ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 32,No. 5,Sep. 2017,pp. 1052－1060
DOI:10.16337/j.1004-9037.2017.05.023

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

基于特征词向量的短文本聚类算法

刘　欣[1]　佘贤栋[2]　唐永旺[1]　王　波[1]

(1.解放军信息工程大学信息系统工程学院,郑州，450002;2.中国人民解放军92899部队，宁波，315200)

摘　要：针对互联网短文本特征稀疏和速度更新快而导致的短文本聚类性能较差的问题,本文提出了一种基于特征词向量的短文本聚类算法。首先,定义基于词性和词长度加权的特征词提取公式并提取特征词代表短文本;然后,使用Skip-gram模型(Continous skip-gram model)在大规模语料中训练得到表示特征词语义的词向量;最后,引入词语游走距离(Word mover's distance,WMD)来计算短文本间的相似度并将其应用到层次聚类算法中实现短文本聚类。在4个测试数据集上的评测结果表明,本文方法的效果明显优于传统的聚类算法,平均F值较次优结果提高了56.41%。

关键词:短文本;特征词;词向量;相似度计算;聚类

中图分类号:TP391.1　　**文献标志码**:A

Short Text Clustering Based on Feature Word Embedding

Liu Xin[1], She Xiandong[2], Tang Yongwang[1], Wang Bo[1]

(1. School of Information and Systems Engineering, PLA Information Engineering University, Zhengzhou, 450002, China; 2. 92899 Troops, PLA, Ningbo, 315200, China)

Abstract: Aiming at the problem of poor clustering performance for short text caused by sparse feature and quick updating of short text on the internet, a short text clustering algorithm based on feature word embedding is proposed in this paper. Firstly, the formula for feature word extraction based on word part-of-speech(POS) and length weighting is defined and used to extract feature words as short texts. Secondly, the word embedding that represents semantics of the feature word is gained by means of the training in large scale corpus with continous skip-gram model. Finally, word mover's distance is introduced to calculate the similarity between short texts and applied in the hierarchical clustering algorithm to realize the short text clustering. The evaluation results on four testing datasets show that the proposed algorithm is significantly superior to traditional clustering algorithms with the mean F of 58.97% higher than the secondly best result.

Key words: short text; feature word; word embedding; similarity calculation; clustering

引　言

根据2014年《中国社交类应用用户行为分析研究报告》,即时通信在整体网民中的覆盖率为89.3%,社交网站覆盖率为61.7%,微博覆盖率为43.6%。网民通过上述应用相互联系和自由地发表

基金项目:国家社会科学基金(14BXW028)资助项目。
收稿日期:2015-11-19;**修订日期**:2016-04-06

言论,从而产生了海量短文本信息。面对海量短文本信息,如何通过计算机技术挖掘与分析其中所蕴藏的信息资源具有重要的意义。文本聚类作为自然语言处理中一种无监督的机器学习方法,可以自动对文本进行有效地组织与划分,进而得到对这些杂乱无章的文本信息的整体认知。

已有的聚类方法[1-2]多数是利用传统的向量空间模型(Vector space model,VSM)来表示文本,存在如下问题:(1)依赖词形相似性度量文本间的相似度,没有考虑文本中同义词对句子间相似度的贡献,无法准确计算句子之间深层的语义相似度。比如"喜欢"和"爱"虽然意思相近,但是由于计算两个句子相似度时不会统计这两个词,因此不会提高句子之间的相似度,从而影响了文本聚类性能;(2)由于短文本长度较短,采用传统的 VSM 表示文本时会出现严重的特征稀疏问题。为了解决上述问题,研究人员分别提出了基于知识库的文本聚类方法[3-5]和基于主题模型的文本聚类方法[6-9]。

