◎大数据与云计算◎

基于词向量的微博事件追踪方法

张佳明,席耀一,王 波,唐浩浩,李天彩
ZHANG Jiaming, XI Yaoyi, WANG Bo, TANG Haohao, LI Tiancai

解放军信息工程大学 信息系统工程学院,郑州 450001
Institute of Information and System Engineering, PLA Information Engineering University, Zhengzhou 450001, China

ZHANG Jiaming, XI Yaoyi, WANG Bo, et al. Method of micro-blog event tracking based on word vector. Computer Engineering and Applications, 2016, 52(17):73-78.

Abstract: The traditional methods in micro-blog events tracking do not achieve good performance, because the length of micro-blog text is shorter and the cyber-words emerge constantly. To solve this problem, a method of micro-blog event tracking based on word vector is proposed. By using word vector, semantic similarity between the words can be computed, and the accuracy of semantic similarity between micro-blogs can also be improved. Firstly, the Skip-gram model is trained to get the word vector by using a large dataset. Then, the models for initial event and micro-blogs are constructed by extracting the keywords. Finally, the semantic similarities between micro-blogs and the initial event are computed through word vector, and the task of event tracking is completed according to the decision of pre-defined threshold. The experimental results show that the proposed method can make full use of semantic information contained by word vector, which can effectively improve the tracking performance compared with traditional methods.

Key words: micro-blog; event tracking; short text; Skip-gram model; word vector; semantic information

摘 要:微博文本长度短,且网络新词层出不穷,使得传统方法在微博事件追踪中效果不够理想。针对该问题,提出一种基于词向量的微博事件追踪方法。词向量不仅可以计算词语之间的语义相似度,而且能够提高微博间语义相似度计算的准确率。该方法首先使用Skip-gram模型在大规模数据集上训练得到词向量;然后通过提取关键词建立初始事件和微博表示模型;最后利用词向量计算微博和初始事件之间的语义相似度,并依据设定阈值进行判决,完成事件追踪。实验结果表明,相比传统方法,该方法能够充分利用词向量引入的语义信息,有效提高微博事件追踪的性能。

关键词:微博;事件追踪;短文本;Skip-gram模型;词向量;语义信息
文献标志码:A **中图分类号:**TP391 doi:10.3778/j.issn.1002-8331.1412-0144

1 引言

随着Web2.0的发展,微博作为社交媒体的代表受到了学术界、商业界以及政府部门的广泛关注。微博(Micro-blog)是一个基于用户关系的信息共享、传播及获取平台,用户不仅能够以140字以内的文本发布消息,实现即时分享,还可以通过微博对某一事件发表评论、表达观点并向其他网民施加影响[1]。微博在为网民提供便捷服务的同时,也为不良信息、网络谣言甚至是反动言论提供了传播渠道。微博已经成为舆情监测和情报搜集的重要信息来源,吸引了大批学者对其进行研

基金项目:国家高技术研究发展计划(863)(No.2011AA7032030D);全军军事研究生课题资助项目(No.2011JY002-158);国家社会科学基金项目(No.14BXW028)。

作者简介:张佳明(1989—),男,硕士研究生,研究方向:事件追踪与情感分析,E-mail:ZhangJM0629@163.com;席耀一(1987—),男,博士生,研究方向:基于时间线的事件追踪与摘要;王波(1970—),男,副教授,研究方向:网络协议分析、智能信息处理;唐浩浩(1990—),男,硕士研究生,研究方向:文本倾向性分析;李天彩(1990—),男,硕士研究生,研究方向:文本分割。

收稿日期:2014-12-10 **修回日期:**2015-02-13 **文章编号:**1002-8331(2016)17-0073-06
CNKI**网络优先出版:**2015-06-16, http://www.cnki.net/kcms/detail/11.2127.TP.20150616.1342.017.html

究[2]。因此,研究如何快速准确地追踪事件相关博文,能够辅助有关部门快速了解事件发展状况并及时作出舆情引导,对维护社会和谐稳定具有重要意义。

微博文本长度短,存在特征稀疏的问题;而且句法结构不规范,网络用语和口语化普遍,使得与同一事件相关的微博可能存在多种描述,这些描述在语义上是完全相近的,但是在形式上可能有很大不同,单纯地比较特征之间形式上的相似度而忽略其语义信息,很容易造成大量博文漏检。上述问题影响了传统微博事件追踪方法的性能。

词向量的基本思想是通过训练将每个词映射成 k 维的实数向量,通过词间的距离来判断它们之间的语义相似度。词向量包含丰富的语义信息,可以通过多种方式挖掘词与词之间的语义关系,在词语表示上更加准确[3]。因此,本文提出一种基于词向量的微博事件追踪方法,通过词向量计算词语之间的语义相似度,能够提高微博间语义相似度计算的准确率,从语义层面完成事件追踪,相比传统方法可以有效提高事件追踪的性能。

