**基于Doc2vec的政策文本分析 支撑政府科学决策**

大数据是继互联网、物联网、云计算之后又一伟大的技术革命。作为实施“十四大战略”之一的“国家大数据战略”，推进数据资源开放共享，促进行政决策改革与创新，对提高政府管理能力和公共服务水平有着非同小可的意义。本文以政策文本分析为突破口，借助大数据中较为成熟的Doc2vec算法，在剔除政策文本中冗余成分的基础上，将政策文本“粉碎”成诸多段落，实现了以段落为存储单位的政策文本数据库的创建；借助语义联想和语义搜索帮助用户快速查找相关政策条款；成功组建城市政策血缘关系库。为提高政府决策科学化提供了充分的数据支撑和理论依据。

1. **背景**

随着电子信息技术的飞速发展，各国的政府部门都开始进入大数据时代，对于如何把海量的数据利用于政府日常决策，提高政府决策的专业化、科学化成为提高政府科学决策能力的一个新的研究方向。当前，我国地方政府决策存在不同程度的问题，集中表现为规划缺乏、参与薄弱、成本过高、信息不全、忽视公众诉求等。作为政府领导，决策不能只凭主观经验，而是要掌握客观的海量的信息和资源，通过自然语言分析，对大量的可靠数据进行全面的分析研究后做出科学有效的决策，这样避免了过去政府决策的片面性和延时性。

政策文本是指因政策活动而产生的记录文献，既包括政府或国家或地区的各级权力或行政机关以文件形式颁布的法律、法规、部门规章等官方文献，也包括政策制定者或政策领导人在制定过程中形成的研究、咨询、听证或决议等公文档案。常见的政策文本分析方法有文本分析法、文献法、比较分析法、图示法等。然而在大数据时代，传统的政策文本分析方法不仅需要投入大量的人力物力，而且效果有限。因此采用基于自然语言处理的政策文本分析方法，是能够降低成本、提高决策效率的一种方法。

1. **国内外研究现状**

目前，政策文本分析方法几乎都基于词袋法(Bag-of-Word，BOW)。词袋法将政策文本看成是一些词的集合，在该集合中，每个词的出现是相互独立的，且不考虑词的顺序、语法和语义等信息。它将一篇政策文本表示成与训练词汇集合相同维度的向量，向量中每个位置的值即是该位置所代表的词在文档中出现的次数，并且随着新词汇的增加，文档向量维度也会增加。虽然词袋法在传统分类器上的分类效果不错，比如目前比较成熟的分类技术：回归模型、最近邻分类(KNN)、贝叶斯分类、决策树、RBF神经网络、支持向量机(SVM)等，但它依旧存在几个主要问题：1)维度太高，文本向量的维数与训练数据集中出现的所有单词的数目一样多，这样容易出现所谓的“维度灾难”现象，而且如果某一个词汇在训练集中没有出现过，则该词汇在测试集中出现时就无法成为该文本的特征；2)一篇普通文档只有1000个词左右，词向量的维度却能达到10万，利用率仅为1％，所以基于BOW表示的文档向量非常稀疏，不利于一些自然语言处理任务；3)词袋法无法很好地表示一篇政策文本的语义，它假设词与词之间相互独立，并不考虑词与词之间的关系，如“人才”与“能人”这两个词在用词袋法所表示的文档向量计算相似度时的值为0，但是“人才”与“能人”是同一种意思；4)词袋法很难区分同一个词在不同语境中的意义，如“先生”，根据上下文，它可能是对男性的称呼，也可能是古代对老师的称呼，但在词袋法中，其文档向量计算相似度为1。

随着深度学习的发展，基于神经网络的自特征抽取的词向量表示方法越来越受工业界和学术界的关注。基于前人的研究，Mikolov等人在2013年提出了word2vec模型用于计算词向量(即下文的Distributed Representation，后面均简称为词向量)。word2vec模型利用词的上下文信息将一个词转化成一个低维实数向量，越相似的词在向量空间中越相近。将词向量应用于自然语言处理非常成功，已经被广泛应用于中文分词、情感分类、句法依存分析等。

Le和Mikolov提出了Doc2vec方法作为一个处理可变长度文本的总结性方法。除了在增加一个段落向量以外, Doc2vec几乎等同于word2vec。相比于Word2vec，Doc2vec 优势在于**训练出每一篇文本的向量，更能全面理解文本的语义特征**。基于此，本文采用Doc2vec对政策段落进行向量化，同时，为了帮助业务人员理解聚类或分类的结果，本文还采用IT-IDF提取每个段落的主题词。

