SOCIAL EVENT DISCOVERY BY TOPIC INFERENCE

Xueliang Liu, Benoit Huet

EURECOM
Sophia Antipolis, France

ABSTRACT

With the keen interest of people for social media sharing websites the multimedia research community faces new challenges and compelling opportunities. In this paper, we address the problem of discovering specific events from social media data automatically. Our proposed approach assumes that events are conjoint distribution over the latent topics in a given place. Based on this assumption, topics are learned from large amounts of automatically collected social data using a LDA model. Then, event distribution estimation over a topic is solved using least mean square optimization. We evaluate our methods on locations scattered around the world and show via our experimental results that the proposed framework offers promising performance for detecting events based on social media.

1. INTRODUCTION

In recent years, online media sharing websites are playing a growing and important role in our daily life. These online services make it possible to upload and share users' life record, in the form of microblog, photos or videos. Nowadays, with the development of mobile communication, we can easily and instantly capture and share content as we experience it, such as in a concert, a party or during a journey. Meanwhile, the tremendous popularity of social media also brings lots of new challenges. The research community has noticed the importance of mining valuable information from large amount of media data available on crowded social sharing platforms.

Events are a natural way for us, human, to organize and browse through our media collection. Event recognition has gained significant interest in the past decade. To address the problem, Quack *et al.* [1] presented methods to mine events and object from community photo collections by clustering approaches. A similar problems is also studied in [2] where Firan *et al.* focused on building a Naive Bayes event models which classify photos as either relevant or irrelevant to given events. In [3], an approach is proposed to detect specific events based on the analysis of uploading behavior along time.

It is well known that social media data collecting is a much easier task than labeling. For this reason unsupervised methods operating on unlabelled data have attracted considerable attention. In particular, the topic models such as LSA, pLSA or LDA[4] have shown encouraging results. In [5], Latent Dirichlet Allocation (LDA) has been employed in an incremental learning framework to create well labeled dataset automatically. In this paper, we employ the LDA scheme to model the topics in a city for event detection.

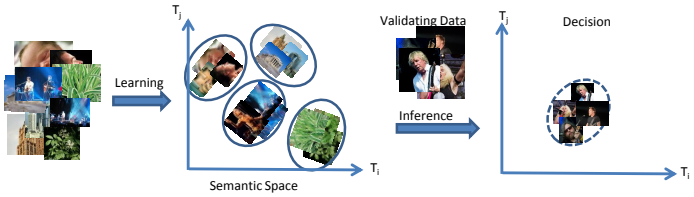

Fig. 1: The proposed framework

In this paper, we propose a method for discovering specific event from social media data. In details, we take the events as a special conjoint distribution over latent topics. As shown in Figure 1, first the topics are learned from large quantities of data captured at a given location. Then, we use the least mean square algorithm to estimate the events distribution on a group of validated data samples. We detect the events, from a test dataset, if they fit the distribution over latent topics well. Importantly, unlike the some previous work on the event classification, the object of this paper is to discover specific events such as Lady Gaga concert, the wedding of Prince William, etc... To the best of our knowledge, this is the first attempt to discover specific events from latent topics point of view.

The remaining of this paper is structured as follow. In Section 2, we detail our approach for detecting and identifying events in social media. In Section 3, we briefly describe the dataset and present experimental results. Finally, we conclude and outline future work in Section 4.

2. EVENT DETECTION

The goal of the proposed work is to detect events from social media data for a given location automatically. To solve the problem, we consider inferring latent topics existing in lo-

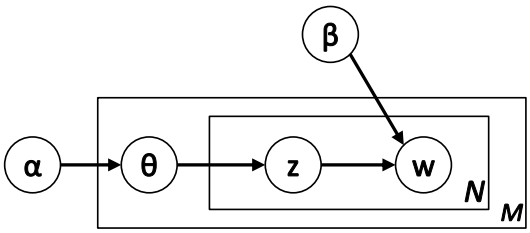

Fig. 2: LDA model

calised media data. These topics are stationary with respect to a location and can be learned from a large scale dataset. Documents originating from an event have a specific distribution over these topics. Our framework consists in two steps: (i) the distribution for the given social media documents, which are solved by the LDA model. (ii) Estimating the distribution of events over the topics, which could be solved by the least mean square optimization on the KL divergence measure. All of the details will be given in the following section.

