A Fresh Look at Understanding News Events Evolution

Longtao Huang, Shangwen Lv, Liangjun Zang, Yipeng Su
Institute of Information Engineering, Chinese Academy of Sciences
Beijing, China
{huanglongtao, lvshangwen, zangliangjun, suyipeng}@iie.ac.cn

Jizhong Han, Songlin Hu
[1]University of Chinese Academy of Sciences
[2]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
husonglin@iie.ac.cn

ABSTRACT

This paper proposes a novel approach to retrieve news articles related to a specific event and generate a storyline to help people understand the event evolution. First, a similarity calculation method is proposed to retrieve news articles related to the specific event, which combines textual similarity, temporal similarity and entity similarity. Then a multi-view attribute graph is constructed to represent the relationship between retrieved articles. Finally, a community detection algorithm is developed to segment and chain subevents in the graph. Experimental results on real-world datasets demonstrate that the proposed approach achieve better results than existing methods. [1]

KEYWORDS

event evolution, storyline, document understanding

1 INTRODUCTION

Motivation: Information overload has become a headache for people as the overwhelming number of news articles makes it difficult to find their desired results directly. Furthermore, only collecting the related articles to an event is not adequate. People are willing to get the knowledge of the overall view of the specific event after they have read some news. Thus, discovering the evolution process from massive and unorganized news articles is helpful for people to get an instant understanding of the event.

Problem Statement: This paper focuses on the problem of retrieving all the news articles related to a given event and generating a storyline to describe the evolution process. The challenges of this problem include the following aspects:

1) It is hard to distinguish a specific event from others with similar topics since different events with similar topics are always described with common words. This can hardly be differentiated by most existing methods based on textual similarity.

2) The topics in an event might evolve over time. For most existing methods, subevents with different topics might be identified as different events, since they can hardly recognize the connections by considering textual similarity.

3) In order to describe the evolution process of an event, it needs to segment the related news articles to several stages. The challenge is how to decide the segmenting points, between which the subevents have a significant change.

Contribution: In this paper, we propose a novel approach aiming at tackling the above challenges in understanding the event evolution. First, we propose Event-Oriented Similarity that combines textual similarity, temporal similarity and entity similarity to retrieve news articles related to the specific events. Then a multi-view attribute graph is constructed to represent the relationship between retrieved news articles. Next, a community detection algorithm is developed to segment subevents and link the subevents into a storyline. Finally, we select the most representative news article in each subevent to show the evolution process of the event. The experimental results show that we can get better performance than other methods. This paper involves the following contributions:

- We design a novel approach from starch to Understand News Events Evolution (UNEE) for a specific event according to user input news articles.
- We propose a similarity calculation method (Event-Oriented Similarity) combining textual similarity, temporal similarity and entity similarity to help distinguish similar events and identify the topic evolution within an event.
- We conduct comprehensive experiments on real-world datasets to verify the proposed method.

2 METHODOLOGY

This paper targets at retrieving all related news articles according to some user input news articles about the specific event and generating a storyline for the event. Formally, given a set of news articles $D = \{d_1, d_2, \dots\}$, and some user input news articles $Q = \{q_1, q_2, \dots\}$ about the specific event e. The goal is to find all news articles C about e and generate a storyline S, where $C \subseteq D$ and S is a sequence of articles chosen from C. To address this task, we propose a method of understanding news events evolution, which consists of two steps: event retrieval and storyline generation.

In event retrieval stage, we firstly propose a similarity calculation method (Event-Oriented Similarity), which combines textual similarity, temporal similarity and entity similarity. It can help a lot in distinguishing similar events and recognizing the topic evolution within an event. Then, we can retrieve high-quality news articles related to the given news articles based on Event-Oriented Similarity.

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04. https://doi.org/10.1145/3184558.3186913

Given two documents i and j, the *textual similarity* is measured by cosine metric:

$$sim_{tex}(i,j) = cosine(\boldsymbol{v}_i, \boldsymbol{v}_j) \quad (1)$$

where $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ are the feature vectors corresponding to i and j. This paper adopts Paragraph Vector model [1] to learn a distributed representation for each news article.

The *temporal similarity* of i and j is measured as follow:

$$sim_{tem}(i,j) = -\frac{\log((|t_i - t_j|+1)/H)}{\lceil \log H \rceil} \quad (2)$$

where t_i and t_j are the publication time of i and j, and H is the time horizon of the whole corpus and $\lceil \log H \rceil$ is used to compress the temporal similarity between 0 and 1.

