杭州电子科技大学

本科毕业设计（论文）

（2023届）

题　　目	服装名称知识图谱构建
学　　院	自动化学院
专　　业	自动化
班　　级	19062813
学　　号	19061345
学生姓名	卓逸挺
指导教师	韩志敏老师
完成日期	2023 年 6 月

诚 信 承 诺

我谨在此承诺：本人所写的毕业设计（论文）《服装名称知识图谱构建》均系本人独立完成，没有抄袭行为，凡涉及其他作者的观点和材料，均作了注释，若有不实，后果由本人承担。

承诺人（签名）

年　　月　　日

摘　要

近年来，随着服装电子商务和人工智能技术的高速发展，出现了许多新型应用，如服装商品检索和个性化推荐等。然而，随之而来的也是海量的服装数据。面对这些数据，目前市面上所存在的传统意义上的存储技术对于其价值并不能够进行很好的开发和利用，为了充分利用大数据环境下产生的海量服装数据，并提高检索效率，知识图谱被作为一种人工智能前沿技术而受到广泛关注，为我们提供了一种在结构化的方式下储存和管理数据的方法。因此，本文基于目前主流电商网站的服装商品属性，构建了服装名称知识图谱。本文核心内容包括以下两部分：

（1）基于目前主流电商网站的服装商品属性，以天猫、京东、拼多多等电商网站的服装商品数据作为主要的数据来源，通过爬虫技术获取服装商品数据。在 Python 环境下，借助于 Request、Selenium 等 Python 第三方工具库，通过模仿网上的现实用户在电商平台浏览界面中所进行的操作，利用关键字定位和路径选择来找到符合要求的服装商品数据，并访问到服装商品所相对应的商品数据详情页获取记录 Url，通过 XPath、Css.selector 等特殊元素定位方法对网页进行解析，定位所需的服装商品数据在 HTML 文件中的位置，最后以纯文本形式存储表格数据。

（2）从知识图谱的研究动态的角度出发，结合本体概念和图数据库技术相关理论内容后，基于服装名称的商品属性，提出了一种基于“七步法”的本体模式层构建方法的改进方法，通过使用 Python 的第三方工具库 py2neo 对服装名称知识图谱的本体模式层进行构建，随后对所生成的 RDF 三元组数据使用语义插件，完成 RDF 三元组数据到图数据结构的映射，最后将服装商品数据存储到 Neo4j 图数据库中，完成对于女装、男装、童装与鞋子的服装名称知识图谱的构建。

关键词：selenium，爬虫，Neo4j 图数据库，服装名称知识图谱

ABSTRACT

In recent years, many new applications have emerged, such as garment product retrieval and personalized recommendations. However, this has also resulted in massive amounts of clothing data. Traditional storage technologies cannot effectively utilize and develop the value of such massive clothing data. To fully utilize the massive clothing data generated in a big data environment and improve retrieval efficiency, knowledge graphs have been widely studied as a cutting-edge technology in artificial intelligence. They provide a method for storing and managing data in a structured way. In response to the above-mentioned problem, this article constructs a clothing name knowledge graph based on garment attributes to solve the aforementioned problem. The core content of this article includes the following two parts:

（1）Based on the attributes of clothing products on mainstream e-commerce websites, such as Tmall, JD.com, and Pinduoduo, crawler technology is used to obtain clothing product data. In the Python environment, with the help of third-party libraries such as Request and Selenium, simulated user operations in the browser are performed by using the keyword search method to search for matching clothing products. The corresponding URL of the product is obtained and recorded by accessing the details page of the product. XPath, CSS.selector and other special element positioning methods are used to parse the webpage and locate the position of the required clothing product data in the HTML file. Finally, table data is stored in plain text format.

（2）From the perspective of research progress in knowledge graph, combined with ontology concepts and graph database technology, an improved method based on the ”seven-step method” ontology pattern layer construction method is proposed for commodity attributes based on clothing names. The clothing domain ontology is constructed using the third-party Python library py2neo, and RDF data is generated. After mapping RDF triples to a graph data structure using a semantic plugin, clothing data is stored in the Neo4j graph database, completing the construction of a knowledge graph for clothing names in women’s wear, men’s wear, children’s wear, and shoes.

