeks, and the number of months, from the end of the time period covered by the training data. These results show that the

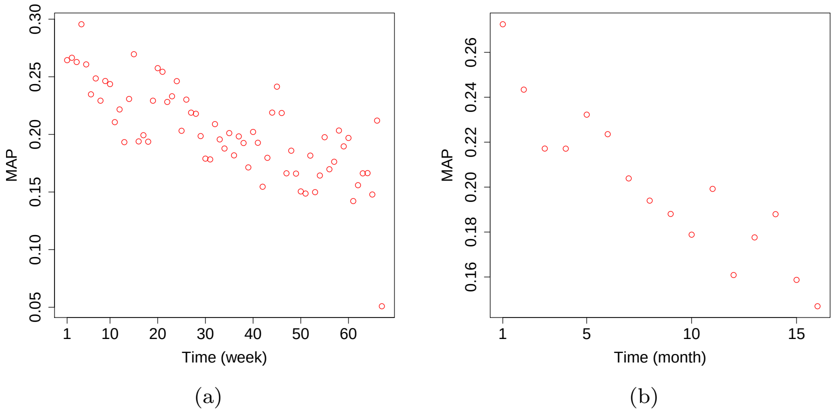


Fig. 9 The effects of time on prediction accuracy, measured by (a) weeks or (b) months after the end of training period. The prediction accuracy decreases as the temporal gap between testing events and training events becomes larger.

Table 6 Representative words of each topic

Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
soccer	baseball	sport	sport	president	film	fashion	music	sport	dress
ball	run	team	round	press	actor	people	arts	hockey	hair
goal	base	match	world	ceremony	actress	attend	performance	goal	shoe
team	field	stadium	golf	minister	director	culture	singer	game	black
coach	league	rugby	club	vice	premiere	portrait	entertainment	scoring	earring

freshness of the training data is indeed crucial for this task: those events in the month immediately after the training data have MAP of 0.2724, compared with 0.1587 MAP for events 15 months after the training period. Indeed, the performance in the first month after the training period represents a massive 33% improvement over the results reported in our main evaluation for the entire dataset, and is much closer to the level of performance that would be required for real-world applications.

5.2.3 Topic

Stock photo collections, such as the one we are studying, include photos from a diverse number of categories, but tend to be dominated by Sports, Entertainment, and Politics. In this section we analyse the differences in events belonging to different categories, in terms of the ability to predict their attendees. To discover the latent topics related to our test events, we apply the Latent Dirichlet Allocation (LDA) topic modeling algorithm to them [35]. We run the LDA algorithm on the test events, with the number of topics set to 10 (based on the work of Griffiths & Steyvers [36] and empirical knowledge, we set the other parameters as follows: $\alpha = 50/K$ and $\beta = 0.1$, iterations=1000).