1. **关键技术**

政策文本分析过程中涉及到两个关键技术，词的向量化和文档的向量化。词的向量化认为一个词表示成一个向量，文档的向量化是一篇文档表示一个向量，而文档是许多词组合的综合体，因此，文档的向量是基于词的向量计算。

**1）词的向量化**

词的向量化就是将语言中的词进行数学化，也即把一个词表示成一个向量。词的向量化主要有以下3种表达方式。

(1)one-hot representation方式

这是一种最简单的方式，用一个很长的向量来表示一个词。向量的长度为词典的大小(通常达到10^5)，向量的分量只有一个1，其余全为(0，1)的位置对应该词在词典中的位置。比如，“人才”表示为[0 0 0 0 1 0 0 0 0 0 0 0 0 0⋯]，而“能人”表示为[0 l 0 0 0 0 0 0 0 0 0 0 0 0 0⋯]。这种方式虽然可以简单明了地表达一个词语，但是却无法有效表达它们的语义信息。“人才”和“能人”虽然是同一种意思，但利用常规的向量距离公式，比如欧几里德距离或者余弦距离公式，都无法有效计算它们的相似度，显然这种方式不能很好地表达词之间的相似性。

(2)Distributed representation(词向量)

这种方式能很好地克服one-hot representation方式的缺点，最早由Hinton提出，它是将词映射到一个低维、稠密的实数向量空间中(空间大小一般为100或者200)，使得词义越相近的词在空间的距离越近。上面的例子可以类似地表达如下，“人才”可以表示为：[0.843 0.125 0.734 0.345 0.654 ⋯]，而“能忍”为[0.923 0.231 0.698 0.633 0.743 ⋯] ，显然，这种表示方式有利于使用距离向量公式比较词向量之间的相似度。

(3)word2vec模型训练词向量

通过借鉴Bengio提出的NNLM(Neural Network Language Model)以及Hinton的Log Linear模型，Mikolov等提出了word2vec语言模型。word2vec可以快速有效地训练词向量。

word2vec模型有两种，分别是CBOW模型以及Skip-gram模型。其中CBOW模型利用词加(f)前后各c(这里c=2)个词去预测当前词；而Skip-gram模型恰好相反，它利用词w(t)去预测它前后各c(c=2)个词。

图1：CBOW模型 图2：Skip-gram模型

**2）文档的向量化**

(1)BOW模型

文档的向量化就是将一篇文档表示为一个向量，主要是基于词的向量化。将文档向量化之后，就可以利用常规的距离向量公式比较两篇文档之间的相似度。传统的BOW可以看作是词的one-hot表示向量的叠加，比如“人才”的词向量为[0 0 0 0 0 1 0 0 0 0 0 0 0 0 0⋯]，“能人”的词向量为[0 1 0 0 0 0 0 0 0 0 0 0 0 0 0⋯]，而一篇仅包含“人才”和“能人”这两个词的文本就可以表示为[0 1 0 0 1 0 0 0 0 0 0 0 0 0 0⋯]。显然这种表达方式存在的问题与one-hot一样，由于其特征向量的高维性和稀疏性，很难利用常规的向量距离公式有效地计算两篇文档之间的相似度。当然传统的BOW也有许多优化的方法，利用TF-IDF加权是其中一种，即将文本向量中出现非“0”的值替换为TF-IDF权值，这样的特征向量比传统BOW在文本分类方面更有效。

(2)Doe2vec模型

Doc2vec模型的训练与word2vec模型类似，在利用词的上下文对当前词进行预测的训练过程中添加了一个文档特征向量，其预测模型和训练过程分别如图3和图4所示。



