UNSUPERVISED SOCIAL MEDIA EVENTS CLUSTERING USING USER-CENTRIC PARALLEL SPLIT-N-MERGE ALGORITHMS

Minh-Son Dao[⋆] *Anh-Duc Duong*[⋆] *Francesco G.B. De Natale*[†]

[⋆] mmLab - University of Information Technology - Vietnam National University of HCM city
Quater 6, Linh-Trung Ward, Thu-Duc District, HCM City, Viet-Nam
{sondm, ducda}@uit.edu.vn
[†] mmLAB - University of Trento
Via Sommarive, 5 I-38123 Povo, Italy
denatale@ing.unitn.it

ABSTRACT

Social Networks have been developed dramatically just in decades. People now have a convenient way to interact with both social media and other people by making the most of using these social networks. Nevertheless, there is still lack of useful tools that can help users (both consumers and providers) managing such social media under events perspective. In order to meet one of these emerging requirements, a user-centric parallel split-n-merge framework applied for unsupervised clustering social media events is introduced. The purpose of this framework is to cluster social media to events they depict by exploiting and exploring the role of users (who) and the way users interact with data (where, what, when) and others (what, who). The output of the proposed framework can be used for event organization/summarization, and as pre-processing stage for event detection and tracking. Major advantages of the proposed framework are (1) low computational solution w.r.t large-scale data, (2) parallel running, and (3) unsupervised clustering with no training data and third-party information requirements. The comparison between the proposed framework and up-to-date methods with MediaEval2013[1] test-bed and evaluation tools shows a very competitive result.

Index Terms— Social media events clustering, user-centric, split-and-merge, user-time image

1. INTRODUCTION

In [1], the authors define Social Media Network is an application that can unite users of the World by enabling the ability to create and exchange media via Internet. Recently, an interesting proposal of Social Life Network (SLN)[2] where all people are always keep up to data information of the real-life situations by connected in a giant social network, is introduced with the emphasis of a multimedia problem. In [3], the high correlation among different types of social streams (e.g., Twitter, Youtube) w.r.t. the same topic related to a real event is seen as a major hint to annotate and predict event in video domain. These researches pointed out the challenges and opportunities of controlling a large-scale social media ever-increased dramatically. One of these challenges is how to manage social events where social events are defined as "events that are organized and attended by people and are illustrated by social media content created by people"[4]. In [5],[6],[7], several problems related to social media analysis such as event detection and classification, tracking, summarization, and association were introduced and discussed. They had the same agreement that finding digital content related to social events is the utmost challenge due to large-scale volume of data coming from different sources and sites, and with a lot of noise in annotations tagged by users from different communities.

MediaEval 2013 recently called for the solution for one of these utmost requirements in social event detection: cluster the entire dataset for all images included in the test set according to events they depict. This problem is vey necessary not only for users who want to organize their data but also for providers who want to analyze data to offer the better tools for their customers under social events perspective. Several groups have accepted this challenge and introduced their up-to-date solutions. In [8], K-means clustering (with value of k parameter is deduced from training data) and document ranking are used as a semi-supervised method to cluster event-related data. They use only text information. In [9], a data-driven three steps approach is applied with text and visual information. This method calculates inter-correlations among clusters to verify the final result. In [10], both text and visual information are used with variety of classifiers (SVM, Decision Trees) to cluster data. In [11], Factorization Machines is used to learn similarity between two time-ordering documents. This method requires a lot of tuning parameters. In [12], propagating geographic locations are applied to com-

[1] http://www.multimediaeval.org/mediaeval2013/

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

pensate the lack of exact location information. Text and visual features are concatenated with weight ratio to feed a linear support vector classifier. In [13], Lucern filter and affinity matrix are constructed with text and visual information. Nevertheless, they recognized at last that visual information makes their result worse.

In general, these methods need to use the whole data set for analyzing. Besides, most of them are supervised methods that require a ground-truth for training. Both of these conditions are very difficult to be met in reality. In order to cope with the curse of ground-truth and volume of data, the unsupervised parallel clustering framework that exploits and explores the most interesting characteristic of social media: user role, is introduced. The contributions of the proposed framework are: (1) low computational solution w.r.t large-scale data, (2) parallel running, and (3) unsupervised clustering with no training data and third-party information requirements.

