Mobile Netw Appl (2017) 22:625–633
DOI 10.1007/s11036-016-0804-7

Discovering Social Bursts by Using Link Analytics on Large-Scale Social Networks

Jai E. Jung[1]

Published online: 29 December 2016

Abstract Social Network Services (SNSs) have been regarded as an important source for identifying events in our society. Detecting and understanding social events from SNS has been investigated in many different contexts. Most of the studies have focused on detecting bursts based on textual context. In this paper, we propose a novel framework on collecting and analyzing social media data for i) discovering social bursts and ii) ranking these social bursts. Firstly, we detect social bursts from the photos textual annotations as well as visual features (e.g., timestamp and location); and then effectively identify social bursts by considering the spreading effect of social bursts in the spatio-temporal contexts. Secondly, we use these relationships among social bursts (e.g., spatial contexts, temporal contexts and content) for enhancing the precision of the algorithm. Finally, we rank social bursts by analyzing relationships between them (e.g., locations, timestamps, tags) at different period of time. The experiments have been conducted with two different approaches: i) offline approach with the collected dataset, and ii) online approach with the streaming dataset in real time.

Keywords Social media · Big data · Events · Social bursts detection · Spatio-temporal reasoning · Location-based services

This paper is significantly revised from earlier version presented at the 2nd NAFOSTED Conference on Information and Computer Science (NICS 2015) held in HCM, Vietnam in September 16–18, 2015.

✉ Jai E. Jung
j3ung@cau.ac.kr

[1] Department of Computer Engineering, Chung-Ang University, Seoul, Korea

1 Introduction

Recently, discovering and understanding social bursts from SNSs (e.g., Facebook, Instagram and Flickr) are the most important tasks [3, 4, 12]. In this study, we are particularly focusing on social media data, because of the billions of images and videos on SNSs. This data has been shared between a large number of users (e.g., among friends and among communities) [5, 8], and has covered interesting topics [9, 13]. In fact, user comments on the contents in forms of tags, ratings, preferences give these data sources an extremely dynamic nature that reflects these events [7, 15]. There have been a lot of studies using tags as well as other attributes for predicting system [18], recommendation system [2, 20], clustering algorithm, and classification algorithm, etc.

Besides, the non-textual features such as the geotags, the time, and the visual content have been involved in the social burst discovery recently [3, 12]. Focusing on the use of these geotagged resources from SNSs will find out many useful information for social network applications (e.g., for traveling problem [5, 11, 18]). Usually, social bursts can be categorized into three different types.

Regular social burst This kind of social bursts is organized many times in many places in a regular period (e.g., World Cup, Miss World, Grammy, Oscars). We make a list of ordered locations (these places where organized this social burst) following its famous to find out the place that people feel the most interesting with a given social burst.

Regional social burst This kind of social bursts is organized in a particular place. By discovering *regional social burst*, we can answer the question: "Which is the most interesting social burst that was happened at a location?" (e.g., the question: "Which is the biggest social burst that was organized at Seoul last month?").

Real-time social burst This kind of social bursts appears in a short period of time (e.g., one hour, one day, or one week). To solve this issue, we have to find out: "What did the most interesting social burst happen all over the world yesterday?". With *real-time social burst*, we have to face with huge online data on many SNSs (e.g., Twitter, FaceBook, Instagram, Tumbr, etc.) to extract the needed information.

In this paper, we focus on the *regular social burst*. Firstly, we collect a set of Flickr photos, with both user-supplied tags and other metadata, including time and locations (consisting of latitude-longitude coordinates) of social bursts. The purpose is to discover a sequence of time and locations corresponding to this social burst. Associated through photos, each tag usage occurrence can be attached with temporal and locational encodings. Secondly, we simultaneously analyze the temporal and locational distributions of tag usage occurrences to discover location-related tags. Following this step, we can built a feature vector (including a set of tags). Finally, we discover ordered periodic social bursts which are extracted from a sequence of social bursts by ranking using tags.

The workflow is shown in the Fig. 1. Its components are described as follows.

- *Collecting social data*: datasets are collected from SNSs by using Flickr API. Due to the privacy of Flickr, we have to add some tips for collecting more than 4000 photos with a specific tag;
- *Analyzing data*: the tags are first pre-processed by removing stop-words, conducting stemmings and then the term frequency (TF) is calculated;
- *discovering social bursts*: this step aims to determine the occurrence among social bursts by using the tags and their relationships with locations and time;
- *Ranking social bursts*: HITS algorithm is applied [6, 10] for ranking using the list of social bursts that were discovered from the above step;

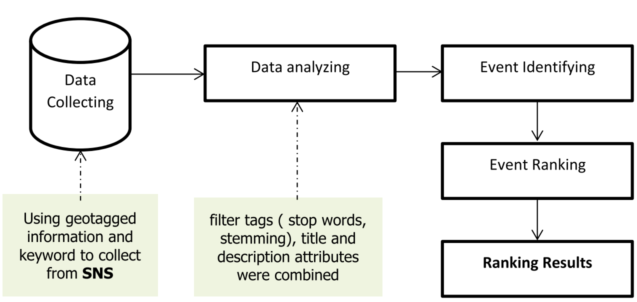

Fig. 1 The proposed workflow of ranking social bursts

- *Results of the ranking*: the final results are the ordered list of social bursts.

