一种基于时空分析的事件抽取方法*

梁月仙[1,2,3] 郭 智[1,2]
(1.中国科学院 空间信息处理与应用系统技术重点实验室,北京 100190;
2.中国科学院电子学研究所,北京 100190; 3.中国科学院大学,北京 100190)

摘 要:新闻媒体及社交网站每天呈现大规模的时空文本数据,人们难以从中获取有价值的事件信息。针对前人方法依赖大量的标注数据,同时以孤立的方式考虑事件的时间要素和空间要素等的问题,提出一种基于时空分析的事件抽取方法,该方法首先引入数据立方体结构存储事件信息,用户可基于不同的时空粒度抽取出重要的事件;然后提出一种基于语义相似性的实时事件聚类算法,该聚类算法采用 GloVe 模型训练词的语义关联性,使聚在同一事件类的事件元素具有强的语义关联性。在大量未标注的网络文本中,该方法取得了 77.4%的 F_1 值,表明了该方法能够实现时空分析下的事件抽取任务。

关键词:事件抽取;时空分析;实时聚类;可视化
中图分类号: TP391.1 **文献标识码**:A **国家标准学科分类代码**: 520.2020

Approach of event extraction based on spatio-temporal analysis

Liang Yuexian[1,2,3] Guo Zhi[1,2]
(1. Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China;
2. Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China;
3. University of Chinese Academy of Sciences, Beijing 100190, China)

Abstract: News media and social networking sites present a great volume of spatio-temporal text data every day. It is difficult for people to get valuable event information. Aiming at the existing methods which depend on a large number of annotation data and consider the temporal elements and spatial elements of events in an isolated way, this paper proposes an event extraction approach based on spatio-temporal analysis. Firstly, a data cube structure was introduced to store event information. Users can extract important events based on different temporal and spatial granularity. Secondly, a real-time event clustering algorithm based on semantic similarity is proposed. In the clustering algorithm, a GloVe model was adopted to learn the vector space of words, so that the event elements which collected in the same event cluster have strong semantic relevance. According to a large number of unlabeled network text, the approach achieves a 77.4% value of F_1 which indicates that it can achieve event extraction task based on spatio-temporal analysis.

Keywords: event extraction; spatio-temporal analysis; real-time clustering; visualization

1 引 言

随着大数据时代的到来,每天都有大规模的时空文本数据产生,例如新闻媒体、微博等及时地呈现发生于世界各地的事件信息。这些文本数据具有非结构化、涉及面广、数据量大、杂乱零散等特点。人们难以从中获取有价值的事件信息。因此研究者们提出了事件抽取的任务,事件抽取旨在从非结构化的文本中抽取出有用的事件信息,并以结构化的形式呈现。事件抽取是一种有效的数据组织方式,在信息检索、问答系统和推荐系统等领域有着广泛的应用[1]。一个事件通常包括时间、地点、人物、事件内容等信息。另外,事件的时间特性和空间特性在事件抽取任务中扮演着重要的角色,因此如何充分利用这些文本数据的时空特性,挖掘出有价值的事件信息,已经成为一个

收稿日期:2017-03
* **基金项目**: 国家高新技术研究与发展(2015AA7013033)项目资助

研究热点。

传统的事件抽取方法常常采用两步策略：事件类型的识别以及事件元素的发现，前人工作通常采用监督机器学习方法实现事件抽取任务[2-5]。但是监督机器学习方法存在着不足，首先，它们需预先指定预处理的领域，存在领域移植性的问题；第二，这些监督机器学习方法依赖于大量的标注数据，但是标注数据的获取需要耗费大量的人力资源；最后，已有的标注数据无法涵盖各个领域的文本数据，难以对新兴的文本数据实现抽取任务。为了解决传统方法中存在的问题，研究者们提出了开放领域的信息抽取技术[6-8]。开放领域事件抽取旨在从开放领域的网络文本中抽取出有用的事件信息。但是他们采用传统的离线聚类方式对事件元素进行聚类，不能有效的实时处理在线的动态网络数据流，同时这些方法尚未考虑事件的时空属性。

近年来，研究者们越来越关注文本数据中的时空信息。本文将一个事件视为 TDT 任务中给出的定义，即“发生于某时某地，有一定人物参加的事情”，同时，本文认为时空分析下的事件抽取旨在抽取出事件的发生时间、发生地点以及事件元素，并挖掘出事件间的时空关联性。由于结合事件的时间特性以及空间特性能够最好的描述发生于某时某地的事件信息，因此综合考虑事件的时空属性并且挖掘出他们的时空关联性，是一个值得研究并且充满挑战性的问题。但是，前人工作主要以孤立的方式来考虑事件的时间属性以及空间属性。例如文献[9-11]研究基于时间维度上的事件抽取。而文献[12-14]专注于抽取事件的空间特性。

