Future Generation Computer Systems 100 (2019) 10–18

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Social event decomposition for constructing knowledge graph

Hoang Long Nguyen, Jason J. Jung *

Department of Computer Engineering, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul, Republic of Korea

HIGHLIGHTS

- Proposing a novel method using ICA for decomposing and discovering social events.
- Proposing the SKG model for representing social events and their relationships as knowledge graph.
- It was verified that our method can efficiently provide people with a high understandability and traceability of social events.

ARTICLE INFO

Article history:
Received 29 January 2019
Received in revised form 8 April 2019
Accepted 5 May 2019
Available online 11 May 2019

Keywords:
Social event decomposition
Event-driven knowledge graph
SocioScope framework
Independent component analysis

ABSTRACT

Given the large amount of data collected from social media, it is very difficult for users to identify social events and understand their societies. In this paper, we propose a novel method for *i*) decomposing and discovering social events and *ii*) representing social events and their relationships as a knowledge graph. In particular, the proposed method is based on Independent Component Analysis (ICA) and the SocioScope Knowledge Graph (SKG) model. To demonstrate the actual performance, the proposed method has been evaluated with the support of the SocioScope framework (Nguyen and Jung, 2018). Then, it was verified that the system can efficiently provide people with a high understandability and traceability of social events.

1. Introduction

We are living in the age of Internet of Knowledge (IoK) [1], wherein knowledge is constantly and promptly updated through human activities. Understanding social events, which are extracted from social data [2], is an effective method for keeping our knowledge up-to-date. Hence, social event detection is an important research topic for discovering physical events and occurrences, and it could be applied to various areas including, but not limited to, news tracking [3–5], emergency monitoring [6–8], trust updating [9], and traffic surveillance systems [10,11].

Various studies have been conducted to detect social events using document-pivot [12,13] and feature-pivot [14,15] approaches. However, to the best of our knowledge, none of the existing studies have addressed social event disambiguation for keyword-driven data collection. In order to construct this type of dataset, we first need to use a specific keyword to collect data using Application Programming Interface (API). For example, assuming that we want to know the social events related to *summit* in our society, *#summit* can be used as a keyword for data crawling. Fig. 1 shows an example of two tweets sourced from Twitter. Although both tweets are related to the topic *summit*, they are actually references to different events (i.e., the G7 summit and the Trump-Kim summit). Assuming that we have no prior knowledge of these two events, how can we decompose summit event to obtain the G7 and Trump-Kim summit events? In this work, we solve this problem by applying ICA. ICA is a popular method for solving various types of blind source separation (BSS) problems, such as reducing noise [16], separating signals [17], or representing the linguistic details of words [18]).

Further, social events themselves cannot yield any knowledge unless we use a descriptive model for disambiguating them and representing their relationships. We select the knowledge graph because it is a formal, semantic, and structured representation of knowledge. This helps users and computers efficiently and unambiguously comprehend information. In this study, we modify the simple event model (SEM) [19] to create the SKG model. The SKG model is appropriate for our requirements and definitions, as detailed in a later section, and can be embedded into the SocioScope framework [1] to automatically construct event-driven knowledge graphs representing social events.

In this section, we briefly provided an introduction to the problem under investigation and our motivation. The rest of the paper is organized as follows. In order to provide some background, we describe related work in Section 2. In Section 3, we provide the definitions necessary for solving the problems. Further, Section 4 summarizes the ICA method for decomposing and discovering social events. In addition, we provide an event-driven descriptive model in Section 5 for representing social

* Corresponding author.
E-mail addresses: longnh238@gmail.com (H.L. Nguyen), j3ung@cau.ac.kr (J.J. Jung).

https://doi.org/10.1016/j.future.2019.05.016
0167-739X/

Fig. 1. Social data collected using keyword #summit.

events as knowledge graphs. Then, we carefully evaluate our study in Section 6, and conclude this work in Section 7 with some notes on future work.

2. Related work

ICA has been widely used as an unsupervised learning algorithm for extracting latent features in various applications (e.g., engineering, finance, and neuroscience). Latent feature extraction is the process of analyzing and discovering hidden properties, structures, or relationships between various items or objects. This task is very important for algorithms related to classification, prediction, and recommendation [20]. Further, the features extracted should satisfy the following requirements: (*i*) They should be general, repeatable, and non-redundant. This means that we can apply the method to similar problems and (*ii*) They are unique and distinctive, so that the feature is characteristic of a specific object. In order to apply ICA, the data can be represented as either signals or matrices. ICA is applied to extract features from a digitized sound waveform [21] to obtain a higher-level understanding of audio sources, which is utilized in automatic sound recognition systems [22] for distinguishing multiple speakers [23], extracting speech [24], and classifying singing voices from instrumental sounds [25]. In addition, the authors applied ICA to image processing, leading to notable achievements, including face recognition [26], noise removal from images [27], and iris detection [28]. More specifically, ICA can be used in natural language processing. In [29], the authors proposed a method using ICA to discover the linguistic features of words. From the ICA representation, words are clustered into different categories (e.g., "neural", "computational", and "cognitive" belong to the group of adjectives and "will", "can", and "may" belong to the group of modal verbs). Similar to the work of [29], authors have applied ICA for identifying latent concepts of verb and noun contexts from a Korean monolingual dictionary [30]. These studies demonstrated that is was possible to address the multilingual problem in ICA with linguistic processing. In this work, we apply ICA to extract latent features of hashtags to decompose social events, as will be discussed in detail in Section 4.