The *entity similarity* of i and j is measured as follow:

$$sim_{ent}(i,j) = \frac{|E_i \cap E_j|}{|E_i \cup E_j|} \quad (3)$$

where E_i and E_j are the sets of named entities extracted from i and j. The entity similarity works because the entities in similar events can differ obviously even though the news articles have many words in common. Besides, entities in different subevents of the same event are usually similar or have some relations even when topics vary in different subevents.

Since the above similarities are calculated independently, Event-Oriented Similarity combines them by calculating the product:

$$sim_{EOS}(i,j) = sim_{tex}(i,j) * sim_{tem}(i,j) * sim_{ent}(i,j) \quad (4)$$

Then the articles with sim_{EOS} higher than the threshold θ_1 are retrieved as related articles C according to the given articles Q.

In storyline generation stage, a Multi-View Attribute Graph (MVAG) is generated to represent the relationship between the retrieved news articles, where the nodes are the articles with multiple attributes (title, content, entities, publication time, etc.) in C and the edges between two nodes means the two articles has higher sim_{EOS} than the threshold θ_2. Then the target of subevents segmentation is transferred to a community detection process on the MVAG. We adopt a fast community detection algorithm [2] to acquire subevents. Then the subevents are chained in temporal order and form a storyline.

To make the storyline more understandable, a representative news article is selected from each subevent and can be regarded as a summary of the subevent. For each subevent, the centrality of an article i is proposed to measure the representativeness.

$$centraliy_i = \frac{SP_i}{SP} \quad (5)$$

where SP is the set of all shortest paths between any two nodes in the graph of the subevent that i belongs to. Similarly, SP_i is the set of all shortest paths between i and other nodes in the subevent. Then, an article with the highest *centrality* is selected from each subevent and generate the final storyline.

3 EXPERIMENTS AND ANALYSIS

This section shows the experiment conducted to evaluate the quality of the output storylines, in comparison with different methods.

Wikipedia current news portal[1] provides high-quality manually-edited structures and ground truth for events retrieval and storyline generation. Each subevent contains a set of news that are related. We crawl the outer links cited in each subevent to form the news corpus. In total, we crawl 46, 768 news articles, including 693 events. We compare the proposed method *UNEE* with some classic methods for events clustering (*KMeans*, *Spectral Clustering*, *LDA-based KMeans*) and a state-of-the-art method *TTGPKUICST2* [3] for storyline generation. We adopt the metric *V-measure* [4] to evaluate the quality of subevents segmentation and the metric *F-measure* [5] to evaluate the temporal coherence of subevents.

Table 1: Comparison Result

Method	*V-measure*	*F-measure*
KMeans	0.3477	0.4930
Spectral Clustering	0.4350	0.5999
LDA-based KMeans	0.3539	0.5728
TTGPKUICST2	0.4456	0.6214
UNEE	**0.5267**	**0.7235**

Table 1 shows the comparison result. We can observe that our proposed method outperforms other methods on *V-measure* and *F-measure*. The good result benefits from two reasons. On one hand, Event-Oriented Similarity can get a better representation for news articles, thus helping to retrieve high-quality news articles from the whole corpus and distinguish unrelated articles. On the other hand, the community detection algorithm can represent the inner structure between news articles related to a specific event by considering subevents as different communities in the graph.

4 CONCLUSIONS

This paper proposes a novel approach to understand the event evolution of news events based on user input news articles. First, we propose Event-Oriented Similarity to distinguish similar events and recognize the topic evolution within the event. Then a Multi-View Attribute Graph is constructed to represent the relationship between news articles. Finally, we transfer subevent segmentation to a community detection problem and generate the final storyline. In the future, we will focus on the methods to represent the entity similarity between news articles and get better performance in generating storylines.

REFERENCES

[1] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. ICML 2014:1188–1196.
[2] Tomoya Yamazaki, Nobuyuki Shimizu, Hayato Kobayashi, and Satoshi Yamauchi. Weighted Micro-Clustering: Application to Community Detection in Large-Scale Co-Purchasing Networks with User Attributes. WWW 2016: 131-132.
[3] Chao Lv, Feifan Fan, Runwei Qiang, Yue Fei, and Jianwu Yang. PKUICST at TREC 2014 Microblog Track: Feature Extraction for Effective Microblog Search and Adaptive Clustering Algorithms for TTG. TREC 2014.
[4] Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropybased external cluster evaluation measure. EMNLP-CoNLL 2007: 410-420.
[5] Erdal Kuzey, Jilles Vreeken, and Gerhard Weikum. A fresh look on knowledge bases: Distilling named events from news. CIKM 2014: 1689–1698.

[1] https://en.wikipedia.org/wiki/Portal:Current_events

This research is supported in part by the National Key Research and Development Program of China (No. 2017YFB1010000) and the National Natural Science Foundation of China (No. 61702500).