基于知识库的文本聚类方法利用知网(HowNet)、WordNet 和维基百科等知识库丰富文本的特征,挖掘词语之间的关系提高聚类效果。Chen[3]通过知网的概念属性扩展文本主题的特征关键词,在一定程度上克服了短文本特征稀疏的问题,提高了短文本聚类的效果。Bouras[4]利用 WordNet 挖掘新闻文章中词语之间的关系,提出了一种基于 WordNet 的新闻文档聚类技术。Banerjee[5]借助维基百科来丰富短文本的特征,提高短文本聚类的准确度。网络语言往往具有奇异性和动态性的特点。奇异性是指现有的词语被赋予了新的意义,例如使用"酱紫"表示"这样子","沙发"表示"第一个跟帖的人","表"表示"不要"等;动态性是指网络语言变化快,每年都会不断产生新的词语,例如,2014 年流行的"小鲜肉"、"不造"、"涨姿势"等,2015 年流行的"然并卵"、"习马会"等。由于知识库更新速度慢,很多新词都无法及时收录,所以仅通过知识库对网络中的短文本进行处理,肯定会影响短文本聚类效果。基于主题模型的文本聚类方法的基本思想是利用主题模型对文本建模,将文本从高维特征空间转换到低维语义主题空间,从而克服传统聚类方法中特征向量维度高的问题。文献[6,7]分别采用隐含狄利克雷分配(Latent Dirichlet allocation,LDA)模型和 Biterm 主题模型(Biterm topic model,BTM)对文本建模,得到文本的主题分布,再结合 K-means 聚类算法实现文本聚类。文献[8,9]分别对 LDA 模型进行改进,提高了聚类效果。虽然基于主题模型的聚类方法在一定程度上克服了前两类方法的不足,但是 Mikolov[10]从大量的实验中发现 LDA 和概率性潜在语义分析(Probabilististic latent semantic analysis,PLSA)等主题模型不适用于大规模数据的训练和处理。另外,LDA 等模型都是假设数据服从指数分布,偏重于从高频数据中归纳语义,忽略了低频词的存在,而互联网上的数据服从的却是长尾分布(Long tail)[11],影响了上述模型应用于互联网短文本数据的性能。

词向量(Word embedding)是一种分布式的低维实数向量,由神经网络语言模型训练大规模语料生成,其作用是让相关或者相似的词语在距离上更接近[12]。例如计算"中国"、"人工"、"北京""和"智能"4 个词语两两之间的相似度时,"中国"和"北京"、"人工"与"智能"的相似度值明显高于其他组合。Mikolov[13]提出 Skip-gram 和连续词袋模型(Continuous bag-of-words,CBOW)模型训练词向量,这两个模型只有输入层、映射层和输出层,在训练语料时忽略单词的顺序,其研究表明,利用词向量对词语进行表示比传统的表示方法更加准确。虽然学术界已有很多成熟的词向量训练模型[14-17],但是 Skip-gram 和 CBOW 模型具有简易性和高效性的特点,仍是目前应用最流行的词向量训练模型。由于词向量是神经网络语言模型经过大规模的语料训练而来,所以利用其对文本表示不受短文本的特征稀疏和知识库更新速度的制约。例如在电影评论中出现的两句话"港囧快把我笑死了,赞一个"和"夏洛特烦恼是抄袭来的",从字面上看两句话无共现词,而且"港囧"和"夏洛特烦恼"都属于新词,HowNet 还未及时收录,所以无法得到这两句话的相似度。但是,这两个词在电影娱乐语料中往往与电影共现,所以经过这些语料训练得到的"港囧"与"夏洛特烦恼"的词向量就带有一定的语义相似性,进而可以衡量句子的语义相似度。为了利用词向量更准确地计算文本之间的相似度,Kusner[18]根据词向量的特点,将不同文本包含的词向量欧氏距离之和的最小值作为文本之间的语义相似度,该方法记为词语游走距离(Word mover's distance,WMD)。

考虑到词向量的优点，本文提出了一种基于特征词向量(Feature words embedding)的短文本聚类方法。首先，将每个短文本看作一个类簇，提取每个类簇的特征词代表类簇，这样可以有效地降低特征的维度；然后通过在大规模数据集中训练 Skip-gram 模型获取带有特征词语义信息的词向量，进而将类簇向量化；最后引入 WMD 计算类簇之间的语义相似度，并将其用于层次聚类算法完成短文本聚类。实验结果表明本文提出的方法切实可行，且相较于传统的基于向量空间模型和主题模型的聚类方法，本文方法的聚类效果得到了显著提高。

1 基于特征词向量的短文本聚类方法

基于特征词向量的短文本聚类方法的基本流程如图 1 所示，主要包括类簇向量化、基于特征词向量的短文本相似度计算和层次聚类 3 个部分。

1.1 类簇向量化

类簇向量化包括类簇特征词提取和训练词向量两个过程。

1.1.1 特征词提取

本文采用凝聚层次聚类算法对短文本进行聚类，在短文本层次聚类过程的初始化时将每个短文本看成一个类簇，单个类簇中包含的特征数量较少，当两个类簇合并为新类簇时，新类簇的特征会不断增多。如果将所有特征都用于类簇之间的相似度计算时，容易出现“富者越富”的现象，使大量的短文本聚集在少量类簇中。为此，本文提取特征词对类簇进行表示。一般情况下，名词和动词比其他词性的词重要。另外，词语包括的字数越多，包含的信息量越大。因此本文定义了一种基于词性和词长度的特征词权重计算公式，即

