**基于Doc2vec的文本分析算法及其应用**

传统的文本分析方法大多采用词袋法把文本建模成词向量，利用余弦相似度量等方法来计算文本之间的相似度，进而对文本进行分类，但是这类方法没有考虑文本的语义信息。基于此，本文提出了基于Doc2vec的文本分析方法，该方法首先应用自然语言处理技术对政策文本进行预处理，然后利用Doc2vec方法提取段落向量，并根据段落相似度对所有段落集合进行聚类从而对海量文本进行有效率分类，有效地支撑政府科学决策。最后，详细介绍了该方法的具体应用，分别是政策血缘网络、智能客服、专利检索和技术重点演进分析，依靠政策血缘网络有效地**延缓和避免政策碎片化，**通过智能客服可以明显提升宁波市科技创新云平台的用户体验并节省服务成本，借助语义检索功能向用户返回更加准确的搜索结果，构建技术重点演进路线支撑专利技术创新性。

1. **国内外研究现状**

目前，文本分析方法几乎都基于词袋法(Bag-of-Word，BOW)。词袋法将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。它存在几个主要问题：1)维度太高；2)文档向量非常稀疏，3)词袋法无法很好地表示一篇文本的语义。

为了解决词袋法存在的问题，Mikolov等人在2013年提出了word2vec模型用于计算词向量。该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。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**

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：

One-hot representation方式、Distributed representation和word2vec模型训练词向量，其中word2vec方法能够更好地表达语义。

**2）文档的向量化**

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。目前文档向量化主要有两种方法：BOW和Doc2vec。其中，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。

1. **方法框架**

本文首先**收集文本数据**，文本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；由于直接使用文本内容作为语料库，可能会引起两个类别的文本分到同一类，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文本内容进行分段，通过段落信息**构建语料库**；接着，采用Jieba中文分词对段落信息进行中文分词；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**；然后，进行**特征处理**，利用Doc2vec提取段落向量，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；最后，对所有段落向量进行Kmeans聚类，通过段落向量计算每个段落的相似度得到每个类别的政策信息，研究人员会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。具体过程如图1所示所示。



图1：基于Doc2vec的文本分析方法总体框架

1. **具体应用**

目前，基于Doc2vec的文本分析方法可以应用于政策血缘网络、智能客服、语义分析和技术重点演进分析这四个场景中。

1. **政策血缘网络**

有研究发现**，**目前我国**政策碎片化现象**比较严重**，**主要表现为政策间相互矛盾、政策重叠，政策体系不够完整且缺乏配套、政策缺乏稳定性等现象**。**通过Doc2Vec政策文本语义分析，实现**政策聚类**，进而对政策网络进行层级划分，实现同功能政策的替代或分解，减少平行政策数量，从而**延缓**政策碎片化；对于新政策，通过**语义检索**在语料库中寻找与新政策相似的已存在的政策，进而为政策制定者提供政策定位的辅助信息，从而**避免**产生新的政策碎片。

构建**政策血缘网络，从而形成清晰的政策体系结构。政策血缘关系**指的是目前所有有效或已失效的政策都因与宪法之间的祖孙关系而存在内在的联系，例如，在我国的社会保障体系中，政策之间的政策血缘关系如下图所示。

