doi:10.11959/j.issn.1000-436x.2016018

短句语义向量计算方法

陈福[1]，林闯[2]，薛超[2]，徐月梅[1]，孟坤[2]，倪艺函[1]
（1. 北京外国语大学计算机系，北京 100089；2. 清华大学计算机系，北京 100084）

摘 要：提出了一种基于人工神经网络的短文语义向量放缩算法，结合社交节点自身信息和短文语义，给出社交网络短文语义计算方法和突发话题发现算法。通过文本数值化实现语义距离的计算、比较、节点的分类及社区发现等。通过自行开发的微博采集工具 Argus 采集的大量新浪微博内容对所提模型和算法进行了验证，最后对未来工作进行了展望。

关键词：在线社会网络；主题语义计算；人工神经网络；突发话题发现
中图分类号：TP393 **文献标识码**：A

Vector semantic computing method study for short sentence

CHEN Fu[1], LIN Chuang[2], XUE Chao[2], XU Yue-mei[1], MENG Kun[2], NI Yi-han[1]
(1. Computer Department, Beijing Foreign Studies University, Beijing 100089, China;
2. Computer Department, Tsinghua University, Beijing 100084, China)

Abstract: A vector semantic computing method study for short sentence based on artificial neural network was proposed. And a semantic computational algorithm for social network texts as well as a discovery algorithm for mergencies was provided with reference to the information provided by the social nodes itself and the semantic of the text. Through the numerization of text, the calculation and comparison of semantic distance, the classification of nodes and the discovery of community can be realized. Then, huge quantities of Sina Weibo contents are collected to verify the model and algorithm put forward. In the end, outlooks for future jobs are provided.

Key words: online social networks, theme semantic computing, artificial neural nets, burst topics discovering

1 引言

移动智能终端的广泛使用和无处不在的网络接入能力，使微博、微信等信息传播形式爆发出巨大的社会影响力，对社会网络用语的语义分析具有重要的意义。社交网络的影响力分析很早就得到了世界一流研究机构的关注[1]。Facebook、LinkedIn 及新浪微博等移动在线网络与传统的社交网络的行为特征、传播手段和影响能力均有巨大不同。最明显的外在特征是具有明显瞬态时间特征的巨量短文本流，如微博的 140 个汉字。因此，加强对这种短文本信息的处理具有重要意义。短文本的语义计算对在线社交网络的社区发现、网络结构拓扑分析、节点推荐、广告精准投放、组织结构管理、恐怖组织识别等均具有重要意义。传统的在线社区发现常通过节点之间的关注进行社区网络拓扑结构的识别，而不是通过语义距离的计算[2]。因此，如何采用向量表示在线短文本的信息成为一个重要问题。

本文针对移动网络用语的短小但语义丰富、实时性高的特点，结合微博节点本身的语义，提出了用于描述微博内容的语义度量向量模型。基于该模型可以进行突发话题发现、意见领袖识别、谣言分析和确认及微博内容和节点的推荐。

收稿日期：2015-05-13；**修回日期**：2015-09-30
基金项目：国家自然科学基金资助项目(No.61170209, No. 61173008, No. 61502038, No.61370132)；教育部新世纪优秀人才支持计划基金资助项目（No.NCET-13-0676）；2011 重点课题基金资助项目（No.BFSU2011-ZS04）
Foundation Items: The National Natural Science Foundation of China(No.61170209，No. 61173008，No. 61502038, No.61370132), The Ministry of Education Program of New Century Excel nt Talents(No.NCET-13-0676), 2011 Key Project (No.BFSU2011-ZS04)

2 相关工作

2.1 话题语义建模方法

话题语义建模是获取微博语义内容的基础。文献[3]采用联合概率生成模型进行了社交网络语言上下文感知和话题建模。LDA（latent dirichlet allocation）是一种重要的话题语义建模方法[4~6]。如TwitterRank[7]采用 LDA 模型从 tweets 中提取潜藏的主题信息，然后根据特定的主题排序[8]。LDA 是一种非监督学习的文档主题生成模型，是一个 3 层贝叶斯概率模型。LDA 采用了词袋（bag of words）模型，但是词袋方法没有考虑词与词之间的顺序。

LDA 基本上是以文档集合作为研究对象的潜在语义分析。对新浪微博这样的短文本的在线社会网络，直接应用 LDA 进行语义获取具有一定的局限性。文献[9]通过对文档聚类并结合 tweets 特征和粒度进行主题发现，主要思想是如果某个词或短语在一篇文章中出现的频率 TF(term frequency)高，在其他文章中很少出现，则认为此词或者短语对语义具有较大的识别意义，也就是词在篇章中的重要性与其在文件中出现的频数成正比，与其在语料库中出现的频率成反比。

2.2 突发话题发现

在线社会网络的突发话题的识别和发现在过去几年得到了广泛的关注[10,11]。话题检测和跟踪(TDT, topic detection and tracking)是突发话题识别、趋势预测的基础。使用状态变迁理论，用带有权重的自动状态机理论进行突发话题识别在早期得到充分的重视[12]。传统上的突发话题均以词频为主，但社会网络除了文本外还包括声音、图片和超链接等。如何在社会网络中特别是针对类似于新浪微博这样的短文本进行突发话题发现是一个热点问题，也得到了很多研究人员的高度关注。对于在线社会网络突发话题的发现，从链路流量特征异常检测的角度进行识别得到了高度关注[13]，*N*-grams 模型[14]、两阶段消息分类[15]均得到了尝试。对这种短文本的预测问题，国内外很多研究人员均采用了用户为图的顶点、传输路径为边，分析信息的传播和转发概率的形式进行一定程度预测[16]。这种以分析链接及转发路径的形式进行的分析缺乏对文本本身语义的考虑，因而具有一定的片面性。结合语义和链接分析的工作因而逐步得到重视[17]。新浪微博主要内容是中文，而中文话题检测与跟踪的实现与评测得到了国内相关学者的广泛关注[18]。其他的研究包括相邻时段间情感分布语言模型间差异分析微博热点事件发现[19]、微博网络热点相似度和测度、传播路径和用户行为的中心化等网络热点发现、隐含语义分析两阶段聚类话题发现方法的聚类分析[20,21]、迭代式的语义分析和话题热度预测模型。文献[22]采用向量空间模型来表示报道和话题等。

