**基于迁移学习和自然语言处理技术的专题领域QB信息感知分析技术研究与应用**

1. **必要性**
   1. **国内外现状和发展趋势**
   2. **存在的问题**

长期以来，QB信息是以非结构化文本为主要载体进行存储、处理和传递的。作为一种直接面向特定需求、特定用户的信息提供方式，专题领域QB信息可以为QB人员提供更加准确、全面、专业的QB研究资料。随着互联网的发展，大量的网络科技文献不断涌现，这也使专题领域QB发生了很大的变化。经过长期工作实践，传统的人工处理方式在技术和管理上已较为成熟。然而，在大数据和人工智能时代，传统处理方式已难以满足专题领域QB信息的感知和分析处理需求，与当前信息化、智能化发展趋势存在不相协调的弊端。传统的人工处理方式与智能计算机的高速计算、逻辑推理能力相比，不仅作业速度较慢、处理效率较低，而且会消耗大量的人力资源，使相关QB人员陷入低价值的重复劳动当中。海量QB文本蕴含着大量有价值的QB信息，但是由于受到人为因素(例如研究人员的专业素养、工作状态等)的限制，很难实现文本之间的信息关联整合，难以充分挖掘出文本当中的有用QB，从而导致许多新知识、新信息被湮没。传统的感知分析技术已难以有效处理海量信息，存在信息过载的问题；其次，为了保证训练得到的信息感知模型具有可信的效果，传统的信息感知方法通常假设数据的生成机制不随环境改变，即要求源域数据和目标域数据具有相同的分布。在跨领域后，由于源域数据与目标域数据间的分布规律不同，模型往往无法适用于最新的领域数据，导致信息感知的效果较差。但很多情况并不满足这种假设，通常需要众包来重新标注大量数据。

1. **研究目标**

面向专题领域智能化发展的大背景，着眼QB信息自动化处理过程，基于自然语言处理及迁移学习等相关技术，为人工智能时代的QB信息处理带来新的突破口，从专题领域海量QB数据中进行信息过滤、抽取、计算及推理，准确高效的处理、理解、挖掘及展示QB信息。针对QB研究工作中面临的信息过载无法有效处理、QB线索发现不全面、人工智能技术应用水平不足等问题，满足特定专题领域QB信息的自动识别、QB线索发现、智能化分析等需求，开展当前先进的预训练模型技术探索及迁移学习建模技术研究，突破基于融合主题模型与预训练词向量的QB信息自动识别技术、基于预训练模型和迁移学习的粗粒度信息过滤技术、基于无监督的细粒度信息抽取式摘要技术等关键技术，实现对专题领域QB信息的自动识别筛选能力；开展基于迁移学习的自然语言处理技术研究，突破基于迁移学习的跨领域专有实体识别技术、基于远程监督的多实例专业领域关系抽取技术、基于数据增强的专业领域事件抽取技术等关键技术，实现专业领域信息中重要项目、关键技术、相关型号、重点事件等QB线索的自动识别和发现；开展知识图谱及知识关联技术应用方法研究，突破基于本体建模及Neo4j的图谱构建与存储技术、基于VUE+D3框架的交互式图谱可视化技术、基于知识计算的图谱关联关系探索技术等关键技术，实现知识点可视化和关联探索。将专题领域内海量QB信息处理成易于计算机理解的结构化形式，帮助专题领域信息智能识别，隐含QB知识发现、可视化关联分析等作战运用，进一步提高态势QB保障能力。

1. **研究内容**
   1. **基于预训练模型和迁移学习的QB信息自动识别技术**

面向专题领域智能化发展的大背景，QB研究工作面临信息量过载、工作人员知识位势不同，从而导致无法有效利用高复杂、多维度数据信息的问题，项目拟通过主题聚类技术、预训练模型及迁移学习技术，实现基于相关QB信息的领域识别、信息过滤及信息摘要功能，最终达到从海量数据中快速识别及定位特定专题领域QB信息的效果。

**1）基于融合主题模型与预训练词向量的QB信息主题识别技术**

QB主题作为QB信息的重要载体，是QB内容和思想的凝练。项目拟采用融合主题模型与预训练词向量的主题聚类技术实现对相关QB信息的领域识别，完成从海量的QB数据中获取与特定专题领域最相关的数据信息，为后续进一步实现粗细粒度信息过滤提供信息预处理支撑。

