**1. 知识图谱**

**1.1 国内外发展现状及建设意义**

知识图谱与大数据和深度学习一起，成为推动互联网和强人工智能发展的核心驱动力之一。近些年蓬勃发展的人工智能本质上就是一次知识革命，其核心在于通过数据观察与感知世界，实现分类预测、自动化等智能化服务。而知识图谱作为人工智能的一个重要分支，是机器具有认知能力的基石，也是人类知识描述的重要载体，能将人与知识智能地连接起来，对各类应用进行智能化升级，为用户带来更智能的应用体验，在人工智能领域具有非常重要的地位。

知识图谱，即语义互联，是一种基于图的数据结构，由节点和边组成。在知识图谱里，每个节点表示现实世界中存在的“实体”或者“概念”，每条边则为实体与实体之间的“关系”，整个图谱以结构化的形式描述客观世界中概念、实体及其关系，将互联网的信息表达成更接近人类认知世界的形式，即提供了一种更好地组织、管理和理解互联网海量信息的能力。因此，知识图谱是关系的最有效的表示方式，它就像福尔摩斯破案一样，利用大数据采集散步在各个角落的碎片化信息并按标准化思考方式整理，再将所有不同类型的信息连接在一起得到关系网络，提供从“关系”角度分析、推理问题的能力，并挖掘相互之间的关联关系与影响程度。

知识图谱技术也正逐渐改变现有的信息检索方式，如谷歌、百度等主流搜索引擎都在采用知识图谱技术提供信息检索，一方面通过推理实现概念检索（相对于现有的字符串模糊匹配方式而言）；另一方面以图形化方式向用户展示经过分类整理的结构化知识，从而改善了搜索引擎性能，使人们从人工过滤网页寻找答案的模式中解脱出来。另外，在人工智能的蓬勃发展下，知识图谱涉及到的知识抽取、表示、融合、推理、问答等关键问题都得到了一定程度的解决和突破。

从2012年谷歌提出Google Knowledge Graph起至今，知识图谱技术日渐繁荣，已经成为知识服务领域的一个新热点，受到国内外学者和工业界广泛关注。目前，开放知识图谱大致分为常识、医学、金融、城市和旅游等15类，包括有DbPedia、Yago、BabelNET、WikiTaA、ConceptNet、Microsoft Concept Graph、Zhishi.Me、Zhishi.schema和XLOR等等；中文开放知识图谱联盟简称OpenKG，中国很多知识图谱数据库都加入了这个联盟，如Zhishi.me、CN DbPydia和PKUBase。

**1.2 知识图谱在金融行业的应用**

金融风险的复杂性与多样化使风险监管的知识图谱建设迫在眉睫。知识图谱作为金融科技领域语义理解和知识搜索的基础技术，为金融科技服务于实体经济提供了深度技术，提供了金融领域知识的提取和融合，以及进行分析、推断和决策的依据，能促进金融科技信息技术与开放知识领域的融合，并可以为金融行业的风险评估、预测、反欺诈、精准营销、智能搜索等提供技术支撑。因此，越来越多的金融机构及企业在探索和构建金融领域知识图谱，为金融领域应用决策提供更加精准可靠的依据。

2018年证监会关于企业知识图谱的建设，即广泛应用于中央监管信息平台的上市公司画像、拟上市公司画像和中央监管私募机构画像，则是通过关联关系识别面向企业层面的风险防范，如非法集资、财务管理风险等，使金融风险管理适应于金融科技产品、服务、行业的发展成为了可能。

现阶段，知识图谱和人工智能技术在金融行业的应用非常广泛，如智能客服智能搜索及可视化展示、反身份作假和群体欺诈、异常分析、失联客户管理、产业链及担保链分析、用户画像与情感分析以及智能知识库等等。

**1.3 知识图谱构建框架**

如今有多种方式构建知识图谱，人工方式比如Cyc，通过Freebase以及维基数据，或者从大规模、半结构化的数据集类似维基百科、DBpedia、YAGO抽取得来。此外，更多的学者提出了基于结构化或者半结构化的信息抽取系统，在这个基础上产生了NELL、PROSPERA和KnowledgeVault模型。