Keywords: selenium，crawler，Neo4j，Knowledge Graph

目　录

1 绪论

1.1 研究背景及意义

随着电子商务的快速发展，我们正处于一个科技和社会进步的时代。电子商务的兴起已经彻底改变了我们的经济结构和个人生活方式。人们不再需要亲自前往商店购物，而是可以足不出户地在网上购买所需商品。这种便利的购物方式无疑满足了现代人们的需求，并且为我们带来了巨大的变化。现在，人们可以更加方便地购买商品和服务，而且也更容易比较价格和质量。同时，电子商务还创造了许多就业机会，促进了经济发展。总之，电子商务的发展已经成为了社会进步和科技创新的重要推动力，它正在深刻地影响着我们的生活和未来。根据《2021中国电子商务发展总报告》[1] 的数据显示，我国电子商务交易额达 43.2 万亿元，其中实物网上零售额为 10.8 万亿元，服装、鞋帽和针织纺织品占实物商品网上零售额的百分之 27.28，高达总销售额的 1/4 以上，具体数据如图1-1所示。同时我国电子商务市场的规模已经连续 9 年保持全球最大网络零售市场 [2]，如图1-2所示。由此可知，服装、鞋帽和针织纺织品的网上零售市场在电子商务商场中占据着较大的比重，拥有着巨大的发展潜力。

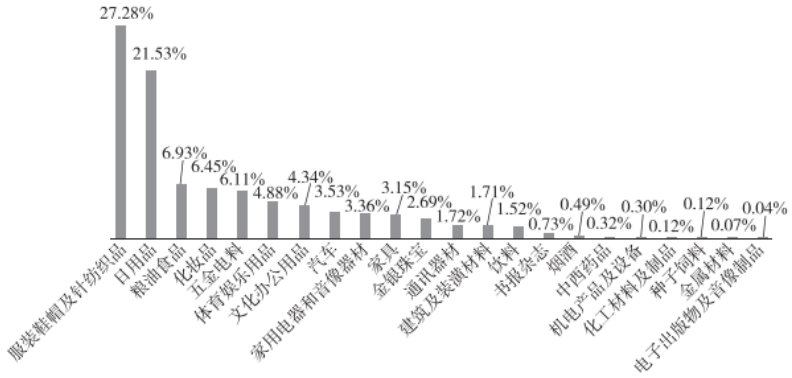

图 1-1 2021 年中国网上零售各类别店铺数量 [1]

虽然在服装商品行业的快速发展中，线上销售发挥着重要作用，成为影响行业发展的关键，但是也可以发现在科技飞速进步所带来的数据爆发性增长下，海量的数据并没有得到重视与有效利用等问题也影响着服装商品线上行业的发展。如何有效利用历史数据来提供足够的有用信息会直接影响商家如何去根据往年的销售情况来做当下的销售调整，如何解决买家从海量的商品数据中根据自己的喜好、场景等信息来高效快速的搜索出自己想要的服装也成为企业和学者关注的关键问题，也是本文研究的主要问题。

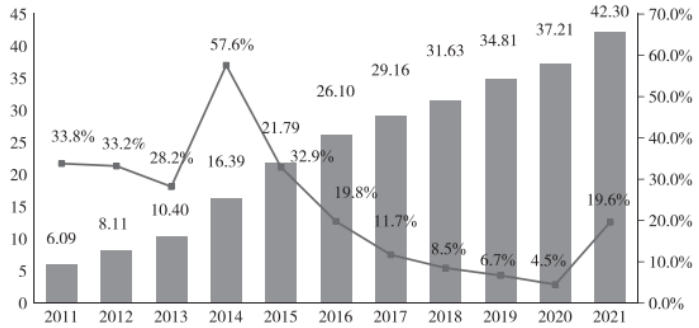

图 1-2 2011—2021 年中国电子商务交易额 [2]

知识图谱的出现无疑是解决这一问题的极好办法，关于如何构建知识图谱的问题的研究，现有研究主要以通用类型知识图谱为主，专门针对服装的知识图谱的研究较少，因此本文将以服装名称知识图谱为研究前景，重点分析和研究服装名称知识图谱构建问题，以期探寻服装名称知识图谱构建的问题原因、机制等，提出对策建议，为后续更加深入的研究提供基础。