传统的主题聚类技术一般采用关键词识别或者关键短语匹配的方式来实现从信息到主题的映射，显然这种单一的、低纬度的映射方式无法满足当前QB业务中信息类别多、共现信息少、更新速度快、精度要求高等特点。因此项目拟采用融合主题模型与预训练词向量的方式，通过构建融合特征向量，提高QB数据信息的语义表征，进而实现对QB信息的有效主题识别。

首先项目拟采集不同领域相关文本数据构建文本语料库，并对文本数据进行清洗、标注、分词等预处理操作，其次从文本层面利用主题模型提取主题向量，并从词汇层面利用预训练模型提取词向量，最后利用TF-IDF计算词向量权重，进而构建文本融合特征，实现对文本浅层主题特征与词汇语义特征的融合，最后通过聚类算法对文本主题进行聚类。在维度层面，低维稠密的加权词向量与文档主题分布向量纵向拼接后仍然是低维稠密向量，解决了短文本数据高维稀疏问题；在语义层面，向量拼接后的融合特征既包含文本全局语义，又包含了词汇顺序信息和深层语义关联信息，词向量加权降低了噪音词汇的干扰。因此，通过融合特征方式表征文本，不但弥补了单一主题模型与词向量二者的缺点，而且丰富了文本向量的语义信息。

项目拟采用融合主题模型与预训练词向量的主题识别方式，充分考虑文本层面与词汇层面的特征，通过联合表征的方式对QB文本信息的主题进行聚类，实现对专题领域海量信息的预筛选工作。

**2）基于预训练模型和迁移学习的粗粒度信息过滤技术**

在已知海量信息所属特定领域主题的情况下，采用自然语言处理相关技术进一步对该领域的信息进行筛选，进而能够迅速定位到该领域相关的QB信息。为了能够捕捉信息的重点，增强对QB信息语义层面的理解，实现对领域相关信息的精确过滤，本项目拟采用预训练模型构建领域内海量信息和领域主题的向量表征，进而通过迁移学习技术实现跨领域信息主题的快速理解与信息过滤。

传统的信息过滤技术往往只分析QB信息表层文本的统计信息，忽视了QB信息深层的语义，或者需要大量的标注数据进行训练才能理解QB信息深层的语义。同时，传统的信息过滤技术难以处理跨领域的问题，而通过预训练模型和迁移学习技术，不仅能够理解QB信息深层的隐含语义，而且能够在小样本数据集上实现信息过滤能力的迁移。因此，项目通过结合预训练模型与迁移学习的技术实现信息过滤能力的跨领域应用。

基于预训练模型和迁移学习的信息过滤技术主要由过滤模块和迁移模块构成。其中，过滤模块首先使用预训练模型对领域内海量QB信息的主题和信息进行统一编码，实现主题和信息的向量化表示；其次对主题向量和信息向量进行相似度比对，相似度得分越高，则主题和信息之间的相关性越高；最后，以相关性为基准，筛选出相关性高的信息作为结果，从而达到领域内海量信息过滤的效果。

仅使用过滤模块往往无法处理跨领域的海量QB信息，进而导致在其它领域内信息过滤效果下降，因此需采用迁移模块对过滤模块所具备的信息过滤能力进行跨域迁移。迁移模块通过获取新领域内的小样本数据对过滤模块相关模型进行训练，并利用训练后的模型快速学习并理解新领域内容，进而实现新领域的信息过滤。

**3）基于无监督的细粒度抽取式信息摘要技术**

经过上述基于预训练模型和迁移学习的粗粒度信息过滤处理后，可以得到与领域最相关的若干信息数据，从细粒度角度来看，单条过滤后的信息中仍可能存在大量重复的冗余信息，导致QB人员难以快速获取业务所需的核心信息。针对上述问题，项目拟采用一种基于无监督的细粒度抽取式信息摘要技术，对单条信息在细粒度层面进行摘要抽取，在不破坏词义、次序和原文内容的基础上，抽取出原文的精简版本，从而达到快速了解关键QB信息的目的，加速对QB信息的处理流程。

传统的信息摘要抽取大多数采用监督学习方式，不仅需要大量的人工标注数据来训练模型，而且不能很好的理解领域内QB信息，导致最终得到的摘要结果质量不高。因此项目拟使用向量表示方法将专题领域内的QB信息中的句子按原文顺序依次编码为句向量，然后采用无监督算法对句子向量进行更为高阶的特征表示，最后生成包含专题领域内关键信息的文本摘要。

