

**学术学位硕士研究生学位论文中期报告**

论文题目： 中文专利的自动分类

|  |  |
| --- | --- |
| 姓 名： | 牛世雄 |
| 学 号： | 21409145 |
| 学科专业： | 计算机应用技术 |
| 指导教师： | 姚念民 |
| 入学日期： | 2014年9月1日 |
| 报告日期： | 2016年10月13日 |
| 报告地点： | 创新园大厦A802 |

# 开题时拟定的研究方案、进度计划

## 1.1 开题时拟定的研究方案

本研究的研究目标为：中文专利文本的自动分类。目前，专利文本的分类主要以人工方式为主，自动分类方式已经开始作为辅助手段，但大规模的专利文本自动分类尚未实现，基于专利文本的语义特征，借助自动分类技术，专利工作员能够对大量的专利文本进行自动且高效地分类，提高工作效率的同时，能够更好地对专利文本中蕴含的丰富的科技信息进行分析和利用。因此，研究专利文本的自动分类具有重要的现实意义。

本研究的主要内容有：爬取中国专利局官网的专利摘要文本作为本研究的数据集，对专利文本进行预处理，去停顿词，分词处理，特征选择，特征加权，文本向量表示，对得到的文本向量统一维度，然后运用机器学习的方法训练模型，用得到的模型进行分类。

具体研究内容包括：

1. 得到数据集，爬取中国知识产权局的10000个某个领域的专利文本，本研究爬取的是飞机领域的专利摘要文本。
2. 预处理，对得到的专利文本预处理，去标点、停顿词、分词。本研究运用的结巴分词器进行中文分词处理。
3. 预处理后的文本进行特征选择与特征加权，专利分类中特征选择最常用的是信息增益法，对特征加权法常用TFIDF法，本研究采用修改的TFIDF，对标题词汇赋予大权重和摘要词汇小权重从而加以区分。
4. 文本向量表示，专利自动分类的关键，本次研究采用了传统的向量空间模型，Google开源的关于Tomas Mikolov的Word2Vec[1]词向量结合向量空间模型以及Doc2Vec[2]文本表示方法
5. 训练模型与分类，本研究运用开源的KNN、随机森林、SVM，并结合传统的向量空间模型和词向量模型以及Doc2Vec模型，进行分类对比。

## 进度计划

针对以上研究内容，原定的进度计划为：

2014年12月至2015年7月，广泛阅读文献，了解国内外研究进展，查阅国内外新方法，新思路，调研课题并提高方案可行性。了解常用的特征选择加权算法和文本表示和分类算法，掌握相关基础知识，初步拟定研究方案，并提高方案的可行性。

2015年7月至2016年10月，阅读文献，下载中国专利局的专利文本，完成对不同的特征选择与加权，文本表示方法使用不同的分类算法，建立初步系统模型，进行分类得到初步结果。

2016年10月至2017年2月，对文本表示模型、和分类算法所得到的结果进行修正完善。

2017年2月至2017年5月，对现有实验结论进行整理及总结，完成毕业论文的撰写、修改工作。

# 学位论文的研究进展完成情况、阶段性成果和创新点论述

## 2.1 研究进展完成情况

根据开题报告确定的研究内容、研究方案、技术路线以及进度安排，现在已完成的内容如下所述：

1. 完成数据采集，首先对中国专利局网站进行分析，然后爬取特定领域的中文专利摘要文本1万篇。
2. 研究专利文本的特点，对专利数据集进行处理，建立初步的自动分类模型。专利文本预处理、专利文本特征提取、专利文本表示、专利文本分类器构建、专利文本分类效果评价。

（3）对专利文本的特征选择加权以及文本表示模型就行修改，运用机器学习分类算法分类，分析特征加权方法与文本表示模型的好坏。

从以上论文覆盖的研究内容方面来讲，已经大致完成了研究内容的前两项内容，依计划，第三项完成了部分工作，正在修改。

## 阶段性成果展示

### 数据集的下载

专利是一种价值巨大，富含大量的技术、法律等多方面信息的一种特殊文本。世界知识产权组织的一项统计显示，全世界每年有 90-95％的发明成果都被写入专利中[3][4]，这些发明创新成果涉及各个领域。有研究表明[5]，绝大多数的发明创造会优先选择申请专利保护，若能充分利用专利文献进行技术创新，不仅可以节约大量科研资金，还可以缩短科研时间。对专利信息的挖掘研究逐渐成为热点问题，人们对专利的重视程度也越来越高。近些年，专利的数量在迅速增长，并且增长速率呈递增趋势。世界知识产权组织的报告显示，世界每年专利的申请量达 200 多万件，全球专利申请量年均增长4.7%左右，目前累计总量已近 4000 万件。