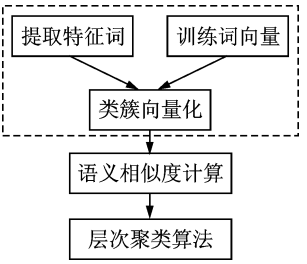

图 1 基于特征词向量的短文本聚类方法流程图

Fig. 1 Flow chart of short text clustering based on feature word embedding

$$\mathrm{Weight}(w_{id}) = \lambda \mathrm{Weight}_{\mathrm{pos}}(w_{id}) + (1-\lambda)\mathrm{Weight}_{\mathrm{len}}(w_{id}) \tag{1}$$

式中：$\mathrm{Weight}(w_{id})$ 表示词语 w_i 在文本 d 中的权重，$\mathrm{Weight}_{\mathrm{pos}}(w_{id})$ 表示 w_i 在文本 d 中的词性权重，$\mathrm{Weight}_{\mathrm{len}}(w_{id})$ 表示 w_i 在文本 d 中的长度权重，λ 和 $(1-\lambda)$ 为加权系数，λ 取经验值 0.6。$\mathrm{Weight}_{\mathrm{pos}}(w_{id})$ 和 $\mathrm{Weight}_{\mathrm{len}}(w_{id})$ 的具体计算公式为

$$\mathrm{Weight}_{\mathrm{pos}}(w_{id}) = \frac{tf(w_i,d)\times \log(N/n_{w_i}+0.01)\times \mathrm{pos}_{w_i}}{\sqrt{\sum_{w\in d}[tf(w,d)\times \log(N/n_w+0.01)\times \mathrm{pos}_{w_i}]^2}} \tag{2}$$

$$\mathrm{Weight}_{\mathrm{len}}(w_{id}) = \frac{tf(w_i,d)\times \log(N/n_{w_i}+0.01)\times \mathrm{len}_{w_i}}{\sqrt{\sum_{w\in d}[tf(w,d)\times \log(N/n_w+0.01)\times \mathrm{len}_{w_i}]^2}} \tag{3}$$

式中：$tf(w_i,d)$ 表示特征 w_i 在文本 d 中的词频；N 表示文本集中文本的总数；n_{w_i} 表示文本集中出现第 i 个词语的文本数；pos_{w_i} 表示词性标注加权值，len_{w_i} 表示词长度加权值，其具体定义为

$$\mathrm{pos}_{w_i} = \begin{cases} 1.5 & w_i \text{ is } 'v' \text{ or } 'n' \\ 1 & \text{其他} \end{cases} \tag{4}$$

$$\mathrm{len}_{w_i} = \begin{cases} 1.5 & \mathrm{len}(w_i) > 2 \\ 1 & \text{其他} \end{cases} \tag{5}$$

式中：当 w_i 为名词或者动词的时候 pos_{w_i} 取 1.5，否则 pos_{w_i} 取 1。当 w_i 包含的字数大于 2 时，$\mathrm{len}(w_i)$ 取 1.5，否则取 1。

在类簇合并过程中,如果将两个类簇的特征词不加区别地合并在一起,就会出现大量的短文本聚集在少数类簇中的情况,因此本文定义以下特征词权重更新公式。

$$c_{\mathrm{new}}=\{\{v(w_i)\}\mid \max_K(\mathrm{Weight}(w_i)),w_i\in(c_l+c_k)\} \tag{6}$$

$$\mathrm{Weight}(w_{i\,\mathrm{new}})=\frac{M\cdot\mathrm{Weight}(w_{ik})+N\cdot\mathrm{Weight}(w_{il})}{M+N} \tag{7}$$

式中:c_{new}表示由 c_l 和 c_k 合并成的新类簇,$v(w_i)$表示新类簇中特征词的向量,$\mathrm{Weight}(w_{i\,\mathrm{new}})$表示 w_i 在新类簇 c_{new}中的权重,M 和 N 分别表示类簇 c_k 和 c_l 中文本的数量。通过式(7)可以增大两个类簇公共特征词的权重,同时减少只在一个类簇中出现的特征词的权重。当新类簇包含的特征词个数大于 K 时,取权重较大的 K 个特征词代表该类簇。

1.1.2　训练词向量

本文使用 Mikolov 提出的 Skip-gram 模型训练词向量,Skip-gram 模型可以通过 Hierarchical Softmax 和 Negative Sampling 两种框架构造实现。本文使用的是基于 Hierarchical Softmax 构造的 Skip-gram 模型,其结构如图 2 所示。

