语义环境下的多维度微博舆情信息关联检测方法

庞海杰，刘春强
（青岛滨海学院 信息工程学院，山东 青岛 266555）

摘　要：针对微博舆情信息的特点，提出基于语义理解的微博舆情信息关联检测方法。从舆情信息表示模型和舆情信息相关度计算方法两个方面展开研究。在信息表示方面，使用微博的评论信息扩充微博信息以期较好地应对数据稀疏现象，基于同义词词林来计算词汇相似度，以应对微博草根性带来的问题，将微博舆情信息表示成多个向量空间模型。在相关性计算方面，提出多维度相关性计算方法。实验证明，所提出的方法对关联检测的准确率和召回率都有较好的提升。

关键词：微博；语义；关联检测；微博舆情；多向量空间模型；草根

中图分类号：TP391.1　　**文献标志码**：A　　**文章编号**：1672-3767(2015)04-0062-05

A Multi-VSM Micro-blog Public Opinion Link Detection Method Based on Semantic

Pang Haijie，Liu Chunqiang
（College of Information Engineering，Qingdao Binhai University，Qingdao，Shandong 266555，China）

Abstract：Based on the features of micro-blog public opinion，this paper proposed a micro-blog public opinion link detection method based on semantic mining. This method focused on the representation model and the similarity computing method of public opinion. In terms of the representation model，responses to a micro-blog post were used to expand this micro-blog post with the aim to deal with the problem of data sparseness，and then based on the Tongyici Cilin，the similarity between words was computed with the aim to deal with the problem of micro-blog's grassroot nature，Finally，the multi-vector space model was established to represent the public opinion. In terms of the similarity computation method，a multi-dimensional similarity computation method was proposed. Experimental results show that our proposed method can improve the precision and recall of the micro-blog public opinion link detection effectively.

Key words：micro-blog；semantic mining；link detection；micro-blog public opinion；multi-vector space model；grassroot

微博是当前深受民众喜爱的新型网络媒介，是一个基于用户关系的信息发布、传播以及获取的平台。随着微博用户群的日益庞大，微博传播突发事件信息的速度越来越快，越来越多的政府部门和企事业单位将其作为舆情监测的重点[1-2]，以期及早发现新出现的舆情信息，掌握舆情发展态势。

面向微博的热点话题检测与跟踪研究越来越受到国内外研究者的关注[3-4]。网络舆情信息关联检测是指判定给定的两篇网络舆情信息是否属于同一对象的技术，其中对象可以是突发事件、个人、企事业单位等，其核心思想是计算两篇微博舆情信息的相关度，这是热点舆情信息识别与跟踪的基础工作。该研究本质上属于话题检测与跟踪(topic detection and tracking，TDT)研究中的报道关系检测(link detection)。传统报道关系检测系统的基本做法是：首先将报道表示成向量空间模型，然后使用 Cosine 函数计算两个向量之间

收稿日期：2015-06-12
基金项目：青岛市科技计划项目(12-1-4-6-(9)-jch)
作者简介：庞海杰(1976—)，男，山东沂水人，副教授，主要从事话题检测与跟踪、自然语言处理方面的研究工作.
E-mail：495131420@qq.com

的相似度;最后通过阈值和相似度之间的比较做出报道是否相关的判断:相似度大于阈值则报道相关,否则报道不相关[5]。传统的报道关系检测是面向书写规范的新闻文本的,而微博文本的草根性、短文本等特性使得采用传统方法难以实现微博舆情信息的关联检测,必须积极探索更加有效的方法。同时,微博本身具有的结构特性可以为微博舆情信息挖掘研究提供新的思路。

基于上述分析,为了更加准确地判定两条微博舆情信息的相关性,提出基于语义理解的微博舆情信息关联检测方法。

1 微博舆情信息关联检测基本框架

本研究用于微博舆情信息关联检测的基本框架如下。

1)对微博舆情信息进行预处理,包括中文分词、去除停用词、去除表情符号及 URL(uniform resource locator,统一资源定位符)。

2)将舆情信息表示成空间向量模型(vector space model,VSM),其中特征项是舆情信息中出现的不同词,特征项的权值为

$$w_k=\frac{1}{L_k}\times T_k。\tag{1}$$

其中:w_k—特征项 k 的权值,T_k—k 在舆情信息中的词频,L_k—k 在舆情信息中的位置序号。

序号标记方法:如果一篇微博舆情信息经过预处理后其结果为“热烈”“祝贺 ”“十八届三中全会”“胜利”“召开”,则“热烈”的序号为 1,“祝贺”的序号为 2,其余单词的序号以此类推。