In addition, we draw inspiration from previous works on applying knowledge graph representations to social events in this section. In [31], the authors proposed an approach for automatically generating Event-Centric Knowledge Graphs (ECKGs) from news articles. Using techniques of natural language processing and semantic web, the extracted knowledge graphs represented various events well (e.g., the FIFA world cup and global automotive industry). The investigators built the knowledge base from

Fig. 2. Example of a tweet that includes terms, mentions, and hashtags.

named events in news articles [32]. EVIN was then developed as a system for automatically capturing social events. Because of the focus on news articles as formal data sources, the ability to respond rapidly to social events is inferior to that of social networking services (SNSs). Over the years, our society has also observed the emergence of many state-of-the-art knowledge graphs (e.g., Cyc and OpenCyc [33], Freebase [34], DBpedia [35], and YAGO [36]); however, these standalone knowledge graphs are of no use unless we integrate them into smart systems. To overcome these limitations, we propose the SKG model, which is able to represent all types of events and establish their relationships. Moreover, this model can be embedded into the SocioScope framework for automatically constructing knowledge graphs of social events from SNSs.

3. Problem definitions

We focus on two problems in this work: (*i*) applying the ICA algorithm for decomposing data and discovering social events and (*ii*) constructing event-driven knowledge graphs for disambiguating social events and representing their relationships. All the necessary definitions related to these two problems are discussed in detail here. The data collected from social media can consist of text or visual media. However, we choose to use text for discovering social events because of its large related volume.

Definition 1 (*Social Text*)**.** Social text t is a type of social data. It is represented by an orderly set of words w_i by order. To discover social events, a set of social texts must be collected from different individuals in the community.

$$t = \{ w_1, w_2, w_3, \ldots, w_i, \ldots, w_n \} \tag{1}$$

$$\gamma(t) = \langle \theta, \sigma, \pi \rangle \tag{2}$$

where w_i is used to denote for the ith word of the social text t, $\gamma(t)$ is the set of attributes that include the topic θ, location π, and time σ.

In the case of social media data, a word can be a term, mention, or hashtag. Fig. 2 shows an example of a tweet from Twitter, including the terms (e.g., thrilled, to, see, our, and review), mentions (e.g, @NatureMedicine, @AlexandreRbcqt, @rbhar90, and so on), and the hashtags (e.g., #DeepLearning, #medicine, and #healthcare). The formula below is used to express the three types of the word w_i.

$$\tau(w_i) \in \{ \mathcal{T}, \mathcal{M}, \mathcal{H} \} \tag{3}$$

where $\tau(w_i)$ is the type of word w_i, and $\mathcal{T}, \mathcal{M}, \mathcal{H}$ are used to denote term, mention, and hashtag respectively. Among these word types, hashtag is important for determining the topic of a social text [37]. People usually use hashtags for categorizing their posts; therefore, hashtags can help social media users to track a specific topic.

Definition 2 (*Topic of Social Text*)**.** The topic of a social text is defined by the set of hashtags contained in this social text.

$$t \supset \theta(t) = \{\, w_i \mid w_i \in t, \tau(w_i) = \mathcal{H} \,\} \tag{4}$$

where the topic of the social text t is denoted as $\theta(t)$. However, the topic of the social text may differ depending on the context in which hashtags are grouped together. Assuming that G_i is a subset of hashtags in t, the topic of t can be determined by the following relation.

$$\theta(t) = \arg\max_{G_i} \mid t \cap G_i \mid \text{ with } G_i \subset t \tag{5}$$

In Fig. 2, if G_i = {#D*eepLearning*, #medicine, and #healthcare} , the topic of t can be *science*. However, *medical* can be the topic of the social text t if G_i = { #medicine, and # healthcare}. Because of this ambiguity, it is difficult to group hashtags to determine the topic of a social text. We overcome this problem by applying ICA in this study. Social events can thereafter be determined by the set of social texts that have the same topic.

Definition 3 (*Social Event*)**.** A social event e is a real-world activity that occurs at location Π in a duration of time Σ. It is defined by the set of social texts having the same topic Θ, as follows.

$$e = \{\, t_i \mid \theta(t_i) = \Theta(e), \sigma(t_i) \in \Sigma(e) \,\} \tag{6}$$

$$\gamma(e) = \langle\, \Theta, \Pi, \Sigma \,\rangle \tag{7}$$

$$\Sigma = [\, \varsigma_s, \varsigma_e \,] \tag{8}$$

where $\gamma(e)$ are the attributes of the social event e, and ς_s, ς_e are used to denote the starting and ending times of this event. It should be emphasized at this point that the location $\pi(t_i)$ of the social text may not be related to the location $\Pi(e)$ of the social event because of the growth of smart devices. People can publish their social texts from different locations without any hindrances. Further, geotagged information can be omitted to guarantee user privacy [38].

In the case of keyword-driven data collection, general keywords are often used for collecting data (e.g., *football*, *summit*, and *disaster*). However, people consider more specific events, for example, *world_cup*, *euro_cup*, and *premier_league*. Therefore, we need a method for decomposing general social events to obtain more specific events.

Definition 4 (*Social Event Decomposition*)**.** Social event decomposition is the process of separating a social event into different sub-social components. It could be expressed as follows.

$$d(e) = \{\, e_i \mid \forall j, \Theta(e_i) \neq \Theta(e_j) \text{ and } \Theta(e_i) \subset \Theta(e) \,\} \tag{9}$$

where $d(e)$ is used to denote the process of decomposing a social event e. In the next section, we will discuss application of ICA to decomposing and discovering social events.

4. Applying ICA for discovering social events

To be able to apply ICA, our dataset must be non-Gaussian. This is appropriate for social text because of its sparse nature. The process of decomposing and discovering social events includes three steps as follows.

1. Generating an m-by-n social text-hashtag (i.e., m is the number of social texts in the dataset and n is the number of unique hashtags) from the set of social texts.
2. Applying the ICA algorithm on the generated matrix to separate n hashtags into k groups, with $k \ll m$.
3. Based on the hashtags, categorizing the set of social texts into k subsets. Each subset is identified as a social event.

Supposing that X is the social text-hashtag matrix as follows.

$$X = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{pmatrix} \tag{10}$$

where each value $x_{i,j}$ is determined using the formula

$$x_{i,j} = \begin{cases} 1 & \text{if } w_j \in t_i \\ 0 & \text{if } w_j \notin t_i \end{cases} \tag{11}$$

Next, we apply the ICA algorithm to extract k features from the contextual matrix X. It includes two processes: (i) whitening X to reduce the number of features that need to be estimated using eigenvalue decomposition (EVD) and (ii) applying ICA on the whitened matrix to obtain the independent component matrix S. Fig. 3 shows these two steps in detail.

