计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(9):210−214 [doi: 10.15888/j.cnki.csa.006554]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

改进特征权重的短文本聚类算法①

马　存[1,2], 郭锐锋[2], 高　岑[2], 孙　咏[2]

[1](中国科学院大学, 北京 100049)
[2](中国科学院 沈阳计算技术研究所, 沈阳 110168)

摘　要: 短文本的研究一直是自然语言处理领域的热门话题, 由于短文本特征稀疏、用语口语化严重的特点, 它的聚类模型存在维度高、主题聚焦性差、语义信息不明显的问题. 针对对上述问题的研究, 本文提出了一种改进特征权重的短文本聚类算法. 首先, 定义多因子权重规则, 基于词性和符号情感分析构造综合评估函数, 结合词项和文本内容相关度进行特征词选择; 接着, 使用 Skip-gram 模型 (Continuous 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, 利用 RWMD 算法计算短文本之间的相似度并将其应用 K-Means 算法中进行聚类. 最后在 3 个测试集上的聚类效果表明, 该算法有效提高了短文本聚类的准确率.

关键词: 特征权重; 情感分析; 词向量; RWMD 距离

引用格式: 马存,郭锐锋,高岑,孙咏.改进特征权重的短文本聚类算法.计算机系统应用,2018,27(9):210–214. http://www.c-s-a.org.cn/1003-3254/6554.html

Short Text Clustering Algorithm with Improved Feature Weight

MA Cun[1,2], GUO Rui-Feng[2], GAO Cen[2], SUN Yong[2]

[1](University of Chinese Academy of Sciences, Beijing 100049, China)
[2](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)

Abstract: Short text research has been a hot topic in the field of natural language processing. Due to the sparseness of short texts and serious colloquialisms, its clustering model has the problems of high dimensionality, poor focus of theme, and unclear semantic information. In view of the above problems, this study proposes a short text clustering algorithm with improving the feature weight. Firstly, the rules of multi-factor weight are defined, the comprehensive evaluation function is constructed based on part-of-speech and symbolic sentiment analysis, and the feature words are selected according to the relevancy between the term and the text content. Then, a word skip vector model (continuous skip-gram model) trained in large-scale corpus to obtain a word vector representing the semantic meaning of the feature words. Finally, the RWMD algorithm is used to calculate the similarity between short texts and the K-means algorithm is used to cluster them. The clustering results on the three test sets show that the algorithm effectively improves the accuracy of short text clustering.

Key words: feature weight; emotion analysis; word vector; RWMD distance

1　相关工作

随着移动终端智能化的发展, 纷繁多样的短文本信息充斥着互联网的各个角落. 由于短文本信息少, 口语化严重, 网络新词多, 使用传统的文档聚类会导致向量空间模型高度稀疏, 缺乏语义信息, 所以需要针对短文本的固有特点寻求一种有效的模型表示和聚类方法.

传统的向量空间模型, 主要通过特征词和权重来表示短文本数据, 它的缺点也很明显, 它忽略了同义词

① 收稿时间: 2018-01-27; 修改时间: 2018-03-07; 采用时间: 2018-03-21; csa 在线出版时间: 2018-08-16

在语义中的贡献并且会出现特征稀疏的问题, 进而造成维数灾难. 为了解决短文本特征稀疏的问题, 一些学者研究了外部信息增强的方法, 对短文本特征进行扩展, 从而提高聚类效果[1−3]. 然而语义扩展方法并没有解决“维数灾难”的问题, 还带来了新的问题, 比如聚类的效果完全依赖于知识库的丰富程序, 无法识别新兴的网络新词, 比如 2016 年流行的“老司机”, “发车了”等. 另有一部分学者通过原始高维特征词空间映射到低维的潜在语义空间或主题空间, 挖掘文本潜在的语义结构[4−6]. 但这种模型忽略了低频词的贡献, 尤其是短文本中贡献度高的低频词, 导致上述模型应用于网络短文本中的效果很差.

