**基于Word2Vec的一种文档向量表示**

**Document Vector Representation Based on Word2Vec**

**TANG Ming ZHU Lei ZOU Xian-chun**

**(School of Computer and Information Science,Science,Southwest University,Chongqing 400715,China)**

**一、科学问题**

**1.1 本文所涉及科学问题**

在文本分类中，将word2vec模型与聚类算法结合形成的doc2vec模型能有效的表达文档信息。但这种方法较少考虑单个词对整篇文档的影响力。为解决这个问题，利用TF-IDF算法计算每篇文档中词的权重，并结合woed2vec词向量生成文档向量，最后应用用于中文文档分类。

**1.2 同行专家如何解决**

合肥学院计算机科学与技术系的张弛提出一种基于词向量和多特征语义距离的文本聚类算法（M-W2-KS），首先使用Word2Vec训练语料库中的所有特征值，以向量的形式进行表征；然后综合考虑特征词的词频。位置等，计算文本间的语义相似度，并应用到K-means算法中，实现文本的聚类。

**1.3 本文所解决的问题**

利用所有词向量求平均值、对词向量聚类、以及doc2vec模型这些方法表示一篇文档，并未重视单个词对整个文档的影响力。针对这个问题，本文在word2vec的基础上，利用TF-IDF算法，对每篇文档的分词进行加权，并在搜狗中文实验语料库上进行测试，测试结果验证了该方法的有效性。

**1.4 本文解决方案效果**

本文提出的基于TF-IDF的加权的word2vec模型相比平均值word2vec模型又有了一些提升，在SVM、KNN以及以及RBF分类器上，宏平均正确率分别提升了2.57%，0.58%以及3.72%；宏平均召回率分别提升了2.73%，1.02%以及3.38%；宏平均F1分别提升了2.65%，0.85%以及3.85%。

**二、研究内容**

**2.1 理论与方法介绍**

针对常见词向量表示一篇文档方法未重视单个词对整个文档的影响力，本文提出结合TF-IDF算法对每篇文档中的分词进行加权。该方法，考虑每类文档集中的每个文档的每个分词，利用TF-IDF算法计算其在该文档中的权重K（t，Di），其表示词t在文档Di中的权重。对于每篇文档Di，wi表示分词t的词向量其文档向量可以表示为如下形式：di = 。所以文档向量d也是一个实数向量。

**2.2 验证分析与实验效果**

验证方法是在搜狗中文实验语料库上进行测试，采用的评估指标包括准确率、召回率、F1指标以及宏平均。将提出的TF-IDF加权的word2vec模型与TF-IDF加权的BOW模型（即将BOW向量中分词相应的位置的值替换成TF-IDF权值）、均值word2vec模型以及doc2vec模型的分类效果进行对比。所有实验采用五分交叉验证，及把数据集随机划分成5份，每次取其中4份进行训练，1份进行测试，然后把5次分类结果的平均值作为最终结果。

本文提出的基于TF-IDF的加权的word2vec模型相比平均值word2vec模型又有了一些提升，在SVM、KNN以及以及RBF分类器上，宏平均正确率分别提升了2.57%，0.58%以及3.72%；宏平均召回率分别提升了2.73%，1.02%以及3.38%；宏平均F1分别提升了2.65%，0.85%以及3.85%。而TF-IDF加权的word2vec模型与doc2vec模型的宏平均值效果相差不大，在几组分类器上的效果相当，可见提出的方法的有效性，其可以作为另一种用于文档分类的有效方法。

**三、论文存在问题及后续研究重点**

**3.1 论文存在问题**

尽管本文将word2vec和TF-IDF结合，提出一种基于word2vec的TF-IDF加权计算文档向量算法，相较于TF-IDF加权的BOW模型与均值word2vec模型有更好的文本分类效果，但与doc2vec模型的宏平均效果相差不大，没有在文本分类的准确率上有比较明显的突破。