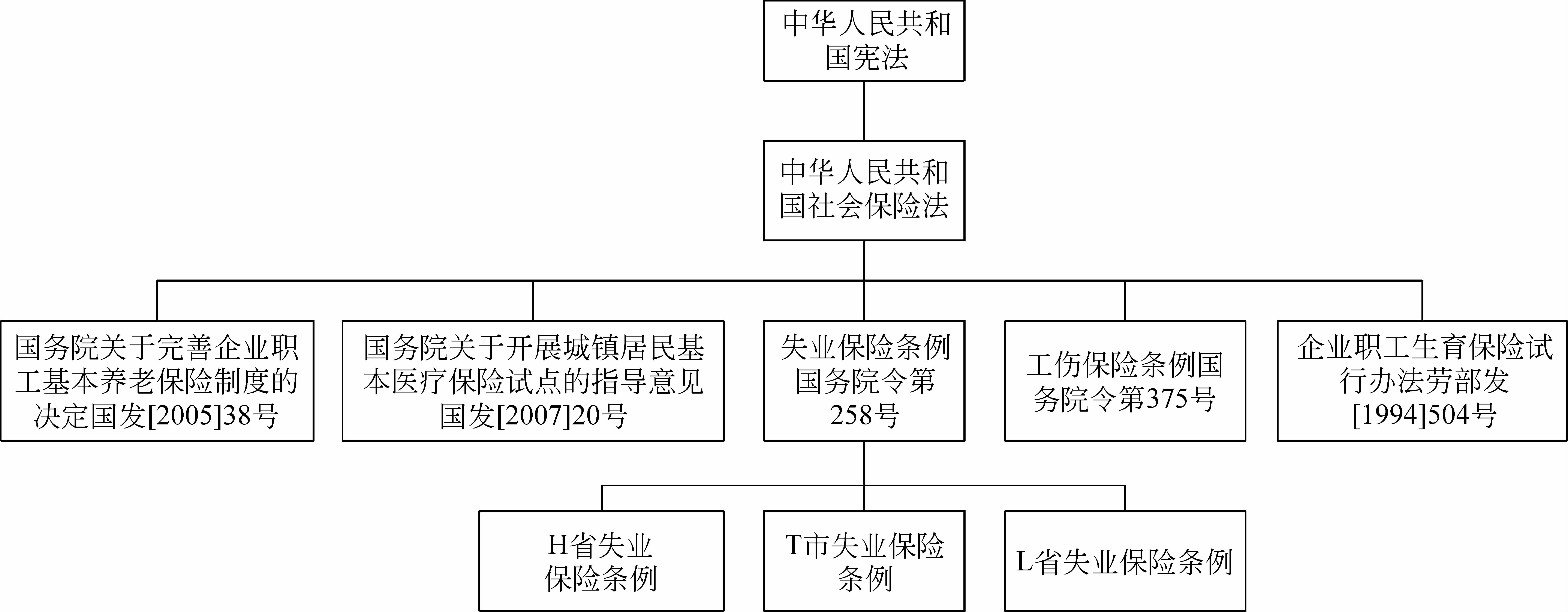


图2：我国社会保障政策族谱树

上图中展示的就是一棵政策族谱树，它描述了图中节点的父子关系。由图可知，树中路径距离较远的节点之间的血缘关系应当较为薄弱，树中路径距离较近的节点之间的血缘关系应当较为紧密。

同时，通过Doc2Vec政策文本语义分析，挖掘出政策间的**隐形血缘关系，从而实现隐形血缘政策网络体系。**以中国社会保险体系为例，该体系将社会保险分为八大类，其组织形式如下图所示。对上述不同类别的政策子系统，其内部又可根据不同的侧重分出不同的子类，每个子类包含若干相关政策。有些政策的组织形式彼此隔离，互不相关。但是**从内容上讲，它们的实施办法相似或相关，这种联系并没能从政策族谱树中体现，这就是政策间的隐形血缘关系**，挖掘这种政策间的**隐形血缘关系。**