参考知识图谱的通用构建框架，本课题的公司主体知识图谱的整体搭建流程如图1所示。

具体实现过程如下：

1. 首先，利用网络爬虫技术，爬取搭建知识图谱所需的原始数据；

2. 其次，基于深度学习技术，以长短时记忆网络（LSTM）为语料特征学习模型，建立命名实体识别方案；

3. 另外，在实体关系抽取中，从公开的上市公司研报和公告或者新闻舆情信息中提取素材，结合领域知识和业务需求找到实体之间的关系，为知识图谱提供理论及数据支撑；

4. 最后，通过Neo4j图数据库，将实体与关系存储其中。

上传图数据库

实体关系抽取

数据信息爬取

命名实体识别

图1 知识图谱通用构建流程图

在公司主体知识图谱的构建过程中，第一步的爬虫所获取的上市公司的信息来源较为复杂，多数是半结构化、非结构化的文本信息，例如自营持仓主体的基本属性包括公司介绍、高管、集团、股东、债务、行业、上游、下游、概念、主营业务等属性；而最重要的是后三步，即知识单元的抽取、知识单元间关系的识别以及知识图谱的绘制：首先将原始数据按照BMEO规则进行标注，使用Python将标注好的数据处理成字标注的形式，将每一个字和标注转换成id的形式，供模型训练使用，并使用Pytorch框架通过命名实体识别，将得到的数字id根据之前的对应关系转换回汉字，稍做处理就可以获得文本中的实体；然后通过定义实体之间的关系，将该段文本转化为词向量表示，计算句子中每个词距离两个实体在句子中的距离，把距离转换成距离向量，与词向量合并，输入Pytorch模型进行训练；最后就可以得到关于上市公司知识图谱的三元组结果，存入Neo4j图数据库中，就得到了上市公司主体知识图谱。

**1.3.1 网络爬虫**

网络爬虫又称为网络蜘蛛或网络机器人，是一个程序或脚本，根据一定的规则自动获取网络资源。它已被广泛应用于互联网领域，例如搜索引擎使用网络爬虫来爬取网页、文档甚至图片、音频、视频等资源，通过相应的索引技术组织这些信息，为搜索用户提供查询。传统网络爬虫的工作流程从一个或多个初始网页的URL开始，并从当前网页中连续提取新的URL并将其放入队列中，直到满足系统的某些停止条件。

本课题通过python爬虫工具对于网页中公开的的数据信息（例如与自营以及融资融券、股票质押等业务相关的股票持仓主体以及其关联方相关的新闻、舆情信息等全网信息）进行实时爬取，梳理其中有效的信息并对梳理后的数据信息进行整合，为知识图谱的构建提供原始信息数据。

**1.3.2 实体抽取**

实体抽取又称为命名实体识别，是指通过自动化手段从非结构化文本中识别具有特定含义的某一类实体概念的技术。命名实体识别的准确性直接决定了关系识别（实体关系抽取）的质量，也就影响到知识图谱构建的后续环节，因此是构建知识图谱的一个首要任务。

命名实体识别的任务被定义为：识别出文本中出现的专有名称和有意义的数量短语并加以归类。而命名实体是文本中最基本的信息元素，是正确理解文本的基础，至于其的确切含义，只能根据具体应用来确定。

利用统计学习方法，或者说机器学习方法来解决命名实体识别这样的自然语言处理问题，多年来已经有了较多的解决方案，其中条件随机场模型（CRF）等监督式学习方法虽然已经有了较为稳定的效果，但是这类方法通常需要配合繁琐的人工特征处理过程，且通常只能学习到与相邻词位置比较近的上下文的特征，而无法获取整个句子甚至更长的上下文的特征。因此，本课题所使用的解决方案则是近几年来较为流行的神经网络学习模型或者说深度学习模型。该方案相比较于传统的机器学习方法改进之处在于能够通过复杂的神经网络结构来学习数据的特征，并且可以由低层次至高层次自动地学习特征的层次结构，省去了需要人工介入的特征提取工程，从而可以获得更优秀的效果。

