

基于需求文档和图神经网络的需求知识图谱构建方法

王红斌¹, 胡永鹏^{1,2}, 侯 湘³

¹昆明理工大学信息工程与自动化学院, 云南 昆明

²昆明理工大学云南省人工智能重点实验室, 云南 昆明

³重庆大学自动化学院, 重庆

Email: whbin2007@126.com

收稿日期: 2021年5月18日; 录用日期: 2021年6月15日; 发布日期: 2021年6月22日

摘 要

知识图谱可以将大量不同种类信息连成一个语义网络, 为人工智能相关的研究提供了从“关系”层面解释分析问题的能力。因此, 构建各领域知识图谱成了近年来的研究热点。然而对于服务需求领域知识图谱的构建研究较少, 针对这一现状, 本文提出了一种基于需求文档和图神经网络的需求知识图谱构建方法, 该方法以100条公司真实需求文档为数据源, 结合传统自然语言处理方法和图神经网络分类模型, 通过知识抽取、人工标注、需求特征编码、需求分类、需求图谱存储和可视化等5个步骤来进行服务需求领域知识图谱的构建。实验表明该方法可以有效地从大量非结构化需求文档中提取到需求语义信息, 并通过图神经网络分类模型较准确地识别需求意图, 从而结合图数据库和可缩放矢量图可视化技术将需求图谱进行轻量级存储和可视化展示。

关键词

知识图谱, 需求文档, 可视化, 图神经网络

Construction Method of Requirement Knowledge Graph Based on Requirement Document and Graph Neural Network

Hongbin Wang¹, Yongpeng Hu², Xiang Hou³

¹Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming Yunnan

文章引用: 王红斌, 胡永鹏, 侯湘. 基于需求文档和图神经网络的需求知识图谱构建方法[J]. 计算机科学与应用, 2021, 11(6): 1725-1737. DOI: 10.12677/csa.2021.116178

²Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming Yunnan

³College of Automation, Chongqing University, Chongqing
Email: whbin2007@126.com

Received: May 18th, 2021; accepted: Jun. 15th, 2021; published: Jun. 22nd, 2021

Abstract

The knowledge graph can connect a large number of different kinds of information into a semantic network, and provides the ability to explain and analyze problems from the “relationship” level for the AI related research. Therefore, the construction of knowledge graph in various fields has become a research hotspot in recent years. But there are few ways to build a requirement knowledge graph. In order to solve this problem, this paper proposes a requirement knowledge graph construction method based on requirement document and graph neural network. The method takes 100 real company requirement documents as the data source. Combined with the traditional natural language processing method and the graph neural network classification model, through knowledge extraction, manual annotation, requirement feature coding, requirement classification, requirement graph storage and visualization, the knowledge graph for the requirements domain will have been successfully constructed. Experiments show that this method can effectively extract requirement semantic information from a large number of unstructured requirement documents, and identify requirement intention more accurately through the graphical neural network classification model. Thus, combined with graph database and scalable vector graph of visualization technology, the requirement graph is lightweight-stored and visualized.

Keywords

Knowledge Graph, Requirement Document, Visualization, Graph Neural Network

Copyright © 2021 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

知识图谱是谷歌在 2012 年提出的一种语义网络,是用图的形式来展示知识和客观世界万物之间的关联关系的一种技术[1],在图书情报界被称为知识域可视化或知识领域映射地图。知识图谱是由边和节点组成。节点是一个概念实体,如苹果、人名、自然语言处理,知识图谱。边则是概念实体的相关属性或者实体之间的关系,如英文名、引用等。知识图谱最早的应用是用来提升浏览器搜索引擎的能力,随着知识表示和机器学习等技术的发展,知识图谱也得到了突破性的进展,近年来被应用于一些前沿领域。例如知识图谱被广泛应用于人机问答交互中,在我们日常生活中常见到的小爱同学、siri、天猫精灵等都有海量的知识图谱作为支撑。知识图谱技术也被应用于辅助进行大数据的分析,被当作先验知识从海量文本中抽取实体和关系,用来增强数据之间的关联性,使得我们可以以知识图谱这种直观的方式对数据进行关联的数据挖掘和分析。此外,知识图谱还被用于物联网设备的互联、可解释性人工智能等多个研究点,体现出了知识图谱具有丰富的应用价值。

然而目前各领域知识图谱构建存在困难,某领域的行业知识图谱的构建一般会交给对该领域有更深研究的专家来承担。尽管专家对于该领域知识十分擅长,但他本身对于知识图谱概念的理解以及构建方法并不熟悉,这导致构建某领域知识图谱的时候,花费的成本会非常高,无法快速地构建出符合该领域规范的知识图谱。此外,越来越多的知识图谱应用场景是以非结构化文档作为知识抽取的直接数据源。针对某个领域,如何设计合理的信息抽取模型,对非结构化文档进行关系抽取也是现在领域知识图谱构建的困难之一。

