Event Detection via LDA for the MediaEval2012 SED Task

Konstantinos N. Vavliakis, Fani A. Tzima, and Pericles A. Mitkas
Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki
&
Information Technologies Institute, CERTH
Thessaloniki, Greece
kvavliak@issel.ee.auth.gr, fani@issel.ee.auth.gr, mitkas@eng.auth.gr

ABSTRACT

In this paper we present our methodology for the Social Event Detection Task of the MediaEval 2012 Benchmarking Initiative. We adopt topic discovery using Latent Dirichlet Allocation (LDA), city classification using TF-IDF analysis, and other statistical and natural language processing methods. After describing the approach we employed, we present the corresponding results, and discuss the problems we faced, as well as the conclusions we drew.

Categories and Subject Descriptors

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

Keywords

Event Detection, Latent Dirichlet Allocation (LDA), Topic Identification, Flickr, MediaEval

1. INTRODUCTION

This paper discusses our approach for the MediaEval2012 Social Event Detection (SED) task. SED organizers provided a collection of 167,332 photos taken from Flickr, captured between the beginning of 2009 and the end of 2011 by 4,422 unique Flickr users. This year the SED task comprises three challenges: **a)** Challenge 1: Find technical events that took place in Germany, **b)** Challenge 2: Find all soccer events taking place in Hamburg (Germany) and Madrid (Spain), and **c)** Challenge 3: Find demonstration and protest events of the Indignados movement occurring in public places in Madrid. In all three challenges, social events are defined as events planned by people, attended by people and illustrated by media captured by people [1].

2. METHODOLOGY

Figure 1 depicts the methodology we employed for all three challenges, consisting of five steps: 1) preprocessing, 2) city classification, 3) topic identification, 4) event detection, and 5) event optimization. Since our methodology does not involve training on past data, we have not used SED's development kit. Next, we discuss each step in more detail.

MediaEval 2012 Workshop, October 4-5, 2012, Pisa, Italy

Preprocessing. First, the textual metadata (title, description and tags) of all available photos were cleaned by removing stop words and html tags. Next, non-English terms were translated using the Google Translate web service. Finally, the terms were stemmed using an implementation of the Porter stemmer. This step created three versions of textual data – *cleaned text*, *translated text* and *stemmed text* – suitable for input to the following steps. We conducted separate experiments for each input, as discussed in Section 3.

City classification. According to the dataset description, longitude and latitude information was available for 20% of photos. By projecting this information in Google Tables, we noticed that photos had been taken in 5 cities: Köln, Hamburg, Hanover, Madrid, and Barcelona. Since the challenges were specified by city, we built a city classifier, in order to reduce the size of the required dataset for each challenge. We extracted the terms appearing in the geolocated photos of each city and calculated their TF-IDF values. Then, we classified photos with no geolocation information to the closest city, in terms of TF-IDF values, by summing up and normalizing the TF-IDF values of their terms.

For photos with no textual information, we used information from other photos of the same user, if available. Furthermore, we assumed that each user may have visited up to two cities in the same day, and that traveling from one city to another would require at least two hours. This assumption enabled us to improve our city classifier, and correct the location of some misclassified photos by using a "majority vote" on classified photos. Overall, city classification left 4,149 unclassified photos out of the original 167,332.

Topic Detection. Next, we focused on discovering, for each city, a set of topics that provide quantitative measures and can identify the semantic content of the photos' textual information. To this end, we employed *LDA topic modeling* with Gibbs samplimg, which is based on the assumption that each photo p_i can be described as a random mixture over topics, and each topic as a focused multinomial distribution over words.

After building the topics characterizing each city, our goal was to select the topics relevant to each challenge, so we retrieved photos with characteristic keywords (e.g. *indignados*, *spanish revolution*, *yeswecamp* and *15m* for Challenge 3), ordered their topics by number of appearences, and manually selected the top t topics, judging by their relevancy to each challenge. Apart from building topics using the LDA process, we also created a "manual" topic for each challenge.

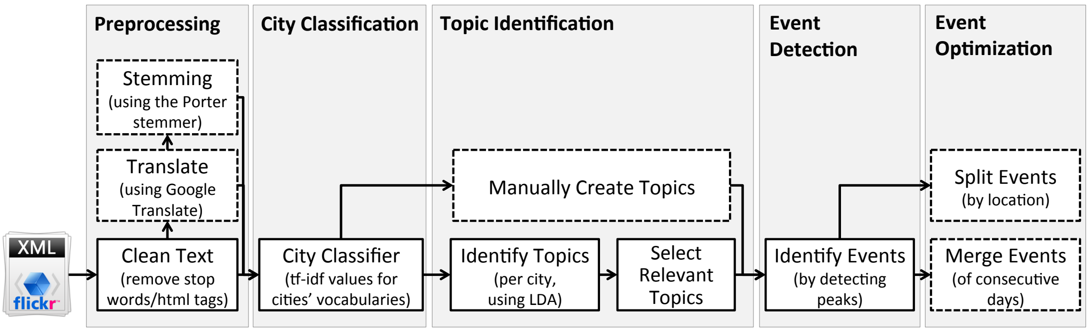
Preprocessing
City Classification
Topic Identification
Event Detection
Event Optimization
XML
flickr
Stemming (using the Porter stemmer)
Translate (using Google Translate)
Clean Text (remove stop words/html tags)
City Classifier (tf-idf values for cities' vocabularies)
Manually Create Topics
Identify Topics (per city, using LDA)
Select Relevant Topics
Identify Events (by detecting peaks)
Split Events (by location)
Merge Events (of consecutive days)

