

**学术学位硕士研究生学位论文中期报告**

论文题目： 基于词向量的短文本聚类研究

|  |  |
| --- | --- |
| 姓 名： | 黄 栋 |
| 学 号： | 21409173 |
| 学科专业： | 计算机应用技术 |
| 指导教师： | 杨志豪 |
| 入学日期： | 2014年9月1日 |
| 报告日期： | 2016年10月 12日 |
| 报告地点： | 大连理工大学创新园大厦 |

研究生院制表

说 明

硕士学位论文中期检查是保证学位论文质量、工作进度和研究生培养质量的重要措施。原则上，要求硕士生应在第4学期末进行中期检查，其中，2年制专业学位硕士生应在第3学期末进行中期检查。

一、考核内容：学位论文内容完成情况、阶段性成果是否正确，开题时方案是否需调整或已做了哪些调整，后续工作思路是否正确、工作进度是否有保障、预期目标能否实现、论文质量是否能够保证以及论文工作存在的问题等。

二、考核时间：原则上，硕士生的中期检查应在第4学期末进行，其中两年制专业学位硕士生的中期检查可在第3学期末进行。

三、报告撰写：中期报告正文字数不少于4000字，正文及参考文献等撰写要求参见《大连理工大学博士学位论文格式规范》。

四、考核办法：由学部（学院）组织，3-5名本学科领域硕导及以上专家组成评审专家组以答辩的方式进行。学生进行口头陈述时间不得少于10分钟。专家组给出考核成绩和是否通过的意见。

五、报告保存：中期报告一式两份，签字后分别由学部（学院）和学生保存。

六、信息登录：中期考核后，学生应及时登录研究生信息管理系统上传中期报告（PDF文档）及考核结果。

**硕士学位论文中期检查报告正文**

撰写大纲：

1. 开题时拟定的研究方案、进度计划；若开题时的研究方案已经调整，应说明调整的原因、调整后该领域的国内外研究状况分析、研究内容、研究方法、进度计划等；
2. 学位论文的研究进展完成情况、阶段性成果和创新点论述；
3. 后续工作的设想、可能遇到的困难和问题及条件保障措施；
4. 已发表、录用的论文和已投稿的论文情况。
5. 参考文献（不占字数）。

# 1 绪论

* 1. 研究背景

随着移动互联网的高速发展，以移动互联网为阵地的微信、微博、电子邮件、论坛、直播平台、点评网站等应用已经越来越普及，其产生的大多数信息都以短文本的形式出现。这些短文本涉及到人们生活的各个领域，也逐渐成为人们广泛使用并且公认的交流方式，改变着人们的生活和沟通习惯。无论是论坛、在线即时聊天工具中包含的关系数据和社会数据，还是电子商务平台中蕴藏的电商数据和信用数据，都具有极大的价值。

文本聚类是将文本集合按照文本内在的联系，将其聚合成多个不同类别文档集合的过程，它假设同一集合内的文本尽可能相似，不同集合间的文本差异尽可能大。文档聚类不需要事先对实验所用文档的类别进行标注，没有训练过程，所以它能很灵活地自动处理文档。传统基于TF—IDF的向量空间模型的短文本相似度计算存在高维、数据稀疏、缺乏语义和维度未归一等问题。基于知识库的文本聚类方法利用知网、WordNet等知识库丰富的文本特征，挖掘词语之间的关系，从而提高聚类效果。Chen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。Bouras提出了一种基于WordNet的新闻文档聚类技术。吴舜尧等人通过 Wikipedia 语义识别文本集的主题，通过文献协同关系发现潜在语义相关性，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。利用知识库的方法可以在一定程度上挖掘词语语义之间的关系，但是网络语言具有奇异性和动态性两个显著特征。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，克服传统聚类方法中特征向量维度高的问题。王春龙采用LDA（Latent Dirichlet Allocation）模型选择在文本集中影响程度最大的主题，并在主题所在的维度上对文本集进行初步聚类找到聚类中心，然后以这些聚类中心对文本集进行所有维度上的聚类。汤秋莲利用BTM模型(Biterm Topic Model)训练之后得到的文档-主题的概率分布矩阵和主题-词概率分布矩阵与传统的利用TF-IDF进行词权重计算的向量空间模型相结合,将主题特征与词特征结合,以此达到改善短文本聚类质量的效果。王少鹏和Yin分别对LDA模型进行改进，提高聚类效果。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，但是新媒体产生的短文本大都由用户自发进行发布，与传统的聚类相比，主题比较分散，孤立点及包含样本点数量较少的簇占很大一部分比例。Mikolov从大量的实验中发现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），影响了上述模型应用于互联网短文本数据的性能。

