计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.191000015

企业风险知识图谱的构建及应用

陈晓军　向　阳
同济大学电子与信息工程学院　上海 201804
(xiaojunchen@tongji.edu.cn)

摘　要　作为语义网的数据支撑,知识图谱在搜索引擎、智能问答和推荐系统等领域发挥着重要作用,成为了人工智能领域的研究热点。知识图谱因其自身的图展示、图挖掘、图模型等计算优势,可帮助企业或金融从业人员进行业务场景的分析与决策。目前已经有公司将知识图谱应用到金融领域,但是这些知识图谱还存在信息缺失、准确度低等问题,并且现有的金融知识图谱构建方法大都只关注构建过程中的某一环节。针对上述问题,对行业知识图谱构建方法进行系统研究,构建一个企业风险知识图谱,从本体构建、知识抽取、知识融合和知识存储4个方面完整阐述了知识图谱的构建流程。最后,基于企业风险知识图谱,构建了一个智能问答机器人,实现了对知识图谱的检索和利用;为了提高问答系统回答问题的准确性,利用基于字级别的BiLSTM-CRF命名实体识别模型。实验结果表明,在样本数量较少时,基于字级别的模型效果更优。

关键词:知识图谱;企业风险;本体;知识抽取;知识融合;问答系统
中图法分类号　TP391

Construction and Application of Enterprise Risk Knowledge Graph

CHEN Xiao-jun and XIANG Yang
College of Electronic and Information Engineering,Tongji University,Shanghai 201804,China

Abstract　In supporting semantic Web,knowledge graphs have played an important role in many areas such as search engine,intelligent question-answering system,and recommender system. Therefore,they have become a hot topic in the field of artificial intelligence. Knowledge graphs have many advantages in graph display,mining,and computing,which can help enterprises or financial practitioners analyze and make decisions on business scenarios. At present,some companies have applied knowledge graphs in the financial domain,but these knowledge graphs still suffer from incompleteness. And most existing methods only focus on certain aspects when building financial knowledge graphs. Aiming at these problems above,this paper engages a systematic study on the domain knowledge graph and construct an enterprise risk knowledge graph. This paper describes the construction process of domain knowledge graph from the aspects of ontology construction,knowledge extraction,knowledge fusion,and knowledge storage. Based on the enterprise risk knowledge graph,an intelligent question-answering chatbot is developed to realize the retrieval and application of KG. In order to improve the accuracy of the question answering system,a character-based BiLSTM-CRF model for named entity recognition is used. Experimental results show that the character-based BiLSTM-CRF model performs better than the baseline when the sample size is small.

Keywords　Knowledge graph,Enterprise risk,Ontology,Knowledge extraction,Knowledge fusion,Question-answering system

1　引言

随着信息技术的发展,人们经历了以网页链接为核心的Web 1.0时代和以数据链接为核心的Web 2.0时代,目前Web技术正朝着Berners-Lee提出的语义网络(Semantic Web)[1]演变。语义网数据规模大,数据来源丰富,类型复杂,传统的数据管理方式受到了一定的制约。知识图谱(Knowledge Graph)的出现,为解决这些问题提供了新的思路。知识图谱的概念最早由Google公司于2012年提出,旨在提高搜索质量,优化用户的搜索体验,目前其已被广泛应用于问答系统、智能搜索和个性化推荐等领域。为了提高信息服务质量,国内外互联网公司纷纷推出了自己的知识图谱产品,如微软的Satori、百度的"知心"和搜狗的"知立方"等。

随着大数据和移动互联网等新兴技术的发展,金融行业每天都会产生海量数据,这为知识图谱的构建与应用奠定了数据基础,因此越来越多的金融机构开始探索构建自己的金融知识图谱。目前,金融知识图谱已被广泛应用于反欺诈、精准营销等业务。例如,在反欺诈场景中,通过将与贷款人相关

到稿日期:2019-10-08　返修日期:2020-03-20　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(71571136)
This work was supported by the National Natural Science Foundation of China (71571136).
通信作者:向阳(tjdxxiangyang@gmail.com)

的多源信息整合到知识图谱中，可以对贷款人可能存在的风险进行全面分析和评估，从而有效防范金融欺诈；在精准营销场景中，企业可以利用知识图谱分析用户行为，为潜在客户构建精准的用户画像，从而对其进行精准推送。因此，研究金融领域知识图谱的构建具有重要的意义和价值。

构建企业风险知识图谱时，通过图谱呈现的上市公司、人物、组织机构、行业等实体信息以及实体之间的关系信息，可以帮助从业者更加直观地了解和检索相关实体和关系信息，帮助他们进行企业风险预测、关联企业等分析与应用。尽管大量的知识图谱日益涌现，但面向企业领域的知识图谱还相对匮乏，且大都只覆盖某一方面。如 Yu[2] 基于部分企业数据集构建了企业法人知识图谱，但其数据量和数据的覆盖面都存在局限性，它们只提供了基本的信息查询功能，没有深入地挖掘企业之间潜在的关系。此外，大多数工作都只关注知识图谱构建过程中的某一环节，如知识图谱中的数据表示、存储和抽取等[3]。对此，本文描述了一个企业风险知识图谱的完整构建过程，从本体构建、知识获取、知识融合以及知识图谱数据的存储 4 个角度重点介绍了企业风险知识图谱构建的完整流程；通过融合企业名称、股东类型、监事、工商登记等信息，结合知识推理技术，构建精准的企业画像；在此基础上，实现了一个基于知识图谱的智能问答系统，并以可视化的形式为用户提供服务。

