**热门事件“影响”态势感知**

李彦江1，孙国梓1，成楚凡2，吉原2

（1.南京邮电大学 计算机学院软件学院网络空间安全学院，南京 210046；2. 南京邮电大学 计算机学院软件学院网络空间安全学院，南京 210046）

摘 要：现今时期，互联网与“互联网+应用”已使得社会各个领域都发生了相应的转变。在这其中，微博是最为典型也是应用最为广泛的新型媒体，其已成为一种基于用户通讯的信息同享和互换、广受用户接待的主流平台。与此同时，对微博平台的海量数据的处理也考验着从事科研的人员分析和筛选信息的能力。通过爬取2019年3-4月份网络新闻数据，在对文本聚类前采用了文本分类方法，由此可得到不同类别的网络新闻。通过Word2vec与LDA结合建模以检测话题模型，并且使用Word2vec与卷积神经网络（CNN）联合进行新闻热点话题的检测。除此之外，通过话题热度、话题指数公式对热点话题的趋势进行研究与描绘。通过采用联合建模的方式，可以从大量的网络新闻中检测出热点话题，对所检测出的话题的趋势进行分析，这对个人、公司、社会等都要较强的实用价值。

**关键词：**热点话题检测; 趋势分析; Word2vec;LDA; CNN; 联合建模

**Popular events "affect" situational awareness**

Li Yan-Jiang1, Sun Guo-Zi1, Cheng chu-fan2, Ji Yuan2

(1.School of Computer Software and Cyberspace Security, NanJing University of Posts and Telecommunications, Nanjing 210046, China;

2. School of Computer Software and Cyberspace Security, NanJing University of Posts and Telecommunications, Nanjing 210046, China)

**Abstract**: In the modern era, the Internet and "Internet + applications" have caused corresponding changes in all areas of society. Among them, Weibo is the most typical and most widely used new media, and it has become a mainstream platform for users to share and exchange information based on user communication. At the same time, the processing of massive data on the Weibo platform also tests the ability of researchers engaged in the analysis and screening of information. By crawling the online news data from April to April of 2019, a text categorization method was adopted before clustering the text, thereby obtaining different types of online news. Word2vec is combined with LDA to model the topic model, and Word2vec is used in conjunction with Convolutional Neural Network (CNN) to detect news hot topics. In addition, through the topic heat, topic index formula to study and depict the trend of hot topics. By adopting the joint modeling method, hot topics can be detected from a large number of online news, and the trend of the detected topics can be analyzed, which has strong practical value for individuals, companies, and society.

**Keywords**: hot topic detection; trend analysis; Word2vec; LDA; CNN; joint modeling

**0 引言**

互联网与“互联网+”已经成为大多数人获得信息的不可或缺的媒介。与此同时，互联网的迅速发展提高了众多网络媒体的影响力。交际网络服务（Social Network Service, SNS）致力于分享信息以及推进信息的发布和流传。当下，微博已成长为集信息公布、交际、及时通信等功能于一体的综合平台。紧接着大数据潮流的涌入，如何让人们在海量的数据中进行数据的分析、获取热点话题、提取出对自己有用的信息已成为人们日益重视的话题。主要原因是随着网络新闻数量的增加，新闻数量过大，除此之外，也有很多虚假和无用的新闻。对于话题的检测与趋势分析可以帮助人们从大量的新闻中获取新闻热点，从而给用户浏览信息带来了极大的便利。随着人工智能（AI）的到来，机器学习（machine learning）、自然语言处理（NLP）应用到了网络新闻话题检测中，并取得了良好的成绩。微博用户急剧增加，网民通过微博参与话题互动、以及对相关话题的讨论度也在不断上升。本文研究目标采用数据挖掘和数据分析算法从大量的网络新闻中找出热点话题，并用相关算法对新闻数据进行分类，进一步对分类好的数据进行建模，得到各个热门话题的热度随时间变化的趋势，具有重要的研究意义。

**1 国内外研究现状**

微博热门话题检测研究的目标在国内与国外基本一致，二者都是基于150或160个字左右的短文本的社交网络平台。比如：Twitter和微博，可以从两个方面分析其具有的相同特性。数据方面：处理对象均是短文本，包含的信息量较少且数据量较大。数据具有一定的时效性并且包括众多社会关系。研究方面：短文本的内容和结构都被海内外研究者们所广泛关注。但是，国外对于热门话题的钻研与国内相比，比国内更早、更成熟。除此之外，国内外微博在文化背景、发展历程、基本国情等方面差别较大，再加上中文与英文在文本方面的差距，对研究人员从事相应的研究增加了分析的难度。

