Knowledge-Based Systems 138 (2017) 105–112

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Adaptive online event detection in news streams☆

Linmei Hu*, Bin Zhang, Lei Hou, Juanzi Li

Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China

ARTICLE INFO

Article history:
Received 7 June 2017
Revised 24 September 2017
Accepted 30 September 2017
Available online 9 October 2017

MSC:
00-01
99-00

Keywords:
Word embedding
Adaptive online clustering
Event detection

ABSTRACT

Event detection aims to discover news documents that report on the same event and arrange them under the same group. With the explosive growth of online news, there is a need for event detection to facilitate better navigation for users in news spaces. Existing works usually represent documents based on TF-IDF scheme and use a clustering algorithm for event detection. However, traditional TF-IDF vector representation suffers problems of high dimension and sparse semantics. In addition, with more news documents coming, IDF need to be incrementally updated. In this paper, we present a novel document representation method based on word embeddings, which reduces the dimension and alleviates the sparse semantics compared to TF-IDF, and thus improves the efficiency and accuracy. Based on the document representation, we propose an adaptive online clustering method for online news event detection, which improves both the precision and recall by using time slicing and event merging respectively. The resulted events are further improved by an adaptive post-processing step which can automatically detect noisy events and further process them. Experiments on standard and real-world datasets show that our proposed adaptive online event detection method significantly improves the performance of event detection in terms of both efficiency and accuracy compared to state-of-the-art methods.

1. Introduction

The rapidly-growing amount of electronically available information threatens to overwhelm human attention, raising new challenges for information retrieval technology. Traditional query-driven retrieval is useful for content-focused queries, but is not proper for generic queries like “What happened?” or “What’s new?”. Therefore, there is a need to build automated, unsupervised methods which can simplify the operation of keeping abreast with new events that occur in the news [1–3]. Event detection is the task of discovering news documents that report on the same event and arranging them under the same group. Event detection provides a conceptual structure of the news stories and facilitates better navigation for users in news spaces.

While event detection has been studied for many years, it is still an open problem [4,5]. The most prevailing approach for event detection was proposed by Allan et al. [6] and Yang et al. [4], in which documents are processed by an online system. In such online systems, when receiving a document, the similarities between the incoming document and the known events (sometime represented by a centroid) are computed, and then a threshold is applied to determine whether the incoming document is the first story of a new event or a story of some known event. Modifications to this approach could be summarized from two aspects: better representation of contents (e.g., using named entities) [7–9] and utilizing time information [4,10]. The dominant technique for document representation in these previous work is standard TF-IDF scheme[5]. TF-IDF vector representation suffers the problems of high-dimension and semantic sparsity, leading to high computation cost and low accuracy. In addition, existing event detection methods can not automatically detect noisy events which need further processing.

In this work, we represent documents based on word embeddings and propose an adaptive online clustering algorithm for event detection. Specifically, we first learn word embeddings, and then cluster the words into different semantic classes via K-means algorithm. Words in the same cluster share the same or similar semantics. Next, we represent each news document as a distribution over the semantic classes. This representation reduces the dimension of traditional document representation (TF-IDF) and alleviates the semantic sparsity. Based on the new proposed document representation, we propose an adaptive online clustering method for event detection. Specifically, we take full advantage of the time information of news documents and arrange the news documents in chronological order. Then we divide the news documents into slices with a fixed time window size. For each time slice, we apply the single-pass online clustering method to detect events. Actually,

☆ Fully documented templates are available in the elsarticle package on CTAN.

* Corresponding author.
E-mail addresses: hlm12@mails.tsinghua.edu.cn (L. Hu), zb16@mails.tsinghua.edu.cn (B. Zhang), lijuanzi@tsinghua.edu.cn (J. Li).

https://doi.org/10.1016/j.knosys.2017.09.039
0950-7051/

this step can be done in parallel for all the time slices, which further improves the time efficiency. Since news documents reporting on the same event are usually close in time distance, time slicing ensures the precision of event detection. On the other hand, to ensure the recall, we merge events in different time slices which may refer to the same event based on similarities. Finally, we propose an adaptive method to automatically detect noisy events and further process these events. Therefore, our proposed event detection method improves event detection in terms of both efficiency and accuracy. Our contributions can be summarized as follows.

(1) We present a novel document representation method based on word embeddings, which reduces the dimension of traditional TF-IDF representation and alleviates the semantic sparsity, thus improving efficiency and accuracy of event detection.
(2) We propose a novel adaptive online clustering method, which can automatically detect noisy events and further process these events. It improves the performance of event detection in terms of both efficiency and accuracy significantly.
(3) Experiments on standard and real-world datasets show that our proposed adaptive online event detection method significantly outperforms state-of-the-art event detection methods.