由图 2 可知,Skip-gram 模型包含输入层、投影层和输出层 3 层结构,其原理是在已知当前词 $w(t)$ 的前提下预测其上下文,该模型的目标函数 L 为

$$L=\sum_{w\in V}\log p(\mathrm{Context}(w)\mid w) \tag{8}$$

$$p(\mathrm{Context}(w)\mid w)=\prod_{u\in\mathrm{Context}(w)}p(u\mid w) \tag{9}$$

式中:V 表示数据集对应的词典,$\mathrm{Context}(w)$表示与 w 距离小于 C 的上下文,C 一般取 5 到 10 效果较好。经过该模型的训练,就可以得到带有丰富语义信息的词向量,具体训练过程参见文献[13]。

1.2　基于特征词向量的短文本相似度计算

采用 WMD 计算类簇间的语义相似度,WMD 将一个类簇的特征词向量全部"流向"另一个类簇的特征词向量所经过的距离总和的最小值作为两个类簇之间的语义相似度,其示意图如图 3 所示。

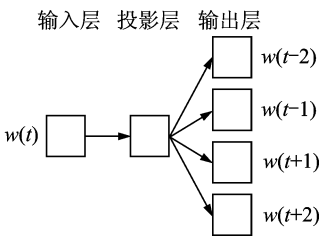

图 2　Skip-gram 模型

Fig. 2　Skip-gram model

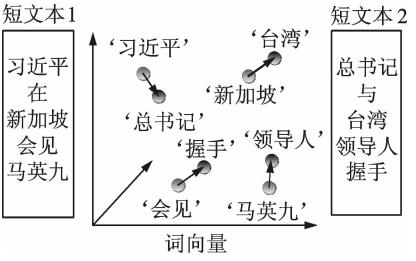

图 3　WMD 计算类簇相似度示意图

Fig. 3　Illustration of the calculation of cluster similarity with WMD

例如计算上图中两个短文本的语义相似度,需要将两个短文本中的非停用词全部映射到同一个向量空间中,只有当'习近平'流向'总书记','新加坡'流向'台湾','会见'流向'握手'和'马英九'流向'领导人'时,所有词向量需要经过的距离之和最短,则将该距离之和作为短文本之间的语义相似度。

1.2.1　特征词之间的语义相似度计算

利用 WMD 计算类簇间语义相似度需要先计算特征词之间的语义相似度。给定 n 个特征词的词向量矩阵 $\boldsymbol{X}\in\mathbf{R}^{d\times n}$,$\boldsymbol{x}_i\in\mathbf{R}^{\,d}$表示特征词 w_i 在 d 维空间的词向量,则可以用欧氏距离来衡量特征词之间的语义相似度,即

$$L(w_i,w_j)=\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2 \tag{10}$$

式中：$L(w_i, w_j)$表示 w_i 和 w_j 的语义相似度，$\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 分别为 w_i 和 w_j 的词向量。$L(w_i, w_j)$的值越小，说明两个特征词的语义相似度越大。

1.2.2 短文本之间语义相似度的计算

利用词向量得到特征词之间的相似度后，WMD 通过计算短文本 d 中所有特征词“流向”短文本 d' 中所有特征词的距离和的最小值来度量 d 和 d'之间的相似度。记 d 中 w_i 的出流总和为

$$d_{w_i} = \frac{tf(w_i, d)}{\sum_{j=1}^{n} tf(w_j, d)} \tag{11}$$

式中：$tf(w_i, d)$表示特征词 w_i 在 d 中出现的次数，n 表示 d 中所有词的个数。首先，令 d 中的特征词 w_i 流向 d'中的任意特征词，矩阵 $\boldsymbol{T} \in \mathbf{R}^{n \times n}$作为分流矩阵，$T_{ij}$代表 d 中 w_i 转化为 d'中 w_j 的程度，在转换结束后要保证从 w_i 出流总和为 d_{w_i}，即 $\sum_{w_j} T_{w_i w_j} = d_{w_i}$。同样 d'中 w_j 的入流总和为 d_{w_j}，即 $\sum_{w_i} T_{w_i w_j} = d_{w_j}$，则 d 中所有的特征词与 d'中所有特征词的欧氏距离之和的最小值为短文本之间的语义相似度，即

$$\mathrm{sim}(d, d') = \min_{T \geqslant 0} \sum_{i,j=1}^{n} T_{w_i w_j} L(w_i, w_j) = \min_{T \geqslant 0} \sum_{i,j=1}^{n} T_{w_i w_j} \left\| \boldsymbol{x}_i - \boldsymbol{x}_j \right\|_2 \tag{12}$$