2. USER-CENTRIC PARALLEL SPLIT-N-MERGE FRAMEWORK

In this section, set of `user-centric parallel split-n-merge` algorithms and the framework to cluster data crawled from social networks into different groups according to events they depict are presented. The whole framework is illustrated in Fig.1. There is assumption that the data should have following properties: `user-id`, `datataken`, `dataupload`, `title`, `description`, `tags`, and `URL` of photo. Except `user-id`, other properties could be NULL (but not all of them are NULL at the same time).

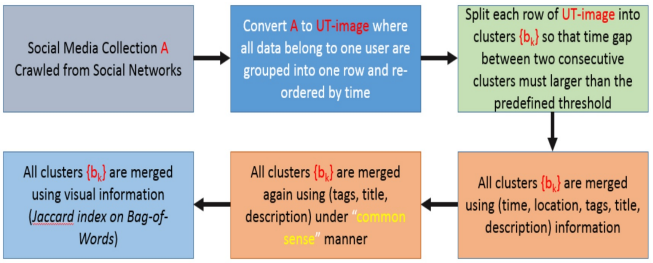

Fig. 1. The proposed framework

2.1. User-Time Images

In order to group data belonging to the same user, the user-time image (`UT-image`) is proposed (see Fig.2). Each row of `UT-image` contains all data belonging to one user, and is ordered by date ascending. Therefore, `UT-image(i, j)` points to data created by i^{th} user at time period j^{th}.

All data whose `time-taken` information is NULL, are grouped together and put at the beginning of each row.

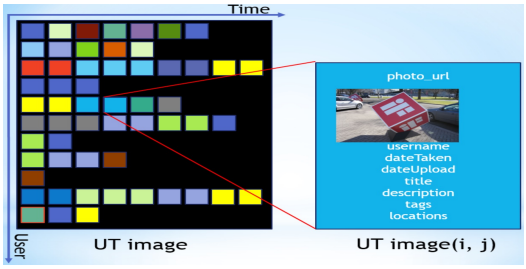

Fig. 2. UT image

2.2. User-time-based Split Algorithm

As mentioned in previous sections, users play an important role in social networks. They generate, upload, and share data related to events they observed or involved in. Therefore, if data crawled from social networks can be grouped by users, events that reported by the same users can be easily detected by clustering data into non-overlap-ordered-time chunks. This can be succeeded due to one user cannot attend two events whose locations are far away each other at the same time. This leads to the fact that the time gap between two consecutive images taken from the same event is usually larger the time gap between two (consecutive) images belonging to two different events, reported by the same user. This observation leads to the first stage of the proposed framework: `user-time-based split` (see Alg.1).

Algorithm 1 user-time-based split algorithm

procedure UTS(*in* A, *in* α, *out* B)
  $B \leftarrow \emptyset$;
  `convert` the original data A into `UT-image`;
  $r \leftarrow$ number of row of `UT-image`;
  for `i=1 to r` **do**
    $c \leftarrow$ number of column of row i^{th} of `UT-image`;
    for `j=1 to c` **do**
      $t_j \leftarrow$ time-taken-of-UT-image(i,j);
      $t_{j+1} \leftarrow$ time-taken-of-UT-image(i,j+1);
      if $| t_j - t_{j+1} | \leqq \alpha$ **then**
        `split` data at column j^{th};
        $B \leftarrow B \cup$ `new-cluster`;
      end if
    end for
  end for
  return B;
end procedure

For each row, any data whose `time-taken` information is NULL, is split as one cluster.

2.3. Time-Location-Tag-based Merge Algorithms

Social networks is a good environment where users in the same community can share and exchange their data. Since

people in the same community (e.g. culture, language, education, hobby, etc.) can give the same "sound and prudent judgment based on a simple perception of the situation or fact"[2], they could tag same words for the same event. Besides, with support of high-tech materials (e.g. camera, smartphone, etc.), most of data, especially images, taken from these materials have time-stamp and location (e.g. GPS) information. These observations are good clues to build the second and third stages of the proposed framework for merging those clusters that could depict the same event: `time-location-tag-based merge` (see Alg.2) and `common-sense-based merge` (see Alg.3).