The remainder of this paper is organized as follows. In Section 2, we formulate the event detection problem. In Section 3, we detail our proposed method. Section 4 describes our experimental results. In Section 5, we review the related literature, followed by conclusion and future research directions in Section 6.

2. Problem definition

In this section, we first define some concepts as well as the problem of event detection.

News stream. News documents from various sources usually form a stream in chronological order. A news stream $D = \{d_1, \ldots, d_m, \ldots\}$ is a sequence of documents. d_i is associated with a pair (d_i, t_i), where d_i is a document comprising a sequence of words and t_i is the publishing time in non-descending order, i.e. $t_i \leq t_{i+1}$.

Event. An *event* is a particular thing that happened at a specific time and place [6,11], and an event is usually composed of a set of news documents reporting on it. We consider an event $E = \{d_1, \ldots, d_M\}$ as a sequence of news documents.

For example, "2010 Chile earthquake" is an event. It consists of sequential news documents describing different aspects of the event such as *rescue efforts, damages, chaos*, and so on.

Event detection. The task of event detection is to discover news documents reporting on the same event from a news stream $D = \{d_1, d_2, \ldots, d_m, \ldots\}$ and divide them into event-centric clusters $\{E\}$, where $E = \{d_1, \ldots, d_M\}$ according to the events they report on.

Event detection typically has two major components: 1) document representation and 2) cluster analysis [12]. Document representation handles with translating documents into structures appropriate for clustering. This is generally completed by representing documents numerically as vectors and matrices. Cluster analysis contains methods for designing meaningful data clusters from the data structure formed by document representation methods.

Traditional methods usually use TF-IDF vectors to represent documents, which leads to high dimension and sparse semantics. In addition, IDF needs to be incrementally updated when more and more new documents come. Therefore, we propose a new document representation based on word embeddings, which avoids the problems of TF-IDF. Based on the document representation, a lot of clustering methods such as *K*-means, LDA and single-pass online clustering can be applied for event detection. However, *K*-means and LDA need a prior knowledge of cluster number, which is quite hard to determine. The number of clusters has a big influence on the clustering results. On the other hand, it is more convenient to control the similarity threshold of single-pass online clustering. In addition, single-pass online clustering is a one-pass algorithm which is much more efficient. Therefore, we propose a new adaptive online clustering algorithm based on single-pass online clustering. Our proposed clustering algorithm further considers improving the efficiency and accuracy of online event detection via time slicing and adaptive post-processing respectively.

3. Our method

In this section, we detail our proposed adaptive online event detection method to automatically group news documents according to the events they report on. As a result, each event is composed of a sequence of news documents. We first present our document representation based on word embeddings and then describe our adaptive online clustering algorithm for event detection.

3.1. Document representation

To alleviate the problems of traditional TF-IDF representation, we propose a novel document representation based on word embeddings. As shown in Fig. 1, our proposed document representation consists of three steps: word embedding, word clustering and document vectorization.

Word Embedding. Word is the basic element in a document, so we first transform words into continuous low-dimensional vectors. Let $\mathcal{V}$ denote the vocabulary in all the news documents D, we employ skip-gram model [13] to learn a mapping function: $\mathcal{V} \rightarrow R^M$, where R^M is the M-dimensional representation of w_i. Specifically, given a document $d \in D$ associated with word sequence $w_1, w_2, \ldots, w_N$, skip-gram model maximizes the co-occurrence probability among words that appear within a contextual window k:

$$\max_{\mathbf{w}} \frac{1}{N} \sum_{i=1}^{N} \sum_{j=i-k}^{i+k} \log p(w_j|w_i) \tag{1}$$

The probability $p(w_j|w_i)$ is formulated as:

$$p(w_j|w_i) = \frac{\exp(\mathbf{w}_j^T \mathbf{w}_i)}{\sum_i \exp(\mathbf{w}_i^T \mathbf{w}_i)} \tag{2}$$

where $\mathbf{w}_i$ is the vector representation of word w_i. Different from TF-IDF vector representation, word embeddings do not need to be incrementally updated with more and more new documents coming since we can use a large independent news corpus to train word embeddings.

Word Clustering. Traditional TF-IDF representations suffer problems of the curse of dimensionality and feature independence assumption. These methods often ignore the semantic relationships among word features which leads to document sparse representation with many zero features values. If there are two documents describing similar events but using different words, they have difficulty in making a correct decision that they belong to the same event [7]. Thus, traditional text processing based on keyword comparison could not provide good performance. To reduce the semantic sparsity, we use *K*-means clustering to cluster words referring to the same or similar meaning to obtain a latent semantic space.

Document vectorization. At last, for each document, we can replace the words with corresponding word cluster indexes. Then a document can be easily represented by the distribution of word clusters. By representing the documents in the same latent space, we can alleviate the problems of high dimension and semantic