Ranking Based Clustering for Social Event Detection

Taufik Sutanto
Queensland University of Technology
taufik.sutanto@qut.edu.au

Richi Nayak
Queensland University of Technology
r.nayak@qut.edu.au

ABSTRACT

The problem of clustering a large document collection is not only challenged by the number of documents and the number of dimensions, but it is also affected by the number and sizes of the clusters. Traditional clustering methods fail to scale when they need to generate a large number of clusters. Furthermore, when the clusters size in the solution is heterogeneous, i.e. some of the clusters are large in size, the similarity measures tend to degrade. A ranking based clustering method is proposed to deal with these issues in the context of the Social Event Detection task. Ranking scores are used to select a small number of most relevant clusters in order to compare and place a document. Additionally, instead of conventional cluster centroids, cluster patches are proposed to represent clusters, that are hubs-like set of documents. Text, temporal, spatial and visual content information collected from the social event images is utilized in calculating similarity. Results show that these strategies allow us to have a balance between performance and accuracy of the clustering solution gained by the clustering method.

1. INTRODUCTION

The Social Event Detection (SED) task at the 2014 MediaEval Benchmark for Multimedia Evaluation consists of two subtasks: (1) Image clustering based on a given set of events; and (2) retrieval of social events based on predefined queries [4]. The SED task poses challenges to clustering analysis due to the real-world nature of the data such as the large number of dimensions, large data size, multi-domain types of features, and the need to group data into a large and unfixed number of clusters. This paper focuses on proposing a solution to the first subtask, i.e., semi-supervised clustering of social event images based on the metadata and visual content.

Search engine technologies e.g., *Sphinx*, *Lucene* or *Solr* have been successfully implemented to process large sized document collections for information retrieval. Utilizing the concept of ranking scores used in search engines, coupled with using prior knowledge from the learning data, in semi-supervised clustering has shown to be an effective and efficient approach of clustering text data [5, 6]. This type of approaches works fine when the collection size or the number of clusters required is small. Calculating ranking scores for a large number of documents is known to be computationally expensive, as well as, a large size cluster makes the similarity measure between documents ambiguous.

Semi-supervised clustering methods have shown to produce a better result compared to their traditional unsupervised counterpart [2]. In the 2013 SED task, we proposed and used a scalable ranking based semi-supervised clustering approach that produces accurate clusters [5]. However, this method suffers with the communication cost for long documents. To deal with the issue, we utilized the document frequency distribution and exclude the most occurring terms in the query document (i.e. the document to be clustered) if needed.

The use of hubs has been explored and has shown its efficacy in dealing with high dimensional data and clusters with large sizes [3]. However, the *k-NN* calculation of hubs demands a considerable amount of extra computation that is not suitable for large data clustering. In this paper, documents are assigned to clusters based on its distances to cluster patches. These patches are calculated based on the ranking scores from the queries. Document frequencies are used to select a subset of terms from documents to create the queries. These *patches* become data representatives to measure distances for a document, instead of using each cluster centroid. The use of *patches* is expected to enable the clustering method to capture more specific sub-topics within a cluster.

In this paper, we present a method based on cluster *patches* to calculate the distance between a document and the groups of documents inside a cluster (Figure 1). Instead of a single centroid, patches are proposed to represent a large high-dimensional cluster in order to control the significance of similarity measurement.

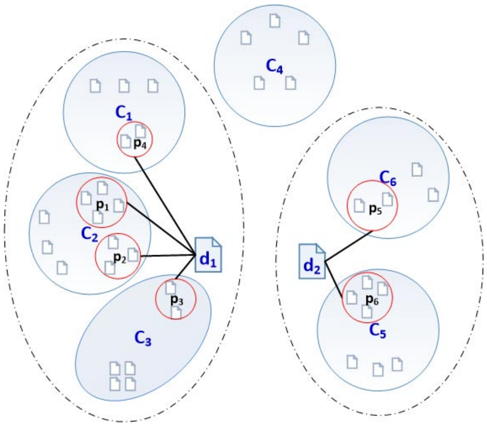

Figure 1: Ranking based document clustering with patches.

MediaEval 2014 Workshop, October 16-17, 2014, Barcelona, Spain

2. PREPROCESSING

All the features of the images were used in the clustering process except of their URL. English stopwords and some symbols (e.g. #,&,@) were filtered. *Title*, *tag*, *username*, and *description* attributes were combined into a short document. No external resources were used in the analysis. The document length normalized tf-idf was used as the term weighting scheme. The time information were transformed into day interval between *date taken* and *date upload*. Spatial information (i.e. latitude and longitude) were used by utilizing a modified Harversine-formula. The modification is done by changing the range of the measure to a unit value as in cosine distance. Feature-based super-pixel segmentation is used to extract compact color and texture representation for small image patches [1]. This representation has smaller dimension compared to the bag-of-visual words (BOVW) approach.