本课题中，命名实体识别的技术路线包括几个部分：

1. 采用嵌入词向量方案进行语言文本表示，使用无监督学习算法从大量未标记文本语料库中提取矢量；

2. 采用的特征学习模型为长短时记忆深度神经网络模型（LSTM）。LSTM是在循环神经网络（RNN）的基础上改进而来的模型，而循环神经网络对于传统神经网络模型的改进之处就在于其对于时间序列输入的处理。在本课题的应用场景中，特征学习模型是构成双向LSTM网络结构的BiLSTM，该模型可以分别从前往后以及从后往前这两个方向来去学习上下文的特征，然后进行序列信号的记忆和传递；

3. 采用BIEO标注方案表进行数据标注。马尔可夫随机域（CRF）是一种用于标注和切分有序数据的条件概率模型，考察给定输入序列的标注序列的条件概率，因此在序列标注时，在顶层用马尔可夫随机域对BiLSTM的结果进行二次约束可得到更好的结果。

命名实体识别的整体模型结构见图2。

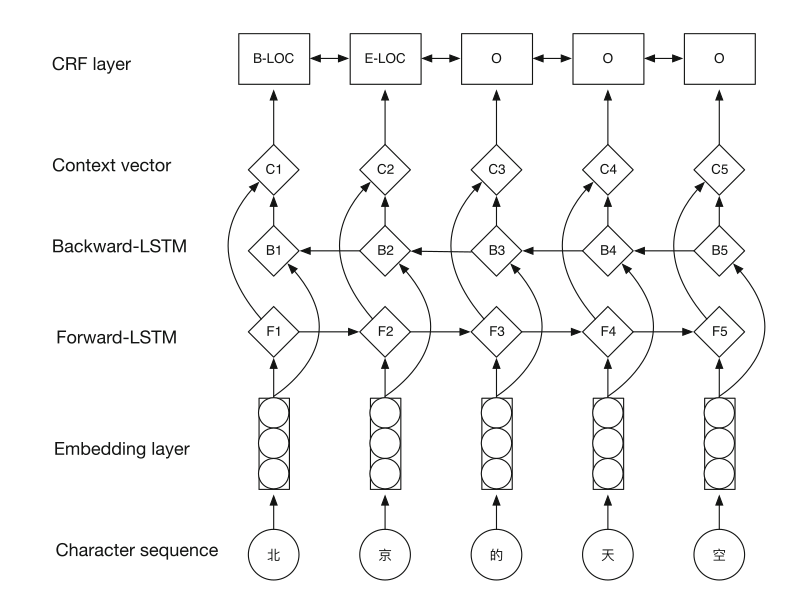


图2 实体抽取总体模型结构

如图2所示，圆形节点是模型的输入层，即Embedding layer（嵌入层）部分；菱形节点代表了长期记忆模型和短期记忆模型的双向网络结构；方块表示马尔可夫随机模型中的随机变量。在模型中要训练的参数包括条件随机场模型中的序列传递矩阵和双向长期和短期记忆模型中的参数，以及每一层中节点和节点的线性权重与偏置的大小。

**1.3.3 关系抽取**

实体关系抽取是自然语言处理（NLP）领域最重要的研究方向之一，其相关工作不仅包含监督模式下的分类学习，也涉及无监督学习的关系发现。其中，大量的工作主要是围绕如何设计出能够更好地描述实体之间句子的特征地办法，从而识别这些关系。

在大多数情况下，关系抽取被视为具有多个分类的有监督问题。有监督的方法能够抽取出有效的特征，然后在准确率和召回率方面是更有优势，因此监督学习方法被证明是解决关系抽取问题的有效方法，并取得了良好的分类效果。然而，传统的监督学习方法严重依赖于特征的质量。近年来，随着代表性学习和深度学习的发展，大量的深度学习技术被应用到自然语言处理中，并取得了许多令人惊叹的成果。