为了解决时空分析下事件抽取任务的挑战，提出了一种基于时空分析的实时事件抽取方法。首先，为了综合分析事件的时空要素并且挖掘事件的时空关联性，引入数据立方体结构存储事件信息；然后，提出一种基于语义相似性的实时事件聚类算法，该算法采用 GloVe 模型训练词向量，能够挖掘出事件元素词的潜在语义信息，使聚在同一事件类的事件元素具有强的语义关联性。本文提出的方法综合考虑了事件的时空要素，并且挖掘出事件间的时空关联性。同时，该方法无需事先指定预处理的领域，不存在领域移植性的问题，能够解决随时间演变的事件信息，可适用于大规模的开源领域网络文本数据流。

2 研究方法

采用非监督学习方式，提出了一种基于时空分析的实时事件抽取方法。事件抽取的系统框架如图 1 所示，整个系统框架包括 3 个组成部分：

1）事件元素的抽取及存储：通过网络爬虫技术抓取大规模的未标注文本数据，并从文本数据集中抽取出时间表达式，地名实体，人名实体以及事件触发词。

2）引入数据立方体结构存储事件元素，以利于综合分析事件的时空要素并且挖掘事件的时空关联性。

3）事件的实时聚类：采用 GloVe 词向量模型训练事件元素的语义关联性，提出基于语义相似性的实时事件聚类算法，抽取出重要的事件类。

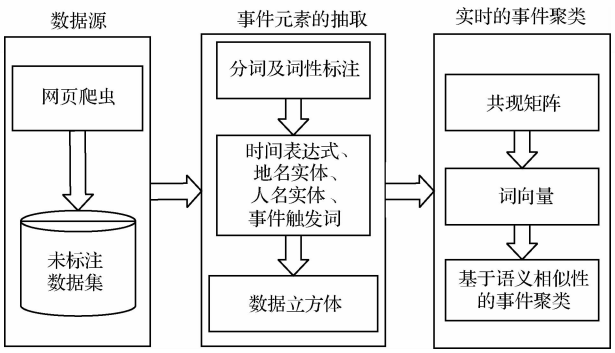

图 1 系统框图

2.1 数据元素的抽取及存储

事件信息通常由发生时间、发生地点、参与人员、事件触发词构成。事件触发词是表达事件发生性质的词，在 ACE 测评会议的标注数据中，事件触发词一般为动词或动名词[15]。而网络文本一般为非结构化信息，为了更好的表示事件信息，需要将非结构化的文本转化为结构化信息，因此需要对文本数据进行数据预处理，即从文本中抽取出时间表达式、地名实体、人名实体和事件触发词。采用 ICTCLAS 分词工具对长文本数据进行分词和词性标注，并抽取出时间表达式、地名实体和人名实体。

为了抽取出事件触发词，本文将事件触发词识别作为一个分类任务，采用条件随机场(CRF)模型抽取出事件触发词。为了训练 CRF 模型，文本借鉴 Timebank Corpus 的标注指导，同时考虑语境特征和词性特征，人工标注 200 条新闻文本的事件触发词，然后运用 CRF 模型对候选事件触发词进行建模，抽取出最能表达事件性质的触发词。

在长文本中，通常一篇文档包含多个时间信息、多个地理位置，即一个文本文档可能描述多个事件，或多个文本文档描述同一个事件信息。本文将距离事件词最近的地名实体作为该事件词的发生地点，然后通过地名消歧义以及同名合并技术，同时将地名实体映射到地名经纬度数据库，最后将地名转化为经纬度坐标。社交网络网站或新闻平台是一个实时报道当天发生事件的平台，因此将文档的生成时间作为事件的发生时间。

为了综合分析事件的时空信息并且挖掘事件的时空关联性，本文创新性地引入数据立方体结构存储事件元素。首先，将事件词的时间信息作为立方体的一个维度，将事件词的经纬维度作为立方体的另外两个维度。然后，基于事件词的时空维度将事件词存储于立方体中。立方体存储事件词的结构如图 2 所示。