2 相关工作

事件追踪作为话题检测与追踪(Topic Detection and Tracking,TDT)研究的子任务,目的是对新闻数据流进行组织和利用,找出关于某个已知事件的后续报道或者相关文档[4]。近年来,随着微博的迅速发展和普遍应用,关于微博的事件追踪也逐渐成为研究热点。

2.1 新闻事件追踪

研究者针对新闻事件追踪提出了很多方法,总体可以分为基于文本分类的方法和基于查询向量的方法。

基于文本分类的方法是利用先验相关报道训练构建事件话题模型,进行有指导的学习,比较成功的有 k-近邻[5]、支持向量机(Support Vector Machine,SVM)[6-7]等。Yang 等[5]根据文档内容的相关性选择与当前报道最相似的 k 个先验报道作为最近邻,然后根据最近邻所属事件类别综合判定当前报道论述的事件;基于支持向量机的方法是通过先验数据训练 SVM 分类器,实现对后续文档的分类。潘渊等[6]等首先利用隐含语义分析(Latent Semantic Indexing,LSI)完成文本特征降维及语义表示,然后利用 SVM 进行事件追踪。Li 等[7]首先使用 KNN 算法提取重要特征,然后训练 SVM 分类器对后续文档进行分类,完成事件追踪。

基于查询向量的方法是根据先验数据构建一个查询向量,然后计算后续文档与该向量的相似度,并根据相似度阈值进行判决,从而完成事件追踪。Zhao 等[8]首先采用向量空间模型(Vector Space Model,VSM)作为文本表示模型,利用 TF-IDF 算法将文本表示成特征向量,然后利用余弦相似度构建文本之间的关系图,并使用图划分的算法对文本进行聚类,完成事件的检测和追踪。Yang 等[9]提出了事件演化的概念,综合考虑事件内容相似性、事件时间接近度、事件报道分布接近度来识别事件间的演化关系。

2.2 微博事件追踪

在微博事件追踪任务中,基于文本分类的方法,需要根据先验数据训练 SVM 分类器,当先验数据不足或者先验数据特征不够丰富时,其追踪效果不够理想。因此,目前使用较多的仍是基于查询向量的方法。Kwak 等[10]指出可以利用微博上的话题标签很好地发现和跟踪事件。然而,对于不含话题标签或者话题标签很多的微博上述问题依然存在。Choi 等[11]提出了一种结合微博时间信息的事件追踪模型,利用后续文档与初始事件之间的时间差来衡量二者之间的相关性,即二者时间越相近越相关。但不同事件有不同的持续时间,该方法中针对不同事件的时间差阈值不易设定。

基于查询向量的方法,简单直观,然而该类方法通常只计算特征向量之间形式上的相似度,忽略了博文之间的语义信息,导致大量“语义相近、形式不同”的博文被漏检。为此,王兰成等[12]提出了一种基于 PageRank 的文本概念图算法(PageRank Text Concept Graph,PTCG)算法,借鉴维基百科的语义相似度计算与词语共现关系,构建词语之间的文本概念图,并利用改进的 PageRank 算法完成事件追踪。然而该方法需要限定领域特征词,不适用于多事件追踪任务。也有很多研究者利用同义词词林[13]、知网(HowNet)[14]和维基百科(Wikipedia)[15]作为语义知识来解决微博短文本特征稀疏和语义计算问题。然而,微博涉及领域广、更新速度快、网络用语普遍,同义词词林、知网均为人工编写,词语数量有限,其覆盖面和更新程度难以满足微博事件追踪的实际需求。例如,关于事件“光大证券乌龙事件”的微博“A 股下午跳水收盘”和“沪深两市也双双高位回落,并先后翻绿”,从字面上看这两句话并没有共现的词语,如果使用传统基于查询向量的方法计算这两句话的相似度,结果会是 0;即使引入外部语义知识库,由于“收盘”和“翻绿”这些词语是特定领域出现的新词,知识库中并没有这些词语,而剩下的词也没有存在明显的语义关系,因此也无法计算这两句话的语义相似度。

词向量通过大规模公开语料训练得到,不仅可以计算词语之间的语义相似度,而且能够提高微博间语义相似度计算的准确率。例如在公开数据集上进行训练,可以从时事新闻和财经点评等报道中的共现关系发现“收盘”和“翻绿”都与“证券”、“股票”一类的词有关,属于语义比较相近的词。相比传统方法,词向量能够较好地识别出“语义相近,形式不同”的两个词语,进而可以计算出这两句话的语义相似度,提高微博间语义相似度计算的准确率。