本文研究将对有关服装名称的知识图谱构建进行深入的剖析与研究，说明与证实服装名称知识图谱构建的作用机制等，研究结果将拓展知识图谱的研究范围，丰富知识图谱的研究内容，为服装知识图谱构建相关的后续研究深入研究提供理论基础与理论依据。

本文研究结果将有助于买家从海量的商品数据中根据自己的喜好、场景等信息来高效快速的搜索出自己想要的服装，满足自己的消费需求，根据自己所搜索的服装名称得知自己所需要的相关信息；有利于商家根据的销售情况与用户的所搜索的信息来做当下的销售调整，帮助制定后续的应对问题的方法；服装名称知识图谱的构建也有利于知识图谱在服装个性化推荐与通过服装名称构建知识图谱方式等方面得到创新的发展；最重要的就是针对服装商品属性对用户检索结果的重要性，通过构建完整的服装知识图谱可以增强用户的检索准确性，从而提升用户的检索效果，为用户的购买决策提供帮助。

1.2 国内外研究现状

针对知识图谱的构建问题，国内外已经有大量的学者开展了相关的研究。目前关于知识图谱的研究主要分为 2 个大方向，分别是从知识图谱本身开始研究与知识图谱的嵌入式应用，下面将从这 2 个方面详细阐述国内外目前的研究现状。

国外在知识图谱的构建方面获得了重大进步，现在的我们有许多途径可以获取知识，其中最常用的知识图谱是谷歌搜索引擎 [3]。谷歌搜索引擎已经成为我们

日常生活中不可或缺的一部分，通过它我们可以快速地获取海量的信息。而这些信息主要来自于维基百科、知识图谱等结构化数据源，例如 Freebase 和 DBpedia[4]。这些数据源具有丰富的实体知识，可以抽取出相互关联的事实，并将相关信息发布出来供大家使用。此外，这些数据源还支持多种不同语言，使得跨语言学习变得更加容易。总之，这些结构化数据源对于人类获取知识和文化交流起到了重要的作用。通用知识图谱 DBpedia 如图1-3所示。

图 1-3 DBpedia 通用型知识图谱项目

国内市场目前涉及了多个知识图谱与应用产品，这些知识图谱包括了通用类型知识图谱和领域知识图谱。其中，百度知心和搜狗知立方是以搜索引擎为基础，提供智能问答服务的应用产品。XLor[5] 则是国内较为知名的预训练语言模型，可以用于文本分类、情感分析等自然语言处理任务。而 Zhishi.me[6] 是一个基于百科全书知识的中文知识图谱，覆盖了数百万实体和关系。MusicBrain[7] 和 IMDB[8] 则是音乐和电影领域的知识图谱，可用于查询相关的艺人、作品信息等。Geo Names[9] 则是地理信息的知识图谱，提供全球各类地理信息的查询服务。而 DBLife[10] 则是健康医疗领域的知识图谱，提供疾病、症状、药品等医学领域的查询服务。这些知识图谱的建设和应用，将促进人工智能技术的发展，助力各行各业的智能化升级。

在上述知识谱图的基础之上，余晓鹏也提出了基于服装商品属性，结合本体概念和图数据库技术，完成的服装知识图谱构建。该知识图谱是一种基于改进 Inception 结构的知识图谱嵌入模型—InceE[11]，具有更强的特征交互信息捕捉能力和普适性。该模型首先使用混合空洞卷积替代标准卷积，以提高特征交互信息捕捉能力；其次使用残差学习网络结构，以减少特征信息丢失。模型 InceE 如图1-4所示。

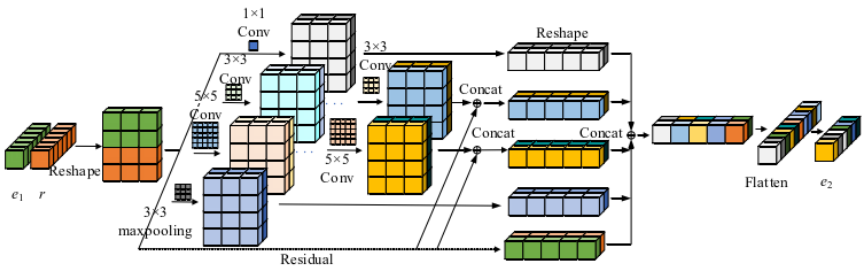

图 1-4 DBpedia 通用型知识图谱项目 [11]