上述基于无监督的信息摘要技术流程可以分为三个主要步骤：首先采用词向量表示方法将文本映射成空间向量，通过词所在位置的上下文信息得到词对应的向量化表示，在生成句向量时，直接累加词向量的方式会损失部分语义信息，因此通过添加辅助权重值的方式来生成句向量；其次通过无监督算法提取输入句向量的高阶特征，并计算句子与主题之间的相关性，最后抽取出与主题相关的关键信息。

项目所采用的无监督信息抽取模型，通过对语义信息的深度理解，实现对专题领域细粒度信息的摘要抽取，最终为后续QB信息的专有实体识别及领域关系抽取等提供数据支撑。

* 1. **基于迁移学习的QB专题领域关键信息自动抽取技术**

由于QB专题领域数据存在信息冗余、知识匮乏、类型众多等现象，情报人员往往难以快速发现信息中的大量实体、复杂关系及完整事件。并且，传统的识别算法往往难以处理跨领域识别的问题。针对上述问题，本项目拟采用基于迁移学习的QB专题领域关键信息自动识别技术，改进传统识别技术，实现QB专题领域的专有实体识别、关系抽取和典型事件提取功能，最终实现QB专题领域关键QB信息的自动识别。

**1）基于迁移学习的跨领域专有实体识别技术**

在专题领域QB信息中，专有实体所占比例较大，通用的实体识别模型无法有效识别输入信息中的关键实体，导致模型识别实体的效果较差。为提高模型的泛化能力，实现实体识别的智能化，以通用实体识别模型为基础，采用迁移学习技术，实现模型在专题领域内的关键实体识别，最终解决目标实体跨领域识别问题。

基于迁移学习的跨领域专有实体识别技术包括两个阶段：专有实体识别和跨领域的知识迁移。其中，专有实体识别流程分为以下三部分：首先获取专题领域信息的专有实体类别及其描述，构造对应的特征矩阵，实现专有知识的特征表示；其次通过知识注入的方法，将专题领域的专有知识以特征向量的形式输入到模型中；最后在注入的特征知识提示下，关键实体识别模型将专题领域的原始语料映射到对应的向量空间，对原始语料中的关键实体实现更精确的特征表示，从而提高关键实体识别的准确性和鲁棒性。

跨领域的知识迁移流程分为以下两部分：首先将源域样本和目标域样本映射到共享特征空间，最小化源域和目标域数据特征之间的分布差异，使得在源域上训练好的模型能适用于目标域关键实体识别。由于不同领域内样本可能存在差别，采用融合贡献度加权的迁移技术，缓解源领域和目标专题领域间存在的领域偏移问题。

通过基于迁移学习的跨领域关键实体识别技术，可以自动识别QB专题领域中的关键知识的实体类别和实体内容，不仅能提高实体识别的准确性和可靠性，而且能提高实体识别模型的可移植性。

**2）基于远程监督的多实例专业领域关系抽取技术**

由于人工标注数据成本昂贵，标注数据较少，进而导致模型抽取特征关系的实际效果较差，本项目拟采用远程监督的方法，扩充原始数据中的标注数据，为专题领域特征关系抽取提供数据支撑，并通过多实例学习技术挖掘QB工作人员关注的专有关系，实现QB专题领域的关系抽取。

针对标注数据较少的问题，本项目拟采用远程监督来扩充标注数据。远程监督技术将通用的结构化外部知识库与专业领域知识对齐，在外部知识的提示下，自动生成专业领域的标注数据。由于外部知识库与QB专题领域知识存在不匹配的情况，远程监督技术通过对外部知识库进行特征表示，选取与QB专题领域特征最相似的外部知识库进行知识对齐，从而提高自动生成标注数据的整体质量和QB领域关系类别的多样性。

由于远程监督生成的标注数据存在噪声，本项目拟采用面向QB专题领域的多实例学习来减少噪声数据的影响，进而提高关系抽取的效果。利用多实例学习技术减少噪声数据影响的具体流程分为两步：首先通过将自动生成的标注数据按特征关系类别进行分组，其次基于多实例学习技术使模型生成最接近目标分布规律的特征表示，从而降低噪声数据对整体训练过程的影响，优化特征关系抽取的实际效果。