2.3 话题相似度计算

文献[3]给出了在线社交网络的测量方法比较全面的综述。文献[23]根据事件的内容相似度、事件和话题的相似度、事件的时间相似度提出了一种计算方法。突发话题确定后，开始计算话题相似性，并不是所有微博用户都会对同一个突发的话题感兴趣。因此，需要计算突发话题与微博兴趣的距离。同时，通过准确度量话题相似性也可以进一步确认话题的突发性。如果 2 个节点的微博语义距离很大，则一般不会被推荐成为相互关联的朋友。对用户而言，及时得到最感兴趣的信息才是最重要的。从语义上对在线短文本信息进行测量，基于语义的信息分类与排序对节点用户而言更有意义。因此度量节点之间的相似性，度量微博内容之间的相似性，度量微博与节点兴趣之间的相似性，是非常核心和重要的研究内容。

很多人从节点和链接的拓扑结构角度做测量，或者从好友数、发文数、跟帖数等度量节点影响力大小[24]。相应的概念包括紧密中心度、介数中心度等度量当前节点对其他节点的影响力或节点的社会关系强度[25,26]。从拓扑结构、用户行为和网络演化等方面对常见的测量方法和典型的网络拓扑在文献[27]中进行了系统的阐述。

从在线网络结构本身的角度进行社团发现、度量节点之间的紧凑程度，由于缺乏语义信息而具有一定的局限性。这种度量方法可以从一定角度上反映节点之间的已经具有的关联或影响关系，但对正在形成或具有潜在影响力的推荐方面没有实质意义。转发关系、回复关系、复制关系、阅读关系及相应关系的随机游走模型下的话题影响力计算可以从一定程度上描述话题直接的关系，但这些关系不能一般性地度量 2 个话题的距离。

2.4 存在问题

从上述内容可以看出，像微博这样的在线短文本的建模、测量和分析得到了高度重视，但仍然存在以下问题。

1) 微博文本简略口语化。基本上，微博这样的在线社交平台都限制了发文的字数，不可能像博客一样可以发表长篇的论述，采用的语言也多是简略甚至口语化的。传统的通过文档集合得到文档、通过文档得到关键词集合的方法，是基于主题单一的篇章结构，而不是口语化的简短的信息描述。使用传统的篇章主题建模方法对微博这样的短句文本进行建模具有语义断层。因此，如何对短文本微博内容进行建模是一个挑战性问题。

2) 语义内容多元离散。一个关注了很多其他节点的节点，他所看到的多条微博内容必然是相互离散的，即使同一个节点在一个时间段所发的微博，其语义必然也是多样的。也就是说，一个节点所收到的信息不能形成一个文档。由于这个原因，采用 LDA 的主题模型从原理上存在矛盾。因为 LDA 要求文档—主题，主题—词语具有内在的关联性。这种关联性在微博这样的以句子为单位的情形下直接使用 LDA 存在一定的不合理性，而且 LDA 在面对大量数据集时需要的计算量过大。

3) 微博数量巨大。上述的瞬态性、简略性和内容离散是针对某一个节点的微博空间而言的。对某一个在线社会网络而言，例如新浪微博，单位时间内涌现的微博数量是惊人的。只有 Twitter 或新浪微博平台本身可以快速、即时得到这些实时发出的微博，其他任何组织或个人无论采用平台提供的 API 还是通过网络爬取均无法全部获取。因此，希望通过微博空间得到即时舆情计算或挖掘都存在很大的偏差。一般而言，针对微博平台整个空间的测量、采集、分析及舆情计算均存在不同程度的时延或偏差。

4) 影响力和内容含量巨大。较短的文本、瞬间即逝的在线社交网络的信息含量大、传播速度快，因而常常具有惊人的影响力。也正是这种惊人的影响力，主流的电视媒体、企业单位、国家部门、名人及普通民众均对微博这样的社交网络表现出了极大的热情。

对在线社会网络的短文信息而言，其内容往往具有瞬时性。例如，通过微博知道了某个事件或某一链接然后通过其他途径进一步深入了解。因此对微博本身的组织、查找和理解与传统对博客、新闻网页等长文本相比，重要性降低。也就是对类似微博这样的在线短文本，基于内容的比较和关键词语义的识别更重要。而且由于短文本的字和词语的数目明显少于长文档，因此对这样的短文进行语义识别，采用传统的如 LDA 这样的方法必然存在局限性。而且针对微博这样的短文本具有诸如口语化这样的特征，需要完善传统的语料库使之具有识别能力。对海量、短文本、多主题、大噪声构成的文本集合进行建模与传统的长文本、主题单一明确、噪声较少的传统媒体明显不同。