Algorithm 2 time-location-tag-based merge algorithm

procedure TLTM(*in-out* B, *in* α, *in* β, *in* γ,)
  for each cluster b_k in B **do**
    `create` time-taken-boundary T_k;
    `create` location-union L_k;
    `create` document D_k from tags, title, and description;
  end for
  do
    with any pair of cluster $(b_k, b_l) \subset B$ do
    `merging` if 2/3 following conditions are hold
    {
    $Tdistance(T_k, T_l) \leqq \alpha$;
    $Ldistance(L_k, L_l) \leqq \beta$;
    $JaccardIndex(D_k, D_l) \geqq \gamma$;
    }
    if did merge **then**
      `update` time-taken-boundary T_k;
      `update` location-union L_k;
      `update` document D_k ;
    end if
  while `(can merge)`
  return B;
end procedure

The `time-taken-boundary` T_k of cluster b_k is created by storing the period of time $(T_k.starttime, T_k.endtime)$ so that

$\forall i : T_k.endtime \geqq b_k.\texttt{time-taken}_i \geqq T_k.starttime$.

The `location-union` L_k of cluster b_k is created by storing all non-empty (longitude, latitude).

The `document` D_k is built by applying several NLP techniques (e.g. Stemming, tokenization, etc.)[3] to filter and store only "meaning" words from tags, title, and description properties of b_k.

`Tdistance`$(T_k, T_l) \leq \alpha$ is TRUE if $(T_k \neq \emptyset \wedge T_l \neq \emptyset) \wedge ((0 \leq T_k.starttime - T_l.endtime \leq \alpha) \vee (0 \leq T_l.starttime - T_k.endtime \leq \alpha) \vee (T_l \cap T_k \neq \emptyset))$.

[2] www.merriam-webster.com

[3] http://nlp.stanford.edu/software/

Algorithm 3 common-sense-based merge algorithm

procedure CMM(*in-out* B, *in* γ)
  for each cluster b_k in B **do**
    `process` **tf-idf** on D_k and `select` the most common keywords to `create` ND_k set;
  end for
  do
    with any pair of cluster $(b_k, b_l) \subset B$ do
    `merging` if $JaccardIndex(ND_k, ND_l) \geqq \gamma$;
    `process` **tf-idf** on ND_k and `select` the most common keywords and `update` ND_k set;
  while `(can merge)`
  return B;
end procedure

`Ldistance`$(L_k, L_l) \leq \beta$ is TRUE if $\exists l_k^i \neq \varnothing \wedge l_l^j \neq \varnothing$: `Haversine-distance`[4]$(l_k^i, l_l^j) \leq \beta$.

The Alg.3 is built based on the fact that there would be major "keywords" that are usually tagged by users who already involved in or interested in the same event (e.g. name of acronym of expo or conferences, name of music bands, etc.). This algorithm will increase the chance of merging those clusters that have a lot of "noise" in tags, cannot be merged by using `JaccardIndex` in Alg.2.

2.4. Visual-based Merge Algorithm

In [14], the authors proved that images depicted one event can share some common visual features that characteristic for that event. Therefore, the third stage of the proposed framework is `visual-based merge` (see Alg.4): merging if two sets of images belonged to two clusters share the subset of common visual features.

2.5. Parallel Split-n-Merge Scheme

In fact, each algorithm of the proposed framework can be divided into several packages that can run independently. For example, Alg.1, each row of `UT-image` can be treat as one thread. Thus, the processing time for splitting will be inversely proportional to the number of threads. For merging, we could divide set B into N subsets B_k, then Alg.2, 3, or 4 can apply for each set B_k. The results of all threads will be merged and divide again to $N/\#threads$ subsets. This progress will loop until cannot merge anymore.