By applying EVD to the covariance matrix $E(XX^T)$, we obtain eigenvector E and diagonal matrix D. The diagonal matrix D is generated by grouping non-zero eigenvalues.

$$E(XX^T) = EDE^T \tag{12}$$

The whitened matrix Y can then be calculated using E, D and X as follows.

$$Y = D^{-1/2}E^T X \tag{13}$$

With $X = A'S$, the ICA problem can be expressed as follows. Our target is to obtain S by estimating A.

$$Y = D^{-1/2}E^T A'S = AS \text{ with } A = D^{-1/2}E^T A' \tag{14}$$

$$S = WY \text{ with } W = A^T \tag{15}$$

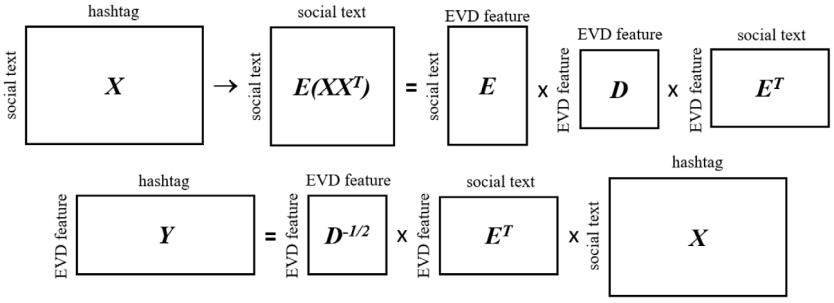
where A, A' are the mixing matrices of the whitened matrix Y and the contextual matrix X, respectively, and W is the inverse matrix of A. After whitening, the new mixing matrix A is also an orthogonal matrix. This can be proved as follows.

$$E(YY^T) = AE(SS^T)A^T = I \Rightarrow AA^T = I \tag{16}$$

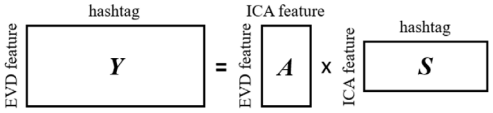
Algorithm 1: FastICA Algorithm

Input: k number of expected features.
Input: whitened matrix Y.
Output: un-mixing matrix $W = [w_1, w_2, ..., w_k]$.
Output: independent component matrix S.
begin
  for *i in 1 to k* **do**
    $w_i \leftarrow$ setting up a random vector of length k.
    $\epsilon \leftarrow$ defined threshold for determining the convergent of w_i
    while $|w_i^+ - w_i| < \epsilon$ **do**
      $w_i^+ = E\{YG'(w_i^T Y)^T\} - E\{G''(w_i^T Y)\}w_i$
      $w_i = w_i^+ / \| w_i^+ \|$
    end
  end
  $S = WY$
end

We choose FastICA [39] as shown in Algorithm 1 for estimating matrix A because of its computational efficiency and lower memory requirement. FastICA solves the problem by maximizing the non-Gaussianity of $w^T Y$. There are various methods for approximation (e.g., based on kurtosis, negentropy, and mutual information); however, we choose to maximize the approximations



(a) Obtaining whitened matrix Y by applying EVD on the contextual matrix X.



(b) Operating ICA on the whitened matrix Y.

Fig. 3. Estimating S to obtain the ICA representations of hashtags.

of negentropy. This approach can achieve fast computation without complex processes [18]. Approximating negentropy relies on the non-quadratic function $G(z)$. There are different methods for choosing $G(z)$, as follows.

$$G(z) = \log cosh(z) \Rightarrow \begin{cases} G'(z) = tanh(z) \\ G''(z) = 1 - tanh^2(z) \end{cases} \quad (17)$$

$$G(z) = -e^{-z^2/2} \Rightarrow \begin{cases} G'(z) = -ze^{-z^2/2} \\ G''(z) = (1 - z^2)e^{-z^2/2} \end{cases} \quad (18)$$

where $G'(z), G''(z)$ are the first and second derivatives of $G(z)$, respectively. The selection of $G(z)$ as provided in Eq. (17) is appropriate for general purposes while the alternative is meant for solving robust problems. After obtaining the matrix S, we determine the topic of a social text based on its hashtags. From that, the dataset can be decomposed into subsets, each of which is determined by a social event.

5. The SKG model for event-driven knowledge graph

The SKG model is a modification of the SEM model aimed at the problems identified in this study. Further, the SKG model is designed for implementation to the SocioScope framework to automatically construct event-driven knowledge graphs from social data. In particular, the SKG model contains two components: `skg:Core` and `skg:Type`. The relationships between the two classes are represented in Fig. 4. There are four types of core classes as follows.

- `skg:Event` for expressing which event happens.
- `skg:Player` for representing who takes part in this event.
- `skg:Location` for specifying where the event occurs.
- `skg:Time` is used for describing when event happens.

Corresponding to each core class, one type of class exists for identifying the types of core entities, namely `skg:EventType`, `skg:PlayerType`, `skg:LocationType` and `sgk:TimeType`. The details of instances of these type classes are as follow.

- `skg:EventType`: to express type of an event, for example, a summit, a football match, an election, or a disaster.
- `skg:PlayerType`: a player has to be a person such as a president, a worker, or an engineer.
- `skg:LocationType`: it can be a small area such as a school, a street, or a house; however, it can be as large as a country, a sea, or a mountain as well.
- `skg:TimeType`: this is very broad, including, but not limited to, a historical period, a year, a month, or a date.

Further, the properties of the SKG model are divided into three different types as follows.

1. Properties for connecting an individual of `skg:Event` to other individuals of the core classes, namely: `skg:hasPlayer`, `skg:hasLocation`, and `skg:hasTime`.
2. Properties for associating instances of core classes to instances of type classes. They are defined as `skg:playerType`, `skg:locationType`, `skg:timeType`, and `skg:eventType`, .
3. Property for linking core and type classes to their individuals, which is `rdf:type`.