* 1. 研究方法

1986，Hinton等人首先提出了词向量（Distributed Representation）的概念，利用它来解决文章中“维度灾难”的问题。2003年，Bengio等人提出了基于前馈神经网络的语言模型（NNLM），该语言利用句子前面出现的词作为上下文信息来预测下一个单词。2013年，Google公司的Mikolov等人提出并开源了语言建模工具 word2vec，并在自然语言处理领域得到了广泛关注。word2vec模型实现了连续bag-of-word模型，以及计算词向量的 skip-gram 结构。词向量是一种分布式的低维实数向量，其基本思想是通过训练将语料中的词语映射到N维实数向量。EMD距离由RUBNER提出并应用于计算视觉问题上，它可以被理解为从一种分布变换为列一种分布的最小代价。在将该距离引入到句子模型后，将词当做物资产地，词频当做某产地的物资供应量，这样求得句子之间的距离来计算句子相似度。

本文采取的研究方案是：考虑到短文本的特点和词向量的优点，本文提出了一种基于特征词向量表示和特征词移动距离的短文本聚类算法。首先使用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；然后利用欧式距离计算特征词相似度，引入EMD距离（Earth Mover’s Distance）对短文本进行建模来计算短文本间的相似度；最后将其并应用到Kmeans聚类算法中实现短文本聚类。

* 1. 进度计划

由于对课题难度估计不够，认识不深，开题报告中所制定的进度计划存在了很多不足之处，因此我对进度计划时间、内容安排进行了修改和完善。目前，实验工作正按计划进行。其中1-6项已经完成，7-9项正在进行中。

表1 进度计划表

|  |  |  |
| --- | --- | --- |
| 序号 | 时间 | 任务 |
| 1 | 2015年09月—2015年10月 | 阅读文献和查阅资料 |
| 2 | 2015年11月 | 方案设计 |
| 3 | 2015年12月 | 配置词向量训练工作环境 |
| 4 | 2016年01月—2016年02月 | 语料采集和预处理 |
| 5 | 2016年03月—2016年06月 | 短文本词向量表示方法研究 |
| 6 | 2016年07月—2016年9月 | 短文本距离计算方法研究 |
| 7 | 2016年10月—2017年03月 | 基于特征词向量表示和特征词移动距离的短文本聚类算法的研究，分别在英文语料集和中文语料集进行实验，并根据实验结果调整相关参数和模型。 |
| 8 | 2017年04月 | 整理实验结论，完成论文初稿 |
| 9 | 2017年05月 | 完成论文终稿 |

# 2 研究内容概述

2.1研究进展与阶段性成果

经过前期的研究与学习，现阶段已经完成了对短文本聚类背景知识的了解，掌握了业内主流的方法与技术，并调研短文本聚类相关数据集。已经完成CNKI爬虫的开发，可实现自定义搜索关键字的论文信息提取。目前已按照学校授予单位为关键词爬取高校硕博士论文信息数据总计52.8 MB，共包含259952条论文信息。

2016年6月至9月，经过大量文献的阅读参考和实验对比，提出了一种基于特征词向量表示和特征词移动距离的短文本聚类算法，制定实验方案步骤。该方法的基本流程如所示，主要包括数据预处理、词向量的训练、短文本的表示，短文本相似度计算、Kmeans聚类五个部分。

为了验证本文提出短文本聚类提取算法的可用性、准确性和鲁棒性，我在三个标准公共英文数据集20newsgroup、3CPhys和7CNetv上对算法进行了评估。本文使用 nltk工具对文本进行去停用词、词根还原等预处理操作，使用的英文词向量为Mikolov公布的使用Google新闻训练出包含大于3百万个单词的词向量，词向量的维度为300维。词向量中没有出现的词，不参与EMD距离的计算（其他对比实验也采用这一标准）。

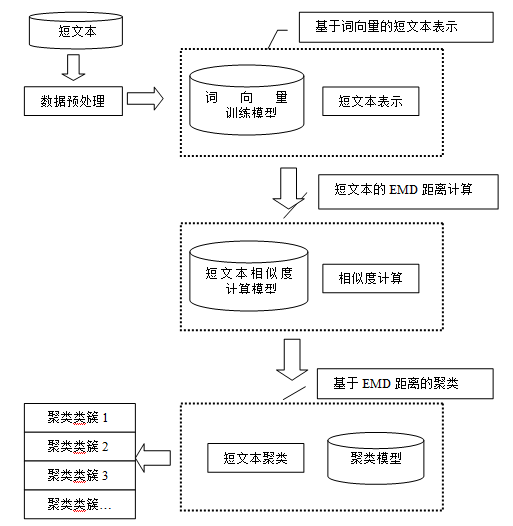


图1基于EMD距离的短文本聚类方法流程图

使用纯度（Purity）和归一化互信息值（Normalized Mutual Information , NMI ）来评价本文算法的性能，纯度和NMI 度量值是用于评价聚类效果的两个最常用的指标，Purity和NMI度量值越大，聚类效果越好。