With the right policy, it is no doubt that the proposed framework can help clustering social media events not only in off-line but also on-line modes. This can help to cope with the emerging problem nowadays: managing social media streams where information are required to be analyzed in real-time.

[4] $en.wikipedia.org/wiki/Haversine\_formula$

Algorithm 4 visual-based merge algorithm

procedure VFM(*in-out* B, *in* θ)
  for each cluster b_k in B **do**
    $BoW_k \leftarrow \emptyset$;
    for each image img_k^i in b_k **do**
      `calculate` dense-RGB-SIFT;
      `generate` bag-of-words BoW_i^k; ▷ 4096 words
      $BoW_k \leftarrow BoW_k \cup BoW_k^i$;
    end for
  end for
  do
    with any pair of cluster $(b_k, b_l) \subset B$ do
    `merging` if $JaccardIndex(BoW_k, BoW_l) \geqq \theta$;
  while `(can merge)`
  return B;
end procedure

3. EXPERIMENTAL RESULTS

The proposed framework is tested and evaluated by using testbed and evaluation tools offered by MediaEval 2013, Social Events Detection task [4]. The proposed framework is applied to cope with the **challenge 1**: *"Cluster the entire dataset of all images included in the test set according to events they depict"*. The major difficulty here is no information of the number of clusters. Another challenge is not all of properties' information are given fully. For example, geographical information (45.9%), tags (95.6%), title (97.9%), and description (37.9%) w.r.t $437,370$ pictures assigned to $21,169$ events.

The proposed framework is compared to methods introduced by nine groups: ADMRG[8], CERTH-1[9], CERTH-2[10], ISMLL[11], QMUL[12], SOTON[13], TUWIEN[15], UPC[16], and VIT[17]. These groups use same testbeds and evaluating tools offered by MediaEval 2013. The comparison result is denoted in Table 1. In general, the proposed framework gained a promising result comparing to others.

Table 2 shows all runs of the proposed framework. The first run does not use any visual as well as third-parties information as compulsory required by MediaEval 2013 -SED task. At the first run, the proposed method did gain a better result ($F1 = 0.9320$) compared to CERTH-1 ($F1 = 0.5698$), CERTH-2 ($F1 = 0.7031$), ADMRG ($F1 = 0.8110$), and QMUL ($F1 = 0.5900$) though most of them using supervised methods with more parameters need to be tuned manually. The first run uses only Alg.1 and 2 with $\alpha = 24$ hours, $\beta = 5km$, $\gamma = 0.2$. The second run uses as the first run but $\alpha = 8$ hours and $\beta = 2km$. The third run uses Alg.1, 2, 3, with same parameters as the second one. The last run is as the third one with additional visual information $\theta = 0.3$ (i.e. Alg.4). The most interesting point is that results (e.g. F1, NMI, Div F1) of the proposed method increase parallel their accuracy after each step meanwhile others do not. For example, in Table 1 CERTH-1 and CERTH-2 cannot get the best F1, NMI, and Div F1 at the same time when changing their parameters.

	F1	NMI	Divergence F1
Proposed Method	0.9320	0.9849	0.8793
(with visual info)	0.9508	0.9931	0.9020
ADMRG [8]	0.8120	0.9540	0.7580
CERTH-1 [10]	**0.7041**	0.9103	0.6333
	0.7031	**0.9131**	**0.6367**
CERTH-2 [9]	**0.5701**	0.8739	0.5025
	0.5698	**0.8743**	**0.5049**
ISMLL [11]	0.8784	0.9655	NA
QMUL [12]	0.7800	0.9400	NA
SOTON [13]	0.9461	0.9852	0.8864
TUWIEN [15]	0.6900	0.8500	NA
UPC [16]	0.8833	0.9731	0.8316
VIT [17]	0.1426	0.1802	0.0025

Table 1. Comparison Results

Run	F1	NMI	Divergence F1
1 - compulsory run w/t visual info	0.9234	0.9829	0.8705
2 - w/t visual info	0.9316	0.9848	0.8788
3 - w/t visual info	0.9320	0.9849	0.8793
4 - with visual info	**0.9508**	**0.9931**	**0.9020**

Table 2. Each run with different parameters

4. CONCLUSION

The `user-centric parallel split-n-merge` framework is introduced for unsupervised clustering social media events. Series of simple algorithms are built based on characteristics of user's role (e.g. common sense, habits of taking, uploading and sharing data) in social networks. Major advantages of the proposed framework are the low computational complexity, easily developing, parallel running, less tuning parameters. The experimental results showed that the proposed framework can beat other methods not only on the accuracy but also the complexity, and the potential ability of on-line processing.