An instance of the `skg:Event` is extracted by selecting the hashtag with the highest frequency, and an instance of the `skg:Time` is identified by the highest peak in data frequency. By contrast, the instances of the `skg:Player` and `skg:Location` are the top named entities found in the social text of an event. The labels of the named entities are used to determine the properties of the knowledge graphs. First, we order named entities by frequency. Then, the set of instances for the `skg:Player` and the `skg:Location` are derived using the following formula.

$$n = \left\{ w_i \mid w_i \in \mathcal{T}, w_i \in \mathcal{NE}, \frac{\Delta(w_i)}{\Delta(w_{i+1})} > \varepsilon \right\} \quad (19)$$

where n is denoted by the set of instances of the knowledge graph, $\mathcal{NE}$ is the named entity, $\Delta(w_i)$ denotes the frequency of the ith named entity, and ε is the defined threshold for filtering out trivial entities.

6. Evaluation result

Our evaluations are conducted with the support of features in the SocioScope framework for automatically collecting and analyzing data. This framework provides necessary features (e.g., social media data crawler, natural language processing, linked data, and visualization tools) for our experiment. From that, we also plan to extend the SocioScope framework to automatically extract the knowledge graph from social data. In order to construct the dataset, we use the batch crawler feature of the SocioScope framework to automatically collect social text from Twitter using the hashtag *#summit* as a keyword. The statistics of this dataset are displayed in Table 1.

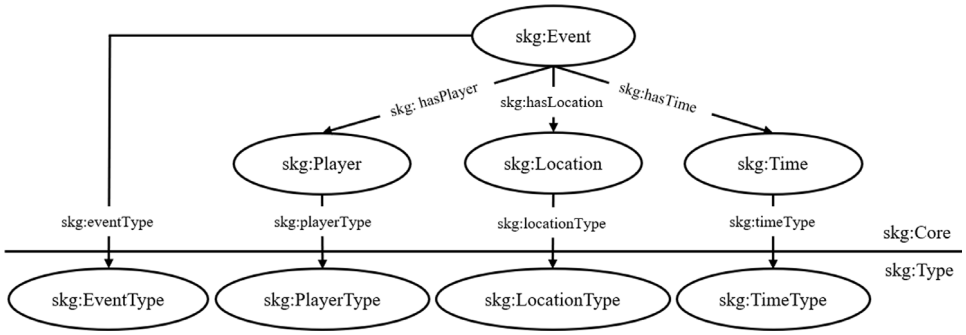


Fig. 4. The SocioScope Knowledge Graph (SKG) model.

Table 1
Statistics of the dataset.

Property	Detail
Number of social text	19,327 tweets
Time duration	9 days
Data size	2.55 GB
Highest frequency by minute	47 tweets
Average frequency by minute	2 tweets

To monitor the social text stream, we use a time series visualization feature in the SocioScope framework, as shown in Fig. 5. We recognize that there are two abnormal time points, namely 2018.08.06 and 2018.08.12. These two days have a significantly higher frequency than the others. Based on the research problem mentioned in Section 1, our task is to decompose the dataset, which is collected using keyword *#summit*, into two subsets with the assumption that we have no prior knowledge. Each subset can be represented by a social event.

We apply the ICA algorithm for separating unique hashtags in the dataset into two groups, as shown in Fig. 6. Here, Fig. 6(a) shows several hashtags in G_1 and Fig. 6(b) shows several hashtags in G_2. For the sake of simplicity, the ICA value is normalized to a [0, 1] range for visualization as a chart. Each group of hashtags can be used to represent a social event. By observing the classified hashtags, we can predict that the social event #1 is related to the G7 summit event ($e_{\text{g7\_summit}}$) because of the appearance of "g7" in the hashtags #g7canada, #g7summit, and #g7charlevoix, and the social event #2 is about the Trump-Kim summit ($e_{\text{trump\_kim\_summit}}$) event due to the appearance of "usnorthkorea" and "trumpkim" in the hashtags #usnorthkoreasummit and #trumpkimsummit. Interestingly, the algorithm can correctly classify hashtags with typographical errors, for example, the hashtag #trumpkimsumit (i.e., missing character "m"). In addition, the other hashtags are related to these two social events as well. In the case of $e_{\text{g7\_summit}}$, these are countries in attendance at the meeting (i.e., #canada and #germany), leaders of these countries (i.e., #trudeau and #merkel), and topics of discussion (i.e., about #russia). Similarly, #northkorea and #kimjongun are the country name and the individual in the $e_{\text{trump\_kim\_summit}}$. This event was held at the #capella hotel, and #nuclearweapons and #china were two of the topics of discussion. Based on this experiment, our proposed method categorizes hashtags effectively. In particular, the event decomposition can be expressed as follows.

$$d(e_{\text{summit}}) = \{\, e_{\text{g7\_summit}}, e_{\text{trump-kim\_summit}} \,\} \tag{20}$$

Based on hashtags, we can easily discover two distinct events by decomposing the dataset into two subsets. Next, we will evaluate the utilization of the knowledge graph using the SKG model. First, the Part-Of-Speech Tagger (POS Tagger) feature of the SocioScope framework is used to determine the value of `skg:Event` for the two summit events. From the two extracted subsets, the two hashtags that have the highest frequencies are *#g7summit* (i.e., appearing 188,121 times) and hashtag *#trumpkimsummit* (i.e., appearing 859,165 times). These are selected as the instances of the `skg:Event`. In addition, *2018.06.08* and *2018.06.12* are the instances of `skg:Time`. To identify the instances of `skg:Player` and `skg:Location`, we use the named entity recognition feature of the SocioScope framework, thereby extracting top named entities, some of which are *kimjongun* and *trump*. In order to identify the instances of the type classes, we retrieve common information from DBpedia,[1] for example, the linked open data of Donald Trump.[2]

For determining the ground truth of this evaluation, we first manually build the knowledge graph of $e_{\text{g7\_summit}}$ and $e_{\text{trump\_kim\_summit}}$ based on our knowledge. This knowledge graph was verified by 100 graduated students in the Department of Computer Science at Chung-Ang University through the following five questions without further explanation.