2. 相关工作

近年来,出现了大量的知识图谱构建方面的研究工作。林杰等人[2]通过以专业社交媒体文本为研究对象,考虑其语料数量巨大、长短不一、专业性强、创作随意性强、口语化的特点及主题图谱内容的结构,提出了一种人工参与少、内容纯净且结构合理的主题图谱构建方法;林单[3]等人立足于生态学理论,引入全景生态的概念,充分考虑微博舆情中主体、客体、主体客体间的关联关系,从主体、客体、全景3个维度对微博舆情图谱进行了知识图谱构建;丁晟春[4]等人立足电商环境下的产品信息组织与挖掘,将产品的静态信息和动态口碑结合,引入知识图谱方法构建产品知识图谱,旨在帮助购物平台改善产品多维信息的组织与关联,为用户提供更好的产品服务;吕华揆[5]等人选择股权结构作为切入点,借助知识图谱工具,从知识关联角度出发,通过考察金融机构间持股关系、持股比例,构建中国金融股权知识图谱;王晰巍[6]等人在知识图谱基础上,围绕分析的主题内容构建大数据驱动的社交网络舆情主题图谱,可以对社交网络中的谣言、欺诈、涉敏等舆情事件的复杂关系网络、舆情事件时间演进过程进行可视化分析;Wang S. [7]等人提出了一种基于众包的新型去中心化知识图构建方法,众包的业务逻辑通过区块链驱动的智能合约实现,以保证透明性、完整性和可审计性。

不同于上述的领域知识图谱构建,本文针对领域服务需求来构建需求图谱,提出了基于需求文档和图神经网络的需求知识图谱构建方法,该方法通过从非结构化需求文档中提取需求三元组,并结合人工标注和图神经网络训练出需求三元组的标注模型,得到所有需求三元组的标签,构建出需求图谱并可视化。进而通过使用构建好的需求图谱,可以很方便进行智能推荐,服务组合等下游相关子任务,也可以作为搭建服务互联的科技服务协同平台的基础构件。

3. 研究思路和方法

本文介绍了基于需求文档的需求图谱构建方法,该方法总共分为5个部分,分别为:非结构化需求文本的三元组抽取,数据整理与人工标注标签,数据特征编码及输入矩阵的构建,图神经网络模型分类,需求图谱的可视化。通过对企业提供的100条非结构化需求文本进行需求三元组的抽取,从而得到100条需求文本中的核心需求三元组表示,随机对这些抽取出来的部分需求三元组进行人工标注,标注出约束、功能等意图标签,利用one-hot编码的方式得到这些三元组的特征编码表示,将数据处理成邻接矩阵和特征矩阵的形式输入到设置好的图卷积神经网络,得到一个可以对需求三元组自动打标签的节点分类模型,标注剩余未标注的需求三元组,最后利用d3.js [8]和Neo4j [9]图像可视化技术和图数据库将需求图谱进行轻量级存储和可视化展示。图1给出了本方法的总体框架图,输入的数据为大量需求文档,通过知识抽取的方法从中提取数据中的需求三元组,提取完成之后随机抽取三元组进行人工标注,将标注完成的数据构建成特征矩阵和邻接矩阵输入到设置好的图卷积神经网络训练,最后得出一个可识别需求三元组意图标签的模型,利用该模型对剩余需求三元组进行自动标注并进行需求图谱可视化和图数据存储。

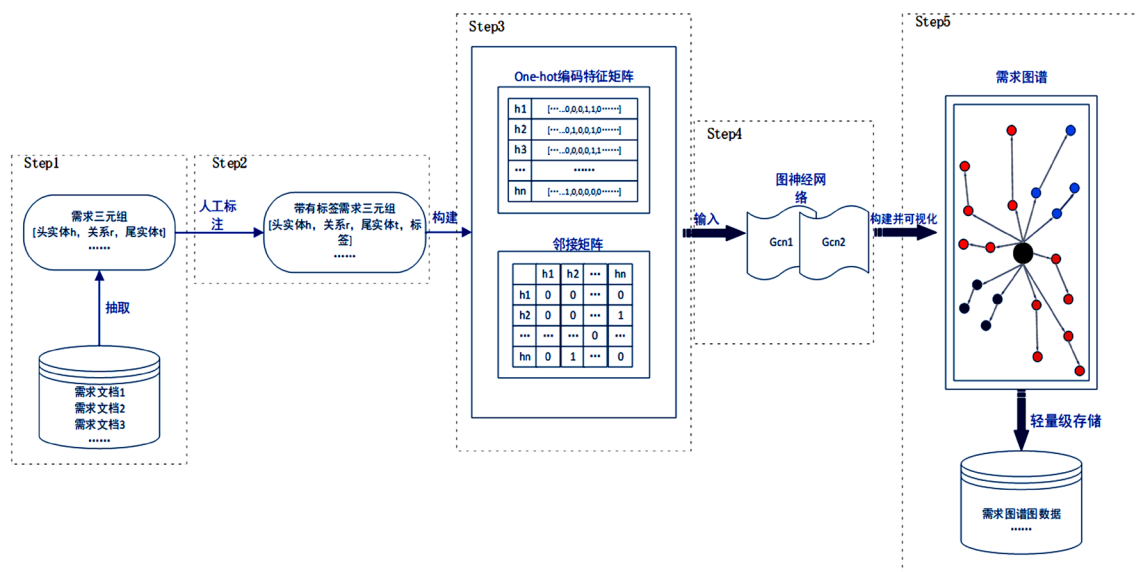


Figure 1. Overall framework of the method
图 1. 方法总体框图

3.1. 非结构化需求文本的三元组抽取

在对于非结构化需求文本的提取方面，本文利用的是传统自然语言处理的方法，即通过依存句法和语义角色标注两种方法来从需求文本中提取我们所需的需求三元组。语义角色标注是根据一个句子中的谓语和论元之间存在不同的语义关系，将论元划分为不同类型，这里的谓语也就是指的是动词，论元则是名词。通过语义角色的标注可以提取一个文本中三元组关系。但有时候只依靠语义角色标注，并不能提取所有句子中的三元组关系，所以当语义角色标注提取失败的时候，本文采用依存句法来提取三元组关系。依存句法是由依存语法和句法分析组成，其中依存语法是基于词间关系的形式语法，它强调以动词为句子中心，其他词汇依存于动词[10]。而句法分析则是用来确定文本中句子的句法结构和句子中词汇之间的依存关系。其主要是完成两方面的内容，首先需要确定一个语法体系，然后根据确定的语法体系，推导出一个句子的句法结构，本文采用自上而下的语法体系[11]构建句法树。