图3：Doc2vec预测模型

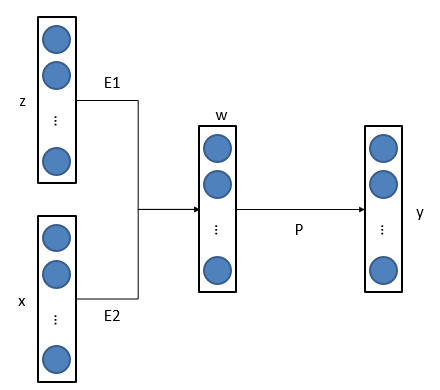


图4：Doc2vec训练模型

从图3可以看出，该模型是利用前几个词去预测当前的词，比如用“我”、“喜欢”、“吃”、“苹果”这个句子中的前3个词去预测第4个词，而且在训练过程中添加了一个文档特征向量D，其余部分与word2vec预测模型类似，可以将该文档特征向量看作是一个表示当前文档中的其余部分信息或者主题信息的向量。通常根据文档特征向量的维度不同，将该文档的特征向量与词向量相结合的方法包括取平均值和串联两种方式。图4中的训练过程是利用串联的方式对文档向量进行训练。其中z代表文档向量，x代表词向量，w代表隐含层权值，y代表输出层的值。与word2vec类似，y中的每一维即代表词典中的一个词，E1、E2和P分别代表它们之间的连接矩阵。其训练过程与word2vec模型类似，并利用BP算法调节参数。

1. **方法框架**

本文首先收集政策数据，通过用户上传政策文件和机器定时爬取网络政策信息；其次，对收集到的政策信息进行分段，构建语料库；接着，采用Jieba中文分词对段落信息进行分词；然后，进行特征处理，包括利用TF-IDF提取段落主题词，利用Doc2vec提取段落向量；最后，对段落向量进行Kmeans聚类，根据聚类结果对政策进行分类，经业务人员质量评估得出最终结果，具体过程如图5所示所示。



图5：政策分析文本方法框架

1. **政策数据收集**

政策数据主要来源于两大类，一类是业务上传数据，包括PDF文件、Word文件、图片、Excel文件和特定URL，另一类是网络爬虫。



图6：政策数据收集

1. **构建语料库**

由于直接使用政策内容作为语料库，可能会引起两个类别的政策分到同一类，即存在误分类情况。为了减少这种情况，本文根据换行符等特殊标记将整个政策内容进行分段，通过段落信息构建语料库。目前已形成人才段落信息10000条，其它类别段落信息30000条，语料库展现形式如图7所示。



图7：语料库

1. **中文分词**

在进行特征提取之前，需要对政策段落进行中文分词，本文采用Jieba中文分词，该软件分词效果较好，效果如两个方框所示。



图8：中文分词效果图

1. **引入停用词**

中文分词之后的段落存在很多无效的词，比如“在”，“等”，还有一些标点符号，不想在文本分析的时候引入，因此需要去掉，这些词就是**停用词**。同时，不同行业的停用词表是略有不同的，因此，本文对每个行业构建**特定停用词表，目前已形成人才、创新、制造三个领域停用词表。**



图9：不同类别停用词引入后效果

1. **特征处理**

特征处理就是提取段落的特征向量，本文主要采用TF-IDF和Doc2Vec两种技术，其中TF-IDF实现简单，但是缺少语义信息，而Doc2Vec包含语义信息，两者相互补充。



图10：TF-IDF算法步骤

TF-IDF是Term Frequency -  Inverse Document Frequency的缩写，即“词频-逆文本频率”。它由两部分组成，TF和IDF。TF指的是各个词在段落中出现的频率，IDF指的是各个词在语料库中出现的频率。

TF-IDF应用包括段落主题词，每个段落获取TF-IDF得分前五的词，还是以之前的段落内容为例，它的段落主题词是制造业、创新、技术、制造、高端，段落主题词如右图所示。



图11：词云图

Doc2Vec利用神经网络计算文本的特征向量，能更深入地表征出段落文本的语义特征。在Doc2Vec中，向量维度和窗口大小直接影响段落文本向量的表示效果，窗口太小会导致目标词的上下文相关词数量过少，容易漏掉与目标词关联度高的词语；窗口过大会引入噪声词，根据实验结果，本文选取窗口大小为4，如下图所示。文本向量每一维的值代表一个具有一定语义的特征，因此高维数的向量可以提升模型的准确率，但是维度过大会引入噪声，根据实验结果，本文选取向量维度为300。

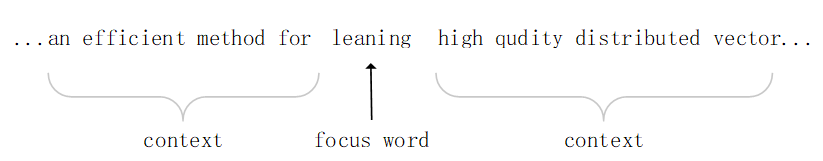


图12：文本的特征向量

1. **建立分析模型**

提取出段落向量之后，本文就可以对所有段落向量进行聚类，实验中采取Kmeans方法，研究人员选取类别数（K值），通过段落向量计算每个段落的相似度，然后得到每个类别的政策信息。接着研究人员对聚类结果进行人工筛选，对于筛选过语料库进行重新聚类，直到获得研究人员满意的聚类结果。