2 相关工作

根据知识图谱的应用领域，知识图谱可以分为通用知识图谱和领域知识图谱。本节将简单介绍通用知识图谱和金融领域知识图谱构建的相关研究。

2.1 通用知识图谱构建的研究现状

在通用知识图谱的构建方面，目前已有相对成熟的技术和知识图谱产品。比较有代表性的知识图谱是 FreeBase[4]，YAGO[5]，DBpedia[6] 和 Wikidata[7] 等。FreeBase 的数据主要是由人工构建的，另一部分数据来自维基百科等网站或语料库。2010 年，谷歌收购 FreeBase 作为其知识图谱的重要组成部分。YAGO 是由德国马普研究所开发的链接知识库，它主要集成了 Wikipedia，WordNet 和 GeoNames 3 个来源的数据。YAGO 将 WordNet 的词汇定义与 Wikipedia 的分类体系进行了融合集成，具有更加丰富的实体分类体系。此外，最新的 YAGO3 知识库为实体赋予了时空维度的信息。截至目前，YAGO3 包含的实体数量超过 1 700 万，三元组数量达到 1.5 亿。DBpedia 是早期的语义网项目，是从 Wikidata 抽取出来的链接数据集，与 Freebase，OpenCYC 和 Bio2RDF 等多个数据集建立了数据链接。DBpedia 采用 RDF 语义数据模型，包含超过 30 亿条三元组数据。Wikidata 是一个自由的协作式多语言百科知识库，支持用户自由编辑，它集成了 Wikipedia 和 Wiktionary 等项目中的结构化知识。目前，Wikidata 包含 2500 多万个实体和 7 000 多万个三元组条目。除此之外，国外发展得较成熟且质量较高的知识图谱还有 BabelNet[8]，ConceptNet[9] 及 Microsoft Concept Graph[10] 等。

上述知识图谱都是基于英文的，即使是多语言知识图谱，也是以英文为主的，其他语言知识通过跨语言知识推理得到。近几年，国内也涌现出了大量以中文为主的知识图谱。2012 年，搜狗推出了国内首个应用于搜索引擎的知识图谱产品“知立方”。“知立方”首先利用百科类数据、结构化/半结构化数据以及用户的搜索日志完成实体抽取；然后利用实体对齐等技术完成异构数据融合，进而完成 Web 规模知识图谱的构建；最后通过知识推理、实体重要性排序和实体挖掘等技术来提高知识图谱的知识覆盖率。同年，百度“知心”知识图谱上线，旨在构建宏大的知识网络，包含世间万物以及它们之间的联系，以图文并茂的方式展现知识的各方面，让人们更加便捷地获取知识。此外，许多高校也构建了自己的知识库，如清华大学构建的跨语言知识图谱 XLore[11]、上海交通大学构建的 Zhishi. me[12] 和复旦大学构建的 CN-DBpedia[13] 等。

2.2 金融领域知识图谱构建的研究现状

相对于通用知识图谱的发展，行业知识图谱尤其是金融领域的研究还有待展开。Pujara 等[14] 将开放域信息抽取技术与主题模型相结合，从财务报表中抽取三元组用于金融知识图谱的构建。Ruan 等[15] 提出了一种基于语义技术，利用多种数据源构建企业知识图谱的方法，其提供了可视化查询功能，并将企业间的潜在关系用于决策支持。Song 等[16] 介绍了 Thomson Reuters 公司推出的首个金融知识图谱框架，其可以定制金融服务，提高了行业服务水平。

众所周知，金融领域每天都会产生海量的数据，而知识图谱可以将这些海量的非结构化信息自动地利用起来，为精准营销、风险预测、金融反欺诈等场景提供可靠的依据。因此，工业界对金融知识图谱的研究与构建工作越来越重视，目前已有多个跨国公司（如 Amazon，Netflix 和 Spotify 等）利用机器学习，结合知识图谱技术开发了金融系统，用于分析客户需求，有针对性地进行产品营销。国内也有一些金融机构开始探索将知识图谱应用到具体业务中，如恒生电子采用知识图谱辅助投资顾问和投资研究，浦发银行等金融机构也将知识图谱用于风险预测、金融反欺诈、量化交易等场景，同时各大互联网公司也在尝试利用其积累的海量信息构建自己的金融知识图谱。然而，以上研究整体仍处于探索起步阶段，距离成熟和完全商用还有很长的路要走，因此金融知识图谱构建的研究迫在眉睫。