式(12)必须满足

$$\sum_{j=1}^{n} T_{w_i w_j} = d_{w_i} \quad \forall i \in \{1, 2, \cdots, n\} \tag{13}$$

$$\sum_{i=1}^{n} T_{w_i w_j} = d_{w_j} \quad \forall j \in \{1, 2, \cdots, n\} \tag{14}$$

式(12)的具体求解过程可参见文献[18]。

1.3 层次聚类算法流程

本文利用基于词向量的语义相似度计算方法(即式(12))构造层次聚类算法，其流程如下所示。

算法：基于特征词向量的层次聚类算法

输入：计算文本集中每个文本 d 中的词语权重，将每个文本看成是一个初始类簇，$c_i = \{v(w_j) \mid w_j \in d_i\}$，所有类簇构成一个聚类集合 $C = \{c_1, c_2, \cdots, c_i, \cdots, c_n\}$，聚类个数 P

输出：P 个类簇集合

$m = n + 1$

while $|C| > P$：

for $(c_k, c_l) \in C \times C$：

$\mathrm{sim}(c_k, c_l) = \min_{T \geqslant 0} \sum_{i,j=1}^{n} T_{w_i w_j} \left\| \boldsymbol{x}_i - \boldsymbol{x}_j \right\|_2$

$(c_{n1}, c_{n2}) = \max\{\mathrm{sim}(c_k, c_l) \mid (c_k, c_l) \in C \times C\}$

$c_m = c_{n1} \bigcup c_{n2}$

for w_i in c_m

$$\mathrm{weight}(w_i, c_m) = \frac{M \cdot \mathrm{weight}(w_i, c_{n1}) + N \cdot \mathrm{weight}(w_i, c_{n2})}{M + N}$$

$C = C \backslash \{c_{n1}, c_{n2}\} \bigcup \{c_j\}$

$m = m + 1$

end

2 实验与结果

2.1 实验数据

实验数据来自于 3 个英文数据集和 1 个中文数据集，具体为：

(1) 20newsgroup_subjects:本文根据 Bora[19] 的数据构造方法,从英文文本分类通用数据集 20newsgroup 的 20 个类别中选取 5 个类别的主题内容作为实验数据集,记为 20newsgroup_subjects,具体信息如表 1 所示。

(2) CLing-2002 数据集:由 Makagonov[20] 等收集计算语言学领域 CICLing2002 会议的 4 个类别的 48 篇文章摘要组成,具体为 Linguistics 领域的 11 个摘要,Ambiguity 领域的 15 个摘要,Lexicon 领域的 11 个摘要和 Text Processing 领域的 11 个摘要。该数据集共有 3 382 个词语,平均每篇摘要 70.45 个词[21]。

(3) KnCr 数据集:由 D. Pinto[22] 收集医学领域的 16 个类别的 900 篇文章摘要组成,本文选取其中 blood,colon,genetic studies,skin 和 stomach 等 5 个类别作为实验数据[21]。

(4) Weibo_topic 数据集:从中文微博情感倾向性分析研究领域的通用数据 NLP&&CC2012 中选取 5 个事件数据作为实验数据,具体信息如表 2 所示。

表 1　20newsgroup_subjects 数据集

Tab. 1　20newsgroup_subjects dataset

数据集	短文本数量	平均长度(词)
talk. politics. mideast	70	8.85
comp. windows. x	60	6.61
rec. sport. baseball	70	6.45
sci. med	78	6.15
soc. religion. christian	71	5.46

表 2　Weibo_topic 数据集

Tab. 2　Weibo_topic dataset

数据集	短文本数量	平均长度(词)
“菲军恶意撞击中国渔船”	86	23.24
“官员财产公示”	82	25.33
“韩寒方舟子之争”	87	29.35
“皮鞋果冻”	82	14.62
“中国教师收入全球几垫底”	81	26.51