**2 关键技术**

**2.1 微博文本特征**

微博的“微”体现在其规定：每条信息不超过150或160个字。因而可知，微博信息的显著特点即“微”，即短文本。微博文本之间的共同信息较少。从文本处理的角度分析，微博中提取的关键词缺少且相互独立，对于处理短文本而言，传统的处理方法：统计并不适用。

除此之外，微博具备社交网络的特点，微博网络中的信息传播迅速，时效性较强，且会随着事件的发展而演变发展。因此，在对微博内容进行分析时除了要注意时效之外，还要保持较低的时间复杂度、保证分析的准确性。

**2.2 文本表示模型**

常见的文本表示模型有语言模型和向量空间模型。[1]除此之外，另有基于LDA的文本模型和基于Word2vec的文本模型。

**2.2.1基于LDA的文本表示模型**

LDA即Latent Dirichlet allocation，隐含狄利克雷散布，用于生成文档主题模型。LDA可以用来识别文档集以及隐藏在语料库中的主题信息[2]。LDA可将文档转化为向量，给出每一个文档的主题几率散布。采用python的机器学习库——gensim的LDAmodel模块对LDA模型进行相应的训练，该模块使用频率较高的参数见表1：

表1 gensim LDAmodel模块常用参数

|  |  |
| --- | --- |
| alpha | 影响文档主题分布的稀疏性超参数 |
| Eta | 影响文档主题词分布的稀疏行超参数 |
| corpus | 训练语料 |
| id2word | 从单词id到字符串的映射 |
| num\_topics | 从训练语料库中提取要求的潜在主题的数量 |

本文截取的某一个文档的主题概率表见表2：

表2 主题概率

|  |  |
| --- | --- |
| 主题 | 概率 |
| 13 | 0.094383 |
| 15 | 0.027599 |
| 18 | 0.014091 |
| 34 | 0.066292 |
| 38 | 0.089933 |

由此散布表可知：由文档主题几率矩阵，可将文档向量化。因而，求出文档类似度可采用文档向量化的方式。

**2.2.2基于Word2vec的文本表示模型**

Word2vec是将词转化为实数值向量的一种高效的算法模型，简而言之，Word2vec通过采用神经网络训练的方法将每个词映射到一个向量上。其词语代表效果较好，因此目前已被广泛应用到NLP（Neural Language Network）上。Word2vec包含两种模型：CBOW和Skip-gram。CBOW是按照语境来推测当前词语的频率，Skip-gram则是按照当前词语频率推测语境，恰好与之相反。CBOW和Skip-gram如图1所示：

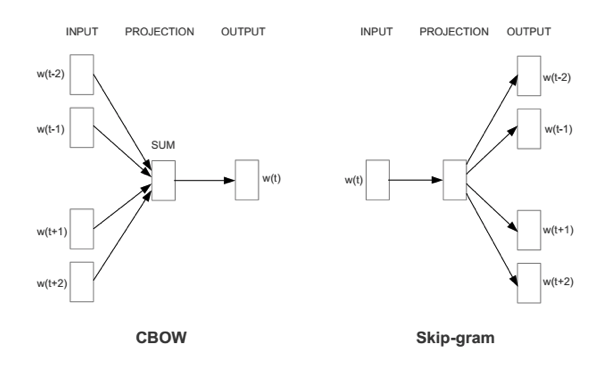


图1 CBOW & Skip-gram模型

通过Word2vec将所爬取的文本训练出词向量，把词向量转化为句向量，再计算文本相似度。式(1)为文档的向量化公式：

本文采用gensim框架训练的word2vec模块，具体参数如表3所示：

表3 gensim word2vec模块常用参数

|  |  |
| --- | --- |
| sentences | 指的是输入的训练语料 |
| Sg | 定义训练算法 |
| Size | 词向量的训练维度 |
| alpha | 初识的学习速率 |
| window | 句子中当前词与预测词之间的最大距离 |
| min\_count | 忽略总频率低于此的所有单词 |
| sample | 配置哪些较高频率的字随机下采样的阈值 |
| workers | 使用多少个工作线程来训练模型 |
| Hs | 分层softmax将用于模型训练、使用负采样 |
| negative | 将采用负例，负例采用int表示采用多少嗓音字 |

在训练word2vec过程中，本文采用的是CBOW模型，sg=1并且hs=0，设置negative = 5，window = 5，alpha = 0.015，min\_count = 2，词向量的维度确定为200。