在权值中引入序号是出于两方面的考虑:①由于微博文本的长度限制,内容大都是以开门见山的方式展开,也就是说位置越靠前的词语越重要;②由于微博文本较短,很多词语在微博中只出现一次,所以仅依据词频信息无法区分词语的重要程度。

3)使用 Cosine 函数计算舆情信息向量之间的相似度。假设 $w_{11},w_{12},\cdots,w_{1n}$ 和 $w_{21},w_{22},\cdots,w_{2n}$ 分别为特征项 $keyword_1,keyword_2,\cdots,keyword_n$ 在舆情信息 S_1 和舆情信息 S_2 中的权值,则 S_1 和 S_2 的基于余弦函数的相似度 $Cos(S_1,S_2)$ 的计算公式为

$$Cos(S_1,S_2)=\frac{\sum_{k=1}^{n}w_{1k}\times w_{2k}}{\sqrt{\sum_{k=1}^{n}w_{2k}^2}\times\sqrt{\sum_{k=1}^{n}w_{1k}^2}}。\tag{2}$$

4)通过 $Cos(S_1,S_2)$ 和阈值之间比较做出舆情信息是否相关的判断:如果相似度大于阈值,那么 S_1 和 S_2 相关;否则二者不相关。

2 基于语义扩展和 Multi-VSM 的微博舆情信息建模方法

2.1 微博舆情信息语义扩展方法

主要从两方面对微博文本进行语义扩展:一个面向微博文本的“短”特性,另一个面向微博的草根性特点。

一方面,微博文本的突出特点是“短”,长度不超过 140 个字,所以其中的数据稀疏现象比较严重。受信息检索领域中查询扩展方法的启发,在实验中探索微博文本的语义扩展方法。微博作为一种基于用户交互关系的信息发布、传播与共享平台,具有独特的结构特性,即一条微博舆情信息发布后会引起其他人的评论,而这些评论基本围绕用户发布的微博展开,所以本研究尝试使用评论信息对微博舆情信息进行语义扩展。

微博的评论信息中通常包含许多主观性信息,而主观性信息中包含的情感词的主题性不强,比如“这个政策太给力了”。为此,提出将评价信息中的情感词(如“给力”)以及对情感词进行修饰的程度副词(如“太”)和否定副词删除,用删除后的评论信息来扩充微博舆情信息。实验结果表明,这一举措可以有效提高语义扩充效果。所以,一条微博舆情信息 S 的语义扩展 $E(S)$ 可以定义为

$$E(S)=\{w|w\in S\}\bigcup\{w|w\in R(S),且\ w\ 不是情感词\}。\tag{3}$$

其中，w 代表词，$R(S)$代表对 S 的回复。

值得说明的是，式(3)中的并操作(⋃)不完全等同于数学集合的并操作，此处的并操作中如果遇到相同词，则只保留一个词，但是该词的词频等于词在原始微博舆情信息及其回复中出现的总次数。

另一方面，微博作为一种草根性极强的信息发布与共享平台，任何注册的用户都可以在该平台上发表自己的微博文本。不同用户的用词习惯不一致，比如：

“这个邮递员工作很认真”

“这个邮差做事一丝不苟”

上述两条微博文本本来是同一个含义，但是如果仅仅依据其中包含的词汇计算相似度会得到二者不相似的结果。为此，本研究基于《同义词词林》计算微博文本的相似度。相似度计算主要依据文献[6]中给出的方法进行，具体如下。

首先，按照下列方法计算两个义项之间的相似度，用 $SenseSim$ 表示：

1)如果两个义项 A，B 不在同一棵树上，则

$$SenseSim(A,B)=f; \tag{4}$$

2)如果两个义项 A，B 在同一棵树上，则

$$SenseSim(A,B)=\begin{cases} a\times\cos\left[(n\times\frac{\pi}{180})(\frac{n-k+1}{n})\right], \text{若 } A,B \text{ 在第 2 层分支上} \\ b\times\cos\left[(n\times\frac{\pi}{180})(\frac{n-k+1}{n})\right], \text{若 } A,B \text{ 在第 3 层分支上} \\ c\times\cos\left[(n\times\frac{\pi}{180})(\frac{n-k+1}{n})\right], \text{若 } A,B \text{ 在第 4 层分支上} \\ d\times\cos\left[(n\times\frac{\pi}{180})(\frac{n-k+1}{n})\right], \text{若 } A,B \text{ 在第 5 层分支上} \end{cases}。 \tag{5}$$