2.2　评测标准

作为准确率(Precision,P)和召回率(Recall,R)的综合考量,F 值是文本聚类领域最常用的评价指标。本文采用平均 F 值作为实验结果的评价指标。

$$P(i,j)=\frac{\text{类别 } i \text{ 在簇 } j \text{ 中的个数}}{\text{簇 } j \text{ 中的文本个数}} \tag{15}$$

$$R(i,j)=\frac{\text{类别 } i \text{ 在类簇 } j \text{ 中的个数}}{\text{类别为 } i \text{ 的所有文本个数}} \tag{16}$$

$$F(i,j)=\frac{2P(i,j)R(i,j)}{P(i,j)+R(i,j)} \tag{17}$$

式中:$P(i,j)$,$R(i,j)$和 $F(i,j)$分别表示类别 i 在类簇 j 中的准确率、召回率和 F 值,取类别 i 在各个类簇 j 中 $F(i)$的最高值作为其最终的 F 值。在聚类评测中,整体聚类效果由平均 F 值进行评价,定义为

$$F=\sum_i \frac{|i|\times F(i)}{\sum_i |i|} \tag{18}$$

2.3　实验对比方法

为了验证本文方法的可行性与有效性,使用 VSM[23],LDA[24],BTM[25] 和成分计算网络(Componential counting grids,CCG)[26] 这 4 种模型对文本进行表示。

(1) VSM 模型:将文档映射为一个特征向量 $\boldsymbol{V}(d)=(t_1,\omega_1(d);\cdots;t_n,\omega_n(d))$,其中 t_i $(i=1,2,\cdots,n)$为互不相同的词项,$\omega_i(d)$为 t_i 在 d 中的权值,用 TF-IDF 计算其权重。

(2) LDA 模型:通过对文档集合建模,得到每个文本的主题分布向量,挖掘出潜在的语义信息,在一定程度上弥补了单纯利用词频信息表示文本带来的信息丢失的不足。

(3) BTM 模型:统计任意两个词语组成的共现词对,以共现词对作为单位建模,获取文档-主题和主

题-词语的概率分布表示文本。

(4) CCG 模型:将文档看作是多个主题的分布,通过网格对文档中主题的位置分布进行建模,利用文档的概率分布和网格的位置表示文本。

利用余弦相似度计算短文本之间的语义相似度,采用目前主流的层次聚类算法进行聚类作为对比实验,LDA 和 BTM 模型中的参数按照文献[7]进行设置,主题数 S 设为 12,超参数 α 和 β 取经验值 $\alpha=50/S$,$\beta=0.01$,迭代次数为 2 000。根据文献[26]设置网格和窗口的大小,评测结果分别记为 VSM,LDA,BTM 和 CCG。

2.4 实验设置与分析

本文中的英文词向量选取 Mikolov 公布的词向量,中文词向量由 Skip-gram 模型在“搜狗实验室全网新闻数据”和“华为实验室微博数据”[27]上训练所得。从词向量中在提取各类簇中所有特征词的词向量后,将类簇向量化,分别利用余弦相似度和 WMD 计算短文本间的相似度,采用层次聚类算法进行聚类,根据数据集的类别数设定聚类个数 P,评测结果分别记为 FWV-COS 和 FWV-WMD。本文设置了 2 组实验,第 1 组实验是在 4 个数据集上对上述 5 种方法进行综合评测;第 2 组实验分析参数 K 对聚类结果的影响,同时检验特征词权重计算公式的有效性。

2.4.1 综合对比实验

以平均 F 值作为评价指标,在以上 4 个数据集中使用上述 5 个方法的实验结果如表 3 所示。从表 3 可看出,基于主题模型的聚类评测结果要优于基于 VSM 模型的聚类评测结果,这说明在数据特征较少时,VSM 这种忽略同义词之间的关系,通过计算 TF-IDF 值来提取特征的方法会受到短文本数据稀疏的严重影响;基于 BTM 模型的聚类评测结果优于基于 LDA 模型的评测结果,这是因为 LDA 直接对文档建模,需要统计文档-主题的概率分布,在短文本特征词较少时,不具备统计意义,而 BTM 模型对文本任意的共现词对建模,弥补了 LDA 模型的缺点;基于 CCG 的方法在 CLing-2002 和 KnCr 两组数据中的评测结果优于基于 BTM 的方法,这是因为这两组数据中的类别同属于一个大领域,类别具有相似性,且数据量较大适合利用网格对文档的主题建模;而 20newsgroup_subjects 和 weibo_topics 数据集数据稀疏,类别也相差很大,适合利用共现词对建模,使基于 BTM 的方法聚类效果高于基于 CCG 的方法。而本文方法的评测结果要远远优于对比方法,在 4 个数据集的实验中,性能较次优结果平均提高了 56.41%。在 4 个数据集中,20newsgroup_subjects 数据集中平均每个文本包含的词数最少,数据严重稀疏,但是 FWV-WMD 相较于 BTM 的平均 F 值提高了 93.18%,这充分说明从大规模数据中训练得到的词向量带有丰富的语义信息,不受短文本数据稀疏的影响。从 FWV-COS 和 FWV-WMD 的评测结果可以看出,WMD 比传统的余弦相似度计算方法更能挖掘词向量之间的语义相似性,从而提高了聚类效果。