**3.2 后续研究重点**

继续研究word2vec与TF-IDF结合是否可以继续完善改进以达到超越doc2vec模型的文本分类准确率。

**四、该问题相关研究成果**

**4.1 相关论文一**

### **（1）题目：**

### **Distributed Representations of Words and Phrases and their Compositionality**

1. **作者介绍：**

**Tomas Mikolov，Ilya Sutskever，Kai Chen**

**（3）摘要**:

最近推出的连续Skip-gram模型是学习高质量分布式矢量表示的有效方法，可以捕获大量精确的句法和语义单词关系。在本文中，我们提出了一些改进，使Skip-gram模型更具表现力，使其能够更快地学习更高质量的向量。我们表明，通过对频繁的单词进行二次采样，我们可以获得显着的加速，并且还可以通过我们的任务来学习更高质量的表示。我们还引入了负抽样，这是噪声对比度估计（NCE）的简化变体，与分层softmax相比，它可以学习更频繁的单词矢量。单词表示的固有局限性是它们对单词顺序的漠不关心以及它们无法表示惯用语。例如，加拿大''和'Air''的含义不能轻易合并以获得“加拿大航空”。在这个例子的推动下，我们提出了一种简单有效的查找短语的方法，并表明它们的矢量表示可以通过Skip-gram模型准确学习。

**4.2 相关论文二**

### **（1）题目**：

### **[The Application of K-Medoids and PAM to the Clustering of Rules](http://www.springerlink.com/index/y7c4ahuqgcm4vx88.pdf" \t "http://xueshu.baidu.com/usercenter/paper/_blank)**

**（2）作者介绍**： **[Alan P. Reynolds](http://xueshu.baidu.com/s?wd=author:(Alan P. Reynolds) &tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=person" \t "http://xueshu.baidu.com/usercenter/paper/_blank)，[Graeme Richards](http://xueshu.baidu.com/s?wd=author:(Graeme Richards) &tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=person" \t "http://xueshu.baidu.com/usercenter/paper/_blank)，[Vic J. Rayward-Smith](http://xueshu.baidu.com/s?wd=author:(Vic J. Rayward-Smith) &tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=person" \t "http://xueshu.baidu.com/usercenter/paper/_blank)**

**（3）摘要**：

早期的研究已经产生了用于数据挖掘的“全规则”算法，该算法产生了上述给定置信度和覆盖阈值的所有联合规则。 虽然这是一个有用的工具，但它可能会产生大量规则。 本文描述了两种聚类算法在这些规则中的应用，以便识别相似规则集并更好地理解数据。

**4.3 相关论文三**

### **（1）题目**：**[新的K-均值算法最佳聚类数确定方法](http://www.cqvip.com/qk/91690x/201016/34089809.html" \t "http://xueshu.baidu.com/usercenter/paper/_blank)**

**（2）作者介绍**：**[周世兵](http://xueshu.baidu.com/s?wd=author:(%E5%91%A8%E4%B8%96%E5%85%B5) %E6%B1%9F%E5%8D%97%E5%A4%A7%E5%AD%A6%E4%BF%A1%E6%81%AF%E5%B7%A5%E7%A8%8B%E5%AD%A6%E9%99%A2;%E6%B1%9F%E5%8D%97%E5%A4%A7%E5%AD%A6%E7%90%86%E5%AD%A6%E9%99%A2;&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=person" \t "http://xueshu.baidu.com/usercenter/paper/_blank)，[徐振源](http://xueshu.baidu.com/s?wd=author:(%E5%BE%90%E6%8C%AF%E6%BA%90) %E6%B1%9F%E5%8D%97%E5%A4%A7%E5%AD%A6%E4%BF%A1%E6%81%AF%E5%B7%A5%E7%A8%8B%E5%AD%A6%E9%99%A2;%E6%B1%9F%E5%8D%97%E5%A4%A7%E5%AD%A6%E7%90%86%E5%AD%A6%E9%99%A2;&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=person" \t "http://xueshu.baidu.com/usercenter/paper/_blank)，[唐旭清](http://xueshu.baidu.com/s?wd=author:(%E5%94%90%E6%97%AD%E6%B8%85) %E6%B1%9F%E5%8D%97%E5%A4%A7%E5%AD%A6%E4%BF%A1%E6%81%AF%E5%B7%A5%E7%A8%8B%E5%AD%A6%E9%99%A2;%E6%B1%9F%E5%8D%97%E5%A4%A7%E5%AD%A6%E7%90%86%E5%AD%A6%E9%99%A2;&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=person" \t "http://xueshu.baidu.com/usercenter/paper/_blank)**