目前，专利文本的分类主要以人工方式为主，自动分类方式已经开始作为辅助手段，但大规模的专利文本自动分类尚未实现，因此研究专利文本的自动分类具有重要的现实意义。对于中文专利文本，现在没有相应的开放数据集，分析和调研了各个网站的专利信息，中国知识产权局的专利信息最广泛，正式。

本研究分析了中国知识产权局的网站，设计相应的专利自动爬取系统，请某个领域内的专利，爬取到这个领域的大量专利摘要文本。相关技术及处理：在某些情况下比较有用，比如IP被封了，或者比如IP访问的次数受到限制等需要设置代理。网站反感爬虫的到访，于是对爬虫一律拒绝请求，这时候我们需要伪装成浏览器，这可以通过修改http包中的header来实现。爬取网页时，服务器要验证客户身份，这时候需要设置Cookie。gzip/deflate支持现在的网页普遍支持gzip压缩，这往往可以解决大量传输时间，需要在专利爬取时做解压处理。由于网络原因，请求会失败，这时候需要设置失败后自动重试，设置超时。

### 文档预处理、特征选择与特征权重计算

专利文本预处理是从专利文本中提取出特征词来对专利文本进行表示的处理过程，它的主要任务是进行专利文本分词和去除停用词。去除停用词即是去除与专利文本自动分类关系不大的符号和词语。一般的做法是设计一个停用词表，

然后将专利文本中出现在停用词表的词或者符号去掉，只保留余下的词汇用来进

一步分词。本研究用了结巴分词器分词。

在文本分类的过程中，需要控制向量空间的维数，如果维数过高，则会产生维数灾难的问题[6]，分类器无法高效地对高维的特征空间进行处理，同时还会导致分类精度降低。而如果特征空间的维数过于稀疏，分类器在训练语料库的过程中无法得到足够的有效信息，也就无法做出正确合理的判断。因此，寻求一种有效的特征选择算法成为文本分类的关键问题。本文用来进行分类实验的专利文本的标题和文摘篇幅较短，但是每一篇简短的专利文本都会包含许多不同的特征词汇，大量的专利文本会导致得到的向量空间维数特别高。而由于专利文本中的语言文字具有相关性，许多特征词汇都是相互关联的，而且不同类别的专利文本或多或少也会包含许多公共特征词汇，对专利文本的分类起不到作用，因此特征提取是专利文本自动分类实验中不可缺少的关键步骤。 专利文本特征提取的难点在于特征词的选择和权值的计算。专利文本的特征空间维数过大将导致分类效率低下且分类准确率不高，因此需要降低专利文本特征空间的维数，选择出对分类贡献比较大的词汇用于分类；词汇的权值在分类过程中被计算处理，权值的大小代表了词汇对分类起到的作用大小。

信息增益法

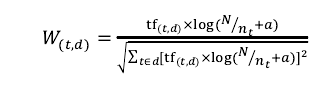
IG 在机器学习领域应用较为广泛，它从信息论角度出发，衡量词的引入能够给分类带来的信息。带来的信息越多，说明该特征越重要。IG 可以用下式表示：



TF/IDF 权值法

TF/IDF 权值法是一种统计方法，是在文本分析中最常用的向量权重计算方

法。TF 的数值越大，说明特征词对文本越重要；IDF 越大，说明该特征词在文本集中的分布越集中，越能够作为文本集的区分词。TF/IDF 权重计算公式如下：



TF/IDF 权值法计算法没有考虑词汇位置对词汇权重的影响，出现在标题相比于摘要中的词对专利文本主题的表达能力更强，将出现在不同位置的词赋予不同的权重并通过加权进行处理[7]。得到如下公式：