通过基于远程监督的多实例专业领域特征关系抽取技术，提取QB专题领域的关键特征关系，识别QB领域实体关系的内在联系，为面向QB专题领域的知识图谱的构建与使用提供数据支撑。

**3）基于数据增强的专业领域事件抽取技术**

针对专业领域结构化数据稀少，多源跨域后原始语料语义模糊，目标事件提示不足等问题，本项目拟采用数据增强技术增加训练数据的多样性，解决QB专题领域中的数据不足问题，并采用联合学习模型解决QB专题领域的多源跨域问题。

基于数据增强的专业领域事件抽取技术包括两部分：面向QB专题领域的数据增强技术和多源跨域的专有事件抽取技术。其中，数据增强技术通过重构QB专题领域中已有数据的方式，进而增加QB专题领域的训练数据量。同时，采用基于混合特征采样的数据增强技术， 在保证原始语义不变的前提下，对QB专题领域的专业数据进行准确合理的扩增，进而确保扩增的数据能够准确涵盖QB工作人员关注的触发词、事件论元和论元角色等事件抽取模型敏感的语义信息。并且，为适应QB专题领域研究工作中多源跨域的业务需求，基于联合模型的事件抽取技术利用张量层捕捉论元交互信息，实现角色论元和事件论元的联合抽取，从而确定QB专题领域的事件类型，并从非结构化文本中提取具有不同角色的论元。

通过基于数据增强的专业领域事件抽取技术，扩充训练数据集，并从非结构化数据中抽取用户关注的专有事件，方便其他下游工作的展开。

* 1. **基于构建知识图谱的可视化知识探索技术**

针对不同专题QB信息知识量庞大、关联关系复杂等特点，传统数据库难以实现对QB数据信息的有效存储与深度探索，因此本项目拟通过本体建模的方式自底向上构建QB信息图谱，实现QB信息的高效存储，并通过基于VUE+D3框架的可视化技术以一种多维度、多视角的方式展示复杂多样的QB信息，最后通过知识计算实现对QB信息关联关系的深度挖掘，最终达到对QB信息高效存储、有效观测及深度探索的目的。

**1）基于本体建模及Neo4j的图谱构建与存储技术**

通过基于迁移学习的专题领域关键信息自动识别技术，经过实体抽取、关系抽取、事件抽取等关键步骤将非结构化、半结构化的QB信息转化成结构化三元组数据，进而通过本体建模自底向上实现图谱相关概念的抽象和定义，将不同种类的QB信息以关系网络的形式连接在一起，实现对QB信息的结构化语义描述，最终形成QB信息图谱，并通过Neo4j图数据库完成对图数据的高效存储与管理。

基于不同专题QB信息，本项目拟采用自底向上本体建模的知识图谱构建技术完成QB信息图谱的构建工作。即首先对于QB数据相关实体信息进行归纳组织，基于现有QB信息行业标准、知识特点及当前QB业务场景需求形成QB相关领域底层概念，正确描述QB知识体系和知识点之间的关联关系，并在基础上逐步向上抽象，进一步形成领域上层概念，实现实体定义、关系定义、属性定义等，从而完成低冗余、可扩展的高质量QB信息图谱的构建。

基于已构建好的QB信息图谱，本项目拟采用Neo4j图数据库实现对于图谱信息的高效存储。Neo4j是一种基于Java的高性能NOSQL图形数据库，不同于以往二维表形式实现数据存储的关系型数据库，Neo4j数据库将结构化数据以网络（图）状进行存储，利用自定义存储格式和基于图相关的概念来描述数据模型。与传统关系型数据库相比，Neo4j图数据库更能满足数量庞大、关系复杂的QB信息存储需求。该图数据库突破以往表结构的关系束缚，将数据以实体、属性和关系的形式进行网状存储，这种数据格式的结构特殊性使得Neo4j数据库更能清晰表达数据本身的粒度层次、数据之间的关联关系，使之成为上层知识应用的基础。因此项目拟采用Neo4j图数据库存储结构化的三元组QB信息。

**2）基于VUE+D3框架的交互式图谱可视化技术**

针对存储在Neo4j图数据库中的结构化QB图谱数据，本项目拟采用VUE+D3框架实现多维度、立体化、交互式的QB信息图谱可视化效果，实现对QB信息图谱的直观观测与立体感知。