本课题涉及的大多数关系类型可以直接从权威的结构化数据中提取：作为资本市场，可以从自己的结构化数据库中获得提取所有上市公司的法定资本结构，如法定的公司投资关系和股东关系；或从市场关系的角度来看，公司产业链上下游的信息可以从公司披露的年度报告和相关公告中获得。这些信息基本上以半结构化数据的形式获得，利用信息抽取技术可以抽取和存储结构化信息。另外人与人之间的社会关系，最主要的提取方法是从个人在资本市场上的相关简历中获取个人经历，通常以半结构化的数据格式显示在文档中。

本课题采用语句级的粒度，研究一句话中任意两个实体之间是否存在联系。

实体关系抽取算法层次结构主要包括五个部分：输入层、词向量模型训练层、双向LSTM层、attention注意力机制层、关系分类输出层。

整体结构见图3。

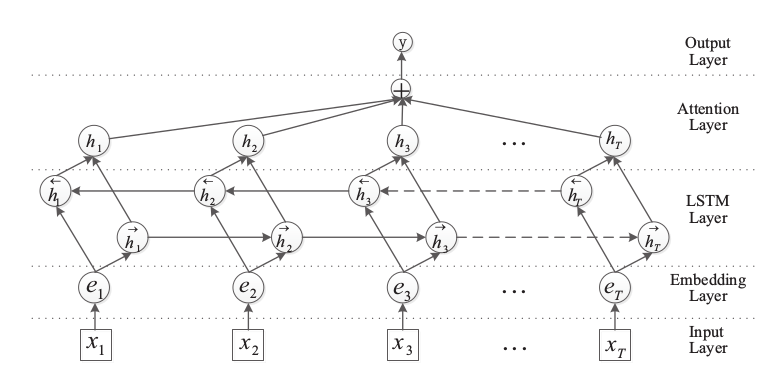


图3 BiLSTM+ATT模型结构图

**1.3.4 知识存储**

知识图谱技术解决了知识的获取与整合，但仍需储存于某一介质，用于后续的查询和可视化展示。本课题选用图形数据库存储整合的知识图谱数据，将非结构化数据与半结构化数据，通过命名实体识别和实体关系抽取后，就可以得到结构化数据，再使用资源描述框架（RDF）对数据进行符号性描述，如[头实体]-[关系]->[尾实体]的三元组形式，最后将其上传到图数据库中。

图形数据库是一种非关系数据库（NoSQL）。在图数据库中，节点间的节点信息和链路信息构成整个数据集。节点信息类似于传统关系数据库中的表结构，每个节点对应表中的一段信息，而关系是图形数据库的唯一组成，关系数据库对应于表之间的数据关联。因此，与传统的关系数据库相比，图形数据库对节点和节点间复杂关系有更加良好的支持，所以在存储和表示大量关系方面具有更多的优势，成为了知识图谱存储的首选。

本课题的图形数据库采用Neo4j图形数据库，它是世界领先的开源图形数据库，以图形结构的形式存储数据信息，能将实体和关系通过图形的方式显示出来。它可以存储节点、节点属性值、节点之间关系值信息，因此可以使用图结构的自然伸展特性来设计免索引邻近节点便利的查询算法，即图的遍历算法设计。图的遍历是图数据结构所具有的独特算法，即从一个节点开始，根据其连接的关系，可以快速和方便的找出它的邻近节点。这种查找数据的方式并不受数据量的大小所影响，因为邻近查询始终查找的是有限的局部数据，不会对整个数据库进行搜索，所以，Neo4j具有非常高效的查询性能。