2.4.2 特征词参数和特征权重参数分析

为了使本文方法的聚类效果最佳,对特征词参数 K 以及特征权重参数 pos_{w_i} 和 len_{w_i} 进行了分析,结果如图 4 所示。其中以 weibo_topics 数据集作为实验对象,令 pos_{w_i} 和 len_{w_i} 同时为 1 或 1.5,K 在[5,100]内以 5 为步长进行聚类,评测结果分别记为 FWV-WMD-1 和 FWV-WMD-1.5。

由图 4 可以看出,平均 F 值随着选取的特征词个数 K 的变化而变化。K 在 50 到 60 范围内时评测结果较好,这说明当特征词个数较少时,无法准确表示类簇从而使聚类效果差;当特征词个数较多时,类簇的主题信息不突出,在聚类时容易出现多数短文本聚集在少数几个类簇中的情况,无法得到最佳聚类效果。另外,FWV-WMD-1.5 的评测结果要普遍优于 FWV-WMD-1 的评测结果,这说明本文提出的基于词性和词语长度加权的特征词提取方法切实可行,能更加准确地提取类簇的特征词。

表 3　4 个数据集的评测结果

Tab. 3　Evaluation results on four datasets

数据集	20newsgroup_subjects	CLing-2002	KnCr	weibo_topics
VSM	0.164	0.263	0.150	0.332
LDA	0.215	0.328	0.325	0.333
BTM	0.323	0.355	0.362	0.412
CCG	0.226	0.362	0.382	0.396
FWV-COS	**0.619**	**0.531**	**0.472**	**0.604**
FWV-WMD	**0.624**	**0.546**	**0.501**	**0.620**

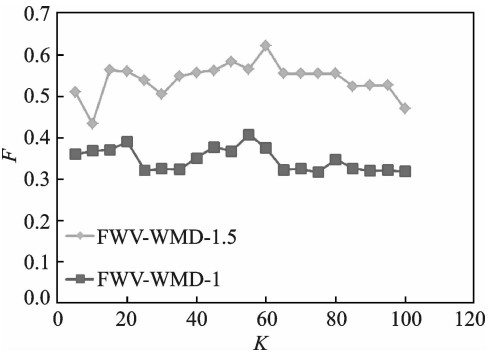

图 4　特征词个数和权重参数分析

Fig. 4　Analysis of feature word number and weight parameter

3　结束语

本文提出了一种基于特征词向量的短文本聚类方法。将词语词性和长度加权可以更加准确地提取类簇的特征词,避免了在聚类过程中出现多数短文本聚集在一个类别中的情况;利用词向量的性质解决短文本数据稀疏和更新速度快对文本聚类带来的问题;引入 WMD 使短文本之间的相似度计算更加准确,提高了聚类性能。实验结果表明,本文方法切实可行,可以显著提高短文本的聚类效果。由于本文只采用 Skip-gram 模型训练词向量且仅将上述方法应用在层次聚类算法中,因此下一步准备利用多种神经网络语言模型训练词向量,并将其应用在其他主流聚类算法中(比如 K-means 和 FCM),以寻求短文本聚类的最佳效果。

参考文献:

[1] 索红光, 王玉伟. 一种用于文本聚类的改进 k-means 算法[J]. 山东大学学报: 理学版, 2008, 43(1): 61-64.
Suo Hongguang, Wang Yuwei. An improved k-means algorithm for document clustering[J]. Journal of Shandong University (Natural Science), 2008, 43(1): 61-64.

[2] 张霞, 王素贞, 尹怡欣. 基于模糊粒度计算的 K-means 文本聚类算法研究[J]. 计算机科学, 2010, 37(2):209-211.
Zhang Xia, Wang Suzhen, Yin Yixin. Research of text clustering based on fuzzy granular computing[J]. Computer Science, 2010, 37(2):209-211.

[3] Chen X, Zhang Y, Cao L, et al. An improved feature selection method for Chinese short texts clustering based on HowNet [J]. Lecture Notes in Electrical Engineering, 2014,277: 635-642.

[4] Bouras C, Tsogkas V. A clustering technique for news articles using WordNet[J]. Knowledge-Based Systems, 2012, 36: 115-128.

[5] Banerjee S, Ramanathan K, Gupta A. Clustering short texts using Wikipedia[C]// Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information retrieval. New York: ACM, 2007: 787-788.

[6] 史剑虹,陈兴蜀,王文贤. 基于隐主题分析的中文微博话题发现[J]. 计算机应用研究,2014,31(3):700-704.
Shi Jianhong, Chen Xingshu, Wang Wenxian. Discovering topic from Chinese microblog based on hidden topics analysis[J]. Application Research of Computers,2014,31(3):700-704.