在使用上述依存句法和语义角色标注方法时，我们采用了 LTP 语言模型(哈工大中文语言技术平台)[12]，该模型主要用于对中文语言进行分词、关键词提取等一系列自然语言处理操作，LTP 语言模型中包含分词模型、依存句法分析模型、语义角色标注模型、词性标注模型等常用的中文处理模型的封装包，具体算法运行步骤如图 2 所示。例如输入一个需求文本：X 公司研发导航系统，导航应用于医院常用影像设备。需求主要在硬件设备和软件，应提供整套设备的配套技术。期望研发时间为 10 个月内，预算金额为 4000 万人民币。其中该文本的第一句依存句法和语义角色标注结果如表 1 所示，如果存在 SVO 结构那么程序会在第 3 部分运行结束后停止，返回 SVO 三元组，否则一直执行到第五部分，提取需求三元组。

Table 1. Requirements text extraction steps
表 1. 需求文本提取步骤

步骤	内容
需求文本	某公司研发导航系统，导航应用环境为医院常用影像设备。需求主要在硬件设备和软件，应提供整套设备的配套技术。期望研发时间为 10 个月内，预算金额为 4000 万人民币。
1.分词	{X, 公司, 研发, 导航, 系统, 导航, 应用, 环境, 为, 医院, 常用, 影像, 设备。}

Continued

2.词性标注	{nz, n, v, n, n, wp, n, v, n, v, n, a, n, n, wp}
3.语义角色标注	{2 A0:(0,1)A1:(3,4), 9 A1:(10,13)}
4.命名体识别	{B-Ni, E-Ni, O, O,, O, O}
5.依存句法分析	{2:ATT,3:SBV,0:HED,5:ATT,3:VOB,3:WP,8:ATT,9:ATT,10:SBV,3:COO,14:ATT 14:ATT,14:ATT,10:VOB,3:WP}

Step1. 加载LTP语言模型（中文语言技术平台）：cws.model（分词模型），pos.model（词性标注模型），parser.model（依存语法模型），ner.model（命名识别模型），pisrl.model（语义角色标注模型）。

Step2. 加载一篇需求文档，对需求文档进行分词处理，去除无用标点符号。

Step3. 对分词处理的结果先调用语义角色标注，查看是否存在标准的SV0结构（主谓宾），如果有则提取出SV0三元组，否则判定为语义角色标注失败。

Step4. 语义角色标注失败后，则调用依存语法来进行三元组抽取。

Step5. 对分词好的每个词生成一个依存句法的子节点，并生成其父子数组的依存结构。

Step6. 循环每个词，找到其中动宾关系、定语后置动宾、介宾的主谓动补关系，并进行提取。

Step7. 对于一些提取主宾的词，找到具有相关依存结构的词，进行无用词剔除。

Step8. 将所有提取的三元组保存至excel表格中。

Figure 2. Dependency syntax and semantic role annotation steps

图 2. 依存语法和角色标注步骤

利用依存句法和语义角色标注完成三元组提取之后，本文从 100 篇公司需求文档中提取出了相关的需求三元组，并分成需求编号、头实体、关系、尾实体进行四列进行存储，以备后续步骤进行使用，部分提取出来的需求三元组如表 2 所示。

Table 2. Partly extracted requirements triples

表 2. 分提取出来的需求三元组

需求编号	头实体	关系	尾实体
65	X 公司	研发	导航系统
66	导航应用环境	应用于	医院常用影像设备
67	需求	提供	整套设备配套技术
68	研发时间	为	10 个月内
139	长期稳定性误差	小于	0.05%
156	X 公司技术	采用	独特指纹低功耗设计

3.2. 需求三元组的数据预处理

3.1 小节中我们对某公司 100 个需求文档进行了需求三元组的提取，得到了由需求编号，头实体，关系，尾实体这四部组成的数据，在这里我们定义需求编号为 $I = (i_1, i_2, \dots, i_n)$ ，头实体为 $H = (h_1, h_2, \dots, h_n)$ ，关系为 $R = (r_1, r_2, \dots, r_n)$ ，尾实体为 $T = (t_1, t_2, \dots, t_n)$ ，那么一个最初的需求三元组，则包含这四部分信息：第一部分是需求编号，即表示提取一段需求文档中需求三元组的序列，第二部分则是头实体，如表 1 所示，头实体可以是时间或者金额这样的约束，也可以是一个技术名称，第三部分是表示头实体和尾实体之间存在的关系，例如限制、小于、采用这样的关系，第四部分则是尾实体，其与头实体之间存在第三

部分的关系。一个需求三元组可以用如下公式(1)进行表示:

$$f(n) = (i_n, h_n, r_n, t_n) \quad (1)$$

其中, $f(n)$ 表示一个提取好的需求三元组数据, n 表示三元组在数据列表中所在的行, i_n 表示三元组的编号, h_n 表示需求三元组的头实体, r_n 表示头实体和尾实体之间存在的关系, t_n 表示需求三元组尾实体。考虑到构建需求图谱以及使用图谱进行下游子任务时我们需要知晓一个需求三元组的意图。本文规定需求三元组的意图标签大概分为功能、目标、约束以及需求根节点四类。因此, 我们需要对提取的需求三元组进行人工标注, 标注出这个需求的意图是目标、约束、功能, 还是需求根节点。我们对得到的需求三元组进行数据预处理, 首先将提取出来的三元组的头尾实体及关系进行组合, 然后对所有组合后的三元组进行标签标注。三元组标签定义为 $L = (l_1, l_2, l_3, l_4)$, 标注完成之后如表 3 所示。