3　基于词向量的微博事件追踪方法

本章首先给出基于词向量的微博事件追踪方法的基本流程,然后详细阐述其中的关键技术。

3.1　方法流程

本文方法的基本流程如图1所示,主要包括训练词向量、初始事件和微博表示模型、相似度计算与事件追踪三个部分。

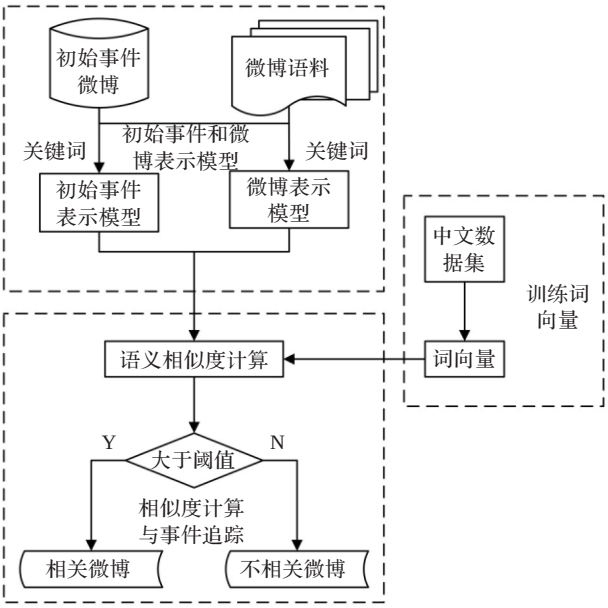

图1　基于词向量的微博事件追踪方法流程

训练词向量:利用Skip-gram模型在大规模数据集上训练,得到能够准确表示词语语义的词向量。

初始事件和微博表示模型:首先,对微博语料进行预处理;然后,定义特征权重计算公式,通过提取关键词建立初始事件和微博表示模型。

相似度计算与事件追踪:利用词向量计算微博和初始事件之间的语义相似度,并以此来衡量微博与事件是否相关,若相似度大于阈值,则可判定为二者相关;否则不相关。

3.2　训练词向量

目前对词向量的研究集中在词聚类、同义词判断和词性分析等任务,主要是利用词向量对词语进行表示,然后计算词与词之间的语义相似度。例如使用词向量计算"北京"、"首都"、"贪官"和"腐败"这4个词语两两之间的相似度时,"北京"与"首都"、"贪官"与"腐败"的相似度明显高于其他组合。

Skip-gram模型(http://code.google.com/p/word2vec)是Mikolov等[3]提出的一种基于神经网络的语言模型,可以快速完成对数十亿词的大规模训练,得到的词向量引入了丰富的语义信息,对词语的表示更加准确。Skip-gram模型存在一个基本假设:相似的单词拥有相似的语境。换言之,特定的语境只有特定的语义才能够与之匹配。通过最大化条件概率,使得单词和语境之间的对应关系最大化,进而满足了基本假设。而满足条件概率最大化的单词向量,也就成为了单词语义的合理表示。Skip-gram模型中的每个词向量表征了上下文的分布,其中Skip是指在一定窗口内的词两两都会计算概率,即使它们之间隔着一些词,例如"白色汽车"和"白色的汽车"很容易被识别为相同的短语。本文使用Skip-gram模型在中文数据集上进行训练得到词向量,如图2所示。

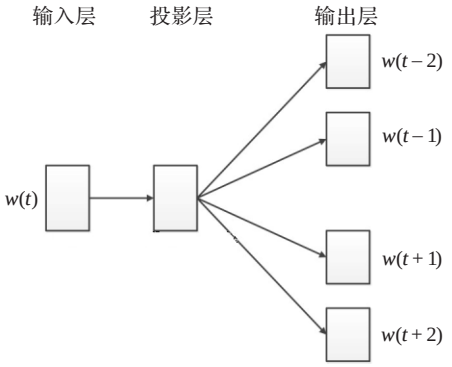

图2　Skip-gram模型

Skip-gram模型包含输入层、投影层和输出层,其原理是假设存在一个 $w_1, w_2, \cdots, w_T$ 的词组序列,在已知当前词 w_i 的前提下预测其上下文,目标函数如下所示:

$$G=\sum \log p(Content(w_i)|w_i) \tag{1}$$

$$p(Context(w_i)|w_i)=\prod_{u\in Context(w_i)} p(u|w_i) \tag{2}$$

其中,$Context(w_i)$ 表示与词 w_i 距离小于 R(R 一般取5到10效果较好)的上下文。Skip-gram模型中使用哈弗曼树(Huffman Tree)对输出层的结果进行表示,树的每一个叶子节点对应一个词,每个非叶子节点包含选择其左右子节点的相对概率,每个词 w_i 都可以通过一条从根节点出发的路径被访问到:

$$p(u|w_i;\theta)=\prod_{j=2}^{L(u)} p(d_j^u|\boldsymbol{v}(w_i),\theta_{j-1}^u) \tag{3}$$

$$p(d_j^u|\boldsymbol{v}(w_i),\theta_{j-1}^u)=[\sigma(\boldsymbol{v}(w_i)^{\mathrm{T}}\theta_{j-1}^u)]^{1-d_j^u}\cdot H[1-\sigma(\boldsymbol{v}(w_i)^{\mathrm{T}}\theta_{j-1}^u)]^{d_j^u} \tag{4}$$

$$\sigma(\boldsymbol{v}(w_i)^{\mathrm{T}}\theta)=\frac{1}{1+\mathrm{e}^{-\boldsymbol{v}(w_i)\theta}} \tag{5}$$