Figure 1: Overview of our methodology regarding MediaEval2012 SED

This topic consisted of selected keywords, based on our observation of topics and aggregate statistics of tags.

Event Detection and Optimization. Having identified the relevant topics for each challenge, we proceeded to event discovery through peak identification in the photos time series. We considered a new event whenever the number of photos was greater than a threshold *D* for a specific day.

The last step of our methodology included the fine tuning of the discovered events, through two additional processes: a) merging of events happening on consecutive days, and b) splitting of events than happen on the same day, but in different places/parts of a city (applied only in Challenge 3).

3. EXPERIMENTS AND RESULTS

In this Section, we present a set of 5 experiments per challenge and the corresponding evaluation results, obtained from the SED task organizers and summarized in Table 1. The first column of Table 1 identifies the run number for each challenge, *T* is the number of total topics detected by LDA (*M* is for the "manual" topic), while *t* is the number of relevant topics selected. The *D* parameter defines the threshold in the event detection process, while the *Text* parameter denotes the text version used as input: *T* means text as extracted from the title/description/tags fields, *TC* cleaned text, *TCE* translated cleaned text, and *TCES* translated cleaned and stemmed text. Finally, *P*, *R*, *F* and *NMI* are for Precision, Recall, F-measure and Normalized Mutual Information, respectively.

Regarding Challenge 1, the difference between runs #4 and #5 is the removal of the keyword *gamescom*. As far as Challenges 2 and 3 are concerned, runs #5 resulted from runs #4 after the event optimization step was performed.

Table 1: Results for Challenges 1, 2 and 3

#	t/T	D	Text	P	R	F	NMI
Challenge 1							
1	2/50	5	TC	**80.98**	19.25	31.10	0.211
2	6/50	2	TC	40.52	19.43	26.26	0.165
3	8/50	5	TC	35.85	19.56	25.31	0.160
4	M	2	T	76.29	**94.90**	**84.58**	**0.724**
5	M	2	T	63.35	50.98	56.50	0.578
Challenge 2							
1	1/50	5	TC	75.72	79.71	77.67	0.698
2	1/100	5	TC	86.67	77.42	81.78	0.741
3	1/100	5	TCE	**91.21**	77.85	84.00	0.768
4	M	5	T	88.18	**93.49**	**90.76**	**0.850**
5	M	5	T	88.18	**93.49**	**90.76**	0.847
Challenge 3							
1	5/100	5	TC	88.53	80.43	84.29	0.376
2	5/100	5	TCES	**90.76**	81.91	86.11	0.315
3	3/50	5	TCES	86.59	84.20	85.38	0.330
4	M	5	T	88.91	**90.78**	**89.83**	**0.738**
5	M	5	T	88.91	**90.78**	**89.83**	0.347

4. DISCUSSION

Taking a close look at Table 1, we may draw various conclusions. First, the proposed methodology is effective and can provide good results, although the formulation and selection of the topics can lead to significant variations. In all challenges the "manual" topics give better results. This is, however, expected, since concepts in the "manual" topics are derived from selected keywords of the LDA topics, plus tags of high occurrence believed to be useful.

In any case, topics identified automatically with LDA, also provide good results, with the exception of Challenge 1. This is also expected as Challenges 2 and 3 are better defined and simpler, and, thus, can be described by a limited number of topics. On the other hand, Challenge 1 is about technical events (mainly conferences) described by a diverse vocabulary and often comprising relatively few photos, thus resulting in topics that contain concepts from irrelevant photos.

Regarding text preprocessing, we notice that in Challenge 2 (runs #2 and #3), and in Challenge 3 (runs #1 and #2), translation and stemming improves both precision and recall. Finally, by comparing runs #4 and #5 for Challenges 2 and 3, we notice that the event optimization step actually reduces NMI, which means that in the ground truth, events could not span over multiple days and/or locations.

5. ACKNOWLEDGMENTS

The work presented in this paper has been partially supported by the CASSANDRA FP7-288429 Project.

6. REFERENCES

[1] S. Papadopoulos, E. Schinas, V. Mezaris, R. Troncy, and I. Kompatsiaris. Social Event Detection at MediaEval 2012: Challenges, Dataset and Evaluation. In *MediaEval 2012 Workshop*, Pisa, Italy, Oct. 4-5 2012.