微博短文本理解、内容挖掘、用户社区挖掘、意见领袖识别和信息传播模式等研究的最根本的工作是短句、主题多变语境的文本理解及量化问题，这也是本文研究的重点。

结合上述特征和目前工作，本文的贡献如下。

1) 本文通过大量微博短文本建立微博语料库，同时结合一般新闻语料库进行文本数值化、向量化训练。

2) 结合文本短小的特点，对短文的关键词进行语义的“放大”，对非关键词进行语义“缩小”，从而建立短文本语义的更加清晰的轮廓。

3) 为了利用数字向量化的结果进行微博短文本的分类、组织，除了对语义进行放缩处理外，本文建立短文本等价类模型，通过语义闭包扩展，增强分类能力。

4) 结构语义放大和文本语义放大。通过人工神经网络建立短文本深度语义模型获得微博短文本统计语义向量。基于向量距离进行语义放大和通过语义关联词库进行语义放大。

3　话题语义的线性放大

3.1　语义线性放缩

对微博等短文本内容计算的最大困难是文本短小、关键字数目少、文本口语化、网络流行性新词多等问题。与此同时，微博内容量大、噪声繁杂，提取隐含的、有价值的信息更为复杂。通过语义放缩的目的是更有利于分类、比较和查找。例如微博这样的离散短文本，就可以通过放大语义信息从而得到相关微博之间的交集而归于同类。反之，如果不进行语义的放大则对部分含义接近而用词差异较大的句子进行语义归类时候存在较大困难。因此，确保原始语义不变的情况下对语义内容、关键词数量进行一定的放缩是非常必要的。在微博情境下，建立线性变换空间 T。确保 T 变换满足可加性和齐次性

$$\begin{cases} T(\boldsymbol{a}+\boldsymbol{b})=T(\boldsymbol{a})+T(\boldsymbol{b}) \\ T(K\boldsymbol{a})=KT(\boldsymbol{a}) \end{cases} \tag{1}$$

其中，$\boldsymbol{a}$、$\boldsymbol{ß}$ 表示语义单位向量，T 表示某种语义变换，K 表示向量倍数。式(1)的 $T(\boldsymbol{a}+\boldsymbol{ß})$表示对 2 个语义向量单元的加和后的变化，$T(K\boldsymbol{a})$表示对语义向量放大 K 倍后的变换。而式(1)表示的是线性变换的数学条件，本文的工作之一就是得到微博文本关键字的向量表示，并在此基础进行变换，具体的变换方法后面会详细阐述。

3.2 语义向量定义

对微博短文关键词词语，用向量表示是进行语义计算的前提和基础，即微博文本词向量表示问题。One-hot 表示方法因为数据稀疏问题、不能描述词语之间的相似性等而一般不被采用[28]。使用人工神经网络将词表征为实值向量得到了广泛关注[29]，从而实现对文本内容的处理转化为向量空间中的向量运算。如向量空间上的相似度可以表示文本语义上的相似度，即用向量内积空间的夹角余弦值度量语义相似性。通过 N-gram 引入情境影响，使句法和语义相近的词具有近似的词向量。Skip-gram 和 CBOX 模型是 2 种使用简单的人工神经网络结构获得词向量表示的模型。Skip-gram 用于预测或估计相关联词，而 CBOX 则是在给定若干词前提下预测下一词汇[30]。本文首先给出关键词和句子的语义向量定义。为后面行文方便，先给出用到的一些定义。

定义 1 $\boldsymbol{Keyword\_V}=\{x_1\cdots x_n\}$：关键字的向量表示，其中，$x_i$ 是实数表示的向量的某一维。

定义 2 $\boldsymbol{Sentence\_Set}=\{k_1\cdots k_n\}$：某一短句经过分词抽取到的有意义的关键字集合，其中，k_i 表示某一关键字，关键字的个数 n 是由句子的构成决定的，句子较长则关键字个数就可能较多。

定义 3 $\boldsymbol{Sentence\_V}=\{y_1\cdots y_n\}$：表示某一短句的向量表示，其中，$y_i$ 是实数，表示句子向量的某一维。

$$y_i=\frac{\sum_{i=1}^{n}\boldsymbol{Keyword\_V}.x_i}{n} \tag{2}$$

定义 4 径向放缩向量 $\boldsymbol{ScalV}$：表示沿向量各个方向放缩的比例。

$$\boldsymbol{ScalV}=\{k_1,\cdots,k_i,\cdots,k_n\} \tag{3}$$

3.3 径向语义向量放缩矩阵

使用向量空间的线性投影进行语义放缩。首先建立核心集语义模型，然后放缩矩阵实现向量空间映射，并借此找到近邻词。通过上述定义，句子已经表示成了向量，语义的放缩问题就转化为了向量放缩问题。因为向量的维数是可以根据计算资源确定的固定值，这里假定为 n。

设变换前语义向量为 $\boldsymbol{X}$，变换后语义向量为 $\boldsymbol{Y}$，则 $\boldsymbol{Y}=\boldsymbol{K}_n\boldsymbol{X}$，其中，$\boldsymbol{K}_n$ 为 n 阶矩阵。

$$\boldsymbol{K}_n=\begin{bmatrix}k_{11} & \cdots & k_{1n}\\ \vdots & \ddots & \vdots\\ k_{n1} & \cdots & k_{nn}\end{bmatrix} \tag{4}$$

比较简单的放大语义本质上是使向量沿着各个方向的拉伸，其矩阵 $\boldsymbol{K}$ 为

$$\boldsymbol{K}=\begin{bmatrix}k_{11} & \cdots & 0\\ \vdots & k_{ii} & \vdots\\ 0 & \cdots & k_{nn}\end{bmatrix} \tag{5}$$