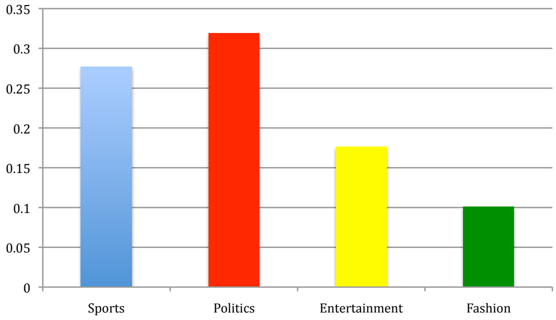


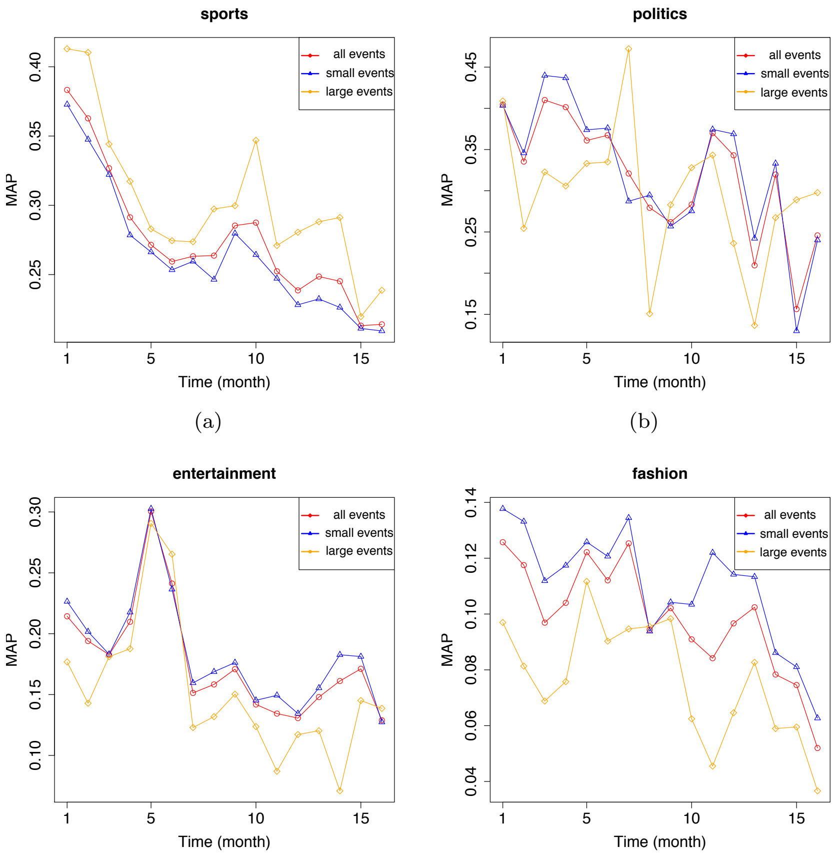
Fig. 10 Prediction precision of different topics. The rankings from the highest to the lowest are politics, sports, entertainment and fashion.

The representative words for each topic are listed in Table 6. From the inspection of the table it seems reasonable to manually cluster the 10 topics into higher level categories. Specifically, Topics 0, 1, 2, 3, 8 are related to *Sports*, Topic 4 is related to *Politics*, Topics 5 and 7 are related to *Entertainment*, and Topics 6 and 9 are related to *Fashion*. From the derived model, given the distribution of latent topics for each event, we take the topic with the maximum probability, and assign the event to its corresponding higher level category. From this, we obtain 661 political events, 2,981 entertainment events, 5,559 fashion events and 7,446 sports events.

Figure 10 shows the MAP for each of the 4 high level event categories. The average prediction performance of political events is the highest, followed by sports and entertainment events, while fashion is particularly difficult to predict, with a MAP of 0.1. This may be due by the fact that political events have often a well-determined structure with rather fixed co-occurrences of figures belonging to the same party, such as during the US presidential elections. Similarly, sports events are also relatively easy to predict. By contrast, entertainment and fashion events are more difficult, perhaps because the scope of candidate attendees are generally much broader than political events, and the form of these events are more diverse. For instance, a fashion show may attract movie stars, singers, athletes or even politicians. These topic based differences in performance are quite large: Political events reach 0.319 MAP, which is 60% better than the overall results reported in the main evaluation, while the Sports events reach 0.2771 MAP, which is 35% better.

5.2.4 Combined Effect of Size, Topic and Recency

While we note previously that the performance for the main evaluation is relatively poor, the better accurate topic-specific performance, combined with the more accurate results for fresher models, suggests that for certain types of events, given fresh enough training data, we can achieve relatively high accuracy for this task. In this section we investigate to what extent the intersection



(a) (b)

(c) (d)

Fig. 11 The joint effects of recency, topic and event size on the prediction accuracy on events of four topics: (a) sports, (b) politics (c) entertainment and (d) fashion. In each plot, events are grouped by the number of months from the testing period and divided into into small (less than 100 photos) and large (more than 100 photos) events. For all four topics, the precision decreases the time from the training period increases. The the exception of sports events, the precision of small size events better than large events, and slightly better than the overall precision for all events, regardless of event size.

of size, topic and recency have a combined effect on performance. In particular, given fresh enough training data on the right type of events, how accurately can we predict event attendees?

