ADMRG @ MediaEval 2013 Social Event Detection

Taufik Sutanto
School of Electrical Engineering and Computer Science, Queensland University of Technology
Brisbane, Australia
taufikedy.sutanto@connect.qut.edu.au

Richi Nayak
School of Electrical Engineering and Computer Science, Queensland University of Technology
Brisbane, Australia
r.nayak@qut.edu.au

ABSTRACT

This paper elaborates the approach used by the Applied Data Mining Research Group (ADMRG) for the Social Event Detection (SED) Tasks of the 2013 MediaEval Benchmark. We participated in the semi-supervised clustering task as well as the classification of social events task. The constrained clustering algorithm is utilized in the semi-supervised clustering task. Several machine learning classifiers with Latent Dirichlet Allocation as feature selector are utilized in the event classification task. Results of the first task show the effectiveness of the proposed method. Results from task 2 indicate that attention on the imbalance categories distributions is needed.

1. INTRODUCTION

The Social Event Detection (SED) task at the 2013 MediaEval Benchmark for Multimedia Evaluation consists of two challenges: (1) semi-supervised clustering; and (2) classification of social events [4]. The dataset consists of images metadata from Flickr and Instagram. It includes text, time, and spatial information. The SED task is to group social event images according to the given initial labels and classify them into one of the given event categories (music, conference, exhibition, fashion, protest, sport, theatrical, other event, or a non-event). We participated in both of these tasks, but our efforts were more concentrated on the semi-supervised clustering task.

The number of initial clusters for the first task in the training data is about 14,000 clusters. This task poses many challenges: (1) the number of initial clusters is large; (2) the events in the test data may be grouped in these cluster labels or form new clusters as stated in [4]; and (3) clusters vary in size. About 2,000 clusters contain just a single member while some clusters contain more than 900 members. We adopted the constrained clustering algorithm [2] for handling large clusters more efficiently with the concept of document ranking and the use of a customized similarity measure dealing with text, time, and space. Memory allocation was suppressed by using a semi-incremental algorithm and by combining in-database and in-memory processing. The experiment results show the efficacy of our proposed method.

In the second task, we apply feature reduction using Latent Dirichlet Allocation (LDA) and train several traditional and more recent machine learning classifiers including ensemble of the classifiers through a consensus function. Results from this task were severely influenced by the imbalanced category distribution within the training and test datasets.

2. THE PROPOSED APPROACH

2.1 Preprocessing

All of the features in SED data were used in the analysis, except the uniform resource locator of the images. The structure of data in task1 and task 2 are similar, except that task 2 data does not contain *date_upload* and *description* attributes. Consequently, the preprocessing steps are the same. All non-text characters (symbols) were replaced by a single white space. English stop words removal and stemming on the text data were applied. All text data such as title, tag, username, and description were combined into a text field and treated as a short document. All the analysis was done using only metadata information provided by MediaEval without the use of additional external resources.

The document length normalized tf-idf was used as the term weighting scheme. Since constrained clustering adopts the spherical K-Means algorithm, each document and centroids vectors were further normalized to unit vectors. The time information in task 1 was transformed into day interval between *date_taken* and *date_upload*, while in task 2 temporal information were calculated as the logarithmic difference between *date_taken* and the *Unix epoch*. Geographical information (*latitude* and *longitude*) were used by utilizing Harversine-formula. Geodistance between a document and a centroid was calculated by first measuring the mean space of the centroid.

2.2 Task 1

The K-Means method compares each document to all cluster centroids in forming the clusters iteratively. *K* was initially set to be the number of clusters in the training data. We conjecture that documents in the same cluster should be relevant to each other. We improvised constrained K-Means by calculating k-nearest neighbor of centroids using data ranking and by only comparing distance of a document to the chosen neighborhood of centroids. Several state-of-the-art document ranking schemes such as BM25, BM25 with proximity, and the Sphinx search engine [1] specific ranking (SPH04) were used for this purpose.

Cosine similarity was chosen to measure the distance between a document and centroids based on the text information. This distance was then combined with spatial and time distance in the proposed linear similarity measure. A threshold γ was used to decide whether a document is assigned to one of the existing clusters or form a new cluster. An experiment with only text similarity measure was used as a benchmark to decide the effectiveness of our proposed multi domain similarity measure.

Terms of documents within a cluster were combined as if it is a document. A term weight in this cluster is the average weight of the term within the cluster. Document information from this cluster were then indexed and stored efficiently in real-time using the in-memory delta index of Sphinx search engine. When calculating similarity measure in all iterations, documents were retrieved incrementally from the database and final distances were stored back in database. Transition of documents between clusters were recorded, centroids were re-calculated only with regards to these changes. This approach is efficient in memory usage and computations, even when full text features were used. An illustration of our approach is given in Figure 1.