Table 3. Triple data is required after manual annotation

表 3. 人工标注后的需求三元组数据

需求编号	需求三元组	需求三元组标签
65	{ ‘X 公司’, ‘研发’, ‘导航系统’ }	目标
66	{ ‘导航应用环境’, ‘为’, ‘医院常用影像设备’ }	功能
67	{ ‘需求’, ‘提供’, ‘整套设备配套技术’ }	功能
68	{ ‘研发时间’, ‘为’, ‘10 个月内’ }	约束
139	{ ‘长期稳定性误差’, ‘小于’, ‘0.05%’ }	约束
156	{ ‘本技术’, ‘采用’, ‘独特指纹低功耗设计方案’ }	功能

3.3. 图卷积神经网络分类模型

在 2.2 部分我们随机标注了 100 个需求文档中提取的需求三元组中 60%的数据作为我们下面模型的输入, 通过构建图卷积神经网络来训练一个可以自动标注剩余需求三元组的模型。因为本文最终目标是构建一个需求图谱, 所以我们需要一个可以对图结构数据进行分类的模型, 近年来图神经网络的应用十分广泛, 有人利用图神经网络来进行文本或者图像分类[13] [14], 并取得了非常不错的效果, 并且图神经网络也拥有非常好的可解释性[15] [16]。据此本文在训练自动标注模型的时候采用 GCN (图卷积神经网络) [13] [17]方法, 下面介绍 GCN 的相关基础知识和原理以及本文如何利用图神经网络进行分类。

3.3.1. 图卷积神经网络的特点

GCN 来源于卷积神经网络, 卷积神经网络是一种具有局部连接、权值共享等特点的深层前馈神经网络[18]。但传统的卷积神经网络仅能够处理欧几里得数据, 也就是具有平移不变性的数据, 常见的如一幅图片, 即 2D 的网格类型数据, 一般用矩阵来存储; 一段文本, 即 1D 的网格类型数据, 一般通过处理后嵌入到一维向量空间来表示。但本文的数据是以知识图谱节点形式构成, 即由很多{ ‘时间’, ‘限制’, ‘3 个月内’ }这种三元组数据构成的图网络, 每个节点的邻居节点数量并不一样, 不具有平移不变性, 属于非欧几里得数据, 不可以通过常规的卷积核去提取相同的结构信息。因此本文采用了图卷积神经网络方法对知识图谱数据进行处理。

GCN 的主要任务是可以从大量节点中学习到结构特征, 从而将不同节点进行分类处理, 如图 3 所示, 图卷积神经网络是一种端到端的学习方式, 一端是图结构数据, 另一端是最终完成任务。本文通过已进行人工标注出意图的需求三元组数据, 通过 GCN 训练出一个可以识别需求三元组的意图的模型, 从而节约构建完成需求图谱的时间。

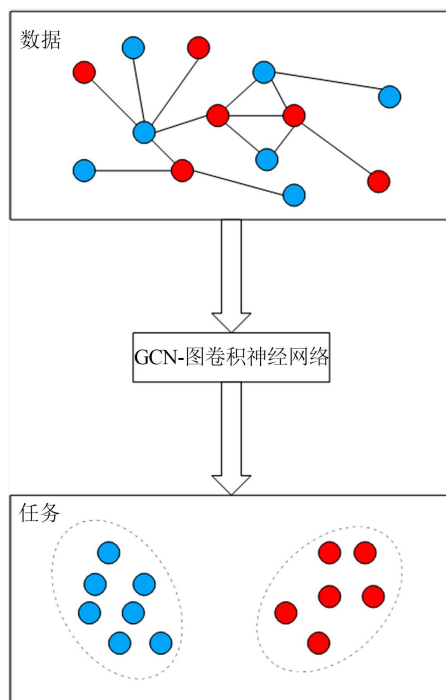


Figure 3. Figure convolutional neural network model classification task

图 3. 图卷积神经网络模型分类任务

3.3.2. 图卷积神经网络架构

GCN 的过程是图卷积的一阶局部近似，可以看多是多层的神经网络，每一个卷积层仅处理一阶邻域信息，通过叠加若干卷积层可以实现多阶邻域的信息传递。通过 GCN 处理后意图相同的需求三元组将会被归类到一起。

GCN 的输入是图结构数据，例如本文输入为一个需求图谱，每个需求节点是由子图的 2 个需求实体节点构成的，里面有 $2n$ 个节点，每个节点表示了一个需求实体，都存在一些特征，那么也就是有 $2n * D$ 维的特征矩阵 F ，节点之间也会存在关系，那么会有一个 $2n * 2n$ 的邻接矩阵 A ，至此 F 和 A 便是 GCN 的输入。如图 4 所示，是图卷积神经网络的模型架构。