**（3）摘要**：

K-均值聚类算法是以确定的类数k和随机选定的初始聚类中心为前提对数据集进行聚类的。通常聚类数k事先无法确定,随机选定的初始聚类中心容易使聚类结果不稳定。提出了一种新的确定K-均值聚类算法的最佳聚类数方法,通过设定AP算法的参数,将AP算法产生的聚类数作为聚类数搜索范围的上界kmax,并通过选择合适的有效性指标Silhouette指标,以及基于最大最小距离算法思想设定初始聚类中心,分析聚类效果,确定最佳聚类数。仿真实验和分析验证了以上算法方案的可行性。

**利用word2vec对中文词进行聚类的研究**

**Research on Chinese word Clustering with Word2vec**

**郑文超，徐鹏**

**（北京邮电大学网络技术研究院，北京100876）**

**一、科学问题**

**1.1 本文所涉及科学问题**

文本聚类在数据挖掘和机器学习中发挥着重要的作用，本文探究了一种新的中文聚类方法。

**1.2 同行专家如何解决**

合肥学院计算机科学与技术系的张弛提出一种基于词向量和多特征语义距离的文本聚类算法（M-W2-KS），首先使用Word2Vec训练语料库中的所有特征值，以向量的形式进行表征；然后综合考虑特征词的词频。位置等，计算文本间的语义相似度，并应用到K-means算法中，实现文本的聚类。

**1.3 本文所解决的问题**

英语的研究中各种词聚类算法已经有一定的的理论成果，但对中文的研究似乎没有英语那么成功，针对这种情况，本文提出的模型算法在中文词聚类方面取得可不错的成果。

**1.4 本文解决方案效果**

词的分类一定程度上可以解决参数空间庞大，训练数据不足，数据稀松等问题，本文使用word2vec工具集，找到了一种较为方便的方法对中文词进行聚类，并且取得了不错的效果。

111111111111111111111111111111111111111111

**二、研究内容**

**2.1 理论与方法介绍**

本文将中文词语看作一系列独立词的“词袋模型”，之后使用深度学习神经网络算法将词转化为n维向量，最后用k-mean算法计算这些向量进行聚类。

**2.2 验证分析与实验效果**

本文使用神经网络算法获取到了词向量表后，通过计算不同词向量直接的余弦距离作为不同词之间的“距离”，最后用K-means算法将近似词聚集到一起。文中列出的表格展示了同“中国”最接近的top10个词与同“计算机”最接近的top10个词，证明模型取得了一定的效果。

**三、论文存在问题及后续研究重点**

**3.1 论文存在问题**

文中展示实验效果时仅列出了部分实验结果没有对实验结果进行分析统计，难以评价模型具体在哪些方面优于传统的方法。

**3.2 后续研究重点**

基于词的语言模型存在着许多问题，如参数空间庞大，训练数据不足，数据稀疏等。后续研究应围绕上问题进行优化与改进使中文词的聚类的算法取得更大的进展。

**四、该问题相关研究成果**

**4.1 相关论文一**

### **（1）题目**：**[初始聚类中心优化的k-means算法](http://d.wanfangdata.com.cn/Periodical/jsjgc200703024" \t "http://xueshu.baidu.com/usercenter/paper/_blank)**