We define *small* events as events with less than 100 photos, and *large* events as those with more than 100 photos. In Figure 11 we plot MAP against the number of months from the training period for small, large and all events, for all four topics. As before, the MAP decreases as the time from the training period increases, verifying that the freshness of training data is important regardless of topic. Although the results show a general downward trend as time from the training period increases, there is much more variation than the general results, which is most likely caused by data sparsity, in that each data point (an intersection of topic, time, and size) is represented by less data.

The results show that, in line with the results discussed in Section 5.2.1, the topic-based MAP scores for small events are generally higher than for large events. On the other hand, the accuracy of Sports events attendees prediction is actually lower for small events: it is not clear why this is the case, and it warrants further investigation.

Focusing on Political and Sports events, which are the event types with the highest prediction performance, we see that the largest Sports events, when they occur recently following the training period, have a MAP close to 0.4130. The downward trend in the accuracy of prediction for Political events is less pronounced; they have a MAP of 0.4045 for the most recent events, but the accuracy for small political events is highest for events 3-4 months after the training period, when it approaches 0.4399 MAP. These results for Sports and Politics exceed twice the best MAP results (0.2019) reported in the main evaluation.

Overall, these results show that, although the overall MAP scores from the main evaluation may be seem quite low, and suggest that this is a difficult task, for right type of events and with fresh enough models are fresh enough the prediction performance can approach the levels of accuracy required for real applications.

6 Conclusions

In this paper, we address the problem of predicting celebrities attendees at future event, using metadata from a public stock photo collection of almost 30 million images. A basic text-based language model uses the textual similarity between a person's metadata and an event description, and incorporates a prior probability of people based on their frequency of occurrence in the collection. To take advantage of the most recent data without ignoring the older data in the corpus, we implement hierarchical temporal smoothing language models. We also build social networks based on photo and event co-occurrence, and use these networks to improve the prediction performance. The experimental results show that combining textual, network and temporal information gives the best performance. Additionally, we analyze the effect of event size, model freshness and topic on prediction performance, and find that with fresh models, the attendees of small Political or large Sports events can be predicted much more accurately than other types of events, with a MAP score of close to 0.45.

These models can be used to predict celebrities attending any events for which we have textual metadata available, which can be very useful in event-based multimedia indexing. We construct textual event descriptions from the metadata of photos depicting the events: in practice, though, any textual description of an event would suffice. These textual descriptions of events and people could also be used, of course, to compare events with events, people with people, and to help discover previously unknown links between events and people. Also, although this paper has focused on predicting attendees at events, the results may be useful even when the people are not actually event attendees, in that the prediction demonstrates a strong relation to the event or event type.

Our work can be further improved in future. First, we need to figure out a robust way to automatically find the known seed persons events for network based prediction. In this paper we assume that a single attendee of an event is already known, but this may not always be the case in practice. We will investigate whether pseudo relevance feedback for this purpose can improve performance. Secondly, we will consider the semantics of social connections in the co-occurrence network. In the present work all social connections are considered equal, but when a person has many connections to other persons, these connections can have different semantic labels based on different types of co-occurring events, and this information can be used to improve the prediction performance. We also plan to explore the use of content based face recognition in combination with our approaches: while the availability of images implies an event that is in the past and we can no longer consider the task strictly a 'prediction' one, the same methods can nevertheless still be applied. Finally, we note that in this paper we have focused on a dataset of stock photos from a professional agency: it is not clear whether the methods could be used as successfully with user generated content, and we would like to explore that in future work.