The outline of this paper is organized, as follows. In Section 2, we refer to some studies related to detecting social bursts and ranking them. The Section 3 introduces backgrounds that relates to ranking issues. Section 4 shows the algorithms for ranking social bursts. In Section 5, the experiments was conducted to evaluate the results. Section 6 draws some conclusion and states some future works.

2 Related work

In order to rank social bursts on SNSs, the first task is to detect them. Social burst detection can be divided into two categories [21], which are retrospective detection and online detection. Retrospective detection refers to discovering hidden social bursts from aggregated timestamped datasets, while online detection entails the discovery of the onset of new social bursts from real-time feeds.

The study of ranking social bursts from these documents is mainly present in automatic summarization [22]. In [22], they applied a new social burst ranking method to a single document based on the association graph among social bursts.

In other issue, to consider the relationships among social bursts and the impact of social bursts to the society, the authors [12] analyzed the relationships among social bursts with an adapted self-exciting point process model and ranking social bursts with random walk algorithm. They introduced to precisely extract social bursts by considering the spreading effect of social bursts in the spatio-temporal space. To capture the triggering relationships among social bursts, the authors adapted self-exciting point process model by jointly considering spatial contexts, temporal contexts and content similarities of the social bursts. They firstly designed the impact of social bursts and estimated it via the random walk algorithm based on the triggering relationships and then ranked social bursts with different histories.

Particularly, in [19], the authors focused on spatial ranking service, which can retrieve a set of relevant resources with these certain tags to find out a list of locations which are collected from geotagged resources on SNSs. We proposed a novel method (called LocHITS) that can analyze an undirected 2-mode graph composed with a set of tags and a set of locations. Thereby, meaningful relationships between the locations and a set of tags are discovered by integrating several weighting schemes and HITS algorithm.

Our approach focuses on understanding social bursts and then ranking them based on geotagged photos which are collected from SNSs (in particular, Flickr).

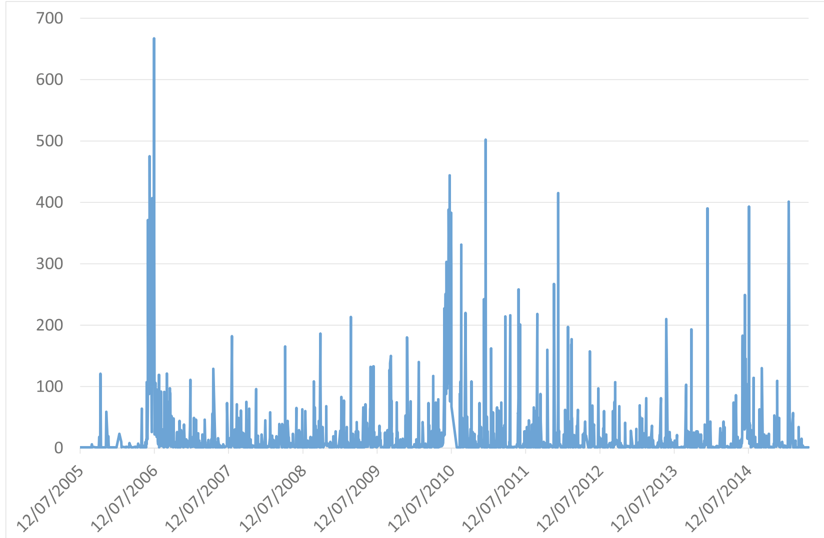

Fig. 2 Statistical distribution of $\mathcal{E}_{WorldCup}$

3 Problem formalization

In this paper, we classify social bursts into the following two types.

- *Social burst (e)* is a specific thing that occurs in a certain place at a certain time [1]
- *Complex social burst ($\mathcal{E}$)* is a sequence of social bursts which are focused on a particular topic

For example, "Miss World 2010" is a social burst and "Miss World" is a sequence of social bursts that refers to the world beauty contest yearly. Therefore, for different *Complex social burst*, the number of social bursts are different (e.g., $\mathcal{E}_{Oscars}$ is annual, while $\mathcal{E}_{WorldCup}$ is organized every four years).

Let $\mathcal{P}$ denote a set of geotagged photos that contains coordinates (latitude and longitude), timestamps, tags and other information. These photos are collected from Flickr with some criteria (e.g., keyword, upload date, privacy level).