2.2 基于语义相似性的事件聚类

基于 2.1 节的数据预处理及存储步骤，已将事件词存

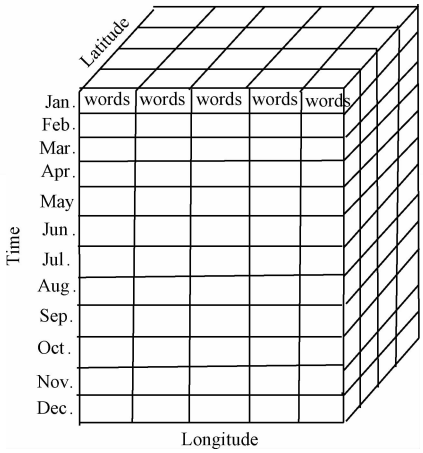

图 2 数据立方体存储事件词的结果示意图

储于立方体中，但是数据立方体里的事件词是杂乱无章的，需要对这些事件元素做聚类，我们定义一个事件类为一个事件，并且进行事件聚类基于以下几方面的考虑：

1）随着数据流的到来，新的事件随时可能出现，因此提出的聚类算法必须不能预先定义聚类的个数。

2）由于新闻信息流是一个随着时间依次到达的动态信息流，因此需要采用 Single-Pass 的聚类方式对事件信息进行聚类，而且应该避免迭代循环计算。

已有的方法研究话题聚类通常采用 K-means 聚类、层级聚类以及 LDA（latent dirichlet allocation）等的改进算法[16-18]，但是这些聚类算法都是批处理形式、而且它们是一种离线的方式，因此这些离线的聚类算法并不适合于动态的网络信息流。近年来，随着网络文本数据的兴起，研究者们提出了许多在线的聚类算法[19-20]。但是这些算法存在着一些不足，当涉及到相似度计算时，通常采用 one-hot 模型表征词的向量空间。然而 one-hot 模型存在着维数灾难的缺点，而且无法表达目标词的潜在语义信息。

2.2.1 事件的实时聚类

提出了一种基于语义相似性的实时事件聚类算法，该聚类算法是一种增量的聚类方式，随着事件信息的到来，聚类结果动态的改变。并且能够基于数据流的到达顺序依次处理数据。基于到来数据与已有事件类的语义相似性，新到达的数据将被聚到已有的事件类或作为一个新的事件类出现。另外，该聚类方法的时间效率是足够高的。

聚类方法的伪代码如下。

Algorithm 1：EVENT-CLUSTER(E,ω)

Input：Word w, Existing event set $E=\{e_1,e_2,\cdots,e_K\}$
Output：Updated event set E
If E is null
$e_1=w, c_1=w$
Else
　For each event e_i in the E **do**
$S_i=Sim(c_i,w)$
　Return the biggest S_b
　If $S_b>$threshold T then
Add w to the existing event e_b
Update the center vector c_b of event e_b
　For word w_i in the e_b do

$$c_b=\frac{1}{k}\sum_{i=1}^{1}w_i$$

Else
　add w to E as a new event

考虑一个新到达的事件词 w，假如 w 是第一个到来的事件词，那么将其作为第一个事件类；否则，将 w 分别与已有的事件类进行相似度计算，然后对所有相似值做降序排序，获得与某一个事件类 e_b 的最大相似值 S_b，如果 S_b 大于阈值 T，w 被聚到事件类 e_b 中，同时更新事件类 e_b 的质心向量 c_b，否则 w 被作为一个新的事件类添加到事件集 E 中，代码中的相似性计算采用余弦相似度公式，如式（1）所示。

$$\cos\theta=w_i^{\mathrm{T}}\tilde{w}_j/(|w_i||\tilde{w}_j|) \tag{1}$$

2.2.2 词向量的构建

上述聚类算法的一个中心环节为词之间的相似性计算，目前，最流行的计算词之间的语义相似性的方式为词向量的方式，前人提出了许多表证词向量的方法，本文采用 GloVe 模型[21]挖掘事件词之间的语义规律。GloVe 模型的优势在于它能在整个数据集上获得真正的全局共现信息。GloVe 模型通过计算词之间的全局共现率建模词的语义关联度。GloVe 认为如果词 i 与词 j 同时出现在一篇文档中，那么它们即为共现。同时，该模型考虑了词的共现距离，即共现距离越近，词的关联度越强。通过采用梯度下降法训练 GloVe 模型，即可获取目标词 w 与语境词 $\tilde{w}$ 的向量空间。本文设计每个事件元素的词向量为 w 与 $\tilde{w}$ 的叠加之和，GloVe 模型的详细推导过程见文献[21]。