References

1. O'Hare N, Aiello LM, Jaimes A (2012) Predicting participants in public events using stock photos. In: Proceedings of the 20th ACM international conference on Multimedia. New York, NY, USA: ACM, MM '12, pp. 1093–1096. doi:10.1145/2393347.2396391. URL http://doi.acm.org/10.1145/2393347.2396391.
2. Xie L, Sundaram H, Campbell M (2008) Event mining in multimedia streams. In: Proceedings of the IEEE. pp. 623–647.
3. Cooper M, Foote J, Girgensohn A, Wilcox L (2005) Temporal event clustering for digital photo collections. ACM Trans Multimedia Comput Commun Appl 1: 269–288.
4. Luo J, Yu J, Joshi D, Hao W (2008) Event recognition: viewing the world with a third eye. In: Proceedings of the 16th ACM international conference on Multimedia. New York, NY, USA: ACM, MM '08, pp. 1071–1080. doi:10.1145/1459359.1459574. URL http://doi.acm.org/10.1145/1459359.1459574.
5. Rattenbury T, Good N, Naaman M (2007) Towards automatic extraction of event and place semantics from flickr tags. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA: ACM, SIGIR '07, pp. 103–110. doi:10.1145/1277741.1277762. URL http://doi.acm.org/10.1145/1277741.1277762.
6. Becker H, Naaman M, Gravano L (2010) Learning similarity metrics for event identification in social media. In: Proceedings of the third ACM international conference on Web search and data mining. New York, NY, USA: ACM, WSDM '10, pp. 291–300. doi:10.1145/1718487.1718524. URL http://doi.acm.org/10.1145/1718487.1718524.
7. Chen L, Roy A (2009) Event detection from flickr data through wavelet-based spatial analysis. In: CIKM. ACM, pp. 523–532. doi:10.1145/1645953.1646021. URL http://doi.acm.org/10.1145/1645953.1646021.

8. Petkos G, Papadopoulos S, Mezaris V, Troncy R, Cimiano P, et al. (2014) Social event detection at mediaeval: a three-year retrospect of tasks and results. In: ICMR 2014 Workshop on Social Events in Web Multimedia (SEWM). Glasgow, UK: ACM.
9. Samangooei S, Hare JS, Dupplaw D, Niranjan M, Gibbins N, et al. (2013) Social event detection via sparse multi-modal feature selection and incremental density based clustering. In: MediaEval'13. pp. -1–1.
10. Brenner M, Izquierdo E (2013) Mediaeval 2013: Social event detection, retrieval and classification in collaborative photo collections. In: Larson MA, Anguera X, Reuter T, Jones GJF, Ionescu B, et al., editors, MediaEval. CEUR-WS.org, volume 1043 of *CEUR Workshop Proceedings*.
11. Zhao M, Teo YW, Liu S, Chua TS, Jain R (2006) Automatic person annotation of family photo album. In: Proceedings of the 5th international conference on Image and Video Retrieval. Berlin, Heidelberg: Springer-Verlag, CIVR'06, pp. 163–172.
12. Zhang L, Chen L, Li M, Zhang H (2003) Automated Annotation of Human Faces in Family Albums. In: MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia. Berkeley, CA, pp. 355–358.
13. Sivic J, Zitnick C, Szeliski R (2006) Finding People in Repeated Shots of the Same Scene. In: Proceedings of the British Machine Vision Conference. Edinburgh, UK, pp. 909-918.
14. Anguelov D, Lee KC, Gokturk SB, Sumengen B (2007) Contextual Identity Recognition in Personal Photo Albums. In: CVPR. IEEE, pp. 1–7.
15. Davis M, Smith M, Canny J, Good N, King S, et al. (2005) Towards context-aware face recognition. In: Proceedings of the 13th annual ACM international conference on Multimedia. New York, NY, USA: ACM, MULTIMEDIA '05, pp. 483–486. doi: 10.1145/1101149.1101257. URL `http://doi.acm.org/10.1145/1101149.1101257`.
16. O'Hare N, Smeaton AF (2009) Context-aware person identification in personal photo collections. IEEE Transactions on Multimedia, Special Issue on Integration of Context and Content for Multimedia Management 11: 220–228.
17. Zickler T, Stone Z, Darrell T (2008) Autotagging facebook: Social network context improves photo annotation. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. CVPRW '08, pp. 1–8.
18. Naaman M, Yeh RB, Garcia-Molina H, Paepcke A (2005) Leveraging context to resolve identity in photo albums. In: JCDL. ACM, pp. 178–187. doi:10.1145/1065385.1065430. URL `http://doi.acm.org/10.1145/1065385.1065430`.
19. Mavridis N, Kazmi W, Toulis P (2010) Friends with faces: How social networks can enhance face recognition and vice versa. In: Abraham A, Hassanien AE, Snasel V, editors, Computational Social Network Analysis, Springer London, Computer Communications and Networks. pp. 453-482.
20. Wu P, Tretter D (2009) Close & closer: social cluster and closeness from photo collections. In: Proceedings of the 17th ACM international conference on Multimedia. New York, NY, USA: ACM, MM '09, pp. 709–712. doi:10.1145/1631272.1631394. URL `http://doi.acm.org/10.1145/1631272.1631394`.
21. Singla P, Kautz H, Luo J, Gallagher A (2008) Discovery of social relationships in consumer photo collections using markov logic. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. CVPRW '08, pp. 1–7.
22. Kim HN, Jung JG, El Saddik A (2010) Associative face co-occurrence networks for recommending friends in social networks. In: Proceedings of second ACM SIGMM workshop on Social media. ACM, WSM '10, pp. 27–32. doi:10.1145/1878151.1878160. URL `http://doi.acm.org/10.1145/1878151.1878160`.
23. Devezas J, Coelho F, Nunes S, Ribeiro C (2012) Studying a personality coreference network in a news stories photo collection. In: Proceedings of the 34th European conference on Advances in Information Retrieval. Berlin, Heidelberg: Springer-Verlag, ECIR'12, pp. 485–488.
24. Ahmed A, Batagelj V, Fu X, Hong SH, Merrick D, et al. (2007) Visualisation and analysis of the internet movie database. In: APVIS. pp. 17-24.
25. Aragon P, Kaltenbrunner A, Laniado D, Volkovich Y (2012) Biographical social networks on wikipedia - a cross-cultural study of links that made history. Arxiv preprint arXiv : 4.