为了体现本文训练词向量的效果，选取4个词进行测试，利用word2vec计算与“特朗普”、“湖人队”、“贾跃亭”、“双11”相近的词语并计算相似度，采用余弦距离计算关联词的度量：假设文档X，Y文档向量化之后分别为向量X（x1,x2,x3,…,xn），Y（y1,y2,y3,…,yn），采用余弦距离计算两个文档之间的距离，则X、Y的余弦距离为式(2)：（其中表示X与Y的内积）[3]

|  |  |  |
| --- | --- | --- |
|  |  |  |

表4 word2vec关联词

|  |  |  |  |
| --- | --- | --- | --- |
| 特朗普 | 湖人队 | 贾跃亭 | 双11 |
| 华尔街日报0.998 | 球星 0.998 | 董事 0.913 | 图文0.910 |
| 总统 0.994 | 科尔 0.998 | 仲裁 0.965 | 天猫0.943 |
| 卫报 0.993 | 贝纳蒂亚 0.998 | 任命 0.922 | 购物0.936 |
| 新潮媒体 0.991 | 海沃德 0.997 | 员工 0.923 | 直击0.917 |
| 民主党人 0.660 | 莱昂纳德 0.997 | FF 0.979 | 专题0.913 |

由表4可知，词向量的训练效果整体不错，选取的测试词语的关联词基本是与其相似度较高的词。与其相似的词越多，则此条新闻的热度越强、影响力越大。

**2.2.3基于LDA与Word2vec联合建模的文本相似度算法**

由上文可知，文本之间的类似度可以被向量之间的类似度所代替，两个文本之间的类似度能够影响两个文本蕴含相同话题的可能性大小。本文采取两种建模方式对文本向量化，这两种建模方式有着各自的优缺点。Word2vec的长处是采用神经网络练习词向量，练习较高效。缺点是转化成句向量后可能会丧失一定的准确度。LDA在话题检测方面是目前较为主流的算法，但它的缺点是需要明确主题的个数。现实文本中的话题数与预设的主题数一般会存在一定程度的偏差，从而导致话题检测效果不佳。因而，采用联合两种向量化文本方式用于计算文本类似度。

采用word2vec建模方式的计算文本相似度公式为式(3)：（）

|  |  |  |
| --- | --- | --- |
|  |  |  |

采用LDA建模方式的文本相似度公式为式(4)：

（为文档X采用LDA向量化后的向量）

|  |  |  |
| --- | --- | --- |
|  |  |  |

设加权因子为，采用word2vec与LDA线性组合建模方式的文本相似度公式为式(5)：

|  |  |  |
| --- | --- | --- |
|  | 1- |  |

**2.3新闻热点话题的检测**

**2.3.1卷积神经网络CNN**

卷积神经网络（Convolutional Neural Networks, CNN）是深度学习（deep learning）算法中的典范之一，可用于文本分类。CNN具有卷积层和池化层是其与一般神经网络的显著不同之处。对于每层的构造而言，将局部感觉的范围作为输入，从卷积核中提取明显特性。

**2.3.2 Word2vec+CNN的文本分类模型**

**2.3.2.1 模型实验过程**

为了验证Word2vec与CNN联合文本分类的有效性，从科技、军事、国际、国内、娱乐、体育、财经、社会，八类话题中分别随机选取5000条共计40000条新闻文本作为训练集和测试集。按照相应的比例将数据集区分为练习集和测试集。采用TensorFlow用于对词向量与CNN模型的组合进行搭建。

**2.3.2.2 模型的结果**

将爬取下来的八类话题数据按照Word2vec+CNN联合分类器的实验结果如表5所示：

表5 分类结果

|  |  |  |
| --- | --- | --- |
| 序号 | 类别 | 文档数 |
| 1 | 国内 | 6500 |
| 2 | 国际 | 6700 |
| 3 | 科技 | 11000 |
| 4 | 体育 | 10500 |
| 5 | 娱乐 | 7000 |
| 6 | 军事 | 7100 |
| 7 | 社会 | 11500 |
| 8 | 财经 | 11700 |

**3 热点话题热度度量与趋势分析**

**3.1 话题的热度度量**

采用如下式(6)计算每日话题的热度，其中ni为话题i包含的文档数，N为总文档数。此公式也适用于更大时间窗口的热度计算。

|  |  |  |
| --- | --- | --- |
|  | H(k)= | (6) |

**3.2 话题的趋势分析**

随着时间的流逝，话题不会持续存在，也会随时间淡去，通过使用如下式(7)求话题随时间的变化趋势：

|  |  |  |
| --- | --- | --- |
|  | Ei(k)= | **(7)** |

其中Hi(k)代表话题i在第k天的热度，H1(k)为话题k在第一天的热度。

**3.3 话题热点排序**

表6为2019年3月1号到2019年3月7号的热点话题热度排序：

表6热点话题热度排序

|  |  |
| --- | --- |
| 热点话题 | 热度 |
| 皇马、巴萨、足球、主场 | 0.012502249 |
| 胜、新浪、彩票、曼联 | 0.008921767 |
| C罗、梅西、大将、进球 | 0.006312186 |
| 决赛、争、负、冠 | 0.004318037 |
| 正式、加盟、中锋、首秀  头条、内容、审查、核实  直播、冲顶、答题、竞赛 | 0.003690315  0.003158454  0.252401752 |