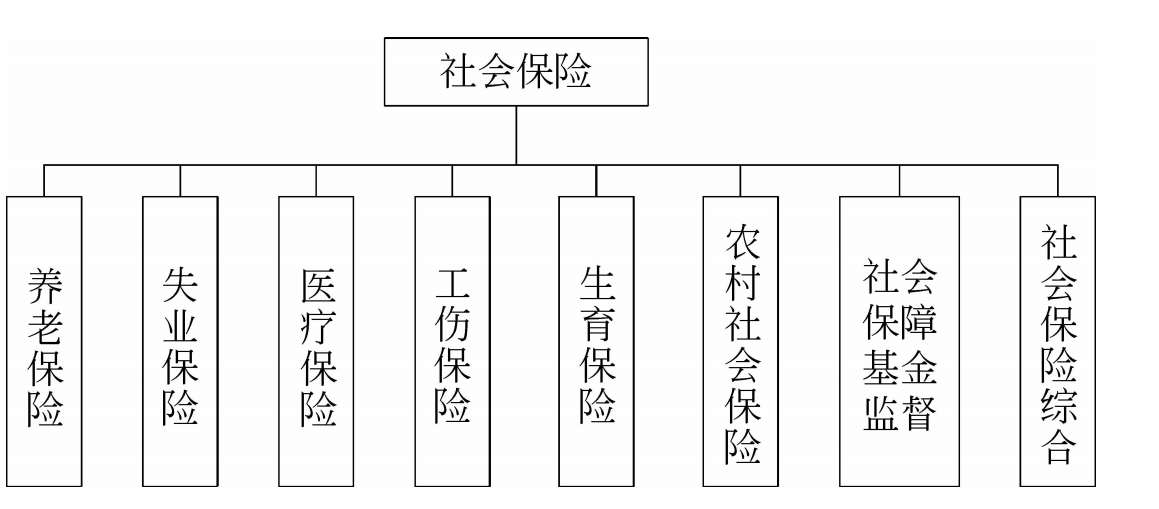


图3：我国社会保险政策分类

1. **智能客服**

随着云平台用户的增多，在使用过程中不可避免产生许多问题。以项目申报系统为例，每临近项目截止日期时许多用户会操作不当或者对申报系统的不熟悉而产生诸多问题。如果系统可以为用户及时的解答疑问，便会极大的提高用户满意度。但是依靠人工客服提供全面的服务，当询问量过多时，明显不能满足需求，且流程繁琐。为解决这一问题，可以利用Doc2Vec文本语义分析形成智能问答系统，迅速反馈给用户准确、简洁的答案。

建立智能问答系统，如图所示，主要由三部分组成，分别是问句分析、信息检索、答案抽取。1、问题分析将用户语言转换成计算机能够识别的语言，主要包括对问句进行中文分词、词性标注、句法分析，为了能够快速准确找到答案，还要对问句进行分类，最后进行关键词提取和拓展。2、信息检索 首先利用Doc2Vec将常见问题、热点问题进行聚类形成问题库，然后对用户提出的问题进行相似度判定3、答案抽取 通过问句相似度计算在问题库中找到与提问问题相似的结果，从高到低进行排序，最终提供答案数据

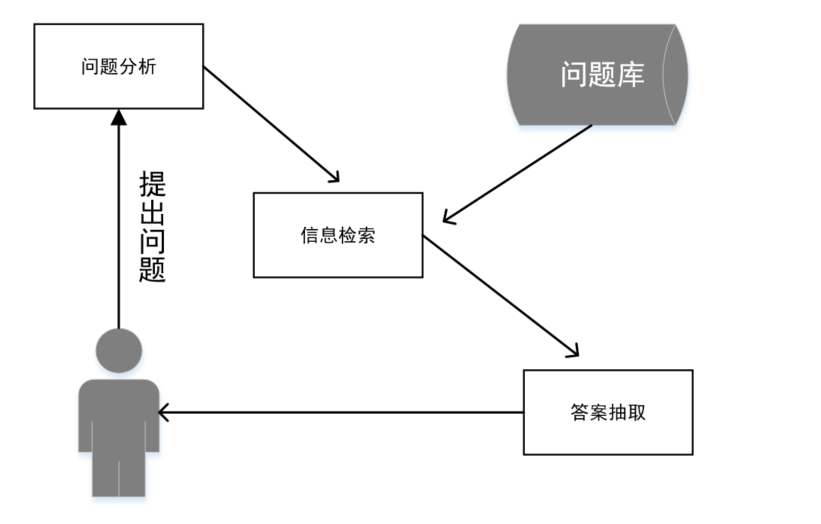


图4 智能问答系统

智能问答系统在云平台为用户服务的过程中起着显而易见的优势，一是提升用户感知，减少用户问题得到解决的难度和复杂度，二是提升服务效率，缩短咨询处理时限，节省服务成本，三是收集用户诉求和行为数据，支撑产品迭代优化。随着技术的成熟，可以运用到知识库智能检索、科技决策等方面。

1. **语义检索**

专利作为技术创新的重要标志和体现，在很大程度上代表着一个国家或企业的技术水平和潜在的技术竞争力。专利文献检索在专利的申请、审查、管理和运用过程中具有重要作用。专利文献检索分为多种方式，包括关键词检索、语义检索等。为了提高查准率和查全率，使用关键词进行检索需要制定合适的检索策略，编写全面的检索式，但是由于检索策略或检索式本身的缺陷，可能导致漏检或者带来极大噪音，使得检索效率下降。因此，语义检索作为一种方便、快捷而又智能化的检索方式，在专利文献检索中得到广泛应用。

专利审查检索是发明专利实质审查程序中的一个关键步骤，每一件发明专利申请在被授予专利权前都应当进行检索。专利审查检索的目的在于找出与申请的主题密切相关或相关的现有技术中的对比文件，或找出抵触申请文件和防止重复授权的文件，以确定申请的主题是否具备新颖性和创造性，或者是否满足同样发明创造只能被授予一项专利的要求。