3 企业风险知识图谱构建

知识图谱主要有两种构建方式：自顶向下（Top-Down）和自底向上（Bottom-Up）[17]。自顶向下的构建方式是指首先为知识图谱定义好本体和数据模式，这种构建方式一般适合于领域知识图谱的构建。在定义本体的过程中，首先从最顶层的概念开始，然后逐步进行细化，形成结构良好的层次结构；在定义好数据模式后，再把实体逐个添加进概念中。这种构建方式通常需要借助百科类知识等结构化数据源。FreeBase 就是采用这种方式构建的，它的绝大部分数据都是从维基百科中得到的。自底向上的方法则相反，首先对实体进行归纳组织，形成底层的概念，然后逐步往上抽象，形成上层的概念。采用这种方式构建的知识图谱的典型代表包括谷歌的 Knowledge Vault 和微软的 Satori。

本文采用了自顶向下的知识图谱构建方法，将企业风险知识图谱构建流程归纳为 4 个步骤，如图 1 所示，包括企业本体的构建、知识抽取、知识融合和知识存储。

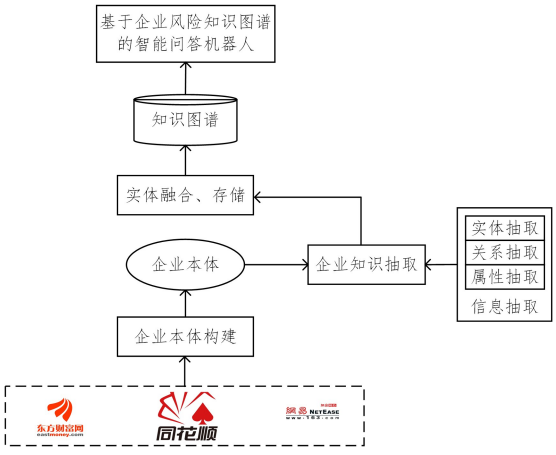

图 1 企业风险知识图谱的构建流程

Fig. 1 Pipeline of enterprise risk knowledge graph

(1)企业本体的构建。通过复用现有的知识本体，半自动化地构建金融企业本体。

(2)企业知识抽取。从结构化、半结构化和非结构化等不同来源、不同结构的数据中抽取信息，并通过命名实体识别、关系抽取等技术对这些信息进行处理，从而得到构建知识图谱需要的实体三元组。

(3)企业知识融合。对异构数据执行实体对齐和实体匹配等步骤，解决数据冲突问题，包括实体属性值不一致、实体属性缺失等。进一步地，借助知识推理技术，丰富和拓展现有知识，为企业决策提供有价值的信息。

(4)知识存储。实现三元组数据的持久化存储，基于已构建的企业风险知识图谱实现智能问答机器人。

3.1 企业本体构建

在知识图谱构建过程中，本体提供了上层的数据模式，是实体存在的形式化描述，是知识图谱的重要组成部分。领域本体构建流程通常包含以下 6 个步骤：本体需求分析、考察可复用本体、建立领域核心概念、建立概念分层次、定义类和创建属性，以及本体评价和进化[18]。针对不同的领域和不同的实际需求，领域本体构建的过程也各不相同。我们研究了金融领域企业本体的实际情况，给出了企业本体的构建方法。

目前，比较被认可的构建本体的方法有：Skeletal 法[19]（又称骨架法）、TOVE 法[20]、七步法[21]等。根据这些方法，深入分析金融领域的相关知识，对已有的结构化数据集进行整体分析，通过分析领域内概念和属性之间的语义关联，构建企业本体的 RDF 图，如图 2 所示。