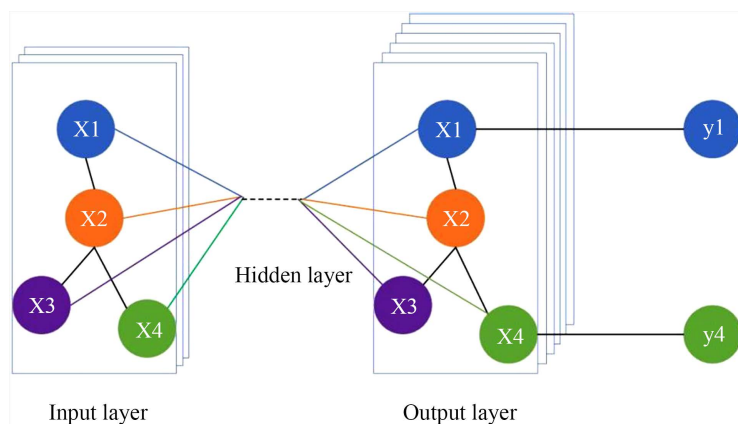


Figure 4. Graph volume neural network model architecture

图 4. 图卷积神经网络模型架构

从图 4 中可以看到主要有输入层、隐藏层、输出层，GCN 也是一种神经网络层，层与层的传播可以用公式(2)来表示，其中 H 代表的是每一层的特征，如果是输入层的话，那么 H 代表的则是上文叙述的特征矩阵 F ， $\tilde{A}$ 代表的是上文叙述的邻接矩阵 A 加上一个单位矩阵， $\tilde{D}$ 代表的是 $\tilde{A}$ 的度矩阵， σ 代表的则是非线性激活函数，在本构建方法中该非线性激活函数为 ReLu 函数。根据 GCN 层与层传递核心公式，我们需要的输入数据为特征矩阵 F 和邻接矩阵 A ，下文将详细介绍如何根据提取出来的需求三元组数据构建特征矩阵和邻接矩阵。

$$H^{(l+1)} = \sigma \left(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)} \right) \quad (2)$$

3.3.3. 图特征矩阵和图邻接矩阵的构建

GCN 需要构建特征矩阵和邻接矩阵作为输入。本文将每个需求三元组看作一个独立的子图，需要构建需求三元组的特征矩阵，我们将每个需求三元组看作一个独立的子图来提取特征，即特征矩阵每行为一个需求三元组的特征。本文这里的特征表示方法采用的是 one-hot 的编码，one-hot 编码首先需要创建一个分词器，并设置只考虑前 2000 个最常见单词，构建整数编码索引，例如其中特征整数编码如下：时间——0，限制——1，技术——2，采用——3。将特征按照从小到大进行排列，[‘时间’，‘限制’，‘技术’，‘采用’]得到的 one-hot 编码表示如下：[[1,0,0,0],[0,1,0,0],[0,0,1,0],[0,0,0,1]]。

根据上述 one-hot 编码特点，我们得到每个需求三元组的 one-hot 特征表示，根据处理得到每个需求三元组的 one-hot 特征表示，定义每个需求三元组的特征为公式(3)所示，其中 $F(f(n))$ 代表了第 n 个需求三元组的特征表示，因为本文设置了特征为 2000 维，所以 $F(f(n))$ 是一个 1 行 2000 列的特征矩阵，其中 m 代表的特征的列数，即第 n 个需求三元组的第 m 维特征。

$$F(f(n)) = (e_{n1}, e_{n2}, \dots, e_{nm}) \quad (3)$$

据此我们可以得到一个关于需求三元组的特征矩阵 F ，每一行代表了一个需求三元组的 2000 维的 one-hot 特征编码表示， F 则代表了 n 个三元组特征的特征集合。

$$F = \begin{bmatrix} e_{11} & e_{12} & e_{13} & \cdots & e_{1m} \\ e_{21} & e_{22} & e_{23} & \cdots & e_{2m} \\ e_{31} & e_{32} & e_{33} & \cdots & e_{3m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ e_{n1} & e_{n2} & e_{n3} & \cdots & e_{nm} \end{bmatrix} \quad (4)$$

设需求图谱 $G_r = (V, E)$ ，其中 $V = (v_1, v_2, \dots, v_n)$ 代表图谱中实体，即 2.1 中的头实体 H 和尾实体 T 的集合。接下来用邻接矩阵 A 描述需求图谱中顶点之间的关联，其定义为：

$$A[i, j] = \begin{cases} 1, & (v_i, v_j) \in V \\ 0, & (v_i, v_j) \notin V \end{cases} \quad (5)$$

其邻接矩阵表示的含义是，如果两个实体存在一条关系，那么认定值为 1，如果不存在关系则值为 0，邻接矩阵代表的是知识图谱中图结构。据此我们得到了一个特征矩阵 F 和邻接矩阵 A ，从而可以表示出需求图谱的特征和结构，为图卷积神经网络训练的实验部分做了数据准备。

4. 实验应用分析

4.1. 实验数据预处理

实验数据预处理主要是读取需求文档，并利用依存句法和语义角色标注提取出其中需求三元组，并

随机抽取需求三元组进行人工意图标注。本文使用的数据为 100 条真实的科技公司需求文档，其中部分需求文档如图 5 所示，例如举例其中一个需求文档：本公司研发导航系统，导航应用于医院常用影像设备。需求主要在硬件设备和软件，应提供整套设备的配套技术。期望研发时间为 10 个月内，预算金额为 4000 万人民币。之后会在这 100 条需求文档中随机抽取 70 条进行数据预处理，对这 70 条随机抽取出来的需求文档进行知识抽取。根据本文上述基于依存句法和语义角色标注的需求三元组抽取方法进行抽取，抽取完成之后通过人工标注的形式对抽取出来的需求三元组进行人工标注，标注出所有需求三元组的意图，其中案例需求文档的数据预处理结果如表 4 所示。接着利用本文上述方法提取出带有意图标注的需求三元组的特征矩阵和邻接矩阵作为图卷积神经网络的输入。

