Mediaeval benchmark: Social Event Detection using LDA and external resources

Mohamed Morchid
Laboratoire d'Informatique d'Avignon
LIA, Avignon, France
mohamed.morchid@etd.univ-avignon.fr

Georges Linarès
Laboratoire d'Informatique d'Avignon
LIA, Avignon, France
georges.linares@univ-avignon.fr

ABSTRACT

This article presents two methods for the automatic detection of social events that were evaluated on the annotated set of pictures as part of the 2011 Mediaeval benchmark [1]. The first method uses a set of web pages and a semantic space obtained by Latent Dirichelet Allocation (LDA, [2, 3]) to classify pictures from Flickr. The second approach uses the query to extract a subset of pictures and classify this subset. Theses approches are compared in the experimental framework of Mediaeval 2011 Social Event Detection task .

Categories and Subject Descriptors

H.3 [**Information Search and Retrieval**]: Indexing

General Terms

LDA, picture categorization

Keywords

LDA, Benchmark, Mediaeval, Classification, Event Detection

1. INTRODUCTION

The search and the browsing of picture collections from sharing platforms requires automatic processing of both content and metadata that are provided by users or owners. Social event detection consists in finding, in a large collection of photos, the ones that are related to a specific event. Our system performs two steps that consist first in extracting all the pictures related to the event category (i.e. *soccer, Barcelona and Roma*), and then to select pictures related to specific events. In the following, the category extraction step is named *subset extraction*, the by-event clustering being named *subset clustering*.

We tested a method based on a semantic representation of pictures by LDA, that is compared to a simple clustering method. These 2 methods are described respectivelly in the Section 2 and 3 of this paper.

2. FIRST RUN: LDA-BASED CONTENT REPRESENTATION

This method relies on an intermediate representation of pictures in a semantic space obtained by LDA. In order to estimate the LDA model, we collect text materials from the Web by using queries related to the event category, represented by a set of keywords.

MediaEval 2011 Workshop, September 1-2, 2011, Pisa, Italy

2.1 Subset extraction

The proposed extraction method relies first on creating a corpus from the web. This corpus is obtained by querying google with the keywords of the challenge query. When we collect this corpus, we create a query-dependent feature vector to evaluate picture/event category similarites. Then, we select the nearest pictures according to a fixed threshold.

→ ***Query representation***

The event category is represented by a feature vector obtained by analysing the related web pages. Web pages are collected by sending query [1] to google and to select the 100 best documents. The query depedent feature vector v is composed by the relative frequency ($p(w|v)$) of each words w of the corpus, divided by the average position of the first occurence of the word $position_w$ in the returned documents. A stop-list based filtering process takes off the meaningless words. The last feature of the vector is the number of seconds since the first january 1970 until the picture taken date.

→ ***Distance between pictures and query***

We want to select the pictures that are related to the query. This is achieved by calculating the distance between the picture and the feature vector by:

$$Dist(pict_k, v) = \sum_{w \in pict_k} v(w) \ , \ v(w) = \frac{TF_w}{position_w} \qquad (1)$$

We create a subset with pictures those distance to the features vector exceeds a fixed threshold.

2.2 Subset clustering

We have to cluster this subset into parts that are supposed to be related to *social events* belonging to the same category. This clustering step is achieved into the LDA space.

→ ***Semantic Space with LDA***

For each challenge, we just have the words in the query to find the events on the pictures set. Here, we locate the query in a topic space estimated by LDA. The 50-topic LDA model is estimated on the dataset obtained for the Subset Extraction step.

→ ***Vector of distance***

We calculate, for each picture, a vector of distance with all topics. We add to this vector another feature: the number of seconds between the date of the picture (dateTaken field) and the 01 jan. 1970.

[1] C1:soccer barcelona rome
C2:may 2009 parc del forum barcelona paradiso amsterdam

$$Dist(pict_k, t_j) = \sum_{w \in ict_k} p(w|t_j).p(t_j|C) \quad (2)$$

We use the prior probability $p(t_j|C)$ of a topic t_j in the corpus C to weight the distance of a picture with a (un)relevant topic.