其中：n—分支层的节点总数；k—两个分支间的距离；a，b，c，d，e，f 为系数，分别为 0.65，0.8，0.9，0.96，0.5，0.1。

其次，计算两个词语之间的相似度。分别两两计算这两个词语的义项，取最大值作为两个词语之间的相似度值。

2.2 基于 Multi-VSM 微博舆情信息建模方法

为了得到高性能的微博舆情信息关联检测系统，必须充分挖掘微博舆情信息中所包含的信息。但是，James Allan 等[7]在研究中发现，把两次不同的火车事故或者爆炸事件区分开是很困难的事情，因为这些事件的报道中所用的词汇大部分相同以至于单一向量空间模型很难区分这些细微差别。所以后续有一些研究者采用多个向量空间模型来表示一条新闻报道[8]。为了更加深入地挖掘微博舆情信息中包含的有效内容，本研究探索了面向微博舆情信息的多向量空间模型表示方法，即 Multi-VSM(multi-vector space model)，基本做法如下：

1)抽取每条微博舆情信息的发布时间 $PubTim$，这是因为在同一时间段发布的微博之间的相关性比不在同一时间段发布的微博之间的相关性要高很多[9]。

2)抽取每条微博舆情信息的主题标签 $HashTag$，即在微博舆情信息中以“#”开头的内容；如果没有 $HashTag$，则抽取微博舆情信息中的命名实体；$HashTag$ 基本能表达微博舆情信息的主要内容[9]。

3)舆情信息正文内容。将除了发布时间以及 $HashTag$/命名实体以外的其他内容组成舆情信息的正文内容。

本研究提出的基于 Multi-VSM 的建模方法主要由以上三个部分形成三元组{$<PubTim>$，$<HashTag_1,HashTag_2,\cdots>$，$<Keyword_1,Keyword_2,\cdots>$}，其中，每条微博舆情信息会包含唯一的一个 $PubTime$，$HashTag$ 基本也是一个，而 $Keyword$ 一般会有多个，其权值仍然采用式(1)计算。

3 多维度的微博舆情信息相关度计算方法

2.2 节将微博舆情信息表示成三元组，为此探索一种多维度的微博舆情信息相关度计算方法：如果两篇

微博舆情信息 S_1,S_2 含有相同的 $HashTag$,则 S_1 和 S_2 的相关度为 $L(S_1,S_2)=1$;否则,

$$L(S_1,S_2)=\frac{Cos(E(S_1),E(S_2))}{|P(S_1)-P(S_2)|}。\tag{6}$$

其中:$E(S_1)$,$E(S_2)$分别为 S_1,S_2 的语义扩展;$Cos(E(S_1),E(S_2))$是指用 Cosine 函数计算的两个语义扩展之间的相似度;$P(S_1)$和 $P(S_2)$分别表示 S_1,S_2 的发布时间,发布时间以天为单位。

式(6)引入发布时间是出于对微博舆情信息传播的时间集中性的考虑,该特性是指随着时间的推移,和某一个话题相关的微博舆情信息会越来越少,即每个话题通常持续一段时间后就不再被人们所关注。从该相关度公式可以看出,两个微博舆情信息的发布时间差距越大,它们相关于同一个话题的可能性就越小。

4 实验设置与结果分析

4.1 语料及评价标准

新浪微博是目前国内规模最大的微博平台,吸引了越来越多用户的关注和喜爱。为了评测本文提出的方法,从新浪微博平台上收集 260 条微博舆情信息,同时收集每篇舆情信息的评论信息,其中 160 条用于建立训练语料,而剩余的 100 条用于建立测试语料。每两篇舆情信息组成一个舆情信息对:与同一个话题相关的两篇舆情信息组成相关信息对,而分别与不同话题相关的两篇舆情信息组成不相关信息对。

经过上述的整理,建立用于舆情信息关联检测的语料,统计结果见表 1。

表 1 用于评测的语料统计信息

Tab. 1 The information of corpus used in the experiments

舆情信息情况	训练语料	测试语料
相关舆情信息对	4 319	2 051
不相关舆情信息对	8 401	2 899
总计	12 720	4 950

实验中采用信息检索领域经典的准确率、召回率、F-Measure 作为系统的评测标准。

4.2 实验与结果分析

首先,将关联检测基本框架在训练语料上进行测试,从测试结果可以看出,当相似度阈值取 0.25 时系统的性能达到最好:精确率为 0.590 5,召回率为 0.460 8,F-Measure 为 0.517 6。

