**基于词向量和多特征语义距离的文本聚类算法**

**Text Clustering Algorithm Based on Word Vector and**

**Multi－feature Semantic Distance**

**--张 弛、张贯虹 重庆科技学院学报(自然科学版)**

**一、科学问题**

**1.1 本文所涉及科学问题**

文本聚类的算法。

**1.2 同行专家如何解决**

文本聚类的算法有基于词频统计的方法、基于主题模型的方法、基于知识库的方法和神经网络的方法。

**1.3 本文所解决的问题**

克服了传统文本特征向量表示方法存在的问题，如数据稀疏和网络词汇更新速度快等因素对文本聚类结果的影响

**1.4 本文解决方案效果**

算法要比其他几种算法的聚类效果 更好，不但能够使特征词向量的表征具有上下文语义和结构信息，保证特征词语义的集中，也能够对文本进行特征降维处理，避免发生维度灾难，能够有效提高文本相似度计算结果的精确性，保证聚类结果的准确性。

**二、研究内容**

**2.1 理论与方法介绍**

使用基于神经网络Word2Vec 工具所依赖的 Skip-gram 模型训练大规模语料，得到语料库中所有特征词的词向量表征; 然后，使用融合词频、词距、位置信息和词向量欧式距离的自定义多特征文本语义距离计算公式，计算文本集中的任意两个文本的相似度; 最后，使用经典的K-means 算法对文本集中的文本进行聚类。）

**2.2 验证分析与实验效果**

利用基于词频的方法、基于主题模型的方法和基于知识库的方法进行对比实验），最终显示，该算法克服了传统文本特征向量表示方法存在的问题( 如数据稀疏和网络词汇更新速度快等因素对文本聚类结果的影响。

**三、论文存在问题及后续研究重点**

**3.1 论文存在问题**

本算法必须指定聚类数量。但是有时候我们并不知道应该聚成多少个类，而是希望算法可以给出一个合理的聚类数量，往往一开始k值很难预先估计并给定。

**3.2 后续研究重点**

**四、该问题相关研究成果**

**4.1 相关论文一**

**（1）题目**：结合半监督学习和LDA模型的文本分类方法

**（2）作者介绍**：韩栋

**（3）摘要**: 针对样本集中具有较少标记样本情况下的文本分类问题,提出一种结合半监督学习(SSL)和隐含狄利克雷分配(LDA)主题模型的标记样本扩展方法(SSL-LDA),并整合朴素贝叶斯(NB)分类器构建一种文本分类方法。使用LDA主题模型生成主题分布,以表示所有样本;根据训练集中已标记样本,通过一种简化粒子群优化(SPSO)算法获得SSL-LDA自训练模型的最优参数;基于SSL-LDA自训练模型对训练集中一些未标记样本进行标记,扩展训练集;基于扩展后的训练集,训练NB文本分类器。在3个数据集上的实验结果表明,该方法能够很好地应对标记样本较少的情况,获得了较高的分类精确度。

**4.2 相关论文二**

**（1）题目**：融合语义资源和关键词的文本聚类

**（2）作者介绍**：吴舜尧

**（3）摘要**：融合关键词形式的属性层知识可有效提高文本聚类的聚类质量,但融合关键词的簇中心初始化仍是一个开放性问题。为此,提出一种融合语义资源和关键词的文本聚类方法。通过Wikipedia语义识别文本集的主题,采用基于资源分配的网络推断策略,通过文献协同关系发现潜在语义相关性,以选择最能代表各主题的重要文档（初始簇中心）,并利用软约束与测度学习相结合的策略融合关键词辅助文本聚类。在20Newsgourp文本集上的实验结果表明,与k-means和仅融合关键词的文本聚类方法相比,该方法可有效提升聚类质量,尤其在NewsDifferent3数据集上标准互信息最多可提升约20%。

**4.3 相关论文三**

**（1）题目**：基于加权语义网的文本相似度计算方法研究

**（2）作者介绍**：张弛

**（3）摘要**：为了更加准确地计算文本特征权重,提高文本相似度计算的精确度,文章提出了一种基于《知网(HowNet)》语义知识库的加权语义复杂网络文本相似度计算方法.该方法首先以特征词为节点,以特征词窗口共现为边,以特征词的TF-IDF值为特征词节点的初始权重,以融合共现频率和概念间语义距离计算特征词节点之间边的权重,构建加权语义文本复杂网络.然后利用综合特征指数作为加权语义网中文本的特征权重.最后基于公开数据集和KNN算法进行文本聚类实验,实验结果表明,在基于F-度量值标准上本文提出的方法要优于传统基于向量空间模型的TF-IDF方法和另一种结合复杂网络权重的方法.