其中,$L(w_i)$ 表示从根节点到词 w_i 的路径长度,$d_j^{w_i}$ ($d_j^{w_i}\in\{0,1\}$) 表示路径中第 j 个节点对应的编码,$\boldsymbol{v}(w_i)$ 表示词 w_i 的词向量,$\theta_j^{w_i}$ 表示路径中第 j 个非叶子节点对应的向量维度,$\sigma(\boldsymbol{v}(w)^{\mathrm{T}}\theta_j^{w_i})$ 表示在第 j 个非叶子节点选择其右边子节点的概率,$1-\sigma(\boldsymbol{v}(w_i)^{\mathrm{T}}\theta_j^{w_i})$ 即为选择其左边子节点的概率。

使用Skip-gram模型训练得到的词向量除了引入语义信息计算词与词之间的相似度之外,一定程度上还满足加法组合运算(Additive Compositionality),在中文数据上进行测试,存在如下的关系:

$$vector(\text{“中国”})-vector(\text{“北京”})+vector(\text{“华盛顿”})\approx vector(\text{“美国”}) \quad (6)$$

$$vector(\text{“中国”})+vector(\text{“男篮”})\approx vector(\text{“姚明”}) \quad (7)$$

3.3 初始事件和微博表示模型

本节通过提取关键词来建立初始事件和微博表示模型,关键词是关于事件的重要特征描述,既能反映事件的主要内容,又能将不同的事件区分开来。

3.3.1 初始事件表示模型

为了避免微博“短文本”造成的特征稀疏问题,准确得到能够描述该事件的关键词,本节将事件包含的初始微博合并成一个文本 d_{event} ,然后按照特征权重提取文本 d_{event} 的关键词,具体计算方法如下:

$$W(c_i, d_{\text{event}})=\frac{tf(c_i, d_{\text{event}})\times c_{i,\text{pos}}}{\sqrt{\sum_{c\in d_{\text{event}}}[tf(c, d_{\text{event}})\times c_{i,\text{pos}}]^2}} \quad (8)$$

其中,$W(c_i, d_{\text{event}})$ 表示词语 c_i 在 d_{event} 中的权重,$tf(c_i, d_{\text{event}})$ 表示词语 c_i 在 d_{event} 中的词频,$c_{i,\text{pos}}$ 表示词性标注加权值,词性标注加权将特征按词性进行加权,具体定义如下:

$$c_{i,\text{pos}}=\begin{cases}P, \text{“v” or “n”}\\ Q, \text{otherwise}\end{cases} \quad (9)$$

其中,当 c_i 为名词或者动词的时候权重取 P ,当 c_i 为其他词性时权重取 Q 。由于虚词没有意义,而实词中的名词和动词可以很好地反映微博的主要内容,因此赋予 P 的值较大,Q 的值较小,分别为1.5和1。

通过计算得到初始事件的关键词后,建立初始事件表示模型,具体表示形式为:

$$\boldsymbol{V}_{\text{event}}=\{c_1, c_2, \cdots, c_i, \cdots, c_L\} \quad (10)$$

其中,L 表示该初始事件包含的关键词个数,c_i 表示第 i 个关键词。

3.3.2 微博表示模型

对于微博数据 $D=\{d_1, d_2, \cdots, d_m, \cdots, d_{|D|}\}$,本节按特征权重提取微博 d_m 中的关键词,具体计算方法如下:

$$W(w_j, d_m)=\frac{tf(w_j, d_m)\times \lg(N/n_{w_j}+0.01)\times w_{j,\text{pos}}}{\sqrt{\sum_{w\in d_m}[tf(w, d_m)\times \lg(N/n_w+0.01)\times w_{j,\text{pos}}]^2}} \quad (11)$$

其中,$W(w_j, d_m)$ 表示词语 w_j 在微博 d_m 中的权重,$tf(w_j, d_m)$ 表示词语 w_j 在微博 d_m 中的词频,N 是数据集中微博总数,n_{w_j} 表示出现词语 w_j 的微博个数,$w_{j,\text{pos}}$ 表示词性标注加权值,设置方法参照3.3.1节。