其中，$k_{ii}>0$，$k_{i-1,\,i-1}>k_{i,\,i}$，当 k_{ii} 为常数时表示沿各个方向等长放大 k 倍。具体放大的倍数根据实际效果和需要而定，这种语义放大的逻辑含义是语义和逻辑结构的线性放大。矩阵 $\boldsymbol{K}$ 的对角线形成了径向放缩向量，根据放缩规模进行设置。

k_{ij} 值表示放缩量的大小，该值的大小是根据时间效果和应用对分类的精度要求所决定的，该值越大分类的精度越低，因此该值的确定需要根据实际需求和分类效果确定。

3.4 语义向量球体放大

上述径向扩展是沿着各维方向的拉伸，并没有法向的扩展。法向扩展可以通过旋转或扩展辖域实现。但由于高维空间的旋转变换较为复杂，本文采用 p-范数表示到向量间距离的概念。

所有到语义向量 $\boldsymbol{Sentence\_V}$ 的 p-范数小于 R 的向量构成的空间，在逻辑上等同于沿向量 $\boldsymbol{Sentence\_V}$ 的各个切面法向量的拉伸。其定义如下。

设变换前语义向量为 $\boldsymbol{X}$，变换后语义向量为 $\boldsymbol{Y}$。则与语义向量 X 距离为 R 的语义扩展向量是一个集合，设该集合为 $Scale\_Semantic\_Set$。

$$Scale\_Semantic\_Set=\left\{\boldsymbol{Y}_i \,\middle|\, \|\boldsymbol{Y}_i-\boldsymbol{X}\|<R\right\} \tag{6}$$

$$R=\|\boldsymbol{Y}_i-\boldsymbol{X}\|=\sqrt[p]{\sum_{j=1}^{n}(y_{ij}-x_i)^p} \tag{7}$$

其中，$\boldsymbol{X}=\{x_1,\cdots,x_n\},\boldsymbol{Y}_i=\{y_1,\cdots,y_n\}$

所有满足上述条件的向量 $\boldsymbol{Y}_i$ 构成的集合 $Scale\_Semantic\_Set$，形成了一个类似于球体的高维封闭曲面，为计算简便，后文的实验采用 2-范数。

对上文提及的词向量表示的实现，本文采用人

工神经网络语言模型，通过无监督学习和领域文本语料库获取相应文本关键字的词向量表示，后文的实验环境会详细介绍。

3.5 语义放缩 SEZOM 算法

根据上文所述的 2 种对语义向量的放缩方法，本文提出了短句的语义向量变换算法。

算法 1　短句向量化及其放缩算法

输入：在线文本语料 *M_CPS*；

在线短句 *M_S*；

放缩距离 *R*；

放缩向量 ***ScalV***；

输出：短句向量表示 ***M_S_V***；

放缩后向量 ***M_S_V_S***；

Scale_Semantic_Set；

步骤：

1) **将语料** *M_CPS* **切分成关键字集合** *M_CPS_Set*;

2) **使用关键字集合** *M_CPS_Set* **训练神经网络模型，得到语言向量集合** *M_CPS_Vec_Set*;

3) **从短句** *M_S* **中得到该句子关键字集合** *MSV_S*;

4) For $\boldsymbol{X}$ in *MSV_S*{

5)　IF ($\boldsymbol{X} \in$ *M_CPS_Vec_Set*)

6)　**取得** $\boldsymbol{X}$ **的词向量表示加入到** *Vec_MSV_S*;

7) }

//**计算句子向量** ***Sentence_V***

8) For $\boldsymbol{Y}$ in *Vec_MSV_S*

9)　while(*i*++ < **词向量维数**)

10)　$\boldsymbol{Sentence\_V}.x_i = \boldsymbol{Sentence\_V}.x_i + \boldsymbol{Y}.x_i$;

11) $n = \|Vec\_MSV\_S\|$;　//**句子关键字个数**

12) while(*i*++ < **词向量维数**)

13) $\boldsymbol{Sentence\_V}.x_i = \dfrac{\boldsymbol{Sentence\_V}.x_i}{n}$;

/* **用** ***ScalV*** **对向量** ***Sentence_V*** **径向放大，得到放大后向量** ***RScal_Sentence_V***.*/

14) $\boldsymbol{RScal\_Sentence\_V} = \begin{pmatrix} ScalV_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & ScalV_n \end{pmatrix} \cdot \boldsymbol{Sentence\_V}$;

//**语义向量球体放大**

15) do{

16) $Y_i = \begin{pmatrix} r & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & r \end{pmatrix} \boldsymbol{Sentence\_V}$;

17) $Scale\_Semantic\_Set = Scale\_Semantic\_Set \cup Y_i$;

18) $r = r + STEP$;

19) $dist = \|Y_i - \boldsymbol{Sentence\_V}\|$;

20) }while($dist < R$)

通过算法 1 实现了关键词的向量化，并将句子向量表示沿着各维的径向做了扩展及沿着各维的法向作了扩展。经过算法 1 的处理使短句的语义向量表示具有一定的外延。

4　实验验证与结果分析

4.1　实验背景分析

本节采用新浪微博数据进行相关算法的测试与实验。新浪微博具有广泛的影响力，基于博文内容对微博节点进行分类可以为用户准确推荐感兴趣的微博节点。在基于兴趣的广告推送，基于内容的舆情分析等情境，均需要对博文内容本身和微博之间进行比较和归类。新浪微博属于典型的短句，因此本文采用新浪微博数据进行验证和测试。

4.2　实验数据集

获取新浪微博相关内容的方法主要包括新浪微博 API **和其他第三方爬取工具，由于诸多原因新浪** API **不能满足一般科研实际需求，而使用其他新浪微博获取工具所得到的数据从内容和格式都太过固定，不能实现定制。因此，本文开发了能够获取多元新浪微博信息的工具** Argus**。通过** Argus **实现了广度优先递归抓取某一节点粉丝** ID **及相关微博内容。**Argus **可以抓取用户的所有微博内容，包括用户的原创微博、转发微博、原微博发起人、转发评论、转发关系等。其体系结构如图** 1 **所示。**