Table 4. Data preprocessing case

表 4. 数据预处理案例

需求编号	需求三元组	需求三元组标签
65	{ ‘X 公司’ , ‘研发’ , ‘导航系统’ }	目标
66	{ ‘导航’ , ‘应用于’ , ‘医院常用影像设备’ }	功能
67	{ ‘需求’ , ‘提供’ , ‘整套设备配套技术’ }	功能
68	{ ‘研发时间’ , ‘为’ , ‘10 个月内’ }	约束
69	{ ‘预算金额’ , ‘为’ , ‘4000 万人民币’ }	约束
65	{ ‘X 公司’ , ‘研发’ , ‘导航系统’ }	目标

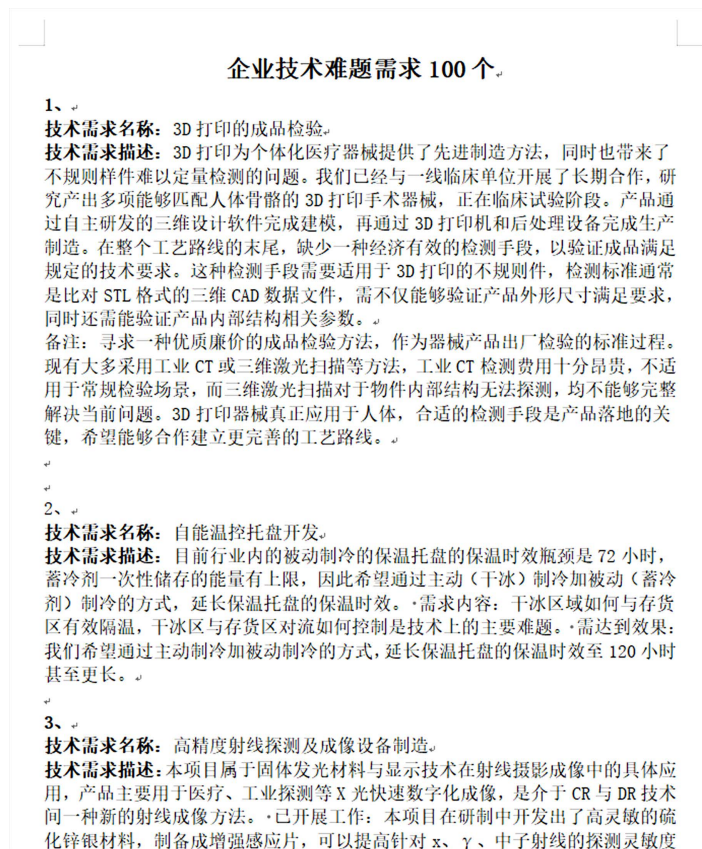


Figure 5. Requirements document data

图 5. 需求文档数据

4.2. 实验参数设置

本文对图卷积神经网络进行了相关配置,所有相关定义如表 5 所示。考虑到图结构网络数据较为复杂训练时间较长,选择 GPU 作为训练环境。根据图卷积神经网络分类研究表明[13],对于正常任务仅需要两层图卷积网络即有很好的效果,所以本文也设置了两层图卷积层,其中输入维度为 2.3.3 节利用 one-hot 编码得到 2000 维的特征维度,隐藏层维度为 16,考虑到最后输出类别有 4 种,所以最后一层图卷积输出维度变为需求三元组意图类别概率,选取概率最大的对应类别作为该需求三元组的意图类别。权重衰减、训练迭代次数、学习率的设置则是在经过多次测试得出来相对最优值,当在学习 1000 次的时候标注准确率不会再提高。本文采用的损失函数为神经网络中经常使用的交叉熵损失函数,优化器为 Adam,因为 Adam 在经过偏置校正后,每一次迭代学习率都有个确定范围,使得参数比较平稳。

4.3. 实验参数设置

实验一是对意图标注进行准确率评估,将预处理完成的数据按 6:2:2 的比例分成训练集、测试集、验证集,在按照 4.2 实验参数设置后进行数据输入,模型最终输出标准结果准确率在 91.7%。本文之所以采用 GCN 来对需求三元组进行意图识别,主要原因有两点,首先无论是抽取的需求三元组还是需求图谱本身是由图数据构成的,分类模型最好能够处理图数据,从而使得方法的使用更为流畅,如若分类模型无法处理图数据,那么构建步骤会变得繁琐,从而降低构建效率。其次,近年来在对图数据分类处理方面,GCN 分类模型表现十分优秀,对于本文的需求图数据仅需两层 GCN 网络即可得到很好的效果,91.7%的高准确率已经满足意图识别功能所需要的结果。

Table 5. Definition of related parameters
表 5. 相关参数定义

名称	定义	参数值
设配	Device	Gpu
图卷积层数	GCN_Layers	2
输入维度	Input_dimension	2000
隐藏层维度	Hidden_Layer_dimension	16
输出层类别	Output_Layer_category	4
权重衰减	Weight_decay	5e-4
训练迭代次数	Epochs	1000
学习率	Learning_rate	0.04
损失函数	Loss	交叉熵
优化器	optimizer	Adam

实验二是对处理完的需求数据进行可视化展示,在分类模型训练完成之后,以新输入一条需求文本为例:本公司研发导航系统,导航应用于医院常用影像设备。需求主要在硬件设备和软件,应提供整套设备的配套技术。期望研发时间为 10 个月内,预算金额为 4000 万人民币。其需求图谱构建流程如表 6 所示。