We use $\mathcal{E}$ to denote a *Complex social burst* which is extracted from $\mathcal{P}$ as follows

$$\mathcal{E}(\mathcal{P}) = \langle \mathcal{S}, \mathcal{L}, \Theta \rangle = \{e_i\} \tag{1}$$

where $\mathcal{S}$ is a set of social burst labels, $\mathcal{L}$ is a set of locations where social bursts occurred, Θ is a set of timestamps

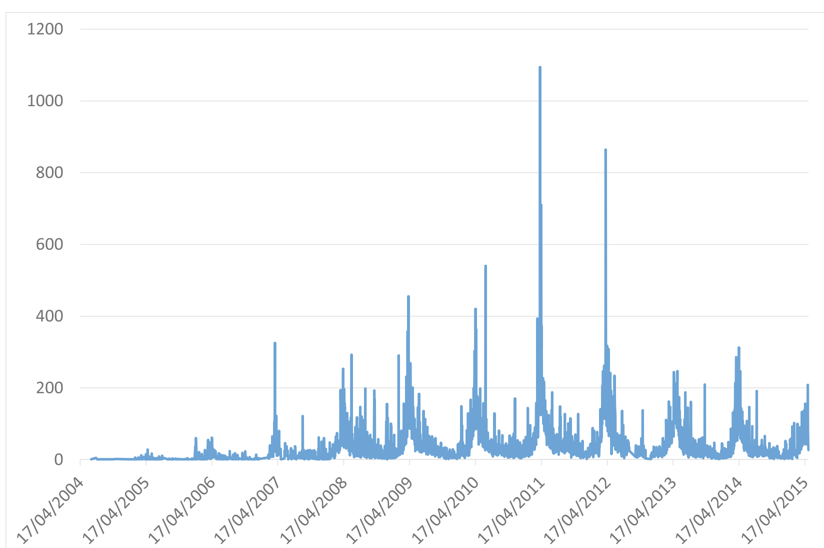

Fig. 3 Statistical distribution of $\mathcal{E}_{Blossom}$

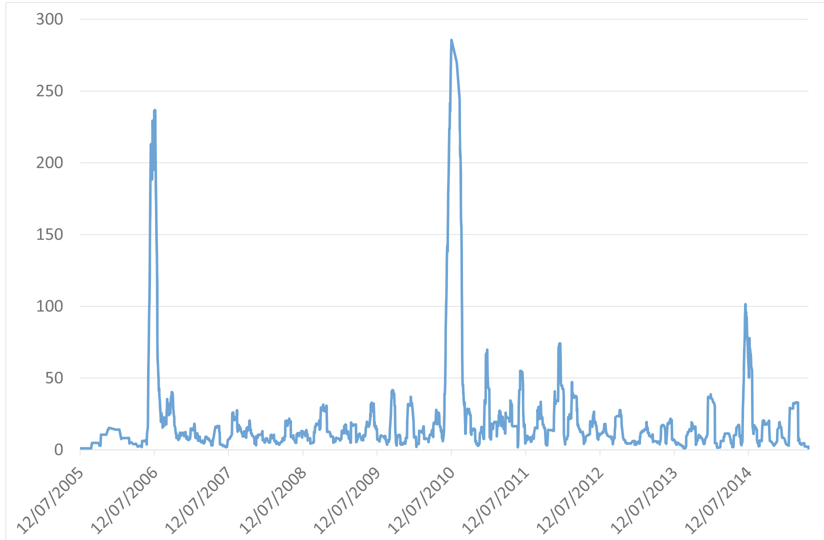

Fig. 4 Statistical distribution of $\mathcal{E}_{WorldCup}$ with smoothing process

when social bursts occurred. They are extracted from a set of photos $\mathcal{P}$.

A social burst e is a triple

$$e_i = \langle s_i, l_i, \theta_i \mid s_i \in \mathcal{S}; l_i \in \mathcal{L}; \theta_i \in \Theta \rangle \tag{2}$$

In our study, we focus on extracting information from certain *Complex social bursts* to discover social bursts and rank them. The ranking results can be used for not only evaluating social bursts in the present but also determining where is the best place for organizing these social bursts in the future.

4 Ranking social bursts

4.1 Discovering social bursts

Giving a set of geotagged photos ($\mathcal{P}$), that is referred to the Section 3, we first have to determine social bursts that belong to a *Complex social burst*. However, this paper only focus on detecting social bursts with a simple approach that has main steps as follows.