知识图谱绘制步骤见图4。

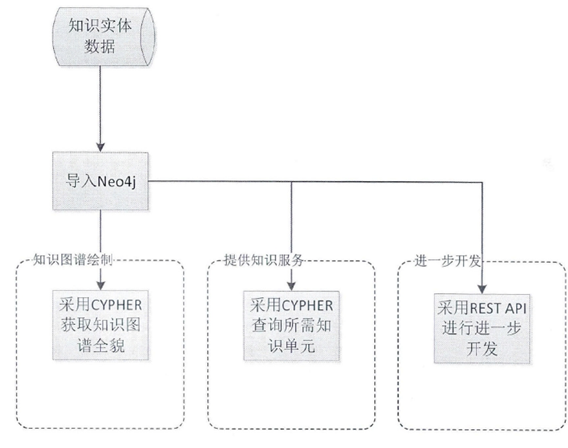


图4 绘制知识图谱流程图

1. 将通过爬虫等方法得到的结构化数据可以直接导入Neo4j数据库；而得到的非结构化数据进行命名实体识别和实体关系抽取后，采用批量导入方法将知识实体与实体之间的关系识别到图形数据库中。

2. 使用CYPHER查询语言对所有节点和关系进行查询，得到知识图谱的全貌。

3. 利用密码语言搜索所需的节点和关系信息，可以为用户提供个性化的知识服务。

4. 通过编程，可以调用Neo4j的REST API接口，从而进一步开发知识图谱接口。

**1.4 知识图谱技术优势**

近年来，随着我国上市公司数量快速增加，金融市场规模持续扩大，风险也在积聚。因此，为了及时发现市场主体的潜在风险，维护市场稳定，切实保护中小投资者利益，金融行业一线监管迫切需要加强监控力度和敏锐度。目前，金融机构已经把公司风险的防范与把控提升到非常重要的高度，实现公司内部风险信息监测预警已经成为金融行业的热点问题。

在风险防范领域，传统模式下事后的、手动的、基于传统机构性数据的风险防范模式需要大量人力的投入并且效率不高，另外其数据也缺乏多样性，所以已经不能满足金融科技新业态的需求。因此，金融机构应与时俱进地在传统监管模式的基础上，借助大数据、人工智能等新兴金融科技进行风险管控，减少人力支出，完成风险的主动识别与控制，有效降低风控合规成本，增强风控合规能力，从而寻求监管突破。

大数据时代的开启为海量数据资源的融合与智能解析提供了完整的新路径。面向不同数据源、不同数据结构和不同数据类型的企业风险信息的集成、融合及其深度挖掘，可以使已知数据中的隐含知识得以最大化的释放，从而对公司的智能监管进行赋能，更好地认识与识别金融行业公司风险。

根据上述要求，知识图谱技术首先将公开的上市公司和互联网信息进行抽取整理，然后研究其中各种实体的概念信息以及相互之间可利用的关联、隶属关系等，实现相关实体概念表述以及关联关系的可视化，从而建立公司主体及关联方主体知识图谱。最终，将建好的知识图谱与深度学习模型相结合，提出一些规则模型和概率模型等分析方法，构建公司内部风险监测预警机制，从而更有效、更准确地、更快速地防范公司风险，为公司提供科学的投融资决策及提升公司的风险管理水平，促进公司业务的健康、快速发展，避免引起系统性风险的事件发生。