3. THE MODEL

A set of patches P are calculated in each iteration based on the ranking score from the document query. Instead of comparing a document with a cluster centroid, the document feature vector is compared with all the *patches*. The patches are calculated based on a certain size (δ) neighborhood of documents based on ranking scores within clusters. Optimal distance from the document and these *patches* is then used to decide the document assignment to a cluster. More detail of the approach is given in Algorithm 1.

```
input : Set of documents $D$, initial clusters
        $C = \{c_1, c_2, \ldots, c_K\}$, neighborhood size $m$,
        patches size $\delta$, and cluster threshold $\gamma$.
output: $K'$ disjoint partitions of $D$.

Index all documents $D$;
for each $d_i \in D_{test}$ do
    calculate a set of cluster patches
    $P = \{0 < |d^{rank}| < \delta_i, i \in I, d \in c_j\}$;
    for each $p \in P$ do
        calculate $p* = max_p\{sim(d_i, p), p \in P\}$;
        if $sim(d_i, p*) > \gamma$ then
            Assign document $d_i$ to a cluster where $p*$
            belongs;
        else
            Form a new cluster c=$d_i$;
        end
        Update cluster labels via the search engine
    end
end
```

Algorithm 1: Incremental ranking based social event images clustering algorithm.

The similarity measure between a document d and a patch p in a cluster c is given by utilizing textual, temporal, spatial and visual information within images:

$$\begin{aligned} sim(d,p) = \beta_1 sim^{cosine}(d,p) + \beta_2 sim^{time}(d,p) + \\ \beta_3 sim^{space}(d,p) + \beta_4 sim^{image}(d,p). \end{aligned} \tag{1}$$

β_i is a weight parameter to combine the effect of various types of attributes. These parameters can be fine tuned manually or calculated from the learning data by using variable importance measures from a decision tree model.

4. RESULTS AND DISCUSSION

We submitted five runs for the supervised clustering task (Table 1). Runs one, three, four, and five used the proposed method on text only, text-time-space, all attributes, and text-images set of attributes respectively. While run two is using the method as described in [5] using the text attribute only.

Table 1: Semi-Supervised clustering results.

	Run1	*Run2*	*Run3*	*Run4*	*Run5*
F1-Score	0.7463	**0.7533**	0.7445	0.7440	0.7456
NMI	**0.9024**	0.9020	0.9017	0.9015	0.9020
Div. F1	0.7447	**0.7516**	0.7428	0.7424	0.7439

The first two runs indicate that the proposed method has comparable accuracy to general ranking method, but an improved cluster quality as shown by NMI. While the remaining runs shows that the usage of image, spatial, and time information is ineffective in this data for the purpose of clustering. The main reason behind this is the dependence of the proposed method on text ranking.

An adaptive weighting where weights of each attribute are dynamic among documents is a priority for future investigation to solve this issue. Future work will also explore on finding the optimal parameter γ and improve the scalability of the method in distributed data and distributed computing environment.

5. ACKNOWLEDGMENTS

We like to thank Dr Simon Denman from QUT Computational Intelligence and Signal Processing lab for providing us the image visual content encoding.

6. REFERENCES

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34, 2012.

[2] S. Basu, I. Davidson, and K. Wagstaff. *Constrained Clustering: Advances in Algorithms, Theory, and Applications*. Chapman & Hall/CRC, 1 edition, 2008.

[3] J. Hou and R. Nayak. The heterogeneous cluster ensemble method using hubness for clustering text documents. In *WISE 2013*, pages 102–110. Springer Berlin Heidelberg.

[4] G. Petkos, S. Papadopoulos, V. Mezaris, and Y. Kompatsiaris. Social event detection at MediaEval 2014: Challenges, datasets, and evaluation. In *Proceedings of the MediaEval 2014 Multimedia Benchmark Workshop Barcelona, Spain, October 16-17, 2014*, volume 1044. CEUR-WS.org, 2014.

[5] T. Sutanto and R. Nayak. ADMRG @ MediaEval 2013 social event detection. In *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop Barcelona, Spain, October 18-19, 2013*, volume 1043, 2013.

[6] T. Sutanto and R. Nayak. The ranking based constrained document clustering method and its application to social event detection. In *Database Systems for Advanced Applications*, volume 8422 of *Lecture Notes in Computer Science*, pages 47–60. Springer International Publishing, 2014.