1. What do you think of the understandability of the knowledge graph provided (e.g., comprehensible naming of entities, types, and relationships)?
2. From the knowledge graph provided, can you comprehend all the information of the two summit events (e.g., event, player, location, and time)?
3. Does the knowledge graph provided contain sufficient information on the two summit events? If the information provided was insufficient, add your own information.
4. Does the knowledge graph provided contain precise information on the two summit events? If any of the information is inaccurate, modify it using your own information.
5. What do you think of the usefulness of the knowledge graph provided?

Based on the majority of responses from the students, we modify our knowledge graph to yield the expected result shown in Fig. 7. We use precision, recall, and F-measure to evaluate the performance of our approach in constructing the event-driven knowledge graph, as follows.

$$Precision = \frac{n_{co}}{n_{co} + n_{in}} \tag{21}$$

$$Recall = \frac{n_{co} + n_{in}}{n_{ex}} \tag{22}$$

$$F\text{-}measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{23}$$

where n_{co} is the number of correct instances, n_{in} denotes incorrect instances, and n_{ex} represents the total number of instances expected in the knowledge graph.

[1] http://dbpedia.org/page/.
[2] http://dbpedia.org/page/Donald_Trump/.

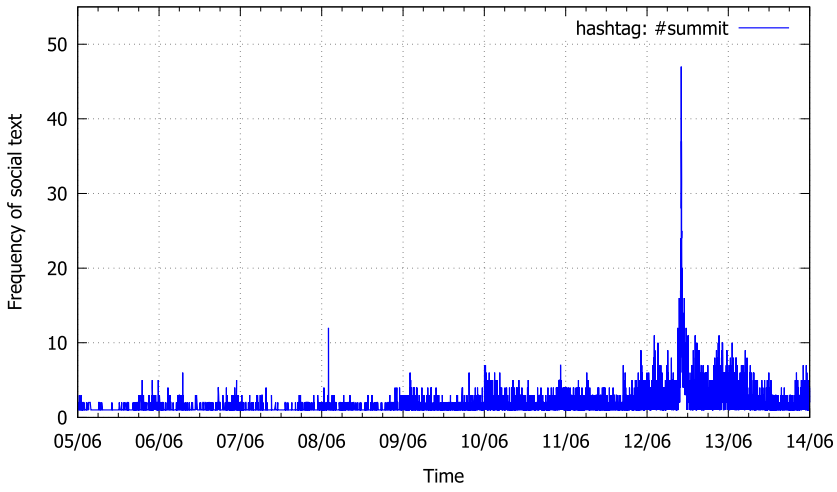
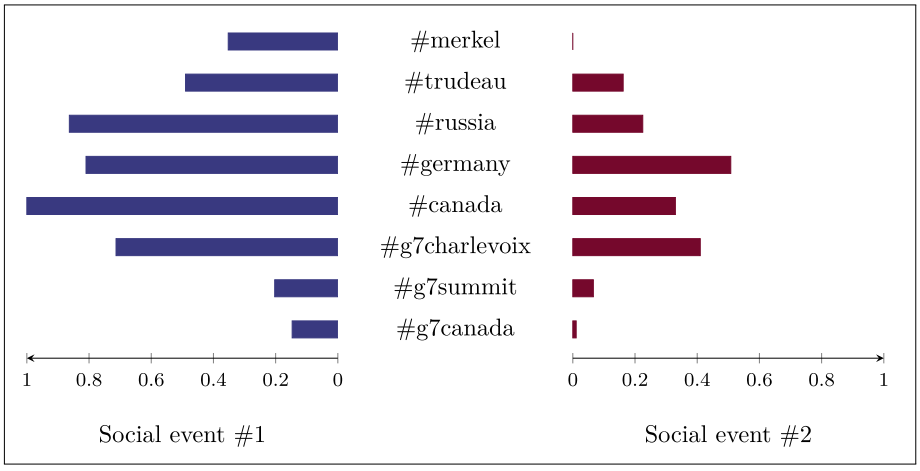
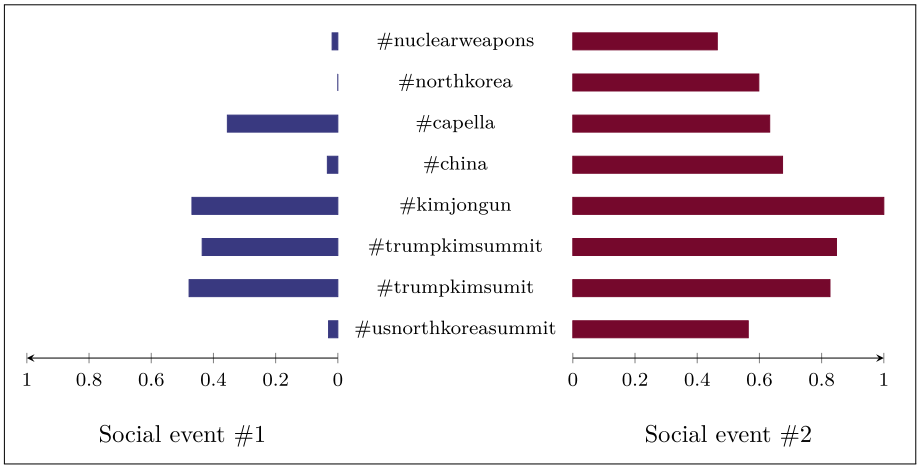


Fig. 5. Frequency of social texts collected using keyword #summit.



(a) ICA representation of hashtags belonging to social event #1.



(b) ICA representation of hashtags belonging to social event #2.

Fig. 6. ICA representation of hashtags.