2.1. Topics Learning

For a given place (a city for example), the set of topics associated with a period of time can be seen as stationary. The semantic events can be regarded as special distribution over those topics. There have already been lots of approaches to infer topics among documents. The method used in this paper is the LDA model [4] which is a generative probabilistic graphical model to discover topics in documents, as shown in Figure 2. The generative rule is that the documents are represented as multinomial distribution over the latent topics, and each topic is characterized by a distribution on the words dictionary. For a given document w_d, it generates the word in the following process:

1. Choose $\theta_i \sim Dir(\alpha)$, where $i \in \{1, \ldots, M\}$
2. For each of the words w_{ij}, where $j \in \{1, \ldots, N_i\}$
   (a) Choose a topic $z_{i,j} \sim Multinomial(\theta_i)$.
   (b) Choose a word $w_{i,j} \sim Multinomial(\beta_{z_{i,j}})$.

Where the M denotes the number of documents, N refers to the number of words in documents, and α is the hyperparameter of the Dirichlet distribution on the latent topics. Given the observed documents, the likelihood of the model can be calculated as

$$P(W, z, \theta|\alpha, \beta) = \prod_{d=1}^{M} \int P(\theta_d|\alpha) \prod_{n=1}^{N} P(z_{d,n}|\theta_d) P(W_{d,n}|z_{d,n}, \beta) d\theta_d$$

The model parameters can be learned by some Bayes inference methods, such as variable inference, Gibbs Sampling or EM algorithm.

In practice, we collect the geo-tagged Flickr photos for a given city (or location), and choose stem words from the title and tags of each photo to represent the social media in the LDA models training. Here, we should argue that the textual feature is not the only feature which could be used, other representations (Bag of visual words[6], for example) also fit our framework. When the models are obtained, they are used to infer distribution over these topics on the validated data to estimate the distribution of event, which will be detailed in the following section.

2.2. Events Estimation

From the LDA model, the distribution of a document d over latent topics can be inferred by equation 1.

$$P(\theta_d|\alpha, \beta, d) = \int \int p(w, z, \theta|\alpha, \beta) dw dz \tag{1}$$

However, there is still a binary classification problem to solve: assign a social media document with an event versus no event. Here, we estimate the distribution of events over the inferred topics using validation data, which are the positive media documents of an event. The details concerning the validation data acquisition will be introduced in the experiment section.

Suppose D is the inference of the validation data over the latent topics, the event distribution $\mathbf{e}$ can be estimated by least mean square optimization theory. The objective is to minimize equation 2.

$$\mathbf{e} = \underset{e \in R^N}{\operatorname{argmin}} \sum_i Dist(D_i, e) \tag{2}$$

Where the function $Dist$ measures the distance between a validating instance D_i and the events estimation $\mathbf{e}$. It is well known that the best measure between two distributions is Kullback Leibler divergence. However KL divergence is not a symmetric measure. We use the following standard symmetric version as the distance measure

$$\begin{aligned} Dist(p, q) &= D_{\mathrm{KL}}(p\|q) + D_{\mathrm{KL}}(q\|p) \\ &= \sum_i p(i) \log \frac{p(i)}{q(i)} + \sum_i q(i) \log \frac{q(i)}{p(i)} \end{aligned} \tag{3}$$

When the event distribution is estimated from equation 2, it can be used to verify if a new document d is event related or not, according to the rule defined in equation 4.

$$d \ is \begin{cases} event, & \text{if } Dist(d, \mathbf{e}) \leq T \\ noevent, & otherwise \end{cases} \tag{4}$$

Where the value T is the threshold of the decision function, which is used to judge if a document is relevant or not in the detection process. In practice, the value of T can be inferred from the validation dataset D as follows:

$$T = k \max_i \{Dist(D_i, \mathbf{e})\} \tag{5}$$

where k is used to suppress the influence of noise contained in the validation dataset. It will be studied in the Experiment section 3.