其中表 6 中用到 d3.js [8]和 neo4j [9]两个技术,d3.js 是一个较为流行的可视化库,它允许绑定任意数据到文档对象模型当中,然后将数据驱动转换应用到文档中,之后通过数组形式数据传递给前端,展示出可缩放矢量图形组成的 svg 图。Neo4j 则是用来存储图结构数据的一个高性能数据库,与常规的 mysql

不同, neo4j 对于图结构数据存储是进行优化过的, 与本文需求三元组格式类似, 它分别存储节点类别、属性、边的类别, 所以非常适合用来存储需求图谱数据。最终该需求文本的可视化结果如图 6 所示。

Table 6. Examples of requirements graph construction

表 6. 需求图谱构建实例

步骤	内容
需求文本	本公司研发导航系统, 导航应用于医院常用影像设备。需求主要在硬件设备和软件, 应提供整套设备的配套技术。期望研发时间为 10 个月内, 预算金额为 4000 万人民币。
知识抽取	[['X 公司', '研发', '导航系统'], ['导航应用环境', '应用于', '医院常用影像设备'], ['需求', '提供', '整套设备配套技术'], ['研发时间', '为', '10 个月内'], ['预算金额', '为', '4000 万人民币']]
意图标注	[['X 公司', '研发', '导航系统', '目标'], ['导航应用环境', '应用于', '医院常用影像设备', '功能'], ['需求', '提供', '整套设备配套技术', '功能'], ['研发时间', '为', '10 个月内', '约束'], ['预算金额', '为', '4000 万人民币', '约束']]
图数据转化	{source: 'X 公司需求', target: 'X 公司', type: '根节点', 'rela': '需求 1'}, {source: 'X 公司', target: '导航系统', type: '目标', 'rela': '研发'}
图数据存储	将 Step3 图结构数据通过数据库添加语句存储到高性能的图形数据库 neo4j 当中。

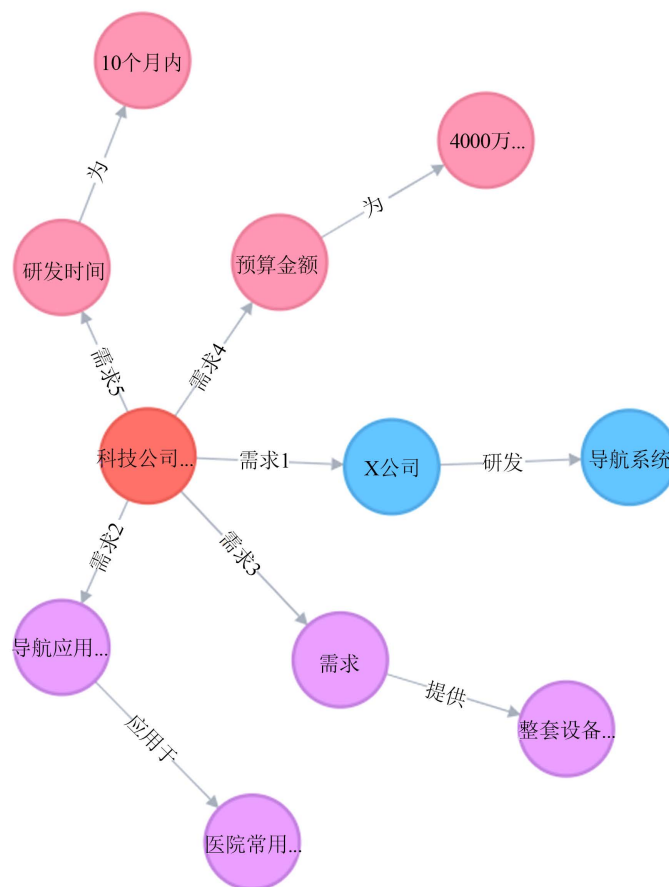


Figure 6. Visualization example of requirements graph

图 6. 需求图谱可视化示例

4.4. 需求图谱应用

构建出的领域知识图谱应用场景十分广泛,例如基于知识图谱的智能问答[19],基于知识图谱的智能推荐,基于知识图谱的热点分析等都是比较常见的应用场景。下面简单介绍服务需求领域的需求图谱在科技文献检索这一个应用场景下的情况。

知识图谱最早是被应用于提高搜索引擎的搜索效率的,早期传统的搜索引擎算法通过从海量的网站链接 URL 中搜索找到网站标题与搜索正文,通过 NLP 相关语义处理层之后得到一个匹配分数,通过分数的从高到底进行排序展示。但这样的搜索需求和搜索结果往往难以匹配,经过会有搜索结果和实际需求不符的情况。而如果搜索引擎增加了知识图谱,通过构建知识图谱可以得到一个庞大的语义网络,为搜索提供丰富且有效地辅助信息,从而提高了搜索的效率和准确率。

本文建立需求图谱同样可以应用于检索方面,以科技文献资源的检索为例,一条案例需求文档:本公司研发导航系统,导航应用于医院常用影像设备。需求主要在硬件设备和软件,应提供整套设备的配套技术。期望研发时间为 10 个月内,预算金额为 4000 万人民币。如果用传统的方法进行检索,无法有效从这么多文字中提取语义信息,搜索出来的导航系统与原需求文档不符。但通过需求图谱检索,搜索引擎会从这条需求的目的、功能、约束这三方面进行综合考虑检索科技文献,例如对案例需求文档进行检索,就会检索导航系统的文献,同时会检索会综合考虑这个案例所提取的功能和约束方面的语义,例如这个导航系统是专门应用于医疗领域的,需要整套软硬件设备等,从而最终检索结果会更符合用户输入的需求文档,检索出来关于导航系统科技文献会更符合用户需求。