为了验证本文方法的可行性与有效性，将本文提出的算法与以下三种模型进行对比，使用Qimin-VSM、Guo-LDA和Yan-BTM三种模型对文本进行表示，采用Kmeans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进行设置，，，Kmeans设置的迭代次数为1000。由于Kmeans算法容易受到初始聚点的影响，导致聚类结果的不稳定，本文采用每次重复进行10次聚类实验，取这十次聚类结果的的平均值来进行对比。评测结果分别记为K-Qimin-VSM、K-Guo-LDA 和K-Yan-BTM，本文方法记为K-W2V-EMD。

使用不同的文本表示模型进行聚类的评测结果统计信息见表2和表3：

表2 不同文本表示的纯度指标评测结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据集 | K-Qimin-  VSM | K-Guo-  LDA | K-Yan-  BTM | **K-W2V-**  **EMD** |
| 20news | 0.430 | 0.495 | 0.566 | **0.645** |
| 3CPhys | 0.486 | 0.532 | **0.612** | 0.605 |
| 7CNetv | 0.308 | 0.369 | 0.375 | **0.402** |

表3 不同文本表示的互信息指标评测结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据集 | K-Qimin-  VSM | K-Guo-  LDA | K-Yan-  BTM | **K-W2V-**  **EMD** |
| 20news | 0.169 | 0.328 | 0.334 | **0.417** |
| 3CPhys | 0.376 | 0.405 | 0.455 | **0.458** |
| 7CNetv | 0.426 | 0.439 | 0.442 | **0.446** |

从上表中可以看出，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会受到短文本数据稀疏的严重影响。本文方法的评测结果要优于基于主题模型的文本表示对比方法，这是因为LDA等主题模型偏重于高频，忽略低频词的问题可以通过大规模语料的训练得到对应的词向量来解决，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。且由于短文本较短，不利于统计词词之间的共现规律，所以LDA等主题模型会出现语义稀疏的问题，使用词向量加WMD距离的句子模型，不能从根本上解决语义稀疏的问题，但是使用大规模语料训练出来的词向量，加上WMD算法保证的是词语的移动是将句子中的词语移动到语义最相近的词语，最大程度上保证了语义的集中，实验结果表明有一定的改进。

2.2创新点论述

本文提出了一种基于特征词向量的文本表示和基于特征词移动距离的短文本聚类算法。利用词向量的性质解决短文本数据稀疏和网络词汇更新速度快对文本聚类带来的问题；引入EMD距离可以更区别短文本之间的相似度，使得聚类结果更加准确。实验结果表明，本文方法可以有效提高聚类效果。

# 后续工作展望

# 3.1后续工作的设想

**（1）数据集的扩充**

目前的实验评测只集中在英文数据集上，下一步将更多考虑在中文数据集上对模型的优化。现阶段的研究方法与技术条件下，数据集的的扩充与数据集质量的提高对于实验结果的提升有着积极的影响，因此后续工作会尝试扩充现有数据集，不仅继续对CNKI论文数据集合进行扩充，也尝试使用国内外论坛、微博平台等数据进行爬取，提高数据集的丰富性。

**（2）实验方法的优化**

现阶段对于数据的清洗与特征词提取方面仍然存在一些错误，需要后续尝试使用更多的工具与方法，提升数据识别与清洗的成功率。此外，目前提出的方法使用特征词的词频权重作为特征词权重，下一步将考虑词语的词性、词序等特征，使模型更好的表达句子的语义，以达到更好的聚类效果和算法效率。

# 3.2可能遇到的困难与问题

中文数据集的处理和英文数据集的处理有很大的差异，尤其是中文语句的分词、词性标注的处理以及特征词的选取需要大量的调研与测试过程，由于不同数据集的领域性差异，在处理的过程中难度也会提升。此外，大量数据集本身内容与结构的复杂性也带来了解析、清洗、识别的复杂度，这也是后续实验中可能遇到的问题。

词向量的训练与语料的数量和质量有很大关系，Word2Vec模型的使用等需要更多的理论知识的学习，且模型的调参大多是超参数，需要不断的实验经验积累，根据实际效果来调优，这也对于后续工作的进行增加了难度。

# 3.3条件保障措施

学校图书馆及电子图书馆可以提供丰富的图书及文献资源，查阅便利；实验室配置有高性能计算机，能够满足数据处理方面的需求；实验室的各位师兄师姐与指导老师在整个研究过程中提供悉心的指导与研究方向的精确掌控。