W（t,d）new=W(t,d)old \* P

### 文本表示模型

专利文本表示就是对文本进行形式化处理，以达到计算机能够理解自然语言

文本的目的。现在采用的一般文本表示模型有空间向量模型、概率模型和潜在语义索引模型。其中，在专利文本自动分类领域中应用最广的是向量空间模型。本文运用了多种文本表示模型，对比试验。提出利用 Google 最新深度学习工具 Doc2Vec和Word2Vec，将每一篇文章都转化为词向量的形式，通过深度神经网络重新训练，结果表明通过将文章转化为词向量的形式可以在某些方面更好反映出文档内在具体含义。

向量空间模型(Vector Space Model，VSM)由 Salton 等人提出，并成功地应用于 SMART 文本检索系统。VSM 的基本思想是将大量的文本表示为特征词矩阵，从而把对文本相似度的比较转化为特征向量在空间上的相似度比较，比较清晰且容易理解。

本文提出的词向量模型，是用深度神经网络对专利文本训练，得到词向量，然后运用得到的词向量建立文本模型。考虑运用了以下几种方式：

1. 直接将某个专利摘要的处理后的文本中的每一个词，将它对应于词向量拼接起来得到如下形式：

Vi=w1w2w3..............wn-2wn-1wn

这种表示形式由于每篇专利摘要处理后得到的词的个数不同，在输入分类器前，需要统一维度，对于短的专利摘要我们需要插值，对于长的文本我们需要采样。

对于短文本的插值采用以下形式：

Vi=w1w2w3..wk1......wk2....wki..wn-2wn-1wn

其中.wk1wk2....wki 为插入的i个向量，每个向量的计算为它前面的m个向量和后m个向量的和的平均值。

对于长文本我们采用采样处理。

1. 直接将某个专利摘要的处理后的文本中的每一个词，将它对应于词向量加权平均得到其对于的文本向量值如下形式：



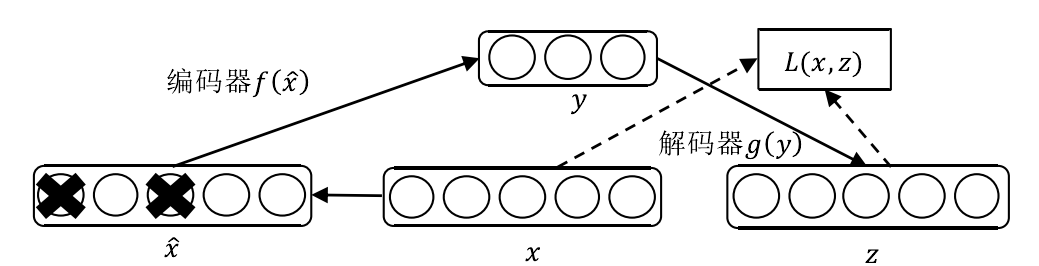
其中Wi为词向量，wi为它对应的TFIDF值，Vi为这篇专利摘要对应的文本向量。

1. 把语料库的单词聚类为N类，本研究使用K-means聚类方法聚类，每vocabulary/N（vocabular/N很小的数，可以考虑大概5~10个词）个词为一类，找到每个词的质心，对短专利摘要文本建立文本向量，维度为N,专利摘要文本中的每个词找到它的质心，把这个维度值加1。最终形成所有文本的向量，建立文本向量模型。
2. 运用词向量建立文本向量模型，此文本向量模型采用二维形式，把训练得到的词向量的维度作为文本向量的一个维度，把当前专利的词数作为另一个维度，把这些词向量连接在一起构成一个二维向量。

Vi=w1w2w3..............wn-2wn-1wn

此时的w 是列向量，得到Vi为二维矩阵形式。由于得到的每篇专利的文本的矩阵形式不同，我们运用深度网络的方法进行降维。

AE[8]构建的深度网络在图像和文本的特征降维方面取得了比传统的特征降维方法更优的效果，使得相关专家纷纷将 AE 应用到特征提取过程中来，并相继提出了稀疏自动编码器(Sparse Auto Encoder, SAE)[9]、降噪自动编码器[10]和卷积自动编码器(Convolutional Auto Encoder, CAE)[11]等不同的改进算法。其中，DAE 在特征提取中的应用相对比较广泛，主要应用于对动态视频纹理、音频、图像的特征提取中，在医学诊断中也有所应用。本文只对 DAE 在专利文本特征提取中的应用进行深入研究。 DAE 主要由两个部分组成，编码器和解码器，结构示意图如下图所示