First, we calculate the distribution of resource based on timestamps. For example, the results of $\mathcal{E}_{WorldCup}$ and $\mathcal{E}_{Blossom}$ are shown in Figs. 2 and 3. Then, to make our

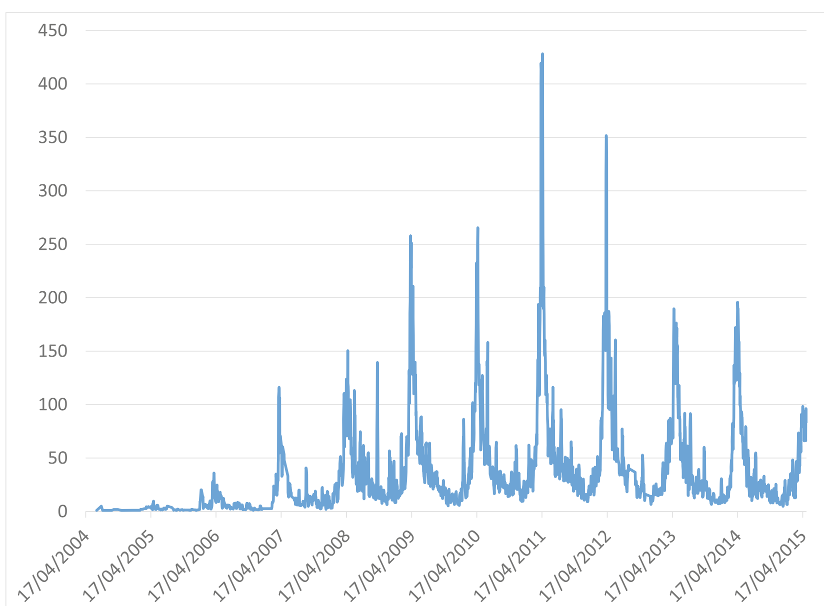

Fig. 5 Statistical distribution of $\mathcal{E}_{Blossom}$ with smoothing process

```
procedure RANKING(ℰ)
Comment: The set of social bursts
    Initialization(Threshold)
    Determine matrix A(m × n)
Comment: m is the number of social bursts; n is the number of tags
    𝒯 ← {t_1, ..., t_n} denote the vector {1,...,1}
    ℰ ← {e_1, ..., e_m} denote the vector {1,...,1}
    Iterations ← 0
    Cond ← max_{i=2,..m-1}(||e_i − e_{i−1}||)
    while Iterations = 0  or  Cond ≥ Threshold do
        j ← 1
        while j ≤ n do
            t_j ← Σ_{i=1}^{m} (a_ij.e_i)
            j ← j + 1
        end while
        i ← 1
        while i ≤ m do
            e_i = Σ_{j=1}^{n} (a_ij.t_j)
            i ← i + 1
        end while
        Normalize(𝒯)
        Normalize(ℰ)
        Iteration ← Iteration + 1
    end while
    return ℰ                ▷ The ranked list of social bursts
end procedure
```

Fig. 6 Pseudocode of Ranking algorithm

time-series smoothing, we use *Moving Average Algorithm (MVA)* [14] for capturing the important data and removing all the unnecessary signal by calculating the mean based on the numbers of previous data.

Smoothing process is to remove insignificant and redundant peaks in a series of measurements. Suppose that we have a time series function $Y = \{y_\theta : \theta \in \Theta\}$ where Θ is a set of timestamps when social bursts occurred (also called the index set). Give the size of the sliding windows n, the value at a point will be smoothed by using the following function.

$$y_\theta = \frac{\sum_{i=1}^{n} y_{\theta-i}}{n} \tag{3}$$

The advantage of *MVA* is that it can work well with irregular fluctuation and easy to apply for different types of data. The data smoothing process makes it easy to discover the social bursts as shown in Figs. 4 and 5. We combine these results with the collected information from the web query to find out these social bursts.

4.2 Ranking social bursts

The ranking approach allows to input a natural language query without boolean syntax to find out the list of ranked records that "answer" the query more oriented toward end-users.

For discovering *Complex social bursts* ($\mathcal{E}$), we rank a list of social bursts $e_i \in \mathcal{E}$ by using the relationships between them based on collected data from Flickr. The algorithm that is used in this issue is formalized according to [10]. The HITS algorithm is used for webpages ranking with the hyperlinks from these webpages form a directed web graph $G = \langle V, E\rangle$, where V is the set of nodes representing webpages, and E is the set of hyperlinks. The hyperlink topology of the web graph is contained in the asymmetric adjacency matrix $L = \{l_{ij}\}$, where $l_{ij} = 1$ if $page_i \rightarrow page_j$ and $l_{ij} = 0$ otherwise. And each webpage p_i has both a hub score p_i^{Hub} and an authority score p_i^{Aut}.

In this study, we use the relationship between tags and social bursts same as hubs and authorities in [6]. However, we use an undirected graph $G = \langle V, E\rangle$, where V is the set of nodes representing tags or social bursts, and E is the set of edges.