We conduct the evaluation by changing the value of the defined threshold ε as mentioned in Eq. (19). Fig. 8 depicts the knowledge graph with the threshold value ε of 0.4. We recognize that several redundant nodes exist compared to the expected knowledge graph (i.e., the nodes Russia, Korea, USA, and Obama). Fig. 9 shows that the value is reached when ε is 0.8 (i.e., the *F-measure* is 86.58%). From this evaluation, we can show that the SKG model with the support of the SocioScope framework is effective at constructing an event-driven knowledge graph.

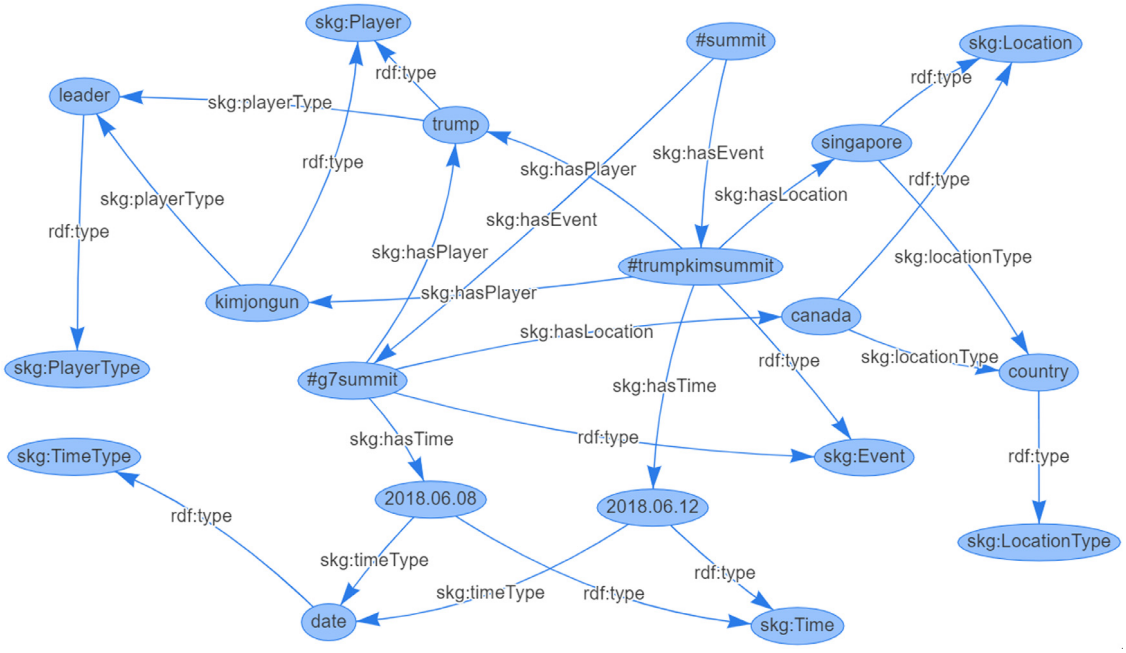


Fig. 7. Expected knowledge graph for the two summit events.

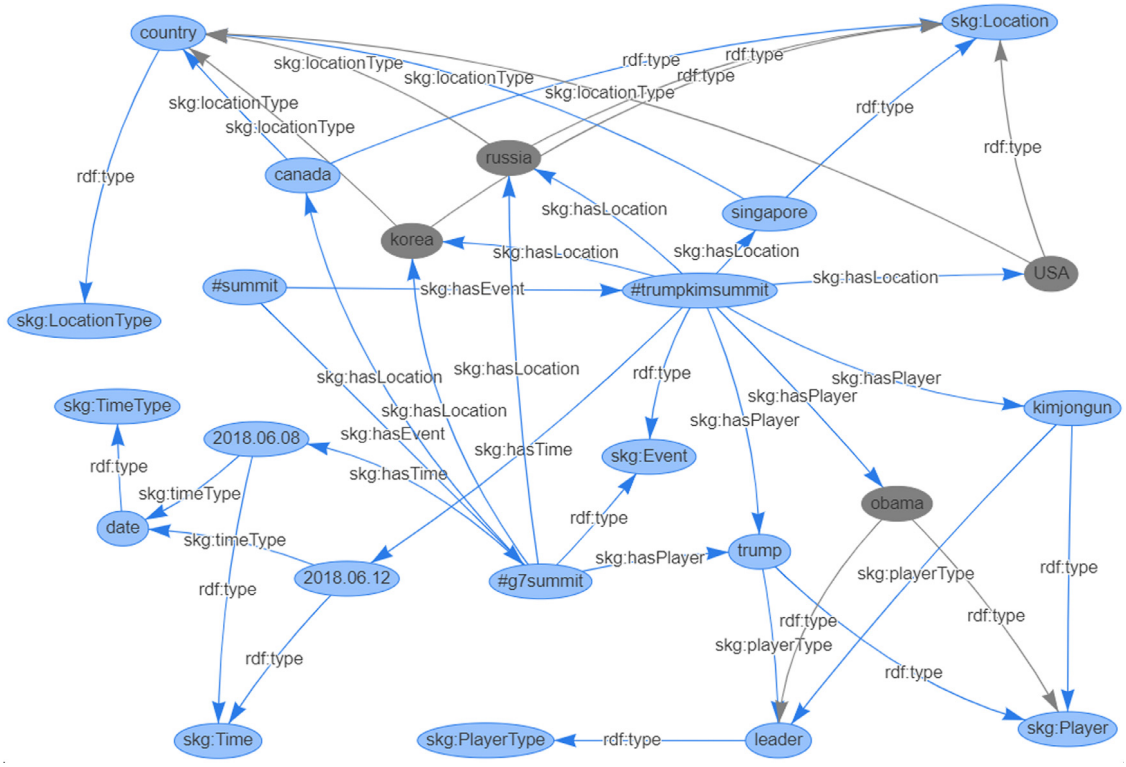


Fig. 8. Knowledge graph for two summit events with $\varepsilon = 0.4$.