本项目通过分析国内外优秀图可视化框架，结合QB业务场景中定制化需求和交互方式，拟采用VUE+D3作为可视化基础框架，具有高灵活度、强拓展性等特点，为QB信息图谱可视化提供有力支撑。D3基于力导向图位置计算算法，通过力导向图模块，在不做过多干预的情况下根据节点与边的关系模拟物理粒子的随机运动。同时D3力导向图还提供了基于向心力、碰撞检测、弹簧力、电荷力、定位力的调参机制，从而更好的适配多场景图谱数据可视化展示。

针对图谱可视化过程中出现的视觉噪声问题，即图谱中文字/节点/边等元素内容混杂在一起，无法实现信息的准确表达和合理呈现，导致图谱体验差、使用效率低，本项目拟采用视觉降噪技术对文字进行遮挡检测，并根据文字的层叠关系，调低底部文字的透明度，使得在多层文字重叠的情况下，顶层文字依然能被快速识别。

针对图谱可视化过程中存在包含大量出（入）边的中心节点的情况，本项目拟采用多边散列排布的算法，通过边夹角偏移量计算和节点半径裁剪，将关系分散排布在节点周围，减少边线重叠的情况，以达到更清晰的视觉效果。此外本项目拟通过参数配置的方式来调整边的样式，比如：贝塞尔曲线控制点、弧度、自旋角度等参数，以满足QB领域各种复杂图谱的可视化场景需求。

针对图谱可视化过程中QB业务人员的交互需求，本项目拟通过画布操作、元素操作、数据操作实现基本交互功能。其中画布操作支持拖动、缩放、动态延展、布局变换、多节点圈选等功能；元素(节点和边)操作支持样式配置、悬浮高亮、元素锁定、单击、双击、右键菜单、折叠/展开、节点拖动、边类型更改等功能；数据操作支持节点的增删改查、边的增删改查、子图探索、数据合并、更新重载等功能。此外，还可以通过路径锁定、聚焦展现等操作实现不同特点场景的交互。

**3）基于知识计算的图谱关联关系探索技术**

在高性能图谱存储与立体化图谱展示的基础之上，本项目拟通过知识计算从结构和逻辑层面对QB信息图谱中所包含的关联关系进行深度探索，实现对QB信息图谱价值的深层次挖掘。知识计算技术通过基于图谱的社区发现和知识推理实现上述两个层面的图谱关联关系探索。

社区结构是复杂网络中的一个普遍特征，通常整个网络可划分为多个社区。而社区发现对研究复杂网络的特性具有重要意义。针对QB信息图谱中存在的复杂网络社区结构，本项目拟采用社区发现相关算法，帮助QB人员理解复杂图谱网络的固有属性和节点之间的关联关系。社区发现特别是实时动态社区识别对于QB信息传播动向、动态演进、实时监测等具有重要价值。

社区发现的过程是将整个网络结构依据网络节点划分成若干小组，使得组内节点连接较为稠密，组间节点连接较为稀疏，其中小组则为深度挖掘到的隐藏社区或子社区。

除上述社区发现外，本项目拟采用基于深度学习的知识推理方式进一步挖掘QB信息图谱中存在的关联关系，完成进一步的知识发现。面向知识图谱的知识推理旨在根据已有的知识推理出新知识或识别出错误知识。不同于传统的知识推理，由于知识图谱中知识表达形式的简洁直观、灵活丰富，面向知识图谱的知识推理方法也更加多样化。面向知识图谱的推理主要围绕关系的推理展开，即基于图谱中已有的事实或关系推断出未知的事实或关系，一般着重考察实体、关系和图谱结构三个方面的特征信息。具体来说，知识图谱推理主要能够辅助推理出新的事实、新的关系、新的公理以及新的规则等，能够预测实体间可能存在的某种特定关系，捕捉实体和关系之间的隐式关联。

通过上述基于QB信息图谱的社区发现与知识推理，从结构和逻辑层面实现对于图谱中关联关系的深度挖掘，扩展QB相关业务人员的知识边界，有力辅助QB相关业务人员进行决策。

1. **技术指标**
2. 选择不少于3个专题领域开发信息识别模型。
3. 对涵盖中英文动态和文献的信息内容进行分析处理，处理信息量不少于100万条。
4. 专题信息识别准确率大于75%，提取项目、技术、型号等QB线索知识点不少于2000个。
5. 实现知识点可视化和知识点关联探索，QB信息与知识点动态关联利用。