Table 1 Dataset

Social burst labels	#Resource photos	#Collected photos	#Data collecting time
Blossom	139,066	134227	1/1/2004-1/5/2015
Grammy	2,117	2,117	1/1/2004-1/5/2015
Missworld	1,046	1,046	1/1/2004-1/5/2015
Oscars	4,901	4,200	1/1/2004-1/5/2015
Bundesliga	27,168	26934	1/1/2004-1/5/2015
Premierleague	10,270	7,263	1/1/2004-1/5/2015
Seriea	5,157	5,143	1/1/2004-1/5/2015
Seagames	348	348	1/1/2004-1/5/2015
FAcup	7,853	7,827	1/1/2004-1/5/2015
Eurocup	2,918	2,918	1/1/2004-1/5/2015
Worldcup	49,688	46,806	1/1/2004-1/5/2015

Table 2 The ranking results of $\mathcal{E}_{WorldCup}$ (using HITS offline algorithm)

Iterations	e_1 (Germany2006)	e_2 (Africa2010)	e_3 (Brazil2014)
1	0.312490018	0.416307299	0.271202683
2	0.315654510	0.406034405	0.278311085
3	0.315424661	0.405311140	0.279264198
4	0.315308777	0.405253552	0.279437671
5	0.315278055	0.405247529	0.279474415
6	0.315270745	0.405246625	0.279482630
7	0.315269053	0.405246450	0.279484497
8	0.315268665	0.405246412	0.279484923
9	0.315268576	0.405246403	0.279485021
10	0.315268555	0.405246401	0.279485043

We denote $\mathcal{T} = \{t_i\}$ is a set of tags, $\mathcal{E} = \{e_j\}$ is a set of social bursts, $\mathcal{E}^{t_i}$ is a set of social bursts which contain tag t_i, T^{e_j} is a set of tags that belong to social burst e_j.

We use a function f in order to determine whether tag t_i is contained in social burst e_j as follows:

$$f(t_i, e_j) = \begin{cases} 1, & \text{if tag } t_i \text{ occurs in social burst } e_j. \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

$$e_j = \sum_{i=1}^{m} \frac{1}{|\mathcal{E}^{t_i}|} t_i; \qquad t_i = \sum_{j=1}^{n} \frac{1}{|T^{e_j}|} e_j; \tag{5}$$

where $|\mathcal{E}^{t_i}|$ is the number of social bursts which contain tag t_i, $|T^{e_j}|$ is the number of tags of social burst e_j.

The formula to compute the value of nodes as follows:

$$\mathcal{T} = A\mathcal{E} \quad and \quad \mathcal{E} = A^T\mathcal{T} \tag{6}$$

where A is an adjacency matrix $(m \times n)$ and a_{ij} is determined by Eq. 4, $\mathcal{T} = \{t_1, t_2, \ldots, t_m\}^T$, $\mathcal{E} = \{e_1, e_2, \ldots, e_n\}^T$.

We can compute Eq. 6 by recursive computing as follows

$$\mathcal{T} = AA^T\mathcal{T} \quad and \quad \mathcal{E} = A^TA\mathcal{E} \tag{7}$$

For each iteration step, the value of nodes are recomputed and normalized as follows.

$$e_i = \frac{e_i}{\sum_{j=1}^{m} e_j} \qquad and \qquad t_i = \frac{t_i}{\sum_{j=1}^{n} t_j} \tag{8}$$

From studying results of [10], we have proposed the *Ranking* Algorithm (as shown in Fig. 6). In which each node is represented by a tag or a social burst.

5 Experimental results

5.1 Result on detecting social bursts

By collecting a set of geotagged photos from Flickr, we discovered some *Complex social bursts* and their resources, as shown in Table 1.

The number of tags are used to represent a *Complex social burst* may grow in large scale. So, we need to employ pre-processing step to find the most valuable feature tags. Firstly, all stop-words are removed from tags and tags are conducted stemming [16]. Secondly, we use the term frequency of tags to create a set of *popular common*