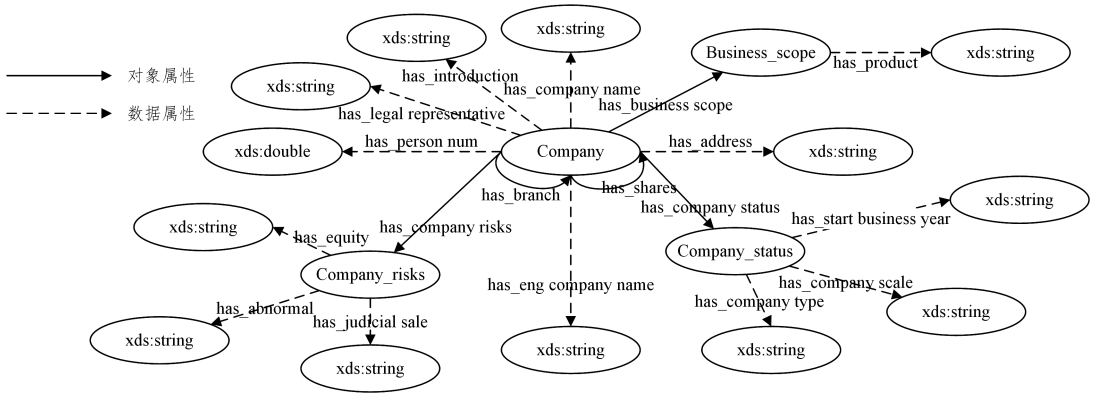

图 2 企业本体的 RDF 图

Fig. 2 RDF graph of enterprise ontology

3.2 知识抽取

企业知识抽取是从不同来源、不同结构的数据中抽取企业相关知识的过程。企业知识抽取一般分为以下几个步骤，如图 3 所示。

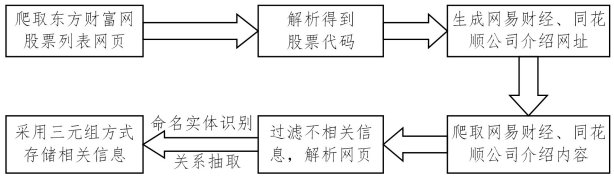

图 3 知识抽取流程

Fig. 3 Pipeline of knowledge extraction

(1)网页解析。该模块首先从东方财富网站爬取“上市公司股票代码列表”，得到上交所 3685 家和深交所 2394 家上市公司和相应股票代码列表；然后用股票代码分别生成网易财经、同花顺公司介绍网页的 URL 地址，并根据 URL 地址，通过关键字找到企业信息，得到公司介绍网页的源码；最后对得到的网页进行解析，并对网页中的冗余信息和不相关信息进行过滤，进而得到较为纯净的语料。

(2)命名实体识别。命名实体识别是指识别一个句子中有特定意义的实体，并将其区分为人名、地名、机构名、专有名词等类别。在半结构化数据和非结构化数据中，很多实体的属性值并没有被识别，且这些数据大多以文本形式存在。这些文本主要有 3 种类别：1)含有超链接信息的文本；2)有明显语义标记的文本，命名实体之间用一致的标点符号分隔，且不存在歧义；3)没有明显语义边界的长文本，命名实体之间没有分隔符[22]。本文主要考虑第 3 种类别。

我们选用 HanLP 作为中文分词工具包。HanLP 是一系列模型与算法组成的 NLP 开源工具包，支持中文分词、词性标注、命名实体识别、关键词抽取、文本分类等功能。HanLP 基于隐马尔可夫模型（Hidden Markov Model，HMM）进行命

名实体识别，对中国人名、音译人名、日本人名、地名、实体结构名等通用类别有着非常好的表现，但对专有名词的识别效果不佳。我们通过自定义词典的方法来提升命名实体的效果。在分词序列的基础上，我们将金融领域专有词表和特殊规则加入到词典中，充分识别未登录词，提高命名实体识别的效果。

(3)实体关系抽取。关系抽取指从文本中识别实体并抽取实体之间的语义关系。针对上市公司的公司介绍，我们使用 HanLP 工具包，基于依存句法分析来抽取开放域中文实体关系。HanLP 提供了两种依存句法分析器，默认采用基于神经网络的依存句法分析器，另一种为基于最大熵的依存句法分析器。依存句法分析就是通过分析语言单位内成分之间的依存关系，揭示其句法结构的过程。如“中信证券股份有限公司成立于 1995 年 10 月 25 日”，句法分析的结果为〈中信证券股份有限公司(主谓关系)成立〉〈成立(动补关系)于〉〈1995 年 10 月 25 日(介宾关系)成立于〉，最后将依存句法分析的结果转化为三元组形式“〈实体 1〉〈关系〉〈实体 2〉”。表 1 列出了一个关系抽取实例。

表 1 关系抽取实例

Table 1 Instances of relation extraction

实体 1	关系	实体 2
中信证券股份有限公司	法人代表	张佑君
中信证券股份有限公司	股票代码	600030
中信证券股份有限公司	所属行业	金融行业
中信证券股份有限公司	注册资金	121.17 亿元
中信证券股份有限公司	员工数量	16161

3.3 知识融合

该阶段需要将从多个数据源抽取的知识进行融合，构建数据之间的关联关系，从而保证知识图谱中的数据一致性和准确性。实体对齐是知识融合过程中的主要挑战，旨在判断两个或者多个不同信息来源的实体是否为同一实体。例如，“中国铝业集团有限公司”和“中国铝业股份有限公司”描述的其实是同一实体，可以合并。

不同的金融网站在展示实体属性信息时会用不同的实体名称表示，这给数据融合造成了极大的困扰，因此需要找到正确的实体名称与之对应，进而将数据融入已有的知识库。实体对齐主要有两个方向：1)实体名的完全匹配；2)实体名的相似度计算。第一种方式主要针对无歧义的实体名称；第二种方式主要针对实体名称简写与实体名称存在差异的情况。实体对齐的常用方法可分为 3 类：基于概率模糊匹配的实体对齐、基于距离度量的实体对齐和基于机器学习方法的实体对齐。基于概率模糊匹配的实体对齐方法主要考虑两个实体各自属性的相似性，但忽略了实体间的关系。基于距离度量的实体对齐方法与概率匹配算法类似，都是基于属性域加权的成对比较方法。这两者的主要区别在于基于距离度量的方法更加注重相似性度量函数的设计，且不需要训练样本。基于距离度量的方法有：欧氏距离、余弦相似度、编辑距离和 Jaccard distance 等。上述两种实体对齐算法过分依赖于相似性度量函数，因此本文采用了可以自动从训练样本中提取特征和捕获上下文语义特征的基于孪生神经网络的实体对齐算法[23]。