通过计算得到微博 d_m 的关键词后,建立微博表示模型,具体表示形式为:

$$\boldsymbol{V}_m=\{w_1, w_2, \cdots, w_j, \cdots, w_M\} \quad (12)$$

其中,M 表示该微博的关键词个数,w_j 表示微博 d_m 中第 j 个关键词。

3.4 相似度计算与事件追踪

本节以 $Sim(\boldsymbol{V}_{\text{event}}, \boldsymbol{V}_{\text{new}})$ 表示新的微博和初始事件之间的语义相似度,下面给出利用词向量计算 $Sim(\boldsymbol{V}_{\text{event}}, \boldsymbol{V}_{\text{new}})$ 的具体方法。

在已知 $\boldsymbol{V}_{\text{event}}$ 的前提下,由于新的微博长短不一,导致两个向量中词语数目不同 $(L\neq M)$ 。为此,本节采用以下方法计算 $Sim(\boldsymbol{V}_{\text{event}}, \boldsymbol{V}_{\text{new}})$ 。

当 $L\leqslant M$ 时,按照下两式计算二者的语义相似度:

$$Sim(\boldsymbol{V}_{\text{event}}, \boldsymbol{V}_{\text{new}})=\sum_{i=1}^{L}MSim_{ij}\cdot W(c_i, d_{\text{event}})\cdot W(w_j, d_{\text{new}}), j\in[1, M] \quad (13)$$

$$MSim_{ij}=\max\{MSim(\boldsymbol{v}(c_i), \boldsymbol{v}(w_1)), MSim(\boldsymbol{v}(c_i), \boldsymbol{v}(w_2)), \cdots, MSim(\boldsymbol{v}(c_i), \boldsymbol{v}(w_j)), \cdots, MSim(\boldsymbol{v}(c_i), \boldsymbol{v}(w_M))\} \quad (14)$$

其中,$W(c_i, d_{\text{event}})$ 和 $W(w_j, d_{\text{new}})$ 分别表示初始事件 $\boldsymbol{V}_{\text{event}}$ 和微博 $\boldsymbol{V}_{\text{new}}$ 中关键词 c_i 和 w_j 的权重,$MSim_{ij}$ 表示关键词 c_i 和 w_j 具有最大的语义相似度,$MSim(\boldsymbol{v}(c_i), \boldsymbol{v}(w_j))$ 表示两个词的语义相似度,$\boldsymbol{v}(c_i)$ 和 $\boldsymbol{v}(w_j)$ 分别表示关键词 c_i 和 w_j 的词向量,本节按照余弦相似度的计算公式计算两个词语之间的语义相似度,即 $MSim(\boldsymbol{v}(c_i), \boldsymbol{v}(w_j))=\frac{\boldsymbol{v}(c_i)\cdot\boldsymbol{v}(w_j)}{|\boldsymbol{v}(c_i)||\boldsymbol{v}(w_j)|}$ 。该计算方法适合长度较长的微博,且没有归一化,提高了重要关键词对事件区分的作用。

当 $L>M$ 时,对于 $\boldsymbol{V}_{\text{new}}$ 中每个关键词,分别计算它与 $\boldsymbol{V}_{\text{event}}$ 中关键词的语义相似度并取最大值 $MSim_{ij}$ $(i\in[1, L])$:

$$MSim_{ij}=\max\{MSim(\boldsymbol{v}(c_1), \boldsymbol{v}(w_j)), MSim(\boldsymbol{v}(c_2), \boldsymbol{v}(w_j)), \cdots, MSim(\boldsymbol{v}(c_i), \boldsymbol{v}(w_j)), \cdots, MSim(\boldsymbol{v}(c_L), \boldsymbol{v}(w_j))\} \quad (15)$$

然后,统计满足 $MSim_{ij}\geqslant 0.01$ 的个数 T ,并按下式计算两个向量集合之间的语义相似度:

$$Sim(\boldsymbol{V}_{\text{event}}, \boldsymbol{V}_{\text{new}})=\frac{\sum_{k=1}^{T}(MSim_{ij})_k\cdot W(c_i, d_{\text{event}})\cdot W(w_j, d_{\text{new}})}{T} \quad (16)$$

该计算方法适合长度较短的微博,特征非常稀疏,通常只有少量的关键词相似。因此只对 $MSim_{ij}\geqslant 0.01$ (此时本文认为两个词语的语义很接近)的关键词进行语义相似度计算,且进行了归一化,避免了无关特征削弱关键词对事件区分的作用。

将微博与初始事件之间的语义相似度作为微博与

初始事件的相关程度,即 $Sim(\boldsymbol{V}_{event}, \boldsymbol{V}_{new})$ 是衡量微博是否与事件相关的标准,若相似度大于判决阈值,则可判定该微博与事件相关,从而完成事件追踪。由于可能存在微博 $\boldsymbol{V}_{new}$ 同时与几个初始事件的相似度都大于阈值的情况,影响追踪性能,因此,在满足大于判决阈值 γ 的基础上,选择相似度最大的一组,并将微博判决到该初始事件中,具体判决方式为:

$$\begin{cases}\max\{Sim(\boldsymbol{V}_{event}, \boldsymbol{V}_{new})\geqslant\gamma\}, \text{相关事件}\\ \text{其他}, \text{不相关事件}\end{cases} \tag{17}$$