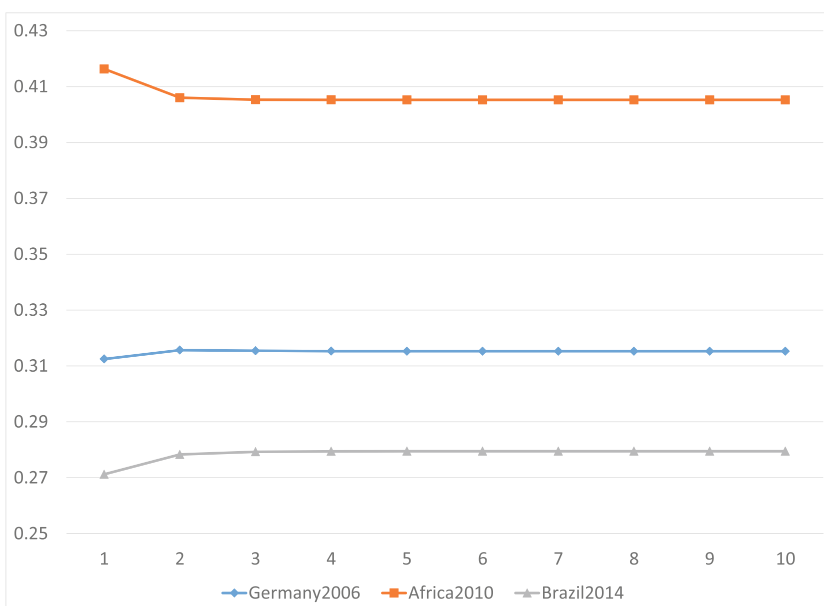

Fig. 7 Ranking result with $\mathcal{E}_{WorldCup}$ using HITS offline algorithm

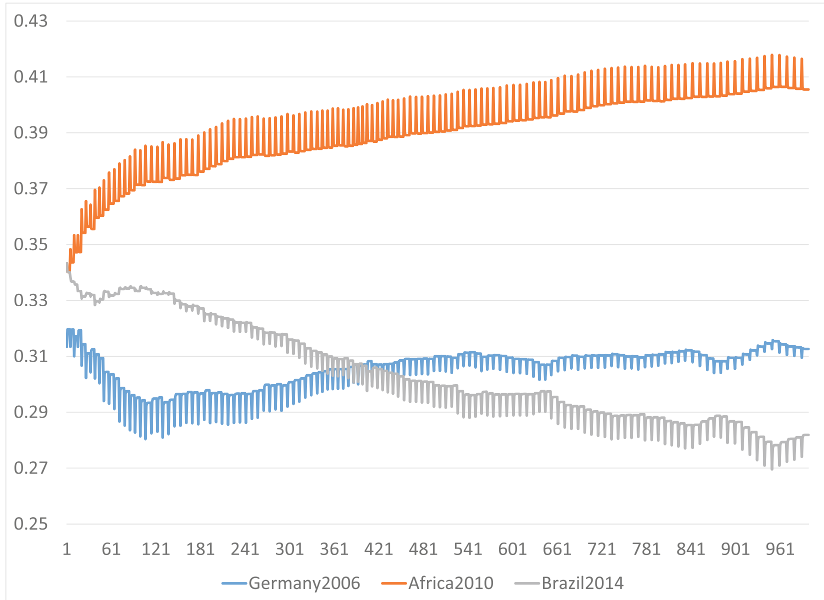

Fig. 8 Ranking result with $\mathcal{E}_{WorldCup}$ using HITS online algorithm

tags. Finally, we filter tags by considering the relationships between these tags and social bursts. We assume that a tag is meaningless in case a tag only appears once in a *Complex social burst* and these kind of tags are removed also.

High results that we get belong to $\mathcal{E}_{WorldCup}$, $\mathcal{E}_{MissWorld}$, and $\mathcal{E}_{Blossom}$. With $\mathcal{E}_{WorldCup}$, we detected the day of social bursts that is the final match, with $\mathcal{E}_{MissWorld}$ is the gala day, and $\mathcal{E}_{Blossom}$ is weeks that blossom grows. $\mathcal{E}_{Seriea}$ is not well detect because this *Complex social burst* is organized for an long time and there are no highlight day during the social bursts. In case of the $\mathcal{E}_{SeaGames}$, the collecting data is small and they are distributed in many days distinguished. Hence, our system could not detect social bursts exactly.

5.2 Result on ranking social bursts

The result of $\mathcal{E}_{WorldCup}$ with the tag search "World-cup" (collected from 2004 to 2015) is shown in Table 2. By using the proposed social burst detection method in Section 4.1, we got three social bursts: World Cup 2006 (Germany2006), World Cup 2010 (Africa2010), and World Cup 2014 (Brazil2014).

After determining social bursts, we try to rank them with both HITS offline and HITS online algorithm.

- With online approach we add 50 pictures for enriching the data at every step. The iteration will be finished when the social burst become stable.