一般地，需要通过确定检索要素（通常是关键词、分类号，也可能是公司名称等字段）从几千万包含标题和文摘的专利文献中，找到与目标文献最相关的文献集合。然而，基于关键词匹配的检索技术只能提供单一角度的查询，如果使用者对于查询领域不是很了解，构造的关键词检索表达式不完整或不准确，很难得到满意的检索结果。同时，由于检索要素的确定是极其困难的（如关键词中，包含很多的同义词、下位词、近义词、俗称、别名等情形），因此多次利用检索要素进行“and”运算，必然会产生漏检。同时还需要花费大量的时间逐篇浏览文摘或全文，以确定与本申请最相关的文献。

语义检索也称语义搜索，是通过对文本中原来的信息进行语义上的处理，将从中获得的各种概念数据组合成知识库，然后根据对用户提问的理解，从知识库中发掘用户需要的信息。其中，语义检索需要分析用户输入的各种形式的查询目标，例如通过关键字匹配或者与用户浏览交互的方法，将查询目标匹配到知识库中的类和实体，通过本体关系推导，发现与用户查询目标相关的概念。

本文提出的基于Doc2vec的文本分析方法，通过训练出每一篇专利文本的向量，全面了解专利文本的语义特征，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。在现有的基础上还可以通过对主题词和停用词库的维护来降低专利文本中的噪声。

此方法不仅可以大量减少投入的人力物力，提高检索的全面性和准确性，同时也可以通过维护主题词库和停用词库来适应不同的场景（例如授权发明专利和实用新型专利、创新政策和人才政策）。语义检索同时适用于专利查重、专利维权等场景。

1. **技术重点演进分析**

目前国内的专利一般采用国际专利分类表IPC（International Patent Classification），除非一项发明的应用本身能决定它的技术特征，通常是依据供功能和内在性质在IPC中查找对应的技术领域类别。由于新技术的不断涌现，专利文献每年增长约150万件，截止2017年约有5000万件。而且各国的科学技术的发达程度差距很大，它并不能够适应每个国家的具体情况。另外，IPC的建立是基于纸件专利文献的管理与检索，在计算机、通讯网络等新技术快速发展的今天，它显现出一些不适应。更重要的是IPC分类在实际研发过程中无法精确的定义技术领域中的技术重点。研究者会把某个技术领域中的技术重点作为主要研究对象，查找相关技术文献和相关研发企业，从而了解该技术重点的基础技术、发展态势和未来发展方向。因此，对专利文献进行技术重点分类是行而有效的解决以上问题的手段。

以专利文献作为研究对象，选择Doc2Vec文本分析方法，剖析专利文献所属技术领域、所属技术重点。本文首先将专利按照“文章—段落—关键词”逐级“粉碎”处理，借助专家的专业能力从诸多技术类关键词中审核出可以表明技术重点的词语，从而组建技术重点语料库。分析流程图如图X所示。



图5 专利技术重点分析流程图

语料库中的数据主要来源于专利/期刊文献，因此语料库中包含多个维度的数据，例如时间维度、区域维度、申请单位维度、发明人维度、基础技术维度、前沿技术等多个维度。如图X所示。



图6 语料库维度图

充分利用多维度数据，获取多角度分析结论。技术重点与其他维度数据结合可以得到很多结论，譬如，以时间为研究对象，可得到技术重点随时间发展趋势；以区域为研究对象，可摸清技术重点的起源地、发展地、需求地等信息；以申请单位为研究对象，则可以获悉掌握技术重点的主要单位等信息。不同维度之间的组合又可以得到更多的信息，例如，以区域和时间为研究对象，可清楚技术重点的发展密集区域和密集期间；以时间和申请单位为研究对象，可清晰的看到单位的技术重点历史演进图；以申请单位和前沿技术为研究对象，可摸清掌握前沿技术的申请单位。诸如此类，举不胜举。因此，以专利为数据来源的技术重点语料库，可以进行多维度数据的多方式结合，从而更加全面的了解技术重点的过去、现在和未来。

全方位数据，多维度数据组合分析，充分摸清技术路线走向和现今发展态势，为市场提供具有理论支撑的可行性建议，从而提高专利的市场转化率，增加技术发展方向准确性，确保专利技术创新性。

1. **总结**

本文提出了一种基于Doc2vec的文本分析方法，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文本进行有效率分类。同时，本文详细介绍了基于该方法的四种应用场景：政策血缘网络、智能客服、专利检索和技术重点演进分析，其中，依靠政策血缘网络有效地**延缓和避免政策碎片化，**通过智能客服可以明显提升宁波市科技创新云平台的用户体验并节省服务成本，借助语义检索功能向用户返回更加准确的搜索结果，构建技术重点演进路线支撑专利技术创新性。