In the future, the parallel stage will be investigated thoroughly and tested on cloud-computing to examine the ability of real-time processing. Moreover, a dictionary of (placename, longitude, lattitude) will be built in order to get better results in location-based merging. Visual information will also be analyzed carefully to discover the optimal scheme to increase the qualification of the proposed framework.

5. REFERENCES

[1] A.M. Kaplan and M. Haenlein, "Users of the world, unite! the challenges and opportunities of social media," *Journal of Business Horizons - Elsevier*, vol. 53(1), pp. 59–68, January-February 2010.

[2] A. Gupta and R. Jain, "Social life networks: A multimedia problem?," in *Int. Conf. on Multimedia*. ACM, 2013.

[3] S.D. Roy, Tao Mei, Wenjun Zeng, and Shipeng Li, "Towards cross-domain learning for social video popularity prediction," *IEEE Transactions on Multimedia*, vol. 15(6), pp. 1255–1267, October 2013.

[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," in *MediaEval 2013*. MediaEval, 2013.

[5] A. Aggarwal and O. Rambow, "Automatic detection and classification of social events," in *Empirical Methods in NLP*. Association for Computational Linguistics, 2010, pp. 1024–1034.

[6] A. Nurwidyantoro and E. Winarko, "Event detection in social media: A survey," in *ICT for Smart Society (ICISS)*. IEEE, 2013, pp. 1–5.

[7] W. Dou, X. Wang, W. Ribarsky, and M. Zhou, "Event detection in social media data," in *VisWwk Workshop on Interative Visual Text Analytics*. IEEE, 2012.

[8] T. Sutanto and R. Nayak, "Admrg@mediaeval 2013 social event detection," in *MediaEval 2013*. MediaEval, 2013.

[9] D. Rafailidis, T. Semertzidis, M. Lazaridis, M.G. Strintzis, and P. Daras, "A data-driven approach for social event detection," in *MediaEval 2013*. MediaEval, 2013.

[10] E. Schinas, E. Mantziou, S. Papadopoulos, G. Petkos, and Y. Kompatsiaris, "Certh@mediaeval 2013 social event detection task," in *MediaEval 2013*. MediaEval, 2013.

[11] M. Wistuba and L. Schmidt-Thieme, "Supervised clustering of social media streams," in *MediaEval 2013*. MediaEval, 2013.

[12] M. Brenner and E. Izquierdo, "Mediaeval 2013: Social event detection, retrieval and classification in collaborative photo collections," in *MediaEval 2013*. MediaEval, 2013.

[13] S. Samangooei, J. Hare, D. Dupplaw, M. Niranjan, N. Gibbins, P. Lewis, J. Davies, N. Jain, and J. Preston, "Social event detection via sparse multi-modal feature selection and incremental density based clustering," in *MediaEval 2013*. MediaEval, 2013.

[14] M.S. Dao, J. Boato, and F DeNatale, "Discovering inherent event taxonomies from social media collections," in *ICMR*. IEEE, 2012.

[15] M. Zeppelzauer, M. Zaharieva, and M. Del Fabro, "Unsupervised clustering of social events," in *MediaEval 2013*. MediaEval, 2013.

[16] D. Manchon-Vizuete and X. Giro-i Nieto, "Upc at mediaeval 2013 social event detection task," in *MediaEval 2013*. MediaEval, 2013.

[17] I. Gupta, K. Gautam, and K. Chandramouli, "Vit@mediaeval 2013 social event detection task: Semantic structuring of complementatry information for clustering events," in *MediaEval 2013*. MediaEval, 2013.