词向量是一种基于大量未标注的语料学习而来的低维分布式实数向量, 充分挖掘了同义词之间的共现关系[7,8]. 基于此, 本文结合短文本的特点和词向量的优势, 提出一种改进的特征词权重并结合松弛词语移动距离 (RWMD) 的短文本聚类算法. 首先, 定义多因子权重规则, 如文本中词性和情感词, 对于情感词的处理主要包括文字和表情符号, 接着使用 Skip-gram 模型基于定义好的权重规则训练特征词向量, 最后引入 RWMD 距离计算文本之间的相似度并以此聚类. 实验结果表明本文提出的方法切实可行, 尤其是在网络短文本中效果明显.

2 改进的特征词向量及聚类模型框架

2.1 改进策略

短文本数据, 尤其是论坛帖子, 商品评论以及微博和微信的聊天记录, 形式复杂多样, 包含各种表情符号, 在数据预处理阶段不能简单的将表情符号当作噪声直接去除, 否则会失去一部分语义信息, 即情感信息; 另外由于数据包含的短文本的长度也大小不一, 因此关键词的位置因素也必须考虑在内; 再者就是词性对短文本的影响[9], 名词、动词、形容词和副词是文本特征的重要组成部分, 因此词性的贡献也不容忽视. 基于此, 本文在文献[8]中提出的特征权重计算法进行了修改, 提出一种融合表情符号、位置因素以及词性信息的多因子加权策略的关键词提取方法:

$$Weight(w)=\alpha Weight_{\text{pos}}(w)+\beta Weight_{\text{len}}(w)+\gamma Weight_{\text{sen}}(w) \tag{1}$$

式中, $Weight(w)$ 表示词语 w 在文本 d 中的权重, $Weight_{\text{sen}}$ 表示单词 w 在文本 d 中情感所占的权重, α, β, γ 为加权系数, 他们之和为 1. $Weight_{\text{pos}}$ 和 $Weight_{\text{len}}$ 的计算公式参考文献[8], $Weight_{\text{sen}}$ 的计算公式为:

$$Weight_{\text{sen}}(W_i)=\frac{tf(w_i,d)\times\log(\frac{N}{n_w}+0.01)\times sen_{w_i}}{\sqrt{\sum_{w\in d}[tf(w_i,d)\times\log(\frac{N}{n_w}+0.01)\times sen_{w_i}]^2}} \tag{2}$$

其中, $tf(w_i,d)$ 表示特征 W_i 在文本 d 中的词频; N 表示文本总数; 表示所有文本集中出现第 i 个词语的文本数量; sen_{w_i} 表示该词的情感加权值, 其具体值需要根据文献[10]的研究内容加以定义, 将表情符号归为 7 个情感类别, 结合实验用数据集, 分别统计每一类情感所占比例, 以此比例作为 sen_{w_i} 的加权值. 定义如表 1 所列.

表 1 情感类别系数

类别	高兴	喜爱	惊讶	焦虑	悲伤	生气	憎恨	No
代号	δ	ε	η	τ	ω	φ	ϕ	
sen	0.29	0.19	0.04	0.16	0.23	0.06	0.03	0

在预处理阶段, 当文本中含有表情符号时, 会根据表 1 中的希腊字母进行替换. 若一个短文本中含有多种表情符号, 则根据多个表情符号的权值综合计算其权重; 若一个文本中不含有表情符号, 则在特征词权重的计算公式中, 第 3 项将为 0. 即:

$$\gamma Weight_{\text{sen}}(w)=0 \tag{3}$$

此时, α 取经验值 0.6. 本文将此模型记为 EFA (Emotion Fusion Algorithm) 算法.

2.2 训练特征词向量

本文使用 Mikolov[11]提出的基于 Hierarchical Softmax 构造的 Skip-gram 模型训练词向量, 它主要包括 3 层结构: 输入层, 投影层和输出层, 目标函数 L 如式 (1) 所示:

$$L=\sum_{w\in V}\log p(Context(w)|w) \tag{4}$$

$$p(Context(w)|w)=\prod_{u\in Context(w)}p(u|w) \tag{5}$$

其中, V 是数据词典, $Context(w)$ 表示单词 w 的上下文窗口, 一般窗口值取 5 到 10 效果较好.