26. Mantrach A, Renders JM (2012) A general framework for people retrieval in social media with multiple roles. In: Proceedings of the 34th European conference on Advances in Information Retrieval. Berlin, Heidelberg: Springer-Verlag, ECIR'12, pp. 512–516.
27. Aiello LM, Barrat A, Schifanella R, Cattuto C, Markines B, et al. (2012) Friendship prediction and homophily in social media. ACM Trans Web 6: 9:1–9:33.
28. Aiello LM, Barrat A, Cattuto C, Ruffo G, Schifanella R (2010) Link creation and profile alignment in the anobii social network. In: Proceedings of the 2010 IEEE Second International Conference on Social Computing. Washington, DC, USA: IEEE Computer Society, SOCIALCOM '10, pp. 249–256. doi:10.1109/SocialCom.2010.42. URL http://dx.doi.org/10.1109/SocialCom.2010.42.
29. Ponte JM, Croft WB (1998) A language modeling approach to information retrieval. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA: ACM, SIGIR '98, pp. 275–281. doi:10.1145/290941.291008. URL http://doi.acm.org/10.1145/290941.291008.
30. O'Hare N, Murdock V (2013) Modeling locations with social media. Information Retrieval 16: 30-62.
31. Smucker MD, Allan J (2005) An investigation of dirichlet prior smoothing's performance advantage. Technical report, The Center for Intelligent Information Retrieval, The University of Massachusetts.
32. Jelinek F, Mercer RL (1980) Interpolated estimation of markov source parameters from sparse data. In: In Proceedings of the Workshop on Pattern Recognition in Practice. Amsterdam, The Netherlands: North-Holland, pp. 381-397.
33. Westerveld T, de Vries AP, Westerveld AT, de Vries AP, van Ballegooij AR (2003) CWI at the TREC-2002 Video Track. In: The Eleventh Text REtrieval Conference (TREC-2002). Gaithersburg, MD, pp. 207-216.
34. Serdyukov P, Murdock V, van Zwol R (2009) Placing Flickr Photos on a Map. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, pp. 484–491.
35. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3: 993–1022.
36. Griffiths TL, Steyvers M (2004) Finding scientific topics. Proceedings of the National Academy of Sciences 101: 5228-5235.