1. 在运用词向量拼接成文档矩阵后，我们定义单词之间的距离[12]为：



对于任意两篇专利摘要，对于文本A中的每一个单词找到文本B中对应的单词，然后计算文本A和文本B之间的距离，它们之间的距离的计算如下：



6.目前已经有研究用深度神经网络的自然语言模型训练文档向量，其原理类似于词向量的训练，本研究计划采用训练得到文档向量，来实现对专利文献的自动分类。

### 分类算法

随机森林

随机森林顾名思义，是用随机的方式建立一个森林，森林里面有很多的决策树组成，随机森林的每一棵决策树之间是没有关联的。在得到森林之后，当有一个新的输入样本进入的时候，就让森林中的每一棵决策树分别进行一下判断，看看这个样本应该属于哪一类（对于分类算法），然后看看哪一类被选择最多，就预测这个样本为那一类。

在建立每一棵决策树的过程中，有两点需要注意 - 采样与完全分裂。首先是两个随机采样的过程，random forest对输入的数据要进行行、列的采样。对于行采样，采用有放回的方式，也就是在采样得到的样本集合中，可能有重复的样本。假设输入样本为N个，那么采样的样本也为N个。这样使得在训练的时候，每一棵树的输入样本都不是全部的样本，使得相对不容易出现over-fitting。然后进行列采样，从M个feature中，选择m个(m << M)。之后就是对采样之后的数据使用完全分裂的方式建立出决策树，这样决策树的某一个叶子节点要么是无法继续分裂的，要么里面的所有样本的都是指向的同一个分类。一般很多的决策树算法都一个重要的步骤 - 剪枝，但是这里不这样干，由于之前的两个随机采样的过程保证了随机性，所以就算不剪枝，也不会出现over-fitting。

按这种算法得到的随机森林中的每一棵都是很弱的，但是大家组合起来就很厉害了。我觉得可以这样比喻随机森林算法：每一棵决策树就是一个精通于某一个窄领域的专家（因为我们从M个feature中选择m让每一棵决策树进行学习），这样在随机森林中就有了很多个精通不同领域的专家，对一个新的问题（新的输入数据），可以用不同的角度去看待它，最终由各个专家，投票得到结果。

KNN 算法

KNN 算法是一种基于类比的分类方法，在训练过程中，KNN 生成所有训练

文本的特征向量，在测试过程比较测试文本的特征向量与所有训练文本特征向量

的相似度，从中找出 K 个最接近的训练文本，然后将测试文本分到这 K 个近邻

中所处最多的类别中去。

KNN 算法依靠周围有限的邻近的样本来确定所属类别，较其他方法更为适

合于类域重叠较多的待分样本集，在专利文本的分类中得到了广泛的应用，日本

国家科学咨询系统中心(NACSIS)策划主办的 NTCIR(NACSIS Test Collections for IR)在 theme 以及 F-term 的专利文本分类实验中，KNN 取得了最好的结果；Kim等利用 KNN 算法进行专利文本分类，达到了 74%的改善性能；Richte 等用 KNN 算法设计了两个(一个注重元数据，另一个忽略元数据)相似的专利文本分类系统并对其进行比较，关注元数据的系统取得了更高的准确率；国内这方面研究也较多，季铎等提出基于共享最近邻的 KNN 专利文本自动分类方法，在 NTCIR-8 专利分类评测任务中充分验证了其有效性。