3. EXPERIMENTS AND RESULTS

To validate the proposed approach, we collected a large dataset from the Internet. The events we aim to detect in this study are concerts. There are two reasons for this choice. On one hand, concerts are popular and important social activities of our life and there are significant amounts of photos taken during concerts and shared on Flickr. On the other hand, it is easy to generate the ground truth on convert events, thanks to event directories such as LastFM, Upcoming or directly from the venue's agenda. In this paper, we concentrate on 7 venues that are located in 6 cities around the world during May 2010 for concert event detection. Other types of events could also be detected with our approach given the appropriate dataset / ground truth combination.

First, we introduce the data collection used in the experiments, and then show a walk through for how our approach works on event detection.

3.1. Dataset

(1) **Training Data** The training photos set is crawled from Flickr based on its public API. We have chosen 6 cities, which are located in Europe and America. The geographic information of these cities (such as the cities geo-coordinates and size) is obtained from Wikipedia. The cities shape is approximated to a circle for simplicity reasons. Although such assumption is reasonable since previous research has studied the distribution rule of social media and shown that most photos are taken in the center of cities [7]. Using geo-coordinates based queries, we gather a collection containing about 49K photos during the month of May 2010 in these 6 cities.
(2) **Validation Data** The validation set includes the photos taken during events which have been held in a target venues in the past. They are used to estimate the event distribution as described in section 2.2. The validation data is collected by Flickr API with event machine tags, as proposed in [3].
(3) **Test Data** The test data includes the social documents that we aim to mine events from. This dataset is collected using the Flickr API with queries combining location (venue coordinate) or text (event title) and time (May 2010), as in [3].

A summary of the photos on the three dataset can be found in table 1. Since the training set is collected at the city level, the venues "Koko" and "HMV Forum" share the same training set but different validation set.

Besides the dataset, we also create the ground truth on these venues for May 2010, based on the events that are listed in the agenda of the venues' website. However, it is important to mention that not all of events are represented in social media data; It is likely that some for some events no photos were captured or shared on Flickr. In order to estimate the subset events from the ground truth that could be inferred from testing data, a manual process is performed to label if a photo is relevant to concert or not. Finally, the labeled subset obtained is used as the ground truth for evaluation. The number of ground truth events for each venue is reported in table 3.

Table 1: Photos Collections over the Venues.

Venue	City	Training Set	Validating Set	Testing Set
Melkweg	Amsterdam	3786	179	355
Koko	London	23384	194	724
111 Minna Gallery	Chicago	11725	175	313
Ancienne Belgique	Brussel	2120	321	496
Rotown	Rotterdam	1575	71	118
Circolo degli Artisti	Rome	6551	107	167
HMV Forum	London	23384	189	97

3.2. Results

To evaluate the proposed approach, we learn LDA model on the training data and employ the trained models to infer the topics distribution on both validation data and test data. Then, the decision rule can be learned after the inference process on the validation data. In the events distribution estimation, the ratio k in equation 5 plays an important role in balancing the recall and precision rate. Obviously, lower k value will lead to high recall but lower precision, and vice versa.

To obtain the accurate value of k, we calculate the ratio

$$k = \frac{Dist(D_i, e)}{\max(Dist(D_i, e))}, D_i \in D$$

Figure 3 illustrates the effect of k on the validation data. From the figure, it is clear that the number of photos decreases as k increases and a noticeable drop occurs for $k \geq$ 0,3. Based on this result, we choose $k = 0.3$ in equation 5.

The final event detection process is performed on the test data and the results are manually and individually checked, based on the matching between the textual descriptions of images and ground truth event. Since we detect events at the media document level, more than one document is inferred to the same event. Therefore, we evaluate precision at the documents level and the recall at events level respectively. Table 2 report the statistics of media data on event detection for the 7 venues. In this table, the number of documents that are detected as event-related is represented. In total, 265 out of the 2270 photos are identified as event-related and 160 photos (out of the 265) are assigned to the right event, leading to an average precision of 0.60.

Table 3 reports the performance of our event detection approach in terms of recall. In total, out of the 99 events available in the dataset (according the meta data), 63 of them are detected by our approach. This corresponds to a recall of 0.64.