**1.5 知识图谱实现效果**

**1.5.1 单个企业信息展示**

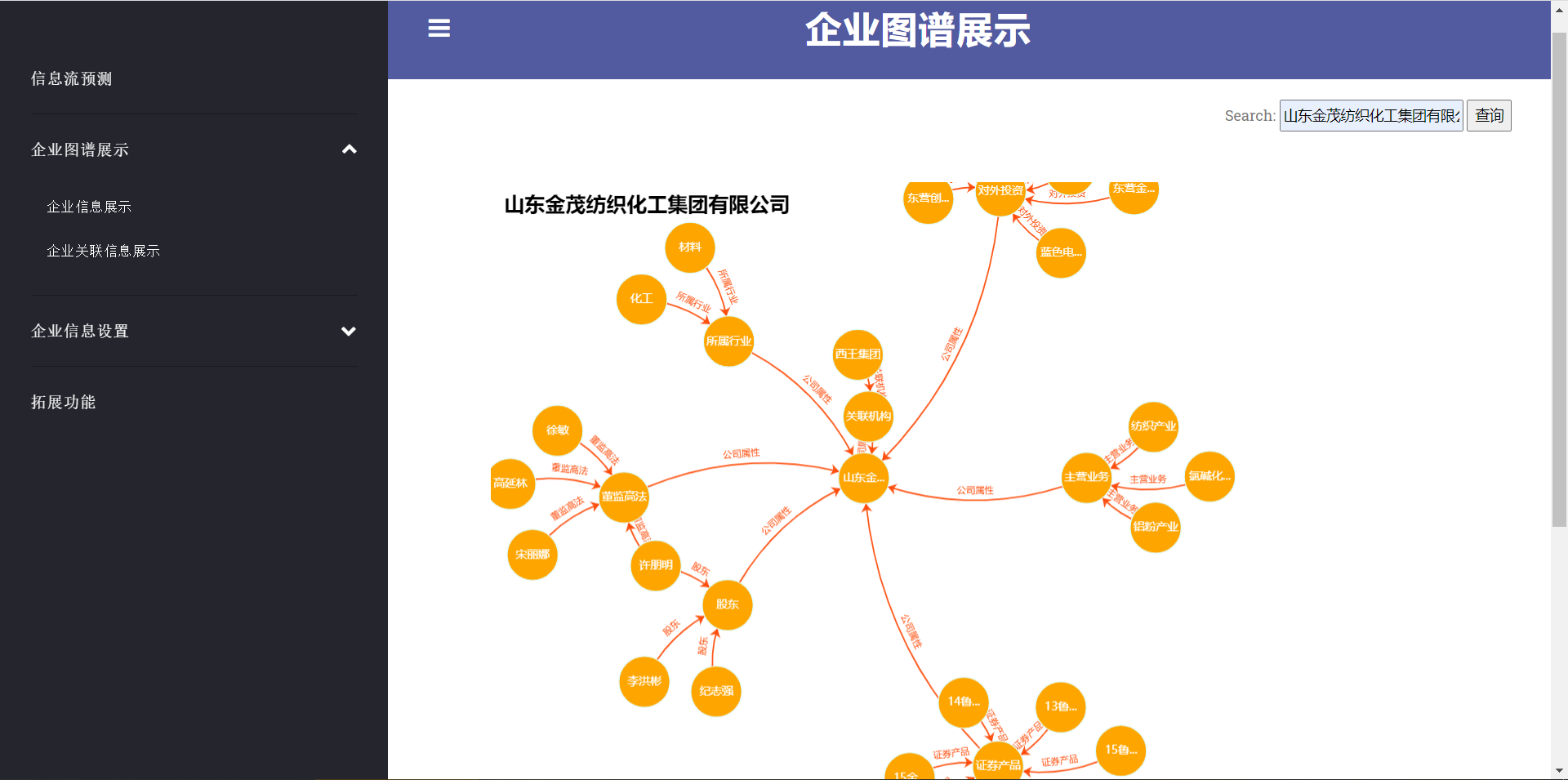


图5 单个企业信息展示

在企业信息展示页面，在搜索框中输入某企业名称，在当前页面得到该公司主体的知识图谱，如图5所示，包括“山东金茂纺织化工集团有限公司”的公司主体、公司属性（例如董监高法、股东、对外投资、证券产品、主营业务、所属行业、关联机构等）以及其属性内容（例如徐敏、化工、西王集团等）。