3 实验结果与分析

3.1 数据集与实验设置

新闻媒体是一个实时报道当天事件的平台，本文采用网络爬虫技术随机抓取 2015 年 3 月 1 日至 2015 年 8 月 30 日的 121 157 条新闻文档，并过滤掉视频和图片。这些文档覆盖政治，经济，体育，娱乐以及环境等领域. 根据前文 2.1 节所述，通过数据预处理，获取 184 个事件的发生时间，7 494 个地名实体，20 665 个人名实体，10 022 个事件触发词，如表 1 所示。本文将人名实体以及事件触发词作为事件词，然后基于事件词的时空特征构建数据立方体存储事件信息。当涉及到基于语义相似性的实时事件聚类时，基于整个数据集的相似性统计分析，设置相似性阈值为 0.66。当训练 GloVe 模型时，设置初始率为0.05，词向量的维度为 300 维，并进行 100 次迭代计算，这些参数的选取是根据 GloVe 模型介绍中选择最优的参数。

表 1 事件元素的统计分析

时间	地名实体	人名实体	事件触发词
184	7 497	20 665	10 022

3.2 事件抽取结果

3.2.1 定性评价

本文抽取出发生在 2015 年 3 月 1 日到 2015 年 8 月 31 日事件抽取结果，获取了 143 个事件类，并人工标注这些事件类型，例如犯罪事件、股票等。并将这 143 个事件类归类到 ACE 测评会议所指定的 8 大类别中，并统计了这 8 大类所占的比例，如图 3 所示。由于获取的新闻语料主要是关于经济类、社会类的事件信息。在时空事件抽取结果中，经济类、政策类、社会类的主题事件占了较大的比例。表明了本文提出的时空事件抽取结果与实际数据相符合。表 2 中，展示了在 2015 年 3 月份，分别发生在北京市、上海市和广州市的 top-4 事件，根据表中展示的结果表明，可以自动发现这些事件的事件类型，例如：体育竞技、政治、环境保护、刑事案件、招生考试类事件。

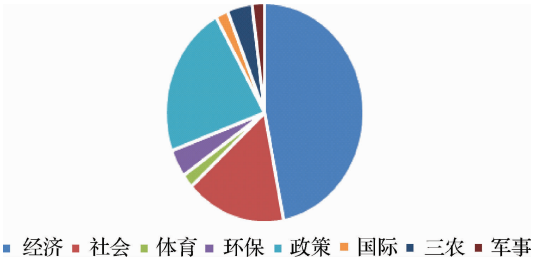

图 3 事件类型分布

表 2 2015 年 3 月份的 top-4 区域事件

时 间	位置	事 件	事件类型
2015 年 3 月	北京	拼搏，攻防，高敬刚，备战，李毅，交锋	体育竞技
		维护，程立峰，编制，出台，调试，法制化	政策出台
		破产，挂牌，出售，沈萌，中标，自营	企业破产
		巡查，开除，惩处，侵害，记过，失职	刑事案件
	上海	高铁，导航，小型化，沈晓明，商机	交通运输
		高考，俞敏洪，考试，办学，录取，授课	招生考试
		谢亚轩，浮动，李大霄，按揭，暴涨，购汇	经济指数
		投标，履约，房贷，出贷，王茹，电子化	企业改革
	广州	促进，数字化，网络化，李庆萍，巨变，城市化	经济发展
		评议，任命，聘用，议事，审定，廖小波	职位变迁
		泄漏，濒危，停产，致癌，污染，净化	环境保护
		交锋，失利，高洪波，发挥，复出，角逐	体育竞技

3.2.2 定量评价

为了定性的评价事件抽取结果，与 StreamCube 方法[22]和 DTM 方法进行对比实验，同时引入 3 个评价聚类质量的指标：NMI（normalized mutual information）、RI（rand index）以及 F1 值。实验结果如表 3 所示，通过测评结果，本文的方法胜于现有最优方法。因为 StreamCube 方法采用 one-hot 模型表征词向量，词向量空间易产生稀疏且冗余的缺点。然而，本文采用 GloVe 模型训练词向量，聚为一类的事件元素词具有更强的语义信息，另外，本文聚类算法可以采用较少的维度（如 300、400 维）来表征词向量，因此占用较少的内存空间以及较少的聚类时间。DTM 方法是一种动态话题模型，旨在模拟话题随时间变化的演变过程，它是一种离线的话题聚类方式，无法很好有效处理动态信息流，另外，它忽略了话题的空间特性，未能识别出话题的地域差异性。

表 3 聚类结果

	NMI	RI	F_1
StreamCube	0.550	0.677	0.541
DTM	0.725	0.768	0.687
RSTEvent	0.803	0.835	0.774

4 结 论

提出了一种基于时空分析的事件抽取方法。为了综合分析事件的时空要素并且挖掘事件的时空关联性，引入数据立方体结构存储事件信息。然后提出一种基于语义相似性的实时事件聚类算法，该算法采用 GloVe 模型训练词向量，使聚在同一事件类的事件元素具有强的语义关联性。本文方法无需事先指定预处理的领域，不存在领域移植性的问题，能够解决随时间演变的事件信息，可适用于大规模的开源领域网络文本数据流。下一步工作将研究事件的时空突发特性。

参 考 文 献

[1] 杜腾飞，毛建华，刘学锋. SOS 支持下的年报文本事件获取、管理与可视化[J]. 电子测量技术，2016，39(5)：61-65.