基于孪生神经网络的实体对齐模型由两个对称的双向 LSTM 网络通过权重共享的方式组成，如图 4 所示。其中，(C_i, C_j)为训练样本；x_i 和 x_j 为 C_i 和 C_j 的嵌入向量表示，并被输入双向 LSTM 中。在每个时间步 t，双向 LSTM 经过计算会产生两个隐藏状态，f^t 为前向状态，b^t 为后向状态；然后将前向状态和后向状态提取的语义信息拼接在一起，得到双向 LSTM 的最终输出 o_i 和 o_j；最后利用相似度计算公式得到实体的相似度得分，进而判定 C_i 和 C_j 是否匹配。为了更好地处理上下文语义较弱的文本，我们采用了以字符(character)为粒度的双向 LSTM 网络。该模型首先让输入字符串经过嵌入(embedding)层，得到实体的字符级词向量，再将该词向量输入 BiLSTM，并将最后一层隐藏状态向量拼接，得到输出向量，最后使用余弦相似度计算实体匹配的概率。

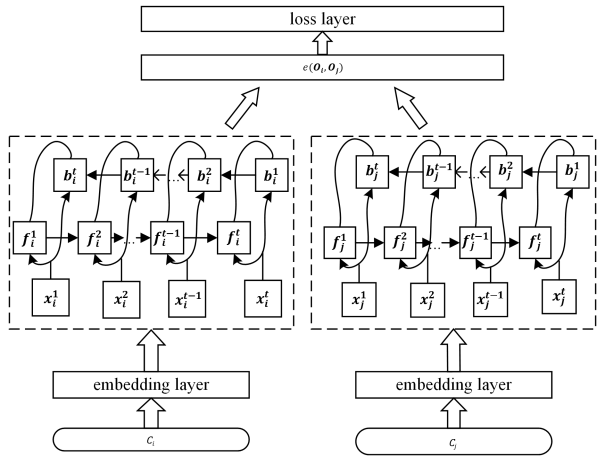

图 4 基于孪生神经网络的实体对齐模型的结构[22]

Fig. 4 Architecture of entity alignment model based on siamese recurrent neural network[22]

由于金融领域缺乏高质量的有标注的语料库资源，我们对前文抽取的文本进行标注，最终在标注好的数据集上对上述模型进行测试，并将其与以单词为粒度的 MaLSTM 模型进行了对比，具体实验结果如表 2 所列。从实验结果可以看出，两种模型在实体对齐任务上都取得了不错的效果，但与 word-based MaLSTM 模型相比，Character-based BiLSTM 更适合处理上下文语义较弱的文本。

表 2 实体对齐的实验效果

Table 2 Experimental results on entity alignment

Algorithm	*Accuracy*	*Recall*	*F*1
word-basedMaLSTM	0.7853	0.8734	0.8382
character-basedBiLSTM	**0.8221**	**0.9775**	**0.8852**

3.4 知识存储

知识存储是知识图谱构建过程中非常重要的一环，在对不同来源、不同形态的数据实现融合后，需要将这些数据存入数据库，用于支撑知识推理、知识计算等上层应用。由于知识图谱中存储的大都是关联密集型的数据，而图数据库能够方便地存储这一类型数据，因此图数据成为了主流的存储方式。图数据库以“图数据结构”来表现和存储数据，并实现了快速查询。它将节点与节点之间的关系以键值对〈Key，Value〉的形式进行组织、索引和存储。与传统的关系型数据库相比，图

数据库的这种数据结构在大规模知识图谱上的查询和搜索效率得到了显著提高。此外，图数据库的设计非常灵活，当加入新的属性或数据时，不需要重构网络。常见的图数据库包括 Neo4j，OrientDB 和 Titan 等。

本文采用 Neo4j 图数据库对知识图谱进行存储。Neo4j 数据库具有高可用性和可拓展性，支持 ACID 完整的事务管理。Neo4j 图数据库中的节点表示实体，边表示实体间的关系，每个实体的属性用〈Key，Value〉键值对的结构进行存储。在导入数据时，使用官方提供的 Neo4j-import 工具进行批量导入。

3.5 系统展示

本文最终构建了一个包含 90 多万企业实体和 300 多万关系的企业知识图谱，企业风险知识图谱的界面如图 5 所示。图 5 中，节点代表实体，在知识图谱中用圆形表示；带标注的连边表示实体与实体之间的关系。以“华纺股份”为例，当用户搜索“华纺股份”时，会展示与之相关的实体和关系，包括所属概念、所属行业、董事会成员等。