# 已发表、录用的论文和已投稿的论文情况

论文《基于词向量和EMD距离的短文本聚类》 已被第22届全国信息检索学术会议（CCIR2016）录用。

参考文献

1. Chen X, Zhang Y, Cao L, et al. An Improved Feature Selection Method for Chinese Short Texts Clustering Based on HowNet[M]//Computer Engineering and Networking. Springer International Publishing, 2014: 635-642.
2. Bouras C, Tsogkas V. A clustering technique for news articles using WordNet[J]. Knowledge-Based Systems, 2012, 36: 115-128.
3. 吴舜尧, 邵峰晶, 王金龙,等. 融合语义资源和关键词的文本聚类[J]. 计算机工程, 2014, 40(4):223-227.
4. 夏云庆, 黄锦辉, 张普. 中文网络聊天语言的奇异性与动态性研究[J]. 中文信息学报, 2007, 21(3):83-91.
5. 王春龙, 张敬旭. 基于LDA的改进K-means算法在文本聚类中的应用[J]. 计算机应用, 2014, 34(1):249-254.
6. 汤秋莲. 基于BTM的短文本聚类[D].安徽大学,2014.
7. 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(09): 129-134.
8. 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. ACM, 2014: 233-242.
9. Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv， 2013:1301.3781.
10. Kevin P. Murphy. Machine Learning-A Probabilistic Perspective[M] Cambridge, Massachusetts London, England: The MIT Press,2012 2-39.
11. Hinton G E. Learning distributed representations of concepts[C]// In Proceedings of CogSci. 1986.
12. Bengio Y, Schwenk H, Senécal J S, et al. Neural probabilistic language models[M]//Innovations in Machine Learning. Springer Berlin Heidelberg, 2006: 137-186.
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. 2013: 3111-3119.
14. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.
15. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of NIPS, 2013.
16. Rubner Y, Tomasi C, Guibas L J. The earth mover's distance as a metric for image retrieval[J]. International journal of computer vision, 2000, 40(2): 99-121.
17. Lang K. Learning to filter netnews[C]// Proc. of 12th International Conference Machine Learning (ICML '95). 1995:331-339.
18. Nibir Nayan Bora, Bhabani Shankar Prasad Mishra, Satchidananda Dehuri. Heuristic Frequent Term-Based Clustering of News Headlines [J] Procedia Technology, 2012,6:436-443.
19. Zelikovitz S. Using background knowledge to improve text classification[C]// Rutgers University, 2002.
20. 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.
21. Guo L, Chun-Jie Z, Zhi-Yuan Z. A text clustering method based on weighted LDA model[J]. Journal of Chemical & Pharmaceutical Research, 2015, 7(3).
22. 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. International World Wide Web Conferences Steering Committee, 2013: 1445-1456.

**大连理工大学学术学位硕士研究生学位论文中期报告评审意见表**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 学 号 | |  | | 学生姓名 | |  | 导 师 | |  | |
| 第一次中期检查 □ | | | | | | 第二次中期检查 □ | | | | |
| 导师考核意见（对学位论文内容完成情况、阶段性成果、论文进度、后续工作思路、预期目标、及论文工作存在的问题等进行考查，给出考核成绩和具体改进意见和建议）：   1. 考核成绩：□ 优秀，□ 良好，□ 中等，□ 及格，□ 不及格 2. 是否通过：□ 通过，□ 不通过 3. 关于开题报告撰写质量及学位论文工作的具体意见（可加页）：   导师签字：  年 月 日 | | | | | | | | | | |
| 评  审  专  家  组 |  | 姓名 | 职称 | | 学科专业 | | | 是否博导 | | 签字 |
| 组长 |  |  | |  | | |  | |  |
| 成员 |  |  | |  | | |  | |  |
|  |  | |  | | |  | |  |
|  |  | |  | | |  | |  |
|  |  | |  | | |  | |  |
|  |  | |  | | |  | |  |
|  |  | |  | | |  | |  |
| 专家组评审意见（对学位论文内容完成情况、阶段性成果是否正确、研究方案和进度是否合理、开题时方案是否需调整或已做了哪些调整，后续工作思路是否正确、工作进度是否有保障、预期目标能否实现、论文质量是否能够保证以及论文工作存在的问题等进行考查，给出考核成绩，投票表决是否通过，并给出具体改进意见和建议）：   1. 考核成绩：□ 优秀，□ 良好，□ 中等，□ 及格，□ 不及格 2. 是否通过：□ 通过，□ 不通过 3. 具体意见（可以加页）：   组长签字：  年 月 日 | | | | | | | | | | |
| 点长意见：  点长签字：  年 月 日 | | | | | | | | | | |