**1.5.2 关联企业信息展示**

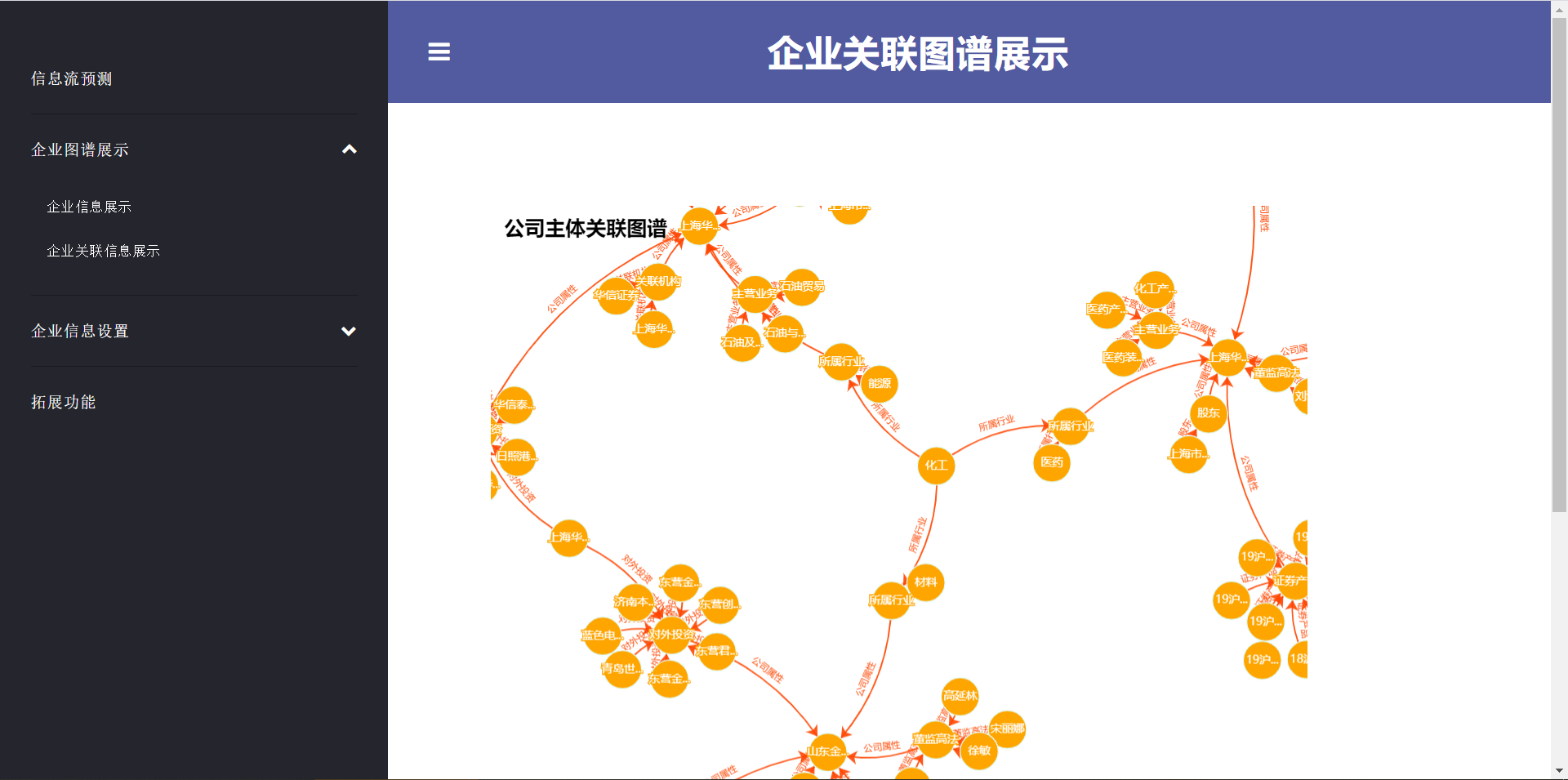


图6 关联企业信息展示

在企业关联信息展示页面，在当前页面得到不同公司主体相互关联的整体知识图谱，如图6所示，不同公司主体“山东金茂纺织化工集团有限公司”、“上海华信国际集团有限公司”以及“上海华谊（集团）公司”共同属于化工行业，且华信和金茂都投资了“上海华投资产管理合伙企业”这个公司。