5. 结论

本文针对科技服务领域的需求文档进行了需求图谱的构建,为服务推荐、服务关系挖掘以及服务互联的科技服务协同平台的搭建提供了基础的构件。本文对 100 条非结构需求文本进行命名实体识别以及关系抽取,从中提取到构建需求图谱所需要的基础需求三元组,之后对基础的三元组进行数据预处理和人工标注,标注出每个需求三元组在一个服务中的意图,将数据处理成图结构数据,并根据图结构数据构建邻接矩阵,同时将每个需求三元组看作是一个子图,利用 one-hot 编码的方式得到了每个需求三元组的 2000 维向量表示需求三元组的特征矩阵,将邻接矩阵和特征矩阵输入到已经设计好的两层图卷积神经网络中,通过图卷积神经网络学习到一个可以用于识别需求三元组意图的模型,对剩余未进行人工标注的需求三元组进行意图识别,将所有进行过意图识别的需求三元组再进行图数据处理,最后利用 d3.js 和 neo4j 技术将需求图谱进行可视化展示并进行存储。

未来随着自然语言处理、深度学习相关技术的不断发展,知识图谱的构建方法也会变得越来越成熟,知识图谱的构建也会渗透到各行业各领域,根据每个领域存在差异性,也会在不同领域的知识图谱构建上也会有领域化独特的构建流程,对于需求图谱领域的构建方法也会进一步发展。

参考文献

- [1] Singhal, A. (2012) Introducing the Knowledge Graph: Things, Not Strings. Official Google Blog. <https://blog.google/products/search/introducing-knowledge-graph-things-not>
- [2] 林杰, 苗润生. 专业社交媒体中的主题图谱构建方法研究——以汽车论坛为例[J]. 情报学报, 2020, 39(1): 68-80.
- [3] 王丹, 张海涛, 刘嫣, 等. 全景生态视角的微博舆情多维图谱构建研究[J]. 情报学报, 2019, 38(12): 1275-1285.
- [4] 丁晟春, 侯琳琳, 王颖. 基于电商数据的产品知识图谱构建研究[J]. 数据分析与知识发现, 2019, 3(3): 45-56.
- [5] 吕华揆, 洪亮, 马费成. 金融股权知识图谱构建与应用[J]. 数据分析与知识发现, 2020, 4(5): 27-37.
- [6] 周莉娜, 洪亮, 高子阳. 唐诗知识图谱的构建及其智能知识服务设计[J]. 图书情报工作, 2019, 63(2): 24-33.

-
- [7] Wang, S., Huang, C., Li, J., *et al.* (2019) Decentralized Construction of Knowledge Graphs for Deep Recommender Systems Based on Blockchain-Powered Smart Contracts. *IEEE Access*, 7, 136951-136961. <https://doi.org/10.1109/ACCESS.2019.2942338>
 - [8] Iglesias, M. (2019) Pro D3.js. Apress, Berkeley. <https://doi.org/10.1007/978-1-4842-5203-1>
 - [9] Vukotic, A., Watt, N., Abedrabbo, T., *et al.* (2014) Neo4j in Action. Manning Publications Co., Shelter Island.
 - [10] 袁文宜. 依存语法概述[J]. 科技情报开发与经济, 2010(18): 158-160.
 - [11] 王一宾, 陈文莉, 陈义仁. 语法分析方法研究评述及其应用[J]. 计算机工程与设计, 2007(13): 3063-3065.
 - [12] Che, W.X., Li, Z.H. and Liu, T. (2010) LTP: A Chinese Language Technology Platform. *Proceedings of the COLING 2010: Demonstrations*, Beijing, 13-16 August 2010, 13-16.
 - [13] Zhang, H.Q., Lu, G.Q., Zhan, M.M. and Zhang, B.X. (2021) Semi-Supervised Classification of Graph Convolutional Networks with Laplacian Rank Constraints. *Neural Processing Letters*. <https://doi.org/10.1007/s11063-020-10404-7>
 - [14] Wu, Z., Pan, S., Chen, F., *et al.* (2019) A Comprehensive Survey on Graph Neural Networks.
 - [15] Pope, P.E., Kolouri, S., Rostami, M., *et al.* (2020) Explain Ability Methods for Graph Convolutional Neural Networks. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, 16-20 June 2019, 10764-10773. <https://doi.org/10.1109/CVPR.2019.01103>
 - [16] Navarin, N., Tran, D.V. and Sperduti, A. (2019) Universal Readout for Graph Convolutional Neural Networks. 2019 *International Joint Conference on Neural Networks (IJCNN) IEEE*, Budapest, 14-19 July 2019, 20243-20249. <https://doi.org/10.1109/IJCNN.2019.8852103>
 - [17] Yan, S., Xiong, Y. and Lin, D. (2018) Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI 2018)*, New Orleans, 2-7 February 2018, 7444-7452. <https://arxiv.org/pdf/1801.07455.pdf>
 - [18] Zhang, Q., Zhang, M., Chen, T., *et al.* (2019) Recent Advances in Convolutional Neural Network Acceleration. *Neurocomputing*, 323, 37-51. <https://arxiv.org/pdf/1807.08596.pdf>
<https://doi.org/10.1016/j.neucom.2018.09.038>
 - [19] 陈璟浩, 曾桢, 李纲. 基于知识图谱的“一带一路”投资问答系统构建[J]. 图书情报工作, 2020, 64(12): 95-105.