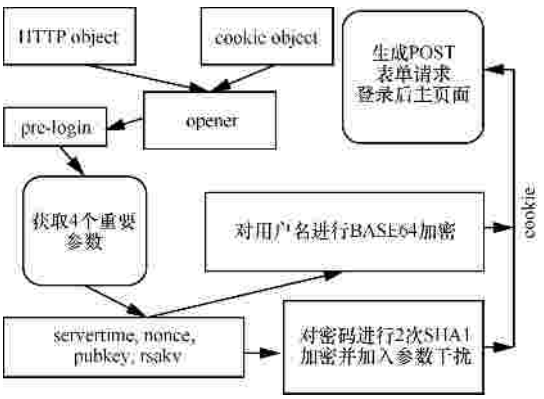

图 1　**新浪微博信息采集工具** Argus **体系结构**

本文关注的是短文分类问题，因此按条目所列的新浪微博内容是本文所需数据。根据吴军等给出的结论，机器学习的数据与问题域实际场景越接近，实验所取得的效果越好。因此本文主要以新浪微博所涉及的词汇为主进行模型训练。由于word2vec 训练需要以空格进行分词，因此本文对所有微博内容进行了分词处理。在突发话题发现时，由于微博的内容主要与本条微博的关键词相关，因此在突发话题发现算法过程中，本文对对微博内容进行了关键词提取，本文采用 jieba 分词进行了关键词提取[31]。

4.3 话题语义线性实验

4.3.1 实验设置

通过使用 Argus 采集大量数据，经过抽取得到每个节点的标签数据和微博内容，然后进行分词和关键字提取。从微博用户解析得到的标签数据和微博内容分词后的语料合并作为训练语料，即算法 1 中的 *M_CPS_Set*。然后使用 *M_CPS_Set* 训练 word2vec 得到语言向量集合 *M_CPS_Vec_Set*。下面将使用 *M_CPS_Vec_Set* 进行句子向量的计算等。

4.3.2 微博内容分词和关键字抽取

本节随机抽取 6 条微博短句，分别记为 T1~T6。为使算法具有一般性，本文抽取各个短句的关键词如表 1 所示。

为简单起见，仅取 10 个关键字，根据算法 1 中的步骤 1)~步骤 13)、步骤 8)~步骤 14)和步骤 15)~步骤 20)分别计算各个短句的向量值，得到 T1~T6 向量表示短句及其放大表示。为了直观展示向量放大的效果和实际的意义，分别计算了 T1~T6 向量与某一个微博节点的各个标签的语义距离。本文以某新浪微博节点为例，其标签为：下一代、动力学、服务平台、管理、互联网、计算、网络服务、微博。

分别使用径向放缩向量和球体放缩放量对微博向量 T1~T6 进行了放大处理，然后分别计算 T1~T6 与上述 8 个标签的语义距离。为了从整体上观察这种放大效果，把 T1 和 T6 这 2 个微博向量的变化情况通过与 8 个标签的语义距离表现出来，其中图 2 表示微博 T1 按不同球体放大倍数放大后的向量与 8 个标签的语义距离变化情况，图 3 表示 T6 按径向的放大后的向量与标签的语义距离变化情况，可以看出语义距离在一定的幅度内变化，从图 2 和图 3 只能看出语义距离的变化，但看不出对径向放大和球体放大效果的区别。

4.3.3 语义向量放大及分析

为了进一步观察径向放大和球体放大效果的区别，将 T1~T6 分别使用径向和球体的放大向量放大 4 个不同倍数，然后再分别求与 8 个标签的语义距离，观察两者的改变情况。例如将向量 T1~T6 径向放大 1~1.5 倍、1.5~2 倍、2~2.5 倍和 2.5~3 倍，将向量 T1~T6 按球体放大向量放大为 1~1.5 倍、1~2 倍、1~2.5 倍和 1~3 倍等。这样每个微博 T*i* 得到了 7 个不同的放大向量，本文随机取样了 T1~T6 共 6 条微博，放大后的向量为 42 个向量。然后分别计算 42 个向量与 8 个标签的语义距离。

表 1 新浪微博及其关键字

微博	短句	关键字
T1	分布式机器学习的故事（四）：Rephil 和 MapReduce——描述长尾数据的数学模型	Rephil，MapReduce，分布式，长尾，数学模型，描述，机器，故事，学习，数据
T2	大数据分析正逐步影响着我们个人的生活、企业的运作，如何有效的应用内部数据和外部数据并且保证用户隐私的问题也将会浮出水面	数据分析，数据，隐私，浮出，外部，运作，用户，水面，应用，保证
T3	大数据征信之全球三巨头公司”，大数据在金融的核心应用是风险控制和信用评估征信公司的核心技术有 2 个…	征信，数据，特征，信用，个人信用，风险，甄别，配对，技术，某人
T4	继上期推出克里斯坦森教授《你要如何衡量你的人生》得到大家广泛热议后，本期哈佛君为大家推出《平衡计分卡战略实践》…	大家，高管，计分卡，热议，坦森，君为，推出，克里斯，哈佛，前三名
T5	瞄准哪儿创新？被誉为驱动我国自主创新发展“三驾马车”的“863”计划、“973”计划和支撑计划，新公布了 2015 年度项目申报指南…	计划，宽带网，位置服务，973，863，2015，创新，第五代，三驾，第三代
T6	明天 Geoffrey Hinton 会到 Sherbrooke 做报告，同时接受 University of Sherbrooke 大学的荣誉博士学位，以表彰他在 AI 领域中的杰…	Hinton，Sherbrooke，明天，博士学位，AI，领，University，Larochelle，Hugo，Geoffrey