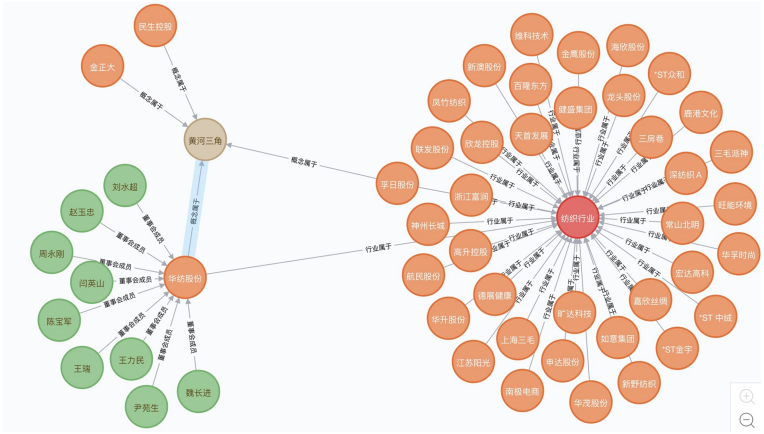

图 5 企业风险知识图谱界面

Fig. 5 Screenshot of enterprise risk knowledge graph

4 基于企业风险知识图谱的智能问答系统

基于前文构建的企业风险知识图谱，用户还不能方便、迅速地获取信息。因此，本文利用深度学习技术，结合知识库构建了基于企业风险知识图谱的智能问答机器人，允许用户以自然语言问句的形式进行提问。问答系统接收到用户的提问时，能够精确地理解用户搜索意图并返回准确的答案，从而帮助用户简单、高效地获取感兴趣的信息。

本节所介绍的基于企业风险知识图谱的智能问答系统的结构如图 6 所示。

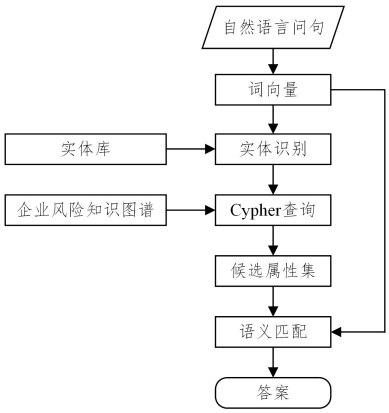

图 6 基于企业风险知识图谱的智能问答模型的结构

Fig. 6 Architecture of question-answering system based on enterprise knowledge graph

智能问答模型在接收到用户的提问后，首先对问句进行预处理，包括中文分词、词性标注、去停用词等操作；然后利用 Word2vec 将用户问题转换成低维向量，并通过命名实体识别技术从问句中提取人名、企业名、机构名等实体信息；最后将自然语言问句转化成 Cypher 查询语句并返回结果。例如，输入问句“斯太尔的董事有哪些”，在对问句进行分词、纠错、词性标注等预处理后，智能问答系统首先采用命名实体识别算法识别出实体“斯太尔”，并生成 Cypher 查询语句，然后返回实体的候选属性集，最终利用语义匹配技术得到问题的正确答案“高立用、李晓振、邹书航、孙建设、郭英祥、吴振平、孙钢宏、王德建”。命名实体识别是流程中的核心，为了提升实体识别准确率，本节使用了基于字的 BiLSTM-CRF 命名实体识别模型。

4.1 基于字的 BiLSTM-CRF 模型

长短时记忆网络（Long Short Term Memory，LSTM）是循环神经网络（Recurrent Neural Network，RNN）的一种特殊类型，可以学习长距离依赖信息。LSTM 通过 3 个门结构（输入门、遗忘门和输出门）选择性地遗忘部分历史信息，加入部分当前的输入信息，将其整合到当前状态，并产生输出状态。

将双向 LSTM 模型用于命名实体识别任务时，可以学习到观测序列之间的依赖，但无法学习到状态序列之间的关系。将传统的条件随机场（Conditional Random Fields，CRF）模型融合到 LSTM 中，可以解决这个问题。CRF 模型可以对隐含状态建模，学习到输出标注之间的关系，进一步提高 LSTM 模型预测的准确性。BiLSTM-CRF 模型的具体结构如图 7 所示，其由 Embedding 层、BiLSTM 层和 CRF 层组成。Embedding 层将稀疏的 one-hot 字向量通过 embedding 矩阵映射为低维稠密向量 $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_t\}$。BiLSTM 层通过将正向 LSTM 输出的隐状态序列 $\{\overrightarrow{\boldsymbol{h}_1}, \overrightarrow{\boldsymbol{h}_2}, \cdots, \overrightarrow{\boldsymbol{h}_t}\}$ 与反向 LSTM 输出的隐状态序列 $\{\overleftarrow{\boldsymbol{h}_1}, \overleftarrow{\boldsymbol{h}_2}, \cdots, \overleftarrow{\boldsymbol{h}_t}\}$ 在对应时刻进行拼接，得到完整的隐藏状态 $\{\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_t\}$。CRF 层接收 BiLSTM 层的隐藏状态序列作为输入，通过学习标签间的约束条件提升标签预

测的准确性，从而得到最终的预测标签序列。

将神经网络与 CRF 结合的 BiLSTM-CRF 模型结合了二者的优势：BiLSTM 可以有效标注文本序列和标签之间的关系；CRF 能够利用上下文信息，有效预测标签与标签之间的关系。BiLSTM-CRF 模型已成为目前命名实体识别方法中的主流模型。