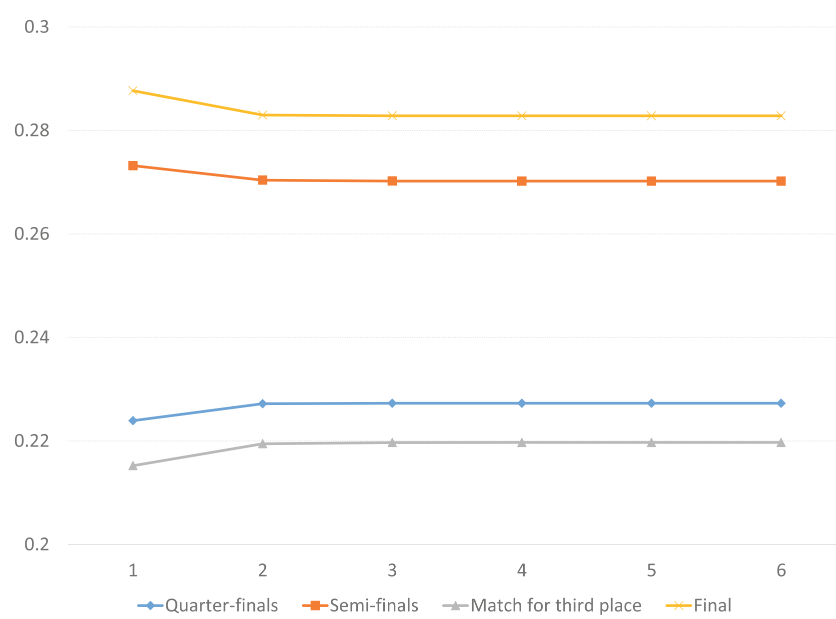

Fig. 9 Ranking result with $\mathcal{E}_{WorldCup2010}$ using HITS offline algorithm

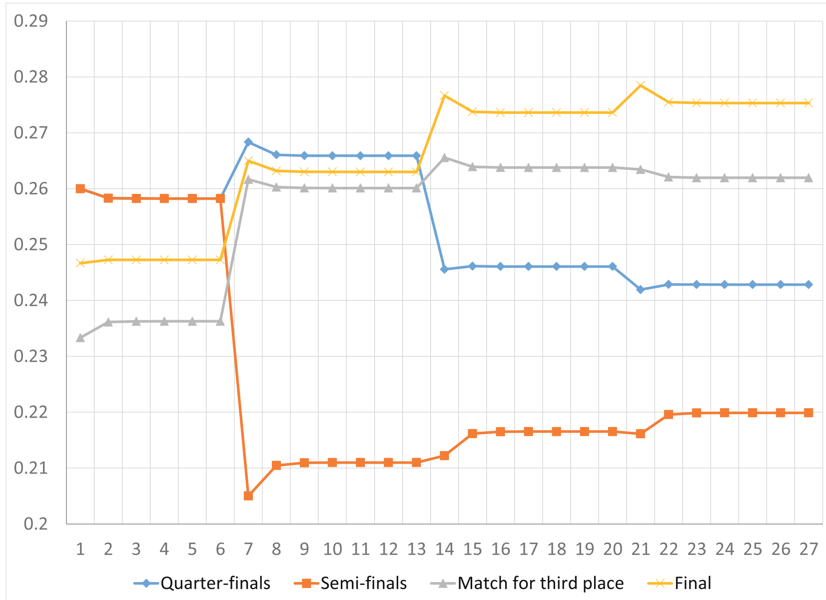

Fig. 10 Ranking result with $\mathcal{E}_{WorldCup2010}$ using HITS online algorithm

– The offline method is applied for *Complex social burst* that has completed dataset at first iteration. The algorithm is stopped when the ranking value has changed less than the threshold (we choose the value of threshold equal to 10^{-8} in our study).

To evaluate the proposed algorithm, we have conducted the experiment on four particular days of World Cup 2010, which are i) Quarter-finals, ii) Semi-finals, iii) Match for third place and iv) Final match. Table 2 shows the results on ranking social bursts. We found that this results are very reasonable and the proposed approach can detect and rank the social burst with high precision.

5.3 Discussion

The *Complex social bursts* that happen in the short time will be easily detected due to the concentration of data. For example, blossom only grows in two, or three weeks of April (as shown in Fig. 5). So the data of $\mathcal{E}_{Blossom}$ will increase quickly in these days. The number of data getting from Flickr is imbalance due to the kind of *Complex social bursts*. For example, people join $\mathcal{E}_{Blossom}$ for taking the photo. Conversely, people will focus on the match and less spend time for taking the pictures with $\mathcal{E}_{WorldCup}$.

Social bursts on Flickr are not real time. The discovered social bursts will have the time gaps (about two days) compared to the real social bursts due to the activities of users. With Instagram, people often take the pictures by smart phone and then they upload immediately to the SNSs. With Flickr, the pictures are almost taken by the camera. Because of this reason, people need some days for processing and then public to Flickr later (we can see that most of the data have the upload date different from the taken date).

Comparing the results of two empirical methods (online, offline) we found that ranking offline method is usually complete earlier than online with the proposed "Ranking" algorithm (as shown in Figs. 7, 8, 9, and 10). This reason are explained by the value of each node (either tag or social burst) in offline method do not increase in each iteration but the number of tags in online method are increased at each steps, it depends on the collected data and also the given threshold.