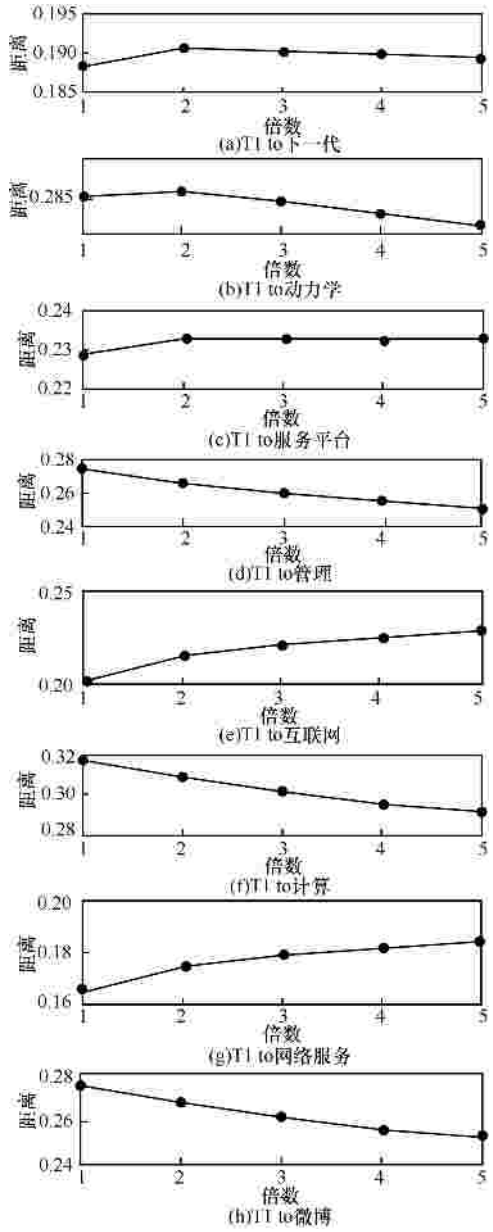

图 2　T1 球体放大与标签语义距离变化

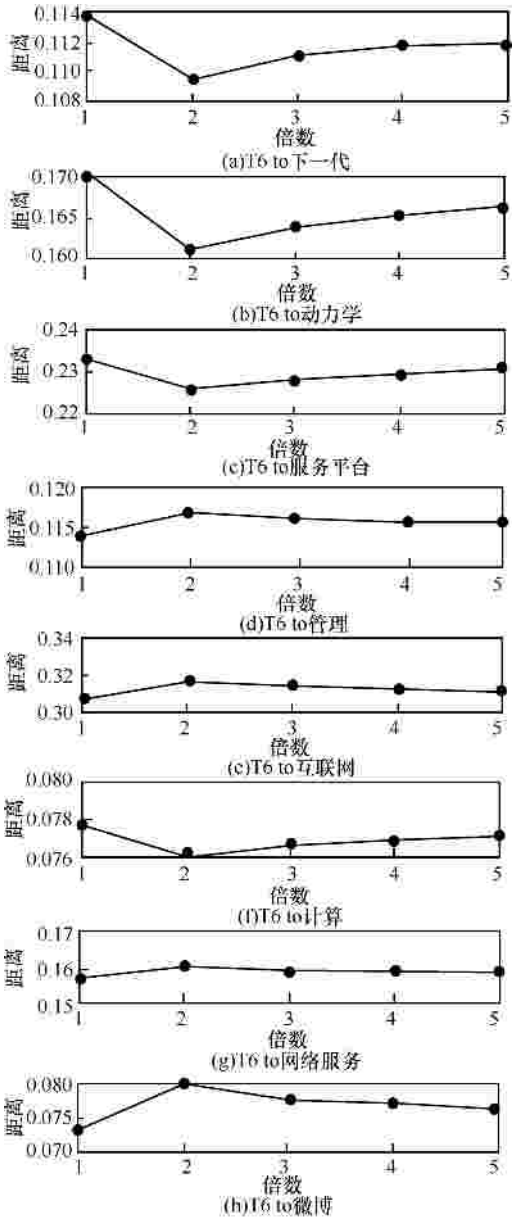

图 3　T6 径向放大与标签语义距离变化

图 4 展示了微博 T5 与各个标签在语义放大不同倍数后的变化情况，可以看出语义向量放大后与各个标签的语义距离变小，但与某些标签的语义距离变化幅度很小。如 T5 与标签 6 基本没有改变。实验中也观察了其他微博的变化情况，发现语义距离变小，但变化的幅度不大。

语义距离越小，语义差异越大。将微博语义放大后与各个标签的语义距离变小容易理解，这是因为将微博语义向量放大后，必然偏离原来的语义位置，从而使该语义变量在原来的基础上偏离。因为径向放大是在基本保持语义向量分量的基础上的调节，而不是整个径向的改变。

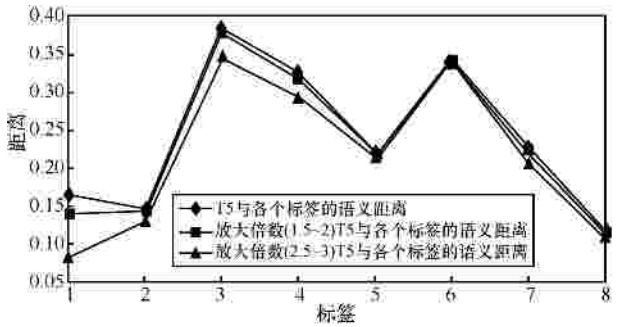

图 4　微博 T5 语义放大后与各个标签的距离

图 5~图 7 分别给出了微博 T5、微博 T1 和微博 T3 的语义向量球体放大后，与各个标签语义距离的变化情况。从图 5 和图 6 可以看出经过放大后语义

向量变化幅度与前面的径向相比明显增加。这是因为语义向量的变化范围是沿着各个分量的整体放大，如图 7 所示的 1~1.5 和 1~3 倍放大，而不是沿着各维径向的扩展放大。