2.3 以特征词表征的短文本相似度计算

文本采用 RWMD 距离算法来计算文本之间的语义相似度, RWMD 算法是基于 WMD 算法放松限制条件来降低算法的复杂度[12]改进而来. RWMD 算法是将一个短文本的特征词向量全部流向另一个短文本的特征词向量所经过的距离总和的最小值作为两个短文本

之间的语义相似度.

2.3.1 特征词之间的语义相似度

RWMD 算法在计算文本的相似度之前需要先计算特征词之间的相似度, 衡量两个特征词之间的相似度使用欧式距离来计算, 即:

$$L(W_i, W_j) = \left\| x_i - x_j \right\|^2 \tag{6}$$

L 的值越小, 说明两个词越相近.

2.3.2 短文本之间的相似度计算

使用 RWMD 距离计算短文本 d 中所有特征词流向短文本 d'中所有特征词距离和的最小值作为短文本 d 和短文本 d'之间的相似度. 假设允许短文本 d 中的每个特征词可以流向 d'中的任意一个特征词, 矩阵 $\mathbf{T} \in \mathbf{R}^{n \times n}$是转移矩阵, 其中$T_{ij} \geqslant 0$, 表示词语 i 有多少转移到了词语 j, $C(i, j)$表示词语 i 和词语 j 之间的语义相似度, 目标函数为:

$$sim(d, d') = \min_{T \geqslant 0} \sum_{i,j=1}^{n} T_{ij} C(i, j) \tag{7}$$

约束条件为:

$$\sum_{j=1}^{n} T_{ij} = d_i, \ \forall i \in \{1, \cdots, n\} \tag{8}$$

2.4 K-means 聚类算法流程

输入: 实验所用的短文本数据集. 通过数据预处理, 并加权计算融合情感词权重的特征词集合, 并由 Softmax 模型训练而成的特征词向量.

输出: 具有 K 类的短文本集合.

Step 1. 指定聚类数目 K, 以及 K 个初始聚类中心.

Step 2. 指定 RWMD 算法为距离函数.

Step 3. 计算每个文本向量 d 与 K 个初始聚类中心的 RWMD 距离, 将每个文本向量 d 分配给距离最小的聚类中心.

Step 4. 重新计算新的 K 个聚类中心.

Step 5. 重复 Step 3 及 Step 4, 直到聚类中心小于阈值.

3 实验与结果分析

3.1 实验数据

本文采用了 3 种类型数据集: 微博数据、文本分类通用数据和 QQ 群聊天数据. 其中文本分类通用数据集从中选取 5 个类别的标题; 聊天记录数据人工标注出若干个聊天片段. 具体描述如表 2 所示.

表 2 数据集描述

通用	数量	长度	微博	数量	长度	聊天	数量	长度
交通	5342	8.23	类 1	8272	25.24	群 1	7417	36.23
医药	5725	8.67	类 2	8499	27.13	群 2	9282	33.76
军事	6191	8.28	类 3	8152	24.65	群 3	8737	27.81
法律	5568	8.17	类 4	8397	17.22	群 4	7259	41.51
教育	5915	9.49	类 5	8161	23.81	群 5	8726	37.29

3.2 评价指标

为了使结果更有对比性, 本文采用了文本聚类常用的准确率、召回率、和宏平均作为实验结果的评价指标:

$$P_{ij} = \frac{C_{ij}}{C_i}; \ R_{ij} = \frac{C_{ij}}{C_j}; \tag{9}$$

$$F_{ij} = \frac{2P_{ij}R_{ij}}{P_{ij} + R_{ij}}; \ F_{\text{macro}} = \frac{1}{m} \sum_{i=1}^{m} \max_j (F_{ij}) \tag{10}$$

其中, P_{ij}、R_{ij}和F_{ij}表示类别 i 在类簇 j 中的准确率、召回率和 $F1$ 值, C_i表示正确类别 i 中的文本数, C_j表示结果中类簇 j 中的文本数, C_{ij}表示结果中类簇 j 中原本属于类别 i 的文本数, 对于类簇 j 取各个类别中F_{ij}最高的作为类别 i 的 $F1$ 值, F_{macro}表示宏平均的结果, m 表示原始类别的个数.