4　实验与分析

4.1　实验数据及评价方法

实验语料由上海交通大学信息内容分析国家工程实验室提供(ftp://ftp.socialysis.org/pub/SJTU/dataset/microblog/microblog.rar)。从该微博语料中选择不同领域的10个事件作为实验数据,共包括2 326条事件相关微博(如表1所示)。此外,再从该微博语料中随机选取49 471条数据作为反例。

表1　实验数据

序号	事件名称	数量
1	中国好外婆	108
2	北京楼顶别墅主人	359
3	斯诺登棱镜事件	131
4	郑钧刘芸马尔代夫婚礼	132
5	七夕,一起看英仙座流星雨	233
6	光大证券乌龙事件	480
7	埃及部队清场事件	153
8	林丹vs李宗伟	349
9	泰国白龙王病逝	157
10	扮丑大赛	224

本文利用准确率、召回率和 $F1$ 值对微博事件追踪方法的性能进行评价。对于多事件追踪任务,实验采用加权平均值评价综合性能,计算方法如下:

准确率:

$$P_{avg}=\frac{\sum_i|e_i|\times P(e_i)}{\sum_i|e_i|}\times 100\% \tag{18}$$

召回率:

$$R_{avg}=\frac{\sum_i|e_i|\times R(e_i)}{\sum_i|e_i|}\times 100\% \tag{19}$$

$$F1_{avg}=\frac{2\times P_{avg}\times R_{avg}}{P_{avg}+R_{avg}} \tag{20}$$

其中,$|e_i|$ 表示第 i 个事件的微博个数,$P(e_i)$ 和 $R(e_i)$ 分别表示第 i 个事件的准确率和召回率。

4.2　实验设置与结果分析

为了验证本文方法的有效性,分别从以下三个方面对本文方法进行综合评价,包括不同方法的综合性能对比、不同初始数据量的性能对比和不同相似度判决阈值的性能对比。

(1)不同方法的综合性能对比实验

实验选取每个事件中最早出现的5条微博作为初始训练数据,将其余微博作为后续实验数据,并实现三种方法进行对比实验,包括:

①SVM方法:利用TF-IDF方法分别将初始训练数据和反例数据表示成词语向量,训练SVM分类器;然后利用训练好的分类器对后续实验数据进行分类,完成事件追踪。

②TF-IDF方法:利用TF-IDF方法将初始训练数据表示成词语向量,并通过计算余弦相似度值判断新的微博所属事件。

③PTCG方法:实现文献[12]提出的PTCG方法,利用中文维基语义图实现语义相似度计算,完成事件追踪。

本文方法中利用Skip-gram模型训练得到词向量,训练数据来源于"搜狗实验室"的全网新闻数据(http://www.sogou.com/labs/dl/ca.html),包含3.79亿个词。在总体性能对比实验中,三种对比方法中的阈值参数均取最优值(最优值的选取方法参照文献[12]),其中,TF-IDF相似度判决阈值取0.3,PTCG相似度判决阈值取0.35,本文方法相似度判决阈值取0.05,初始事件关键词个数 L 和微博关键词个数 M 分别取8个和3个效果较好,结果如表2所示。

表2　微博事件追踪总体性能对比结果　%

方法名称	准确率	召回率	$F1$值
SVM	66.60	84.64	74.54
TF-IDF	97.57	68.92	80.78
PTCG	80.65	66.78	73.06
本文方法	84.98	86.45	85.71

由表2看出,SVM方法的 $F1$ 值仅为74.54%,性能较差,是因为训练数据较少时,该方法难以训练出较好的分类模型,影响追踪性能。TF-IDF方法的 $F1$ 值为80.78%,高于SVM方法,是因为微博短文本的特性,对事件的描述往往更加直接,因此取得了较好的性能。但以上两种方法均未引入语义信息,追踪性能受限。PTCG方法虽引入维基百科语义信息,但需要限制指定领域的特征词,使得其不适用于跨领域的多事件追踪任务,导致在对比实验中综合性能最差。本文方法得到了最高的 $F1$ 值85.71%,综合性能明显优于几种对比方法。通过对比实验,本文方法能够充分利用词向量引入的丰富的语义信息,追踪性能达到最优。

(2)不同初始数据量的性能对比实验

微博事件追踪任务中,经常选取最早出现的少量报道作为事件的初始数据。为了验证不同方法对初始数

据量的要求各不相同,分别利用其他两种方法在不同初始数据量上进行对比实验(由于PTCG不适用于多事件追踪任务,实验不再考虑PTCG方法)。三种方法的参数均按照实验(1)设置,结果如图3所示。