本文选取4个话题趋势进行分析，具体如表7所示：

表7选取四个话题

|  |  |
| --- | --- |
| 话题1 | 皇马、巴萨、足球、主场 |
| 话题2 | 决赛、争、负、冠 |
| 话题3 | 头条、内容、审查、核实 |
| 话题4 | 胜、新浪、彩票、曼联 |

话题1一周的趋势如图2所示：

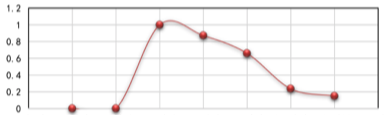


图2 话题1一周趋势

话题1在百度指数2019年3月1号到2019年3月7号的趋势如图3所示：

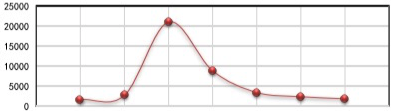


图3话题1百度指数一周趋势

对比图2和图3，两图的变化趋势大致相同，话题在3月3号达到最高点后呈现下降趋势。

话题2一周的趋势如图4所示：

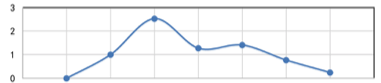


图4 话题2一周趋势

话题2在百度指数2019年3月1号到2019年3月7号的趋势如图5所示：

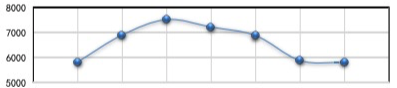


图5 话题2百度指数一周趋势

对比图4和图5，本文的话题2的热度趋势在3月1号达到峰值，随后呈下降趋势，与百度指数的趋势图变化一致。

话题3一周的趋势如图6所示：

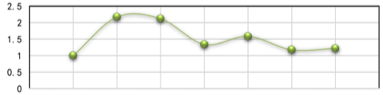


图6 话题3一周趋势

话题3在百度指数2019年3月1号到2019年3月7号的趋势如图7所示：



图7话题3百度指数一周趋势

对比图6和图7，本文实现的话题3在2号达到顶峰，2号到3号有轻微的下降，随后总体呈下降趋势，在5号小幅度的波动，两图结果大体一致。

话题4一周的趋势如图8所示：

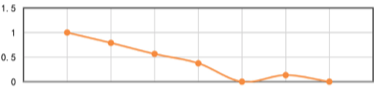


图8 话题4一周趋势

话题4在百度指数2019年3月1号到2019年3月7号的趋势如图9所示：

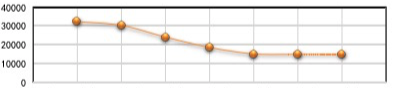


图9 话题4百度指数一周趋势

对比图8和图9，两图的话题热度变化趋势大致相同，在1号到达峰值，随后呈现下降趋势，在5号有轻微的波动。

通过选取的4个话题进行趋势分析并与百度指数所给的趋势图进行对比，可知本文所分析的热点话题趋势与权威搜索引擎基本一致，经过验证说明本论文分析的话题热度趋势的可行性和可靠性。从而可以推导出某话题在某时间段的趋势变化，对于舆情分析具有重大意义。

**4 总结**

总之，本文基于传统的话题检测与趋势分析的基础上，提出了一些改进的方法，并用实验结果证明是可行的。通过本文的网络新闻话题检测与趋势分析研究的方法，可从大量的网络新闻中获取不同类别的热点话题，并得到该话题在一段时间的热度以及其发展的趋势。本文所设计的研究方法与模型具有良好的应用价值，主要体现在能较快速、准确地获取自己关注的话题、可以帮助相关人士掌握社会舆情信息，关注其发展动态、了解同行业的最新信息，可以帮助相关企业实时做出决策。

参考文献：

[1] 尚鸿运.中文微博的热点话题检测及趋势预测算法研究[D].天津：天津大学

[2] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J].Journal of Machine Learning,2003(3):993-1022.

[3] 刘妍.基于Lucene的余弦距离检测文档相似度的研究[J].信息系统工程，2014(4)：128-129

作者简介：李彦江（1997-），男，南京邮电大学计算机学院、软件学院、网络空间安安全学院本科生，研究方向为自然语言处理与信息安全；孙国梓（1972-），男，博士，计算机学院、软件学院、网络空间安全学院教授，主要研究方向为网络空间安全、电子数据取证。

通讯信息：

通信作者：孙国梓 E-mail：[sun@njupt.edu.cn](mailto:sun@njupt.edu.cn)

通信地址：江苏省南京市栖霞区文苑路9号

邮编：210046