7. Conclusion

We propose an approach for decomposing social events and constructing an event-driven knowledge graph by applying the ICA algorithm and SKG model in this study. First, ICA is applied to effortlessly disambiguate social events. The SKG model is then used as the formal structure for representing social events and their relationships. Our method guarantees minimum complexity and simultaneously maximizes efficiency. Based on the event-driven knowledge graph, people can comprehend developments in our society.

However, there are limitations that require future work efforts. We plan to apply ICA to other case studies with longer durations to increase complexity. In addition, other social data sources (e.g., public web data and sensor data) will be considered to enrich the information of the knowledge graph. Finally, we

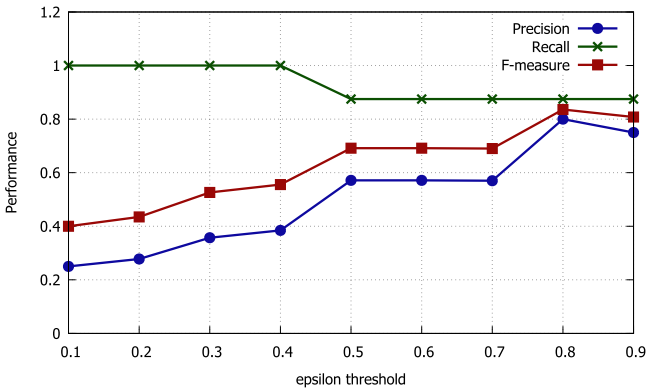


Fig. 9. Performance of the SKG model with different values of the threshold.

will extend the classes of the SKG model to better represent relationships, for example, the cause and effect interactions between social events.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP) (NRF-2017R1A2B4010774).

Declaration of competing interest

No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.future.2019.05.016.

References

[1] H.L. Nguyen, J.E. Jung, Socioscope: A framework for understanding internet of social knowledge, Future Gener. Comput. Syst. 83 (2018) 358–365.
[2] G. Bello-Orgaz, J.J. Jung, D. Camacho, Social big data: Recent achievements and new challenges, Inf. Fusion 28 (2016) 45–59.
[3] S. Petrović, M. Osborne, V. Lavrenko, Streaming first story detection with application to twitter, in: Proceedings of the 11th International Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2010), Los Angeles, California, USA, June (2010) 2-4,, ACL, 2010, pp. 181–189.
[4] S. Phuvipadawat, T. Murata, Breaking news detection and tracking in twitter, in: Proceedings of the International Conference on Web Intelligence and International Conference on Intelligent Agent Technology (WI-IAT 2010), Toronto, Canada, August 31 - September 3, 2010, IEEE, 2010, pp. 120–123.
[5] E. Benson, A. Haghighi, R. Barzilay, Event discovery in social media feeds, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011), Portland, Oregon, USA, June (2011) 19-24, ACL, 2011, pp. 389–398.
[6] T. Sakaki, M. Okazaki, Y. Matsuo, Earthquake shakes twitter users: real-time event detection by social sensors, in: Proceedings of the 19th International Conference on World Wide Web (WWW 2010), Raleigh, North Carolina, USA, April (2010) 26-30, ACM, 2010, pp. 851–860.
[7] F. Abel, C. Hauff, G.-J. Houben, R. Stronkman, K. Tao, Twitcident: fighting fire with information from social web streams, in: Proceedings of the 21st World Wide Web Conference (WWW 2012), Lyon, France, April (2012) 16-20, ACM, 2012, pp. 305–308.
[8] N. Adam, J. Eledath, S. Mehrotra, N. Venkatasubramanian, Social media alert and response to threats to citizens (smart-c), in: Proceedings of the 8th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom 2012), Pittsburgh, Pennsylvania, USA, October (2012) 14-17, IEEE, 2012, pp. 181–189.
[9] H.L. Nguyen, O.-J. Lee, J.E. Jung, J. Park, T.-W. Um, H.-W. Lee, Event-driven trust refreshment on ambient services, IEEE Access 5 (2017) 4664–4670.
[10] R. Kosala, E. Adi, Harvesting real time traffic information from twitter, Procedia Eng. 50 (2012) 1–11.
[11] K. Zeng, W. Liu, X. Wang, S. Chen, Traffic congestion and social media in china, IEEE Intell. Syst. 28 (1) (2013) 72–77.
[12] B. O'Connor, M. Krieger, D. Ahn, Tweetmotif: Exploratory search and topic summarization for twitter, in: Proceedings of the 4th International Conference on Weblogs and Social Media (ICWSM 2010), Washington, DC, USA, May (2010) 23-26, AAAI Press, 2010, pp. 384–385.
[13] J. Sankaranarayanan, H. Samet, B.E. Teitler, M.D. Lieberman, J. Sperling, Twitterstand: news in tweets, in: Proceedings of the 17th International Conference on Advances in Geographic Information Systems (GIS 2009), Seattle, Washington, USA, November (2009) 4-6, ACM, 2009, pp. 42–51.
[14] D.A. Shamma, L. Kennedy, E.F. Churchill, Peaks and persistence: modeling the shape of microblog conversations, in: Proceedings of the International Conference on Computer Supported Cooperative Work (CSCW 2011), Hangzhou, China, March (2011) 19-23, ACM, 2011, pp. 355–358.
[15] J. Lehmann, B. Gonçalves, J.J. Ramasco, C. Cattuto, Dynamical classes of collective attention in twitter, in: Proceedings of the 21st International Conference on World Wide Web (WWW 2012), Lyon, France, April (2012) 16-20, ACM, 2012, pp. 251–260.
[16] Y. Li, Z. Ma, W. Lu, Y. Li, Automatic removal of the eye blink artifact from eeg using an ica-based template matching approach, Physiol. Meas. 27 (4) (2006) 425.
[17] C.-M. Kim, H.-M. Park, T. Kim, Y.-K. Choi, S.-Y. Lee, Fpga implementation of ica algorithm for blind signal separation and adaptive noise canceling, IEEE Trans. Neural Netw. 14 (5) (2003) 1038–1046.
[18] A. Hyvärinen, E. Oja, Independent component analysis: algorithms and applications, Neural Networks 13 (4–5) (2000) 411–430.
[19] W.R. Van Hage, V. Malaisé, R. Segers, L. Hollink, G. Schreiber, Design and use of the simple event model (sem), J. Web Semant. 9 (2) (2011) 128–136.
[20] H.L. Nguyen, J.J. Jung, Utilizing dynamics patterns of trust for recommendation system, in: Proceedings of the 13th International Conference on Intelligent Environments (IE 2017), Seoul, South Korea, August (2017) 21-25, IEEE, 2017, pp. 108–113.
[21] J.-T. Chien, B.-C. Chen, A new independent component analysis for speech recognition and separation, IEEE Trans. Audio Speech Lang. Process. 14 (4) (2006) 1245–1254.
[22] L. Potamitis, N. Fakotakis, G. Kokkinakis, Independent component analysis applied to feature extraction for robust automatic speech recognition, Electron. Lett. 36 (23) (2000) 1977–1978.
[23] R.H. Lambert, A.J. Bell, Blind separation of multiple speakers in a multipath environment, in: Proceedings of the 22nd International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1997), Munich, Germany, April (1997) 21-24, IEEE, 1997, pp. 423–426.
[24] J.-H. Lee, H.-Y. Jung, T.-W. Lee, S.-Y. Lee, Speech feature extraction using independent component analysis, in: Proceedings of the 25th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2000), Istanbul, Turkey, June (2000) 5-9, IEEE, 2000, pp. 1631–1634.
[25] T.-W. Leung, C.-W. Ngo, R.W. Lau, Ica-fx features for classification of singing voice and instrumental sound, in: Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004) Cambridge, UK, August (2004) 23-26, 2, IEEE, 2004, pp. 367–370.
[26] M.S. Bartlett, J.R. Movellan, T.J. Sejnowski, Face recognition by independent component analysis, IEEE Trans. Neural Networks 13 (6) (2002) 1450–1464.
[27] W. Ding, A new method for image noise removal using chaos-pso and nonlinear ica, Procedia Eng. 24 (2011) 111–115.
[28] S.-I. Noh, K. Bae, K.R. Park, J. Kim, A new iris recognition method using independent component analysis, IEICE Trans. Inf. Syst. 88 (11) (2005) 2573–2581.
[29] T. Honkela, A. Hyvärinen, J.J. Väyrynen, Wordica—emergence of linguistic representations for words by independent component analysis, Natural Lang. Eng. 16 (3) (2010) 277–308.
[30] A. Chagnaa, C.-Y. Ock, C.-B. Lee, P. Jaimai, Feature extraction of concepts by independent component analysis, J. Inf. Process. Syst. 3 (1) (2007) 33–37.
[31] M. Rospocher, M. van Erp, P. Vossen, A. Fokkens, I. Aldabe, G. Rigau, A. Soroa, T. Ploeger, T. Bogaard, Building event-centric knowledge graphs from news, J. Web Semant. 37 (2016) 132–151.
[32] E. Kuzey, G. Weikum, Evin: Building a knowledge base of events, in: Proceedings of the 23rd International Conference on World Wide Web (WWW 2014), Seoul, South Korea, April (2014) 7-11, ACM, 2014, pp. 103–106.
[33] D.B. Lenat, Cyc: A large-scale investment in knowledge infrastructure, Commun. ACM 38 (11) (1995) 33–38.
[34] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, J. Taylor, Freebase: a collaboratively created graph database for structuring human knowledge, in: Proceedings of the 34th International Conference on Management of Data (SIGMOD 2008), Vancouver, British Columbia, Canada, June (2008) 10-12, AcM, 2008, pp. 1247–1250.
[35] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, S. Hellmann, Dbpedia-a crystallization point for the web of data, Web Semantics: Sci. Serv. Agents World Wide Web 7 (3) (2009) 154–165.