**（2）作者介绍**：**[袁方](http://xueshu.baidu.com/s?wd=author:(%E8%A2%81%E6%96%B9) %E6%B2%B3%E5%8C%97%E5%A4%A7%E5%AD%A6%E6%95%B0%E5%AD%A6%E4%B8%8E%E8%AE%A1%E7%AE%97%E6%9C%BA%E5%AD%A6%E9%99%A2&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=person" \t "http://xueshu.baidu.com/usercenter/paper/_blank)，[周志勇](http://xueshu.baidu.com/s?wd=author:(%E5%91%A8%E5%BF%97%E5%8B%87) %E6%B2%B3%E5%8C%97%E5%A4%A7%E5%AD%A6%E6%95%B0%E5%AD%A6%E4%B8%8E%E8%AE%A1%E7%AE%97%E6%9C%BA%E5%AD%A6%E9%99%A2&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=person" \t "http://xueshu.baidu.com/usercenter/paper/_blank)，[宋鑫](http://xueshu.baidu.com/s?wd=author:(%E5%AE%8B%E9%91%AB) %E6%B2%B3%E5%8C%97%E5%A4%A7%E5%AD%A6%E6%95%B0%E5%AD%A6%E4%B8%8E%E8%AE%A1%E7%AE%97%E6%9C%BA%E5%AD%A6%E9%99%A2&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=person" \t "http://xueshu.baidu.com/usercenter/paper/_blank)·**

**（3）摘要**: 传统的k-means算法对初始聚类中心敏感,聚类结果随不同的初始输入而波动。为消除这种敏感性,提出一种优化初始聚类中心的方法,此方法计算每个数据对象所在区域的密度,选择相互距离最远的k个处于高密度区域的点作为初始聚类中心。实验表明改进后的k-means算法能产生质量较高的聚类结果,并且消除了对初始输入的敏感性。

**4.2 相关论文二**

### **（1）题目**：**[A Neural Probabilistic Language Model](http://www.researchgate.net/publication/2413241_A_Neural_Probabilistic_Language_Model" \t "http://xueshu.baidu.com/usercenter/paper/_blank)**

**（2）作者介绍**：**[Yoshua Bengio∗](http://xueshu.baidu.com/s?wd=author:(Yoshua Bengio%E2%88%97) &tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=person" \t "http://xueshu.baidu.com/usercenter/paper/_blank)，[Re´jean Ducharme](http://xueshu.baidu.com/s?wd=author:(Re%C2%B4jean Ducharme) &tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=person" \t "http://xueshu.baidu.com/usercenter/paper/_blank)，[Pascal Vincent](http://xueshu.baidu.com/s?wd=author:(Pascal Vincent) &tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=person" \t "http://xueshu.baidu.com/usercenter/paper/_blank)**

**（3）摘要**：统计语言建模的目标是学习单词序列的联合概率函数。 由于维度的诅咒，这本质上是困难的：我们建议用自己的武器来对抗它。 在所提出的方法中，同时学习（1）每个单词的分布式表示（即单词之间的相似性）以及（2）单词序列的概率函数，用这些表示表示。 获得泛化是因为如果由与形成已经看到的句子的单词类似的单词构成，则之前从未见过的单词序列具有高概率。 我们报告了使用神经网络进行概率函数的实验，在两个文本语料库中显示所提出的方法在最先进的三元模型上得到了非常显着的改善。

**4.3 相关论文三**

**（1）题目**：**基于Word2Vec的一种文档向量表示**

### **（2）作者介绍**：**[唐明](http://xueshu.baidu.com/s?wd=author:(%E5%94%90%E6%98%8E) &tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=person" \t "http://xueshu.baidu.com/usercenter/paper/_blank)，[朱磊](http://xueshu.baidu.com/s?wd=author:(%E6%9C%B1%E7%A3%8A) &tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=person" \t "http://xueshu.baidu.com/usercenter/paper/_blank)，[邹显春](http://xueshu.baidu.com/s?wd=author:(%E9%82%B9%E6%98%BE%E6%98%A5) &tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_f_para=sc_hilight=person" \t "http://xueshu.baidu.com/usercenter/paper/_blank)**

**（3）摘要**：在文本分类中,如何运用word2vec词向量高效地表达一篇文档一直是一个难点。目前,将word2vec模型与聚类算法结合形成的doc2vec模型能有效地表达文档信息。但是,这种方法很少考虑单个词对整篇文档的影响力。为了解决这个问题,利用TF-IDF算法计算每篇文档中词的权重,并结合word2vec词向量生成文档向量,最后将其应用于中文文档分类。在搜狗中文语料库上的实验验证了新方法的有效性。