SVM 算法

支持向量机算法的训练过程是要找到一个超平面，使得这个超平面的正反例

分别落在两侧，在所有超平面中与正反例的距离最大且到最近的正反例的距离相

等，然后对未知类别的专利文本，计算其位于超平面的一侧，即为其分属的类别。

支持向量机算法凭借高维灾难问题处理得当、数据稀疏性以及文本特征相关

性不敏感、准确率很高的优势，在专利文本分类中得到了广泛的应用。如 Chen

等利用开放源代码的 LibSVM 对专利文本的分类进行学习和预测，并提出了一个三阶段专利文本分类法，取得了很好的分类效果；Wu等提出一个基于新的遗传算法支持向量机的专利分类系统，在应用不同的内核后均取得了 80%以上的准确率；上海交通大学的吕宝粮教授及其团队基于支持向量机算法，实现了一个改进的并行化最小最大模块化支持向量机(Min-Max Modular Support Vector Machine，M3-SVM)算法，在大规模专利文本分类问题上比 SVMlight算法更加准确，更加节省时间。

## 2.3 创新点论述

目前，本文的创新点及想法主要有：

1. 根据专利的特点，在TFIDF权重的修改，本文的对比用的传统向量空间模型的TFIDF权重的设置，加入了专利本身的特点，如在标题中的题增加权重，领域的词加强权重。
2. 本文运用词向量构造文档向量模型，分别采用如下：

（1）把语料库的单词聚类为N类，每vocabulary/N（vocabular/N很小的数，可以考虑大概5~10个词）个词为一类，找到每个词的质心，对短专利摘要文本建立文本向量，维度为N,专利摘要文本中的每个词找到它的质心，把这个维度值加1。最终形成所有文本的向量，建立文本向量模型。

1. 运用词向量建立文本向量模型，此文本向量模型采用二维形式，把训练得到的词向量的维度作为文本向量的一个维度，把当前专利的词数作为另一个维度，把这些词向量连接在一起构成一个二维向量，然后对得到的矩阵进行降维，使得所有文档维度相同。
2. 直接将专利摘要的处理后的文本中的每一个词，将它对应于词向量加权平均得到其对于的文本向量。
3. 在运用词向量拼接成文档矩阵后，计算文档中对应单词之间的距离，并作为为文档之间的距离。
4. 直接将某个专利摘要的处理后的文本中的每一个词，将它对应于词向量拼接起来得到文档向量后，插值与采样处理为统一维度。

# 后续工作的设想、可能遇到的困难和问题及条件保障措施

对以上提出的文本表示方法想法进行实验验证，比较那种文本表示搭配上相应的分类算法的效果，得到最优的结果。在实验中，对于文本向量的表示与处理，各种统一维度方法需要不断尝试，可能效果不理想，需要不断修改。

# 已发表、录用的论文和已投稿的论文情况

# 参考文献（不占字数）

[1]Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. [Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013](http://arxiv.org/pdf/1301.3781.pdf)

[2] Tomas Mikolov,Quoc Le. Distributed Representations of Sentences and Documents. International Conference on Machine Learning.JMLR,2014

[3] 刘玉琴, 赖院根, 雷孝平. 基于IPC知识结构的专利自动分类模型[J]. 小型微型计算机系统, 2007, 28(12):2295-2298

[4] 李生珍, 王建新, 齐建东, 朱礼军. 基于BP神经网络的专利自动分类方法[J]. 计算机工程与设计, 2010, 31(23):5075-5078

[5] Caterina Camus,Riccardo Brancaleon. Intellectual assets management: from patents to knowledge. World Patent Information,2003,10(25):155-159

[6] Norbert Fuhr, Chris Buckley. A Probabilistic Learning Approach for Document Indexing[J]. Information Systems,1991

[7] 胡冰,张建立.基于统计分布的中文专利自动分类方法研究[J]. 现代图书情报技术,2013(Z1):101-106.

[8] Hinton G E. Reducing the Dimensionality of Data with Neural Networks[J ]. Science, 2006,313(5786):504-507.

[9] Vincent P, Larochelle H, Bengio Y, et al. Extracting and composing robust features with denoising autoencoders.[J].Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML’08, 2008:1096-1103.

[10] Bengio Y, Lamblin P, Popovici D, et al. Greedy layer-wise training of deep networks[J]. Advances in neural information processing systems, 2007,19:153.

[11] Vincent P, Larochelle H, Lajoie I, etal. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion[J]. The Journal of Mach

[12] MattJ.Kusner, YuSun, NicholasI.Kolkin, KilianQ.Weinberger. From Word Embeddings To Document Distances Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 2015. JMLR