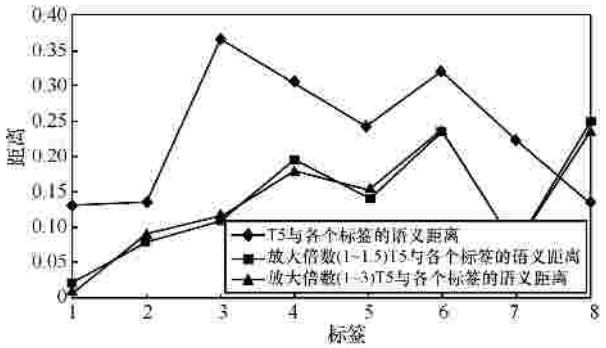

图 5 球体放大 T5 后与各个标签的距离

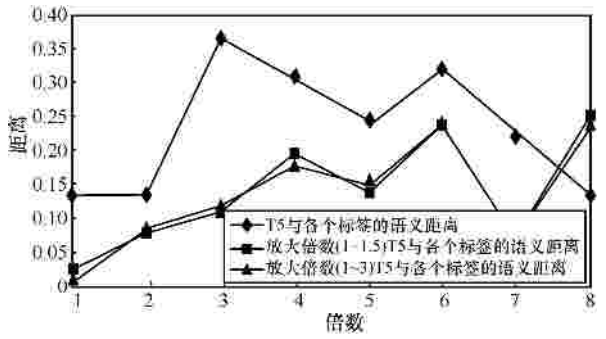

图 6 球体放大 T1 与各个标签的语义距离

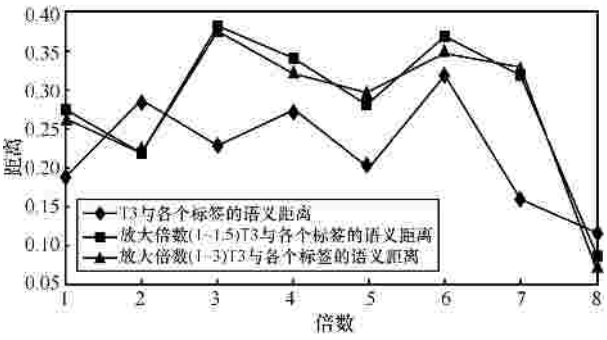

图 7 球体放大 T3 与各标签的语义变化

本文考察了 T1～T6 的不同球体放大倍数的向量，与各个标签的语义距离变化情况。大多数的语义距离值均缩小，其与对应的径向发大倍数相比，变化幅度均比较明显，如图 5 和图 6 所示。但有部分语义距离经过放大后语义距离变大，如图 7 所示。图 7 表示微博 T3 在不同的放大倍数下，与 8 个标签的 6 个语义距离值变大。因为该值越大，表示语义越接近。

综上，可以得出如下结论。

1) 根据算法 1 对短句的向量表示进行各维放大，可以使语义距离发生近域变化，大多数会使放大后的向量表示与其本身的含义越来越远。

2) 球体放大效果略比径向放大变化明显，但均变化有剧烈改变，否则失去通过放大或缩小寻找等价类的意义。

3) 无论是径向放大还是球体放大，本质是在原来的语义范围内形成了一个近邻域。

5 结束语

在线社会网络中的信息采集、处理、分析是社会网络研究领域的一个重要方面，语义分析、比较和数量化测量对舆情监控、广告推送、信息个性化定制等均具有重要意义。本文基于人工神经网络提出短文语义向量放缩算法、综合社交节点自身信息和发文语义给出社交网络短文语义计算算法和突发话题发现算法。通过定义节点微信息语义向量、语义外延闭包扩展建立短句等价关系，进而实现突发话题的发现。

本文提出的算法也可计算微博内容与节点的相关度，进而形成与该节点语义距离远近的相关度排序列表，限于篇幅这一部分内容没有展开。因为近邻与标签语义距离不同，应该加入权重系数进行计算。本文后续工作还包括词向量的数值化方法相关的词向量的语义分类，训练模型的数据排序方法，训练数据的本身分类等问题。这些问题的解决对自然语言的理解、跨语言翻译，文本语义理解和分析均有重要意义。

参考文献：

[1] WASSERMAN S, FAUST K. Social network analysis: methods and applications[M]. Cambridge, U K: Cambridge University Press, 1994.

[2] CHEN K H, HAN P P, WU J. User clustering based social network recommendation[J]. Chinese Journal of Computers, 2013, 36(2): 349-359.

[3] 徐赢, 刘屹, 阴红志, 等. 查询性能预测方法的性能评测研究[J]. 计算机研究与发展, 2013,(S1):70-79.
XU Y, LIU Y, YIN H Z, et al. An empirical study of the performance evaluation of query performance predictors[J]. Journal of Computer Research and Development, 2013,(S1):70-79.

[4] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[5] STEYVERS M, GRIFFITHS T. Latent semantic analysis: a road to meaning[M]. Laurence Erlbaum, 2007.

[6] MILSTEIN S, CHOWDHURY A, HOCHMUTH G, et al. Twitter and the micro-messaging revolution: communication, connections, and

immediacy-140 characters at a time[R]. O'Reilly Report, 2008.

[7] WENG J S, LIM E P, JIANG J, HE Q. Twitterrank: finding topic-sensitive influential twitterers[C]//The Third ACM International Conference on Web Search and Data Mining. c2010: 261-270.