其次,为验证方法的有效性,共设置 7 个实验,其中 Baseline 是基于关联检测基础框架实现的,而组合 1~6是根据所提出方法组合而成的系统,详细设置见表 2。各系统的评测结果如表 3 所示,其中各系统的相似度阈值都设置为 0.25。

表 2 实验设置表

Tab. 2 Experiment Settings

名称	单一 VSM	余弦函数	语义扩展	同义词词林	Multi-VSM	多维度相关度
Baseline	√	√				
组合 1	√	√	√			
组合 2	√	√		√		
组合 3	√	√	√	√		
组合 4		√			√	
组合 5		√	√	√	√	
组合 6			√	√	√	√

表 3 实验结果

Tab. 3 Experimental Results

名称	精确率	召回率	F-Measure
Baseline	0.650 7	0.459 8	0.538 8
组合 1	0.642 3	0.501 7	0.563 4
组合 2	0.662 3	0.460 2	0.543 1
组合 3	0.682 4	0.512 0	0.585 0
组合 4	0.660 1	0.448 6	0.534 2
组合 5	0.675 8	0.480 9	0.561 9
组合 6	0.688 2	0.510 2	0.596 2

从表 3 可以看出,本研究提出的方法对系统性能有不同程度的改善。而加入了多维度相关度计算后,配合 Multi-VSM 表示方法的使用,系统的准确率和召回率都有了不同程度的提升,其中,组合 6 提升最显著。

5 结论

微博是目前发展速度最快且影响力最大的网络媒体之一，是进行舆情监控的重要场所。为了有效判定两个微博舆情信息是否相关，提出基于语义理解的微博舆情关联检测方法。该方法首先利用微博的评论信息对微博舆情进行语义扩展，同时基于同义词词林进行词语的相似度计算以应对草根性带来的问题；在舆情信息的表示方面，提出了基于 Multi-VSM 的舆情信息表示模型；最后采用多维度的相关度计算方法来判定两个微博舆情信息是否相关。实验结果表明，提出的方法可以有效改善微博舆情信息关联检测系统的性能。

参考文献：

[1]Damiano S V. Entity-based filtering and topic detection for online reputation monitoring in twitter[D]. Madrid: Universidad Nacional de Educación a Distancia, 2014: 92-132.

[2]Zhang S L, Luo J Y, Liu Y, et al. Hotspots detection on microblog[C]//4th International Conference on Multimedia Information Networking and Security. Nanjing, Nov. 2-4, 2012: 922-925.

[3]马彬，洪宇，陆剑江，等. 基于线索树双层聚类的微博话题检测[J]. 中文信息学报，2012，26(6)：121-128.
Ma Bin, Hong Yu, Lu Jianjiang, et al. A thread-based two-stage clustering method of microblog topic detection[J]. Journal of Chinese Information Processing, 2012, 26(6): 121-128.

[4]杨亮，林原，林鸿飞. 基于情感分布的微博热点事件发现[J]. 中文信息学报，2012，26(1)：84-91.
Yang Liang, Lin Yuan, Lin Hongfei. Micro-blog hot event detection based on emotion distribution[J]. Journal of Chinese Information Processing, 2012, 26(1): 84-91.

[5]Chen F, Farahat A, Brants T. Multiple similarity measures and source-pair information in story link detection[C]//Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. Boston, Massachusetts, 2004: 313-320.

[6]田久乐，赵蔚. 基于同义词词林的词语相似度计算方法[J]. 吉林大学学报：信息科学版，2010，28(6)：603-605.
Tian Jiule, Zhao Wei. Words similarity algorithms based on Tongyici Cilin in semantic web adaptive learning system[J]. Journal of Jilin University: Information Science Edition, 2010, 28(6): 603-605.

[7]Allan J, Carbonell J, Doddington G, et al. Topic detection and tracking pilot study final report[C]//DARPA Broadcast News Transcription and Understanding Workshop. Lansdowne, USA, 1998: 1-35.

[8]Makkonen J, Ahonen-Myka H, Salmenkivi M. Applying semantic classes in event detection and tracking[C]//International Conference on Natural Language Processing. Mumbai, India, 2002: 175-183.

[9]Guo W W, Li H, Ji H, et al. Linking tweets to news: A framework to enrich short text data in social media[C]//51st Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria, 2013: 239-249.

（责任编辑：吕文红）