1. **具体应用**

**政策分析、执行与监管**

**有助于政策制定者提高政策制定的科学性和有效性，是政策真正成为应对改革风险的有利武器，成为化解矛盾的调节器，成为社会稳定的保障锁，成为社会发展的助推器。**

**语义联想**

语义联想包括同义词和上下文语义词，例如当输入新能源，就会输出“汽车产业”、“乘用车”、“电站”等上下文语义词。政策制定者在做政策规划时就可以

通过语义联想功能，

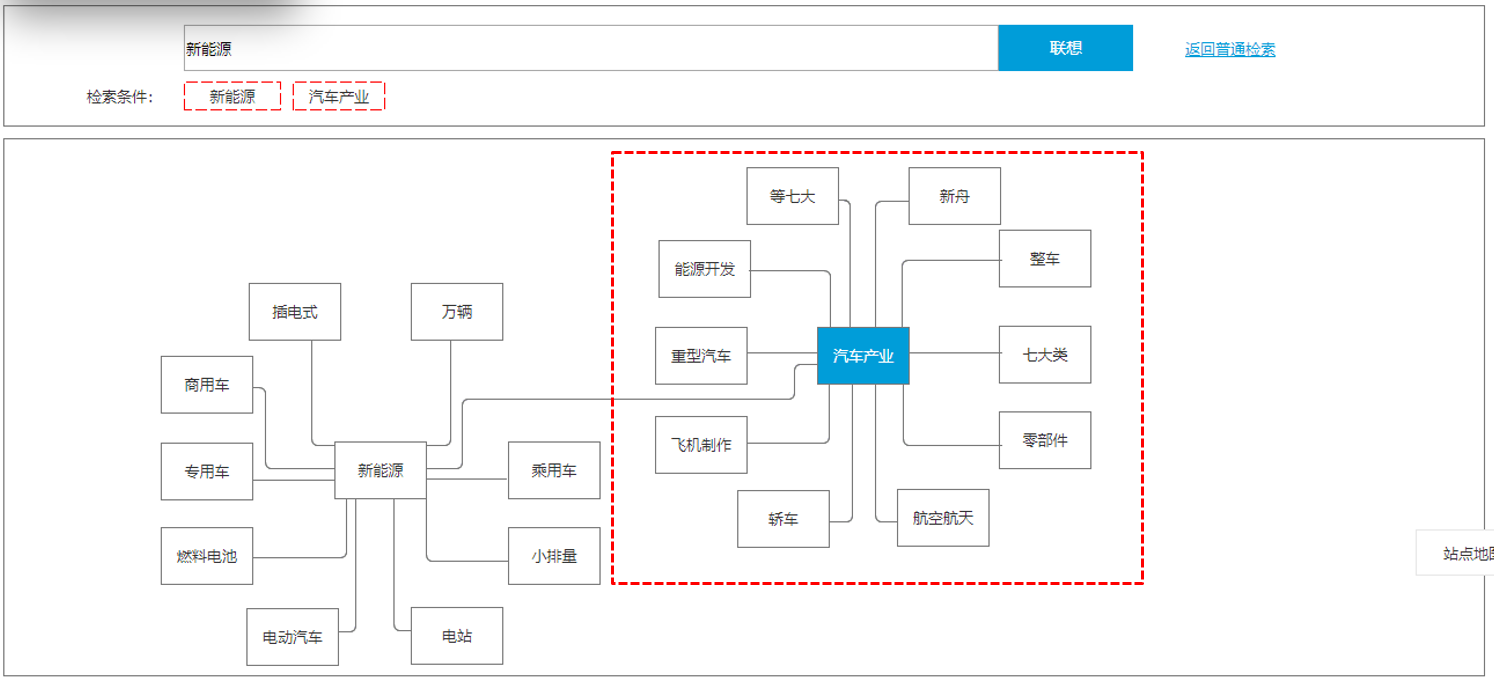


图13：语义联想

Doc2Vec在平台中第二个应用是**语义检索**，寻找同义词，例如输入养老保险，输出包含社会养老保险、基本养老金等段落。

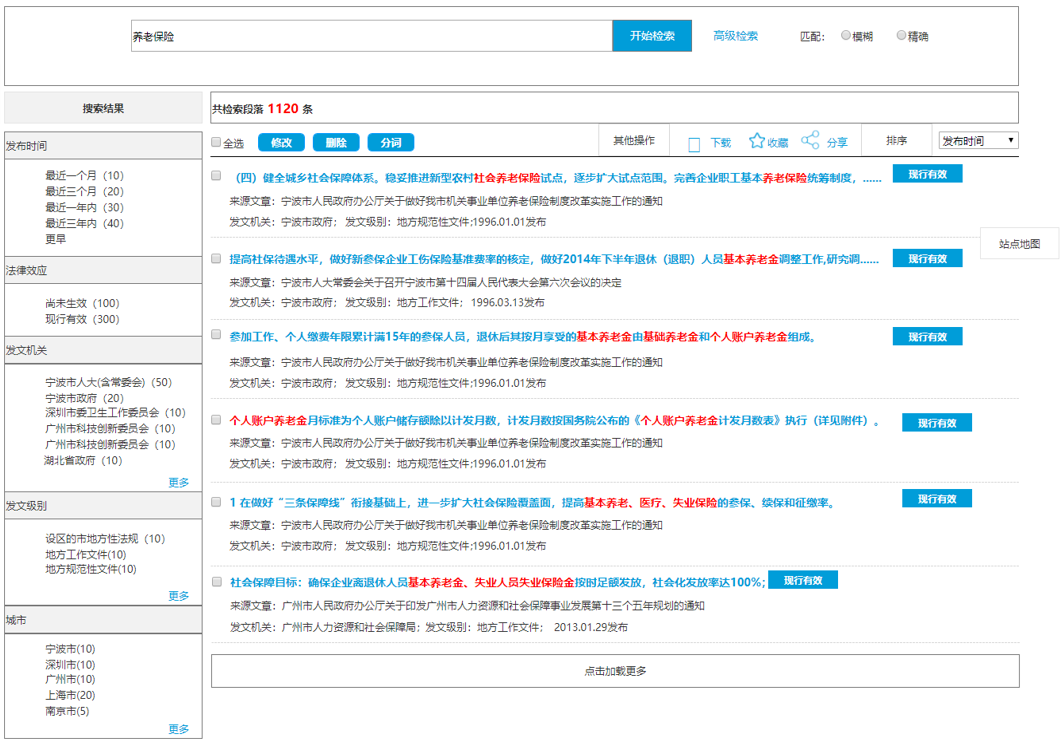


图14：语义搜索

Doc2Vec在平台中第三个应用是**政策血缘关系分析**，通过分析政策发文字号，挖掘出每篇政策的引用和被引用关系，构建政策血缘关系。



图15：某政策血缘关系图

1. **展望**