[36] F.M. Suchanek, G. Kasneci, G. Weikum, Yago: a core of semantic knowledge, in: Proceedings of the 16th International Conference on World Wide Web (WWW 2007), Banff, Alberta, Canada, May (2007) 8-12, ACM, 2007, pp. 697–706.
[37] A. Cui, M. Zhang, Y. Liu, S. Ma, K. Zhang, Discover breaking events with popular hashtags in twitter, in: Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM 2012), Maui, HI, USA, October 29 - November 02, 2012, ACM, 2012, pp. 1794–1798.
[38] N. Hoang Long, J.J. Jung, Privacy-aware framework for matching online social identities in multiple social networking services, Cybernet. Syst. 46 (1–2) (2015) 69–83.
[39] D. Patil, N. Das, A. Routray, Implementation of fast-ica: A performance based comparison between floating point and fixed point dsp platform, Meas. Sci. Rev. 11 (4) (2011) 118–124.

Dr. Hoang Long Nguyen is a Postdoc researcher in Chung-Ang University, Korea since 2019. He received the B.S. in Department of Computer Engineering from Ho Chi Minh City University of Technology, Vietnam in April 2013. And he received MSc and PhD degrees in Department of Computer Engineering from Yeungnam University and Chung-Ang University in Korea, respectively. His research topics include knowledge engineering on social networks by using machine learning, semantic Web mining, and ambient intelligence.

Prof. Jason J. Jung is a Full Professor in Chung-Ang University, Korea, since September 2014. Before joining CAU, he was an Assistant Professor in Yeungnam University, Korea since 2007. Also, He was a postdoctoral researcher in INRIA Rhone-Alpes, France in 2006, and a visiting scientist in Fraunhofer Institute (FIRST) in Berlin, Germany in 2004. He received the B.Eng. in Computer Science and Mechanical Engineering from Inha University in 1999. He received M.S. and Ph.D. degrees in Computer and Information Engineering from Inha University in 2002 and 2005, respectively. His research topics are knowledge engineering on social networks by using many types of AI methodologies, e.g., data mining, machine learning, and logical reasoning. Recently, he has been working on intelligent schemes to understand various social dynamics in large scale social media (e.g., Twitter and Flickr).