3.3 实验结果与分析

本文使用 VSM, LDA 和 BTM 这 3 中模型对文本进行表示来验证模型的可行性和有效性, 分别将结果记为 KM-VSM、KM-LDA、KM-BTM, 本文提出的模型结果记作 KM-EFA. 其中 VSM 中使用 TF-IDF 作为特征权重, LDA 模型和 BTM 模型中主题数设为 15, 超参数α和β取经验值 50/K, β=0.01, 迭代次数为 2000.

3.3.1 对比实验

在上文中介绍的 3 个数据集上分别使用上述 4 种方法进行实验, 使用平均 F 值作为评价指标, 结果如表 3 所示. 从表中可以看出, 基于主题模型的聚类评测结果一般要好于基于 VSM 模型的聚类结果, 说明无法发现同义词之间语义关系的模型会受到短文本数据特征稀疏的影响; 基于 BTM 模型的聚类评测效果优于基

于 LDA 模型的聚类效果, 说明在短文本特征比较少的时候基于主题概率的统计方法统计出的数据意义不大. 其中模型 KM-EFA1 是不考虑情感因素只考虑词性和位置因素的评测结果, 而 KM-EFA2 是考虑了所有因素的评测结果. 对比发现, 本文提出的方法评测结果要优于对比方法, 在 3 个数据集的试验中, 性能比次优的结果平均提高了 13.62%, 从而验证了本模型使用情感加权更能挖掘出词之间的语义相似性, 从而提高聚类效果.

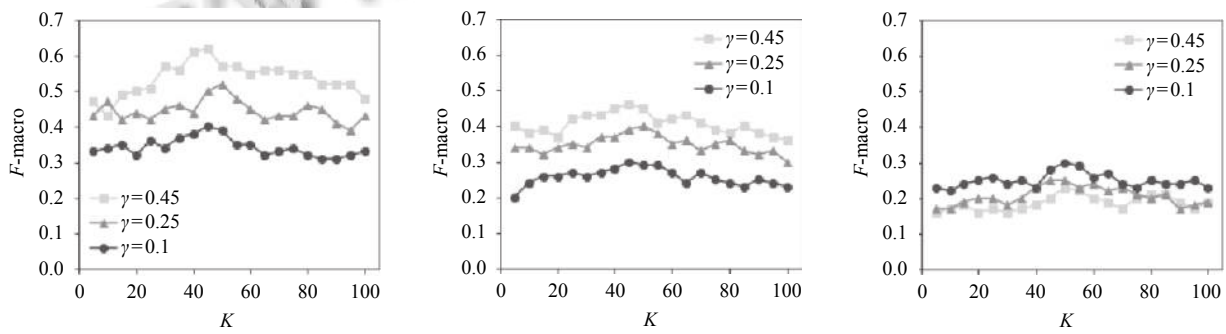
表 3 模型在数据集上的评测结果

数据集	通用	微博	聊天
KM-VSM	0.193	0.216	0.265
KM-LDA	0.324	0.391	0.325
KM-BTM	0.413	0.592	0.570
KM-EFA1	**0.424**	**0.586**	**0.562**
KM-EFA2	**0.462**	**0.613**	**0.611**

3.3.2 特征值参数与权重系数分析

为了校验特征词选择过程的参数 K 以及情感权重加权系数γ对聚类的影响, 本文在 3 个数据集上分别取γ等于 0.1、0.25 和 0.45, 同时对参数 K 在[5, 100]范围以步长为 5, 进行遍历, 结果如图 1 所示.