**基于规则的情感本体和词向量的中文情感分类**

**T基于规则的情感本体和词向量的中文情感分类**

**--郑诚、潮旭、章金平 微电子学与计算机**

**一、科学问题**

**1.1 本文所涉及科学问题**

中文情感分类方法。

**1.2 同行专家如何解决**

Bai等通过语义定位点相似距离 (SO-SD) 模型, 利用Word2vec工具创建出一种新的模型用以建立情感词典, 并使用该情感词典来获取情感倾向。Zhang等[7]使用word2ve对相似的特征进行聚类, 再分别使用基于词典和基于词性的特征选择方法与SVMperf相结合进行情感分类.

**1.3 本文所解决的问题**

识别出用户在评论文本中流露出的情感信息, 将其分为代表肯定和赞扬的正类以及代表否定和批判的负类。

**1.4 本文解决方案效果**

在考虑句法规则的情况下, 本研究的情感分类结果相对于只考虑特征词向量的方法更好, 当加入情感本体后, 情感分类效果最好, 验证了本文方法的可行性, 句法规则的词向量以及情感本体的结合对分类效果起到正面影响.

**二、研究内容**

**2.1 理论与方法介绍**

使用word2vec工具获取词向量, 并利用评论数据自动提取情感种子词, 从而创建领域自适应情感词典, 与规则方法相结合, 构成短文本词向量和情感值相融合情感模型 (VWEO) , 考虑了情感本体和语义信息, 从而达到提高情感分类效果的目的.

**2.2 验证分析与实验效果**

评价标准是准确率 (Precision) 、召回率 (Recall) 和F值 (F-measure) .准确率是测试集中正确分类的评论文本数占分类结果中所有标记为该类文本数量的比例;召回率则为测试集中人工标注的所有属于该类文本数量的比例;F值是两者的调和平均.。

在考虑句法规则的情况下, 情感分类结果相对于只考虑特征词向量的方法更好, 当加入情感本体后, 情感分类效果最好, 验证了本文方法的可行性, 句法规则的词向量以及情感本体的结合对分类效果起到正面影响.

**三、论文存在问题及后续研究重点**

**3.1 论文存在问题**

程度副词词典数量可能不足。

**3.2 后续研究重点**

对被否定的情感词进行更高效的利用, 并对更多的句式进行分析, 而不仅仅只是考虑转折句式, 进一步的使情感分类效果得到提升.

**四、该问题相关研究成果**

**4.1 相关论文一**

**（1）题目**：机器学习与语义规则融合的微博情感分类方法

**（2）作者介绍**：姜杰夏睿

**（3）摘要**: 针对现有文本情感分析方法的不足,设计了一种针对中文微博的基于词典的规则情感分类方法和用于机器学习方法的基本特征模板。提出一种机器学习与规则相融合的微博情感分类方法,将用规则方法得到的多样化情感信息进行转化,扩展并嵌入基本特征模板,形成更有效的融合特征模板。通过3种分类模型集成,提高微博情感分类的性能。

**4.2 相关论文二**

**（1）题目**：Chinese comments sentiment classification based on word2vec and SVM perf

**（2）作者介绍**：Dongwen Zhang; Hua Xu; Zengcai Su; Yunfeng Xu;

**（3）摘要**：Since the booming development of e-commerce in the last decade, the researchers have begun to pay more attention to extract the valuable information from consumers comments. Sentiment classification, which focuses on classify the comments into positive class and negative class according to the polarity of sentiment, is one of the studies. Machine learning-based method for sentiment classification becomes mainstream due to its outstanding performance. Most of the existing researches are centered on the extraction of lexical features and syntactic features, while the semantic relationships between words are ignored. In this paper, in order to get the semantic features, we propose a method for sentiment classification based on word2vec and SVM perf . Our research consists of two parts of work. First of all, we use word2vec to cluster the similar features for purpose of showing the capability of word2vec to capture the semantic features in selected domain and Chinese language. And then, we train and classify the comment texts using word2vec again and SVM perf . In the process, the lexicon-based and part-of-speech-based feature selection methods are respectively adopted to generate the training file. We conduct the experiments on the data set of Chinese comments on clothing products. The experimental results show the superior performance of our method in sentiment classification.

**4.3 相关论文三**

**（1）题目**：A Study on Sentiment Computing and Classification of Sina Weibo with Word2vec

**（2）作者介绍**： Bai Xue , Chen Fu , Zhan Shaobin

**（3）摘要**：In recent years, Weibo has greatly enriched people's life. More and more people are actively sharing information with others and expressing their opinions and feelings on Weibo. Analyzing emotion hidden in this information can benefit online marketing, branding, customer relationship management and monitoring public opinions. Sentiment analysis is to identify the emotional tendencies of the microblog messages, that is to classify users' emotions into positive, negative and neutral. This paper presents a novel model to build a Sentiment Dictionary using Word2vec tool based on our Semantic Orientation Pointwise Similarity Distance (SO-SD) model. Then we use the Emotional Dictionary to obtain the emotional tendencies of Weibo messages. Through the experiment, we validate the effectiveness of our method, by which we have performed a preliminary exploration of the sentiment analysis of Chinese Weibo in this paper.