In addition, our proposed approach is robust when handling the semantic on social data. In our selected venues,

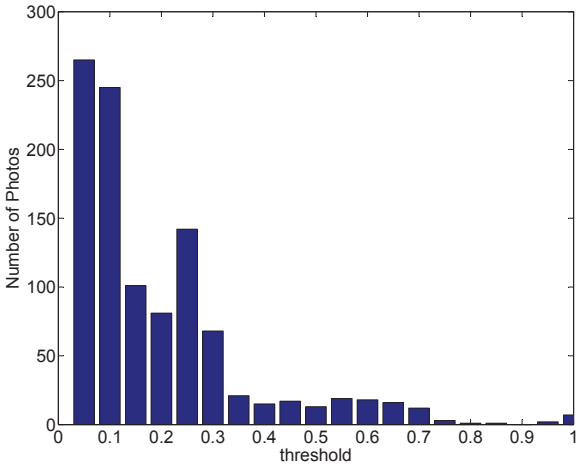

Fig. 3: The histogram of threshold value

Table 2: Social Media Data statistics over Event Detection

Venue	Total Number	Detection	Positive	Precision
Melkweg	355	42	32	0.76
Koko	724	95	44	0.46
111 Minna Gallery	313	26	10	0.38
Ancienne Belgique	496	32	19	0.59
Rotown	118	6	4	0.67
Circolo degli Artisti	167	46	36	0.78
HMV Forum	97	18	15	0.83
Total	2270	265	160	0.60

"KoKo" and "HMV Forum" are located in the same city "London". The results on these two places are obtained from the same topics model, which is trained on the photo documents captured all over London. Nonetheless, acceptable results are achieved on the two places, as shown in Table 2 and 3. Those findings strongly support the assumption that event semantics can be taken as special distribution over latent topics which could be learned from the media collection of an entire city.

4. CONCLUSION AND FUTURE WORK

We presented a novel method for automatically detecting events taking place at given location and time. In this paper, the events are taken as special distribution over latent topics. We mine the topics from a large data collection of shared social media and use the venue specific validation data to infer the event distribution. The experimental results using Flickr photo data demonstrate the effectiveness of the proposed approach. In the current work, latent topics are mined from text. As part of future work, we aim at exploiting multi-modality data(such as visual feature, and EXIF metadata) to improve our social event detection algorithm.

Table 3: Social Event Detection Performance

Venue	GroundTruth	Detection	Recall
Melkweg	27	14	0.52
Koko	15	12	0.80
111 Minna Gallery	4	4	1.00
Ancienne Belgique	19	10	0.53
Rotown	7	2	0.29
Circolo degli Artisti	17	15	0.88
HMV Forum	10	6	0.60
Total	99	63	0.64

Acknowledgments

The research leading to this paper was partially supported by the project AAL-2009-2-049 "Adaptable Ambient Living Assistant" (ALIAS) co-funded by the European Commission and the French Research Agency (ANR) in the Ambient Assisted Living (AAL) program.

5. REFERENCES

[1] Till Quack, Bastian Leibe, and Luc Van Gool, "World-scale mining of objects and events from community photo collections," in *the 2008 international conference on CIVR*, New York, USA.

[2] Claudiu S. Firan, Mihai Georgescu, Wolfgang Nejdl, and Raluca Paiu, "Bringing order to your photos: Event-Driven Classification of Flickr Images Based on Social Knowledge," in *Proceedings of the 19th ACM international conference on CIKM*, New York, USA, 2010.

[3] Xueliang Liu, Raphaël Troncy, and Benoit Huet, "Using social media to identify events," in *the 3rd ACM SIGMM international workshop on Social media*, New Yorkk, USA, 2011.

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, May 2003.

[5] Li-Jia Li, Gang Wang, and Li Fei-Fei, "OPTIMOL: automatic Online Picture collecTion via Incremental Model Learning," *IEEE Computer Society Conference on CVPR*, 2007.

[6] Yu-Gang Jiang, Chong-Wah Ngo, and Jun Yang, "Towards Optimal Bag-of-Features for Object Categorization and Semantic Video Retrieval," in *Proceedings of the 6th ACM international conference on CIVR*, 2007, pp. 494–501.

[7] Livia Hollenstein and Ross Purves, "Exploring place through user-generated content: Using Flickr to describe city cores," *Journal of Spatial Information Science*, vol. 1, no. 1, 2010.