本文提出了一种基于Doc2vec的政策文本分析，与传统的基于词袋法进行文本相似度的计算方法不同，该方法充分考虑政策段落的语义情况，根据段落相似度对所有段落集合进行聚类从而对海量政策文本进行有效率分类。

本文后续的研究将在现有方法的基础上，进一步深入挖掘出政策段落所蕴含特征，包括政策发文机构、政策覆盖面和政策条款分布、政策有效期等特征，更好地提高政策文本聚类的准确性，更直观的展示各区域政策特色或量化得分情况，进而支撑政府科学决策。具体可以从以下三方面开展工作：

1）专利的语义检索：专利的检索方式可从精准检索转向语义检索。借助词语间关联性，检索出与关键词语义相关却词语不同的专利，为检索人员提供更多的可参考资料；

2）技术支持问题分类：技术支持类问题的描述具有描述者的个人特色，合理提取问题中的关键词，并可以区分为“注册”类、“密码”类、“服务质量”类等多种类别，帮助开发人员快速了解平台存在的问题；

3）申报项目技术领域分析：研究机构或企业申报项目时缺少项目涉及的技术领域的信息，历年项目也只能围绕申报单位、项目申报金额、项目获奖等级等信息进行分析。但对代表申报单位当前研究技术领域的关键信息却无从下手。因此，借助本文提供的方法，从申报项目名称和项目正文可以快速的提取申报项目研究技术领域。