目前中文问答领域还没有比较权威的公开数据集，尤其是针对金融领域的问答系统。因此，我们从同花顺、雪球网爬取了金融领域问答语料，并对其进行了筛选，剔除了主观性问题。基于此语料，我们对比了 BiLSTM 和 BiLSTM+CRF 两种命名实体识别模型。具体实验结果如表 3 所列。通过实验结果可知：1)CRF 层能够为最终的预测标签添加约束来确保其合理有效，进而提升模型的性能；2)在样本数量较少时，基于字级别的模型效果更优。

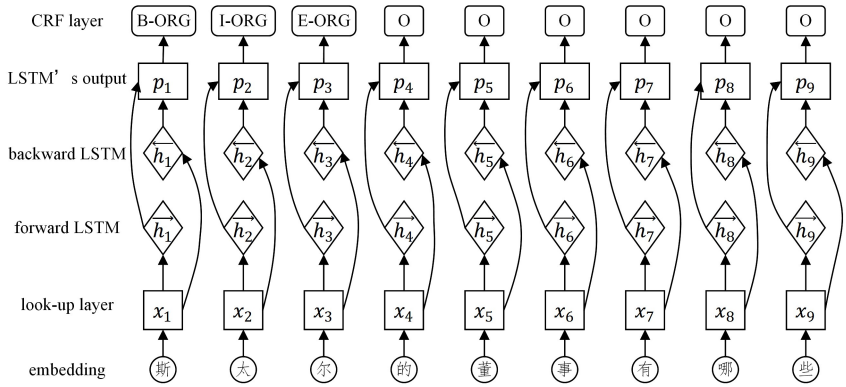

图 7 BiLSTM-CRF 模型的结构

Fig. 7 Architecture of BiLSTM-CRF model

表 3 KGQA 实体识别的实验结果

Table 3 Experimental results on named entity recognition

（单位：%）

Model	*Accuracy*	*Recall*	*F*1
BiLSTM	77.01	81.86	78.90
BiLSTM+CRF	87.12	86.75	87.02
Char-BiLSTM+CRF	**88.32**	**87.56**	**88.18**

4.2 问答系统展示

基于企业风险知识图谱的智能问答系统通过 Flask 实现对问答结果的可视化。Flask 是一个基于 Python 实现的 Web 开发的“微”框架，具备轻巧、简洁、扩展性强等特性。本文首先利用命名实体识别等方法来解析用户问题，然后根据其构造 Cypher 查询语句，通过知识图谱得到实体的候选属性集，最终通过属性链接得到最终的问题答案。图 8 给出了输入“斯太尔的董事有哪些”“华纺股份属于哪个行业”等问题后系统返回的界面。

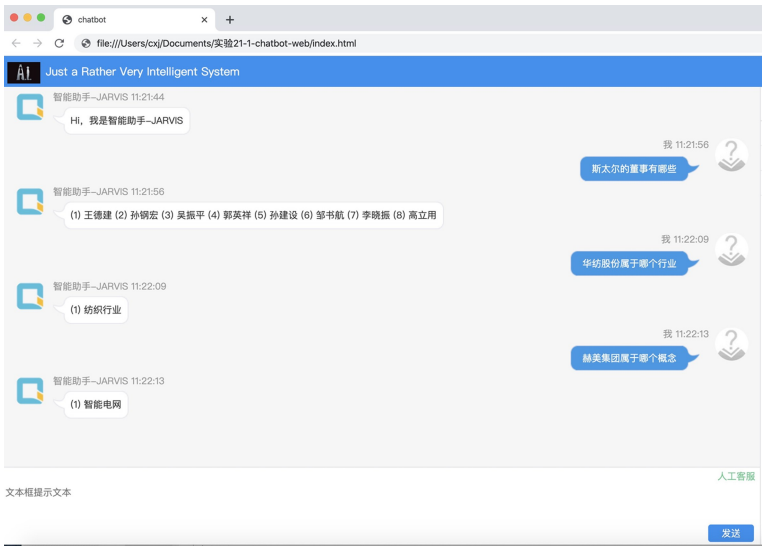

图 8 基于企业风险知识图谱的智能问答系统的可视化界面

Fig. 8 Visual interface of question-answering system based onenterprise risk knowledge graph

结束语 目前知识图谱已经成为学术界和工业界研究的热点，但国内在企业知识图谱构建方面的研究还处于初级阶段，面临着诸多挑战和难题，且相关研究很少。本文详细阐述了企业风险知识图谱的完整构建过程，首先进行领域本体构建；然后通过命名实体识别、关系抽取和实体链接等技术，从不同来源、不同结构的数据中抽取构建知识图谱所需的各种候选实体及其关联属性；接着根据知识图谱的结构特点，采用图数据库对其进行存储，构建企业风险知识图谱。基于构建好的知识图谱，本文结合深度学习和知识库构建了一个问答机器人来提供智能问答服务。

本文的研究将为企业风险知识图谱的构建提供一定的参考价值和借鉴意义,但仍有改进的空间,未来我们将在以下两个方向进行探索。1)知识图谱中的实体和关系可能会随着时间发生改变,例如2018年阿里巴巴的董事长是马云,而2019年阿里巴巴的董事长变成了张勇,因此未来拟在现有图谱的基础上引入时间维度,构建动态知识图谱,以更加有效地挖掘企业风险信息。2)目前本文所构建的智能问答系统还不能回答需要经过多步推理才能获取答案的问题,如"某某公司的法人代表名下还有几家公司"。因此,将知识推理融入智能问答模型中,也是一个值得研究的方向。

参 考 文 献

[1] BERNERS-LEE T,HENDLER J,LASSILA O. The semantic web [J]. ScientificAmerican,2001,284(5):28-37.