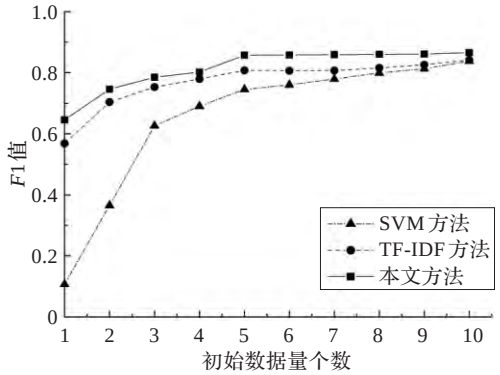

图3 不同初始数据量对 *F*1性能的影响

由图3看出,SVM方法对初始数据量的依赖性较大,当初始数据增多时性能得到较大提升。TF-IDF方法的 *F*1值随着初始数据的增多而缓慢提高。本文方法利用词向量表示词语,一定程度上丰富了语义信息,在初始数据量小的情况下依然具有较高的 *F*1值,并且随着初始数据量的增大该方法仍保持优越的性能。

(3)不同相似度判决阈值的对比实验

相似度判决阈值的设定直接影响事件追踪的性能,为了得到合理的判决阈值,利用本文方法分别在不同相似度判决阈值下进行实验,结果如图4所示(初始数据量为5)。

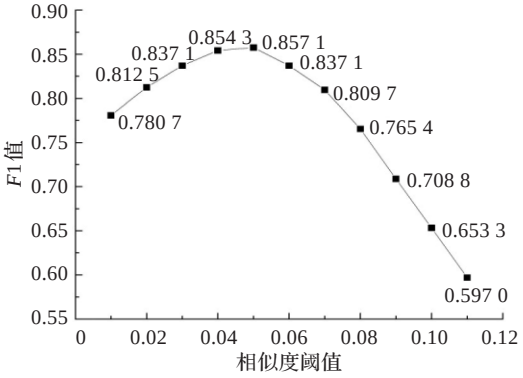

图4 不同相似度判决阈值时 *F*1性能变化

由图4看出,*F*1值随着判决阈值的增加呈现凸抛物线,当相似度判决阈值为0.05时 *F*1值达到最优值;相似度判决阈值在0.04~0.06时,*F*1值较为平衡,获得较好的性能。

5 总结

微博文本长度短,且网络新词层出不穷,使得传统方法在微博事件追踪中效果不够理想。针对该问题,提出一种基于词向量的微博事件追踪方法。词向量不仅可以计算词语之间的语义相似度,而且能够提高微博间语义相似度计算的准确率。该方法首先使用Skip-gram模型在中文数据集上训练得到词向量;然后通过提取关键词建立初始事件和微博表示模型;最后利用词向量计算微博和初始事件之间的语义相似度,并依据设定阈值进行判决,完成事件追踪。实验结果表明,相比传统方法,该方法能够充分利用词向量引入的语义信息,有效提高微博事件追踪的性能。

需要指出的是,微博中新词层出不穷,口语化普遍,使用Skip-gram模型在训练词向量时,会因为数据规模不够庞大,覆盖面不够广泛,影响追踪性能。因此,在实时更新的大规模数据集上训练词向量,提高微博事件的追踪性能,是下一步工作的重点。

参考文献:

[1] 丁兆云,贾焰,周斌.微博数据挖掘研究综述[J].计算机研究与发展,2014,51(4):691-706.

[2] Aiello L M,Petkos G,Martin C,et al.Sensing trending topics in twitter[J].IEEE Transactions on Multimedia,2013,15(6):1268-1282.

[3] Mikolov T,Chen K,Corrado G,et al.Efficient estimation of word representations in vector space[C]//International Conference on Learning Representations,2013.

[4] Ma Nianli,Yang Yiming,Rogati M.Applying CLIR techniques to event tracking[M]//Information retrieval technology.Berlin/Heidelberg:Springer,2005:24-35.

[5] Yang Yiming,Ault T,Pierce T,et al.Improving text categorization methods for event tracking[C]//Proceedings of the 23rd International Conference on Research and Development in Information Retrieval,2000:65-72.

[6] 潘渊,李弼程,张先飞.LS-SVM:一种有效的新闻主题追踪方法[J].计算机应用研究,2008,25(9):2661-2663.

[7] Li Shuping,Zhao Jie,Song Zhichao,et al.Study on topic tracking system based on SVM[C]//Proceedings of 2011 Fourth International Symposium on Knowledge Acquisition and Modeling,2011:83-87.

[8] Zhao Qiankun,Mitra P,Chen Bi.Temporal and information flow based event detection from social text streams[C]//Proceedings of AAAI,2007,7:1501-1506.

[9] Yang C,Shi Xiaodong,Wei C.Discovering event evolution graphs from news corpora[J].IEEE Transactions on Systems,Man and Cybernetics,Part A:Systems and Humans,2009,39(4):850-863.