[2] SAKAKI T，OKAZAKI M，MATSUO Y. Earthquake

shakes Twitter users:Real-time event detection by social sensors[C]. International Conference on World Wide Web,2010:851-860.

[3] KUNNEMAN F, BOSCH V A. Event detection in Twitter: A machine-learning approach based on term pivoting[C]. Benelux Conference on Artificial Intelligence, 2014:65-72.

[4] NGUYEN M T, NGUYEN T T. Extraction of disease events for a real-time monitoring system[C]. Symposium on Information and Communication Technology, 2013:139-147.

[5] LI P, ZHOU G, ZHU Q. Minimally supervised chinese event extraction from multiple views[J]. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP),2016,16(2):13.

[6] LONG R, WANG H, CHEN Y, et al. Towards effective event detection, tracking and summarization on microblog data[C]. Web-Age Information Management. Berlin Heidelberg: Springer,2011:652-663.

[7] WELD D S, HOFFMANN R, WU F. Using Wikipedia to bootstrap open information extraction[J]. Acm Sigmod Record,2009,37(4):62-68.

[8] RITTER A, MAUSAM, ETZIONI O, et al. Open domain event extraction from twitter[C]. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and Data Mining, 2012: 1104-1112.

[9] JULINDA S, BODEN C, AKBIK A. Extracting a repository of events and event references from news clusters[C]. Aha! -Workshop on Information Discovery in Text, 2014:14-18.

[10] KUZEY E, WEIKUM G. Extraction of temporal facts and events from Wikipedia[C]. ACM,2012:25-32.

[11] ALONSO O, SHIELLS K. Timelines as summaries of popular scheduled events[C]. International Conference on World Wide Web, 2013, 1045(1):1037-1044.

[12] QUEZADA M, POBLETE B. Location-aware model for news events in social media[C]. International ACM SIGIR Conference, 2015:935-938.

[13] FOLEY J, BENDERSKY M, JOSIFOVSKI V. Learning to extract local events from the Web[C]. International ACM SIGIR Conference,2015:423-432.

[14] QUERCINI G, SAMET H, SANKARANARAYANAN J, et al. Determining the spatial reader scopes of news sources using local lexicons[C]. Sigspatial International Conference on Advances in Geographic Information Systems,2010:43-52.

[15] 陈自岩,黄宇,王洋,等. 一种非监督的事件触发词检测和分类方法[J]. 国外电子测量技术,2016,35(7):91-95.

[16] LIU H. Internet public opinion hotspot detection and analysis based on kmeans and SVM algorithm[C]. Information Science and Management Engineering. IEEE Xplore,2010:257-261.

[17] SILVA J D A, HRUSCHKA E R. A support system for clustering data streams with a variable number of clusters[J]. Acm Transactions on Autonomous & Adaptive Systems,2016,11(2):1-26.

[18] TSOLMON B, LEE K S. An event extraction model based on timeline and user analysis in Latent Dirichlet allocation[C]. International ACM SIGIR Conference on Research & Development in Information Retrieval, 2014:1187-1190.

[19] SILVA J A, FARIA E R, BARROS R C, et al. Data stream clustering: A survey[J]. Acm Computing Surveys,2014,46(1):13.

[20] YIN J, WANG J. A text clustering algorithm using an online clustering scheme for initialization[C]. The ACM SIGKDD International Conference, 2016: 1995-2004.

[21] PENNINGTON J, SOCHER R, MANNING C. glove: Global vectors for word representation[C]. Conference on Empirical Methods in Natural Language Processing, 2014:1532-1543.

[22] FENG W, ZHANG C, ZHANG W, et al. STREAMCUBE: Hierarchical spatio-temporal hashtag clustering for event exploration over the Twitter stream[C]. International Conference on Data Engineering, 2015: 1561-1572.

作 者 简 介

梁月仙,1981年出生,硕士研究生,主要研究方向为文本数据挖掘。

E-mail: liangyuexian@126.com

郭智,1975年出生,研究员,主要研究方向为文本挖掘,遥感图像处理。

E-mail: guozhi@mail.ie.ac.cn