[7] 汤秋莲. 基于 BTM 的短文本聚类[D]. 合肥:安徽大学,2014.
Tang Qiulian. Short text clustering method based on BTM[D]. Hefei:Anhui University,2014.

[8] 王少鹏, 彭岩, 王洁. 基于 LDA 的文本聚类在网络舆情分析中的应用研究[J]. 山东大学学报 (理学版), 2014, 49(9): 129-134.
Wang Shaopeng,Peng Yan,Wang Jie. Research of the text clustering based on LDA using in network public opinion analysis [J]. Journal of Shandong University(Natural Science),2014, 49(9): 129-134.

[9] Yin J, Wang J. A Dirichlet multinomial mixture model-based approach for short text clustering[C]// Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2014: 233-242.

[10] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. Computer Science, 2013:1301.3781.
[11] Murphy K P. Machine learning—A probabilistic perspective[M]. Cambridge, Massachusetts, London, England: The MIT Press, 2012 2-39.
[12] Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning [C]// Proceedings of the 25th International Conference on Machine Learning. New York: ACM, 2008: 160-167.
[13] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]// Advances in Neural Information Processing Systems. USA: The MIT Press, 2013: 3111-3119.
[14] Huang E H, Socher R, Manning C D, et al. Improving word representations via global context and multiple word prototypes [C]// Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. USA: Association for Computational Linguistics Stroudsburg, 2012: 873-882.
[15] Dhillon P, Foster D P, Ungar L H. Multi-view learning of word embeddings via cca[C]// Advances in Neural Information Processing Systems. USA: The MIT Press. 2011: 199-207.
[16] Levy O, Goldberg Y. Dependency based word embeddings [C]// Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. USA: Association for Computational Linguistics Stroudsburg, 2014, 2: 302-308.
[17] Yi Yang, Jacob Eisenstein. Unsupervised multi-domain adaptation with feature embeddings[J]. NAACL HLT, 2015: 672-682.
[18] Kusner M J, Sun E D U Y, Kolkin E D U N I, et al. From word embeddings to document distances[C]// Proceedings of the 32nd International Conference on Machine Learning. Washington DC: Microtome Publishing, 2015: 957-966.
[19] Bora N N, Mishra B S P, Dehuri S. Heuristic frequent term-based clustering of news headlines[J]. Procedia Technology, 2012, 6: 436-443.
[20] Makagonov P, Alexandrov M, Gelbukh A. Clustering abstracts instead of full texts[C]// International Conference on Text, Speech and Dialogue. [S. l.]: Springer Berlin Heidelberg, 2004, 3206(C): 129-135.
[21] Shrestha P, Jacquin C, Daille B. Clustering short text and its evaluation[C]// International Conference on Computational Linguistics and Intelligent Text Processing. [S. l.]: Springer Berlin Heidelberg, 2012, 7182: 169-180.
[22] Pinto D, Rosso P. KnCr: A short-text narrow-domain sub-corpus of medline[J]. Proceedings of TLH-ENC06, 2006: 266-269.
[23] Qimin C, Qiao G, Yongliang W, et al. Text clustering using VSM with feature clusters[J]. Neural Computing and Applications, 2015, 26(4): 995-1003.
[24] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation[J]. The Journal of Machine Learning Research, 2003, 3: 993-1022.
[25] Yan X, Guo J, Lan Y, et al. A biterm topic model for short texts[C]// Proceedings of the 22nd International Conference on World Wide Web. Switzerland: International World Wide Web Conferences Steering Committee, 2013: 1445-1456.
[26] Perina A, Jojic N, Bicego M, et al. Documents as multiple overlapping windows into grids of counts[C]// Advances in Neural Information Processing Systems. USA: The MIT Press, 2013: 10-18.
[27] Wang Hao, Lu Zhengdong, Li Hang, et al. A dataset for research on short-text conversation[C]// Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. USA: Association for Computational Linguistics Stroudsburg, 2013: 935-945.

作者简介:

刘欣(1990-),男,硕士研究生,研究方向:自然语言处理,E-mail:syliuxinlyl@163.com。

佘贤栋(1977-),男,工程师,研究方向:智能信息处理,E-mail:106386315@qq.com。

唐永旺(1981-),男,讲师,研究方向:智能信息处理、自然语言处理,E-mail:lhlaotang@163.com。

王波(1978-),男,副教授,研究方向:智能信息处理、网络协议分析,E-mail:505781538@qq.com。