[2] YU X Y. Research on the Construction and Application of Knowledge Graph of Enterprise Legal Person [D]. Qingdao: Qingdao University,2018.

[3] E S J,LIN P Y,XIANG Y. Automatical construction of Chinese knowledge graph system [J]. Journal of Computer Applications,2016,36(4):992-996.

[4] BOLLACKER K,EVANS C,PARITOSH P,et al. Freebase:a collaboratively created graph database for structuring human knowledge[C]// Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. New York:ACM,2008:1247-1250.

[5] REBELE T,SUCHANEK F,HOFFART J,et al. YAGO:A multilingual knowledge base from wikipedia, wordnet, and geonames[C]// Proceedings of International Semantic Web Conference. Berlin:Springer,2016:177-185.

[6] AUER S,BIZER C,KOBILAROV G,et al. DBpedia:A nucleus for a web of open data[C]// Proceedings of the 6th International Semantic Web Conference. Berlin:Springer,2007:722-735.

[7] PELLISSIER TANON T,VRANDEI D,SCHAFFERT S,et al. From freebase to wikidata:The great migration[C]// Proceedings of the 25th International Conference on World Wide Web. New York:ACM,2016:1419-1428.

[8] NAVIGLI R,PONZETTO S P. BabelNet:Building a very large multilingual semantic network[C]// Proceedings of the 48th annual meeting of the association for computationallinguistics. Stroudsburg:Association for Computational Linguistics,2010:216-225.

[9] SPEER R,CHIN J,HAVASI C. Conceptnet 5.5:An open multilingual graph of general knowledge[C]// Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. Palo Alto,2017:4444-4451.

[10] WANG Z,WANG H,WEN J R,et al. An inference approach to basic level of categorization[C]// Proceedings of the 24th Acm International on Conference on Information and Knowledge Management. New York:ACM,2015:653-662.

[11] WANG Z,LI J,WANG Z,et al. XLore:A Large-scale English-Chinese Bilingual Knowledge Graph[C]// Proceedings of the International Semantic Web Conference. Berlin: Springer, 2013:121-124.

[12] NIU X,SUN X,WANG H,et al. me-weaving chinese linking open data[C]// Proceedings of the International Semantic Web Conference. Berlin:Springer,2011:205-220.

[13] XU B,XU Y,LIANG J,et al. CN-dbpedia:A never-ending chinese knowledge extraction system[C]// International Conference on Industrial,Engineering and Other Applications of Applied Intelligent Systems. Berlin:Springer,2017:428-438.

[14] PUJARA J. Extracting Knowledge Graphs from Financial Filings[C]// Proceedings of the 3rd International Workshop on Data Science for Macro--Modeling with Financial and Economic Datasets. New York:ACM,2017:5-6.

[15] RUAN T,XUE L,WANG H,et al. Building and exploring an enterprise knowledge graph for investment analysis[C]// Proceedings of International Semantic Web Conference. Berlin: Springer,2016:418-436.

[16] SONG D,SCHILDER F,HERTZ S,et al. Building and querying an enterprise knowledge graph [J]. IEEE Transactions on Services Computing,2019:356-368.

[17] LIU Q,LI Y,DUAN H,et al. Knowledge graph construction techniques [J]. Journal of Computer Research Development,2016,53(3):582-600.

[18] ZHANG W X,ZHU Q H. Research on construction methods of domain ontology [J]. Library and Information,2011,155(1):16-19,40.

[19] USCHOLD M,KING M. Towards a methodology for building ontologies [M]. Edinburgh: Artificial Intelligence Applications Institute,1995:1-13.

[20] FOX M S. The tove project towards a common-sense model of the enterprise[C]// International Conference on Industrial,Engineering and OtherApplications of Applied Intelligent Systems. Berlin:Springer,1992:25-34.

[21] LI W,HAN J,PEI J. CMAR:Accurate and efficient classification based on multiple class-association rules[C]// Proceedings of 2001 IEEE International Conference on Data Mining. Piscataway:IEEE,2001:369-376.

[22] WANG W,WANG Z,PAN L,et al. Research on the construction of bilingual movie knowledge graph [J]. Acta Scientiarum Naturalium Universitatis Pekinensis,2016,52(1):25-34.

[23] LV Y. A Siamese Recurrent Neural Network for Entity Alignment[D]. Nanjing:Nanjing University,2018.

CHEN Xiao-jun, born in 1995, Ph.D, is a student member of China Computer Federation. His main research interests include knowledge graph reasoning and knowledge representation learning.

XIANG Yang, born in 1962, professor, Ph.D supervisor, is a senior member of China Computer Federation. His main research interests include artificial intelligence and natural language processing.