6 Conclusion and future works

Ranking is one of the best ways to find many things out within valuable. We propose an algorithm for not only ranking social bursts but also understanding them in a novel framework using media data from social network. We detect social bursts from the photos textual annotations as well as include visual features and effectively understand social bursts by considering the spreading effect of social bursts in the spatio-temporal space. We discover the relationships among social bursts (e.g., spatial contexts, temporal contexts and contents) for enhancing the precision of detecting social bursts.

Using the similarity measurement between tags for computing the value of social bursts, our approach calculates the term frequency of tags that occur in each social burst to modify the value of tags for ranking. We have shown the experimental results with a set of social bursts from the geo-tagged resources which are collected from Flickr only for

Regular social burst. Therefore, *Regional social burst* and *Real-time social burst* will be considered in our next study.

Also, we will use the data from different SNSs using matching algorithm [17] for enriching the data. We believe that with the feature properties of each SNSs, we can rank social bursts completely. Other issue, we will build a recommendation system based on the ranking results using user's searching keywords.

Acknowledgments This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2014R1A2A2A05007154).

References

1. Becker H, Naaman M, Gravano L (2010) Learning similarity metrics for event identification in social media. In: Proceedings of the third ACM international conference on Web search and data mining. ACM, pp 291–300
2. Belém FM, Martins EF, Almeida JM, Gonçalves MA (2014) Personalized and object-centered tag recommendation methods for web 2.0 applications. Inf Process Manag 50(4):524–553
3. Bello-Orgaz Gema, Jung JJ, Camacho D (2016) Social big data: Recent achievements and new challenges. Information Fusion 28:45–59
4. Chen L, Roy A (2009) Event detection from flickr data through wavelet-based spatial analysis. In: Proceedings of the 18th ACM conference on Information and knowledge management. ACM, pp 523–532
5. Clements M, Serdyukov P, de Vries AP, Reinders MJT (2010) Using flickr geotags to predict user travel behaviour. In: Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval, pages 851–852, Geneva. ACM, Switzerland
6. Ding C, He X, Husbands P, Zha H, Simon HD (2002) Pagerank, hits and a unified framework for link analysis. In: Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval, pages 353–354. ACM
7. Hong M, Jung JJ (2016) MyMovieHistory: social recommender system by discovering social affinities among users. Cybern Syst 47(1–2):88–110
8. Jung JJ (2012) Discovering community of lingual practice for matching multilingual tags from folksonomies. Comput J 55(3):337–346
9. Jung JJ (2013) Cross-lingual query expansion in multilingual folksonomies: A case study on flickr. Knowl-Based Syst 42:60–67
10. Kleinberg JM (1999) Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM) 46(5):604–632
11. Lee I, Cai G, Lee K (2014) Exploration of geo-tagged photos through data mining approaches. Expert Syst Appl 41(2):397–405
12. Li X, Cai H, Zi H, Yang Y, Zhou X (2014) Social event identification and ranking on flickr. World Wide Web:1–27
13. Morrison JP (2008) Tagging and searching: Search retrieval effectiveness of folksonomies on the world wide web. Inf Process Manag 44(4):1562–1579
14. Nguyen DT, Jung JE (2015) Real-time event detection on social data stream. Mobile Networks and Applications 20(4):475–486
15. Nguyen DT, Jung JE (2017) Real-time event detection for online behavioral analysis of big social data. Futur Gener Comput Syst 66:137–145
16. Nguyen HL et al. (2015) KELabTeam: a statistical approach on figurative language sentiment analysis in twitter. In: Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015), pp 679–683
17. Nguyen HL, Jung JJ (2015) Privacy-aware framework for matching online social identities in multiple social networking services. Cybern Syst 46(1-2):69–83
18. Nguyen TT, Nguyen HL, Huwang D, Jung JJ (2015) Pagerank-based approach on ranking social events: a case study with flickr. In: Proceedings of the 2nd NAFOSTED conference on information and computer science (NICS 2015), pp. 147–152, IEEE, HCM, Vietnam, September 16–18, 2015
19. Nguyen TT, Jung JJ (2014) Exploiting geotagged resources to spatial ranking by extending hits algorithm. Computer Science and Information Systems 12(1):185–201
20. Pham XH, Nguyen TT, Jung JJ, Nguyen NT (2014) <A,V>-spear: a new method for expert based recommendation systems. Cybern Syst 45(2):165–179
21. Yang Y, Pierce T, Carbonell J (1998) A study of retrospective and on-line event detection. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pages 28–36 ACM
22. Zhong Z, Liu Z (2010) Ranking events based on event relation graph for a single document. Inf Technol J 9(1):174–178