[8] WENG J S, LIM E P, JIANG J, et al. Finding topic-sensitive influential twitterers[C]//The Third ACM International Conference on Web Search and Data mining. New York, USA, c2010:261-270.

[9] **李劲**, **张华**, **吴浩雄**, **等**. **基于特定领域的中文微博热点话题挖掘系统** BTopicMiner[J].**计算机应用**, 2012, 32(8):2346-2349.
LI J, ZHANG H, WU H X, et al. BTopicminer: domain-specific topic mining system for Chinese microblog[J]. Journal of Computer Applications, 2012, 32(8):2346-2349.

[10] LAVRENKO V, ALLAN J, DEGUZMAN E, et al. Relevance models for topic detection and tracking[C]//The Human Language Technology Conference. San Diego, USA, c2002: 104-110.

[11] **陈友**,**程学旗**, **杨森**. **面向网络论坛的突发话题发现**[J]. **中文信息学报**, 2010,24(3):29-36.
CHEN Y, CHENG X Q, YANG S. Outburst topic setection for Web forums[J]. Journal of Chinese Information Processing, 2010,24(3): 29-36.

[12] JON M. Kleinberg: bursty and hierarchical structure in streams[J]. Data Mining and Knowledge Discovery, 2003, 7(4):373-397.

[13] TOSHIMITSU T, RYOTA T, KENJI Y. Discovering emerging topics in social streams via link-anomaly detection[J]. IEEE Trans Knowl. Data Eng, 2014, 26(1): 120-130.

[14] CARLOS J. MARTÍN D, AYSE G. Real-time topic detection with bursty N-grams[C]//SNOW-DC@WWW 2014. c2014: 9-16.

[15] GEORGIOS P, SYMEON P, YIANNIS K. Two-level message clustering for topic detection in Twitter[C]//SNOW-DC@WWW 2014. c2014:49-56.

[16] ZHAO J J, WU W L, et al. A short-term prediction model of topic popularity on microblogs[C]//The COCOON 2013. c2013:759-769.

[17] DUAN Y, JIANG L, et al. An empirical study on learning to rank of tweets[C]//The 23rd International Conference on Computational Linguistics. Beijing, China, c2010:295-303.

[18] HONG Y, ZHANG Y, LIU T, et al. Topic detection and tracking review[J]. Journal of Chinese Information Processing, 2007,21(6):71-87.

[19] YANG L, LIN Y, LIN H. Micro-blog hot events detection based on emotion distribution[J]. Journal of Chinese Information Processing, 2012, 26(1):84-83.

[20] ZHANG J. Research on the model and platform of hotspot detection based on micro-blog[D]. Wuhan: Huazhong University of Science & Technology, 2010.

[21] YANG G C. Research of hot topic discovery strategy on micro logging platforms[D]. Hangzhou: Zhejiang University, 2011.

[22] SUN J M, TANG J. A survey of models and algorithms for social influence analysis[M]//Social Network Data Analytics, 2011:177-204.

[23] **徐建民**, **张猛**, **吴树芳**. **基于话题的事件相似度计算**[J].**计算机工程与设计**, 2014, 35(4):1193-1197.
XU J M, ZHANG M, WU S F. Event similarity calculation based on topic[J]. Computer Engineering and Design, 2014, 35(4):1193-1197.

[24] RUMI G, KRISTINA L. Predicting influential users in online social network[C]//The Fourth Social Network Analysis. c2010.

[25] SABIDUSSI G. The centrality index of a graph[J]. Psychometrika, 1966, 31(4):581-603.

[26] NEWMAN M E. A measure of betweenness centrality based random walks[J]. Social Networks, 2005, 27(1): 39-54.

[27] DING Z Y, ZHOU B, JIA Y, et al. Topical influence analysis based on the multi-relational network in microblogs[J]. Journal of Computer Research and Development,2013,50(10):2155-2175.

[28] JOSEPH P, TURIAN L R, et al. Word representations: a simple and general method for semi-supervised learning[C]//ACL.c2010:384-394.

[29] YOSHUA B, REJEAN D, PASCAL V, et al. A neural probabilistic language model[J]. Journal of Machine Learning Research (JMLR), 2003, 3:1137-1155.

[30] MIKOLOV T, LE Q V, SUTSKEVER I. Distributed representations of sentences and documents[C]//ICML. c2014: 1188-1196.

[31] Available online[EB/OL]. https://github.com/fxsjy/jieba.

[32] **徐恪**,**张赛**,**陈昊**, **等**. **在线社会网络的测量与分析**[J]. **计算机学报**, 2014,37(1):165-188.
XU K, ZHANG S, CHEN H, et al. Measurement and analysis of online social networks[J]. Chinese Journal of Computers, 2014,37(1): 165-188.

作者简介：

陈福（1973-），男，辽宁朝阳人，北京外国语大学副教授，主要研究方向为下一代互联网及其管理、跨语言网络空间信息采集与分析、进程代数。

林闯（1948-），男，辽宁沈阳人，清华大学教授、博士生导师，主要研究方向为计算机网络、系统性能评价、安全分析和随机 Petri 网。

薛超（1988-），男，陕西渭南人，清华大学博士生，主要研究方向为网络体系结构的性能评价与优化、云计算虚拟资源调度等。

徐月梅（1985-），女，广西梧州人，博士，北京外国语大学讲师，主要研究方向为数据中心网等。

孟坤（1980-），男，河南洛阳人，清华大学助理研究员，主要研究方向为性能评价和随机模型。

倪艺函（1994-），女，江苏连云港人，北京外国语大学博士生，主要研究方向为进程代数。