[10] Kwak H,Lee C,Park H,et al.What is twitter,a social network or a news media?[C]//Proceedings of the 19th International Conference on World Wide Web,2010:591-600.

(下转117页)

6　结束语

本文根据MIMO-OFDM信道特性,提出了一种基于压缩感知改进的新型匹配追踪算法,并将其应用到MIMO-OFDM信道估计中。提出的算法无需预知信号的真实稀疏度,自适应地完成稀疏信号的重构。仿真结果表明,改进的算法有效地降低了信道估计中传统压缩采样匹配追踪算法的复杂度,并提高了抗噪性能和信道估计精度。

参考文献:

[1] Zhou Y, Wang J, Sawahashi M.Downlink transmission of broadband OFCDM systems—part I:hybrid detection[J]. IEEE Transactions on Communications, 2005, 53(4): 718-729.

[2] Barhumi I, Leus G, Moonen M.Optimal training design for MIMO OFDM systems in mobile wireless channels[J]. IEEE Transactions on Signal Processing, 2003, 51(6): 1615-1624.

[3] Maaref A, Aissa S.Impact of spatial fading correlation and keyhole on the capacity of MIMO systems with transmitter and receiver CSI[J].IEEE Transactions on Wireless Communications,2008,7(8):3218-3229.

[4] Li Weichang, Preisig J C.Estimation of rapidly time-varying sparse channel[J].IEEE J Ocean Eng,2007,32(4):927-939.

[5] Donoho D L.Compressed sensing[J].IEEE Transactions on Information Theory,2006,52(4):1289-1306.

[6] Candes E J.Compressive sampling[C]//Proceedings of the International Congress of Mathematics.Madrid, Spain: the European Mathematical Society,2006:1433-1452.

[7] 石光明,刘丹华,高大化,等.压缩感知理论及其研究进展[J].电子学报,2009,37(5):1070-1081.

[8] 方红,杨海蓉.贪婪算法与压缩感知理论[J].自动化学报,2011,37(12):1412-1421.

[9] 何雪云,宋荣方,周克琴.基于压缩感知的OFDM系统稀疏信道估计新方法研究[J].南京邮电大学学报:自然科学版,2010,30(2):60-65.

[10] Qi Chenhao, Wu Lenan.A hybrid compressing sensing algorithm for sparse channel estimation in MIMO OFDM systems[J].IEEE Transactions on Signal Processing, 2011, 58(1):3488-3491.

[11] 王妮娜,桂冠,苏冰涛,等.基于压缩感知的MIMO-OFDM系统稀疏信道估计方法[J].电子科技大学学报,2013,42(1):59-62.

[12] Bajwa W U, Haupt J, Sayeed A M, et al.Compressed channel sensing: a new approach to estimating sparse multipath channels[J].Proceedings of the IEEE,2010,98(6):1058-1076.

[13] Baraniuk R G.Compressive sensing[J].IEEE Signal Processing Magazine,2007,24(4):118-120.

[14] Cai T, Wang L, Xu G W.New bounds for restricted isometry constants[J].IEEE Transactions on Information Theory, 2010,56(9):4388-4394.

[15] Needell D, Tropp J A.CoSaMP: iterative signal recovery form incomplete and inaccurate samples[J].Applied and Computational Harmonic Analysis, 2008, 26(3): 301-321.

[16] Blumensath T, Davies M E.Stagewise weak gradient pursuits[J].IEEE Transactions on Signal Processing, 2009, 57(11):4333-4346.

(上接78页)

[11] Choi J,Croft W B.Temporal models for microblogs[C]// Proceedings of the 21st ACM International Conference on Information and Knowledge Management,2012:2491-2494.

[12] 王兰成,刘晓亮.整合中文维基语义的网络论坛话题追踪方法研究[J].情报学报,2013,32(1):22-27.

[13] Zhang Xujie, Liu Zongtian, Liu Wei, et al.Event similarity computation in text[C]//2011 IEEE International Conferences on Internet of Things, and Cyber, Physical and Social Computing,2011:419-423.

[14] Li Xiaoxian, Li Weijie, Liu Yun.Research based on the HowNet semantic expansion micro-blog hot topic detection system[J].Advanced Materials Research, 2013, 765: 1502-1506.

[15] Medelyan O, Milne D, Legg C, et al.Mining meaning from Wikipedia[J].International Journal of Human-Computer Studies,2009,67(9):716-754.

[16] Mikolov T, Sutskever I, Chen K, et al.Distributed representations of words and phrases and their compositionality[C]//Neural Information Processing Systems Foundation,2013.

[17] 席耀一,林琛,李弼程,等.基于语义相似度的论坛话题追踪方法[J].计算机应用,2011,31(1):94-96.

[18] Chen Xin, Zhang Yuqing, Cao Long, et al.An improved feature selection method for Chinese short texts clustering based on HowNet[M]//Computer engineering and networking.[S.l.]: Springer International Publishing, 2014: 635-642.