从图中可以看出, 当情感权重系数不同时, 随着 K 的变化, F 值也变得有所不同. 综合来说, 当特征 K 在[40, 50]之间时, F 值表现最好, 这是因为 K 太小时, 特征个数不足以表达完整的语义, 当 K 太大时, 句子的主题信息不明显, 会造成“富者越富”的现象, 影响聚类效果. 另外, 当数据集中表现情感的词比较多时, 情感权重的大小会直接影响聚类的好坏. 如微博和聊天数据含有大量情感词, 聚类的效果完全由情感权重决定, 但在普通的分类文本中情感权重越大聚类效果则越差.



(a) 聊天数据集 (b) 微博数据集 (c) 分类通用数据

图 1 特征个数与权重参数分析

4 结束语

本文融合情感加权的方法有效的提高了短文本的聚类效果, 尤其在微博微信等即时聊天的短文本数据中, 效果更好, 这是因为在这类文本中人们使用表情符号的频率相对普通文本较高, 此方法能充分挖掘符号下的语义信息. 但随着深入的研究, 这类文本中也充斥着大量的不规范用语, 如“狗带”, “一颗赛艇”等, 这些不规范用语对聚类结果产生一定的影响, 尤其是一些拆分字没有办法对其准确的表示, 比如“古月哥欠”, 表达的是胡歌, 但经过分词之后, 这几个字会变得毫无意义, 虽然这类词语出现频次较低, 但往往这类词语是短文本的核心语义, 同时用户故意使用这类词语一般均会涉及不正当言论, 是网络监督和舆情管理的重要分析方向. 因此, 对这种现象的研究, 具有重要的现实意义.

参考文献

1 Bouras C, Tsogkas V. A clustering technique for news articles using WordNet. Knowledge-Based Systems, 2012, 36(6): 115–128.

2 冶忠林, 贾真, 杨燕, 等. 基于语义扩展的句子相似度算法. 山西大学学报 (自然科学版), 2015, 38(3): 399–405.

3 Wei TT, Lu YH, Chang HY, *et al*. A semantic approach for text clustering using WordNet and lexical chains. Expert Systems with Applications, 2015, 42(4): 2264–2275. [doi: 10.1016/j.eswa.2014.10.023]

4 Qiu L, Xu J. A Chinese word clustering method using latent Dirichlet allocation and K-means. International Conference on Advances in Computer Science and Engineering. 2013. [doi: 10.2991/cse.2013.60]

5 吴敏. 网络短文本主题聚类研究[硕士学位论文]. 武汉: 华中科技大学, 2015.

6 王鹏, 高铖, 陈晓美. 基于 LDA 模型的文本聚类研究. 情报科学, 2015, 33(1): 63–68.
7 张群, 王红军, 王伦文. 一种结合上下文语义的短文本聚类算法. 计算机科学, 2016, 43(S2): 443–446, 450.
8 李天彩, 席耀一, 王波, 等. 一种改进的短文本层次聚类算法. 信息工程大学学报, 2015, 16(6): 743–748, 752. [doi: 10.3969/j.issn.1671-0673.2015.06.019]
9 韩普, 王东波, 刘艳云, 等. 词性对中英文文本聚类的影响研究. 中文信息学报, 2013, 27(2): 65–73. [doi: 10.3969/j.issn.1003-0077.2013.02.010]
10 刘伟朋, 陈雁翔, 孙晓. 基于表情符号的中文微博多维情感分类的研究. 合肥工业大学学报 (自然科学版), 2014, 37(7): 803–807. [doi: 10.3969/j.issn.1003-5060.2014.07.008]
11 Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. arXiv, 2013: 1301.3781.
12 Kusner M, Sun Y, Kolkin N, *et al*. From word embeddings to document distance. Proceedings of the 32nd International Conference on Machine Learning. Washington DC, USA: Microtome Publishing. 2015. 957–966.