MediaEval 2013 Workshop, October 18-19, 2013, Barcelona, Spain

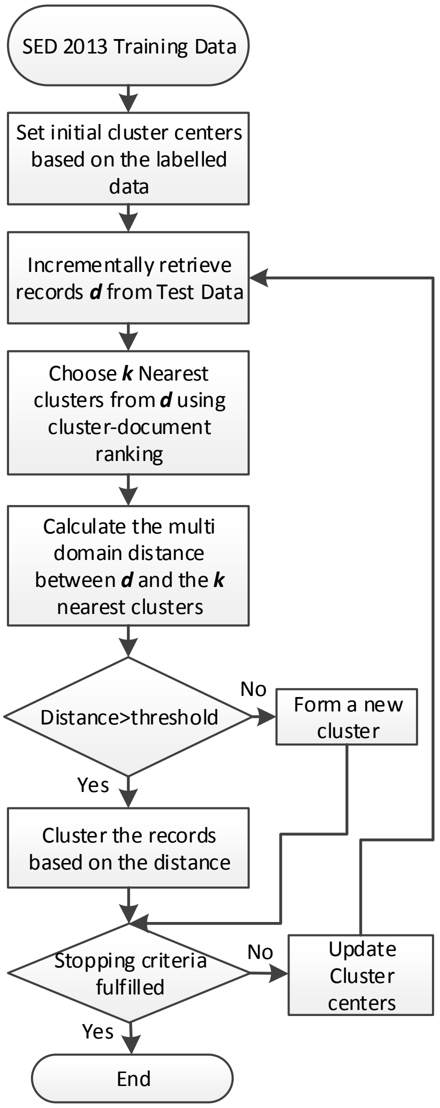
SED 2013 Training Data
Set initial cluster centers based on the labelled data
Incrementally retrieve records ***d*** from Test Data
Choose ***k*** Nearest clusters from ***d*** using cluster-document ranking
Calculate the multi domain distance between ***d*** and the ***k*** nearest clusters
Distance>threshold
No
Form a new cluster
Yes
Cluster the records based on the distance
Stopping criteria fulfilled
No
Update Cluster centers
Yes
End

Figure 1. Proposed clustering approach for task 1

2.3 Task 2

We utilize LDA's Gibbs sampling to automatically form 3,000 topics using the Matlab modelling toolbox [3] from the total of 100,000 text features. Traditional classifiers such as k-Nearest Neighbor (kNN) and decision tree were then used. A more recent classifier (Random Forest) was also used for comparison. An ensemble of the classifiers results were then formed using a consensus function. We used tenfold cross validation on our classifiers by randomly choosing 15% of the training data as validation.

3. EXPERIMENTS AND RESULTS

There are four runs submitted for each task. In task 1, we set threshold to form new cluster γ=0.3 and set the number of nearest clusters k=5. Task 1 run variations were based on different ranking methods and similarity measures. Runs one, two and three in task 1 were using the multi domain similarity measure and using BM25, BM25 with proximity and SPH04 ranking respectively. The last run in this task is used to test the effectiveness of our similarity measure by measuring only text information and using the SPH04 ranking formula.

Results in Table 1 show that the ranking formula positively affects the clustering results and the multi-domain similarity measure effectively improves the clustering quality. We also noted from the result that one of the latest Sphinx ranking formula (SPH04) outperforms the other ranking formula. Furthermore these results confirm the efficacy of our approach in using query ranking to improve scalability of constrained clustering in data with large clusters.

Experiments on task 2 were done by building several classifiers. Random forest, k-Nearest neighbor classifier, and decision tree were used for runs one to three respectively. The last result in task 2 was obtained from the consensus function of the previous classifiers. Since the focus of our experiment was on task 1, the minor attempt on handling the imbalanced category on task 2 has proven to be insufficient.

Table 1. Results for all challenges and runs

SED Challenge	**Results**	**Runs**			
		1	**2**	**3**	**4**
Task 1: Supervised Clustering	F1	0.811	0.802	**0.812**	0.784
	NMI	0.953	0.951	**0.954**	0.943
	Div. F1	0.753	0.745	**0.758**	0.722
Task 2: Classification (non) event/ all-categories	F1	0.475/ 0.105	**0.537/ 0.131**	0.473/ 0.104	0.475/ 0.107
	Div. F1	0.000/ 0.000	**0.035/ 0.021**	-0.01/ 0.001	-0.004/ 0.001
	Overall accuracy	**0.907/ 0.907**	0.825/ 0.817	0.725/ 0.712	0.902/ 0.902

4. CONCLUSIONS AND FUTURE WORK

In this task, we used the constrained clustering algorithm with the customized similarity measure, variable number of clusters, and the use of document ranking. Results show that this method is able to group social events to their corresponding initial labels with higher accuracy. It was also noted that more work is needed to handle the severely imbalanced data of task 2 of classification. Future work will explore the optimal parameter of the similarity measure in the proposed clustering algorithm and investigate further usage of ranking to improve scalability.

5. REFERENCES

[1] A. Aksyonoff, "Sphinx Search," 2.1.1-beta ed: Sphinx Technologies Inc, 2013.

[2] B. Sugato, B. Arindam, and J. M. Raymond, "Semi-supervised clustering by seeding," presented at the Proceedings of the nineteenth international conference on machine learning, San Francisco, CA, USA, 2002.

[3] Griffiths, T., & Steyvers, M., "Finding Scientific Topics," Proceedings of the National Academy of Sciences, 101 (suppl. 1), 5228-5235, 2004.

[4] T. Reuter, S. Papadopoulos, V. Mezaris, P. Cimiano, C. de Vries, and S. Geva, "Social Event Detection at MediaEval 2013: Challenges, Datasets, and Evaluation," Barcelona, Spain, 2013.