→ ***Clustering***

Selected picture set is clustered by using the Expectation-Maximisation, gaussian-based clustering algorithm [4].

3. SECOND RUN: PROPOSED APPROACH

In this run, we use only the information of pictures and the query for the two challenges. These information are constituted by all textual metadata available. The global processing scheme is similar to the one we used for LDA based approach: a first step select a subset of relevant pictures that are clustered in a second step.

Here, similarities are only based on text-level comparison, the query being represented by it's keywords and the pictures being represented by title, description and taggs (if available).

Section 3.1 presents the method to extract a subset of pictures related to the query and Section 3.2 describes the clustering method.

3.1 Subset Extraction

This first selection step relies on an estimation of proximity of a picture to the targeted query. We count the number of occurences of each word of the query in the picture text materials. A specific weighting is applied according to the field in which a word occurs. If the word appears in the title, the weight is 1.0, in the description, 0.75 and 0.25 for a tag.

For each element of the picture, we calculate the f-score[5] between words in query and words in pictures features. If the f-score of a section exceeds a threshold (0.75 seems to give the best result), we apply another boosting rule by multiplying the score of the element by 100. We add to the subset the pictures with a score over 40% of the highest score.

3.2 Subset clustering

We determine the similarity between each pair of pictures:

$$Sim(pict_j, pict_k) = \frac{N_{j,k}}{N_j + N_k}$$

Where $j \neq k$ and N_j is the total number of words in $pict_j$. $N_{j,k}$ is the number of words that belongs to a element of $pict_j$ AND a element of $pict_k$. The system puts each picture with the cluster that contains the picture of highest similarity.

4. EXPERIMENTS

For this task, we used 73269 pictures from Flickr [1]. Each picture is associated to a title, a description, owner nickname and tags.

→ ***Results***

We present in the Table 1 the results in each challenge where *E* is the number of events detected for this challenge, *PA* is the number of pictures accepted for the challenge, *PR* the number of rejected pictures for the challenge and % show the percentage of accepted pictures. In Table 2 we present our evaluation (as evaluated by MediaEval Benchmark organisers). Measures are Normalized Mutual Information (NMI) and F-Score.

5. CONCLUSION

We proposed two methods to cluster a set of pictures from Flickr. In the first run, we use LDA and the Web pages to cluster pictures with topics. We also propose a second method that use the query of the challenge to estimate picture/cluster similarities. Evaluation probably presents technical problem that remains to be clearly understood.

Nevertheless, the results show that high level approach such representation in a semantic space doesn't perform well, probably due to it's complexity and the various possiblity of adding noise at different level of the processing chain (in data collecting, topic modeling, document representation in the topic space, etc.).

6. REFERENCES

[1] S. Papadopoulos, R. Troncy, V. Mezaris, B. Huet, and I. Kompatsiaris, "Social Event Detection at MediaEval 2011: Challenges, Dataset and Evaluation," in *MediaEval 2011 Workshop*, Pisa, Italy, September 1-2 2011.

[2] A. McCallum, "Mallet: A machine learning for language toolkit," 2002.

[3] L. Rigouste, O. Cappé, and F. Yvon, "Quelques observations sur le modele lda," *Actes des IXe JADT*, pp. 819–830, 2006.

[4] T. Moon, "The expectation-maximization algorithm," *Signal Processing Magazine, IEEE*, vol. 13, no. 6, pp. 47 –60, nov 1996.

[5] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation," *Advances in Information Retrieval*, pp. 345–359, 2005.

Table 1: Results each challenges and each runs

	Challenge I				Challenge II			
	E	**PA**	**PR**	**%**	**E**	**PA**	**PR**	**%**
Run 1	10	1223	72046	1,6	9	1373	71896	1,8
Run 2	3	13	73256	0	20	65	73204	0

Table 2: Evaluation for each challenges and each runs

	Challenge I		Challenge II	
	F-Score	**NMI**	**F-Score**	**NMI**
Run 1	10,13	0.0263	12,44	-0.01
Run 2	Un.	Un.	3.53	0.0253