李娜从命名实体识别、实体链接、属性选择和得出答案这四部构建基于知识图谱的问答系统。针对文本序列中的命名实体识别任务，近年来提出了多种有效的模型架构。李娜在命名实体识别模块选择了 BERT-BiLSTM-CRF 模型结合了预训练模型和 BiLSTM 等技术，在该任务上取得很好的效果。而卷积神经网络则在另一个任务——文本分类中表现突出，特别是 MAF-CNN 和交互融合矩阵等方法的结合使用更是在此基础上进一步提升了性能。同时，最大池化层也是常用的处理方式之一。除了单一模型外，还有矩阵融合算法可将不同模型的结果整合起来。这些模型和算法的应用在中文知识图谱问答系统（KBQA）中，通过提取问题和知识图谱实体之间的相似特征，配合以上模型和算法，实现对于问题的自动回答[12]。

在文本分类上由于 Bert-base,Chinese 预训练模型参数巨大，在做分类任务时容易产生过拟合现象，泛化能力弱。针对上述问题李铁飞、生龙、吴迪等人提出了 BERT-TECNN 模型 [13]，该模型使用 Bert-base。Chinese 模型作为动态字向量模型，Transformer-encoder 层再次对数据计算，最后应用 softmax 进行分类。该模型能有效提取文本中字词的特性信息，优化过拟合问题，提高泛华能力。

对于人工数据的处理上 Hochreiter S, Schmidhuber J. 使用的 LSTM[14] 解决了以前递归网络算法从未解决的长时间延迟问题；Weible C L. 也建立了一个互联网电影数据库（IMDd）[15]。

在传统服装领域，推荐系统面临着数据稀疏和冷启动问题。俞逸洋为了解决这些问题构建了服装服饰领域知识图谱 [16]，服装服饰领域知识图谱以服装服饰搭配为切入点，从知识关联的角度出发，通过知识融合技术对于搭配推荐和商品类目进行对齐、消歧。同时，用户历史行为数据可以与实体属性相结合，以发现知识关联并消除歧义。在知识建模方面，利用 protégé 工具和 OWL 数据，采用 Apriori 关联规则算法和 TF-IDF 算法，来挖掘关联关系和语义相似度，结合服装标签和数据表可以提供的隐含关联规则的信息，帮助推荐系统更好地理解用户需求，提供更准确的推荐服务。

潘王蕾构建的基于个性化推荐的服装知识图谱 [17] 将本体技术、知识图谱和推理机构作为构建领域类知识图谱的核心，利用先进的 Jena 推理机构和本体推理模型框架来构建高质量的知识图谱，其中隐性关联关系和自定义规则可以提高知识图谱的自动化能力和图谱质量。对于服装服饰电商平台，基于知识图谱可视化查询推荐系统和个性化推荐算法可以更好地满足消费者购买意愿，同时在线评论情感分析也可以帮助卖家了解消费者反馈。因此，基于个性化推荐的服装知识图谱，将有助于提高电商平台的用户体验和销售业绩。

知识图谱构建是一项复杂的任务，需要对特定领域进行深入研究和了解。研究的深度越深，就能够发现更多的问题和挑战，并且可以掌握更好的构建技术。通过使用自动化能力强大的工具，我们可以更高效地构建高质量的知识图谱。而在构建过程中，我们也需要不断优化图谱质量，确保它能够准确反映领域知识，并为用户提供有用的信息和洞见。虽然目前国内和国外的许多研究人员和研究机构都在知识图谱的研究上取得了一定的进展，但是领域类知识图谱的研究仍然存在着许多的问题，例如知识图谱相关构建技术的自动化能力低下，图谱质量偏低问题等等。

Liu H, Wu Y, Yang Y. 在大规模多关系嵌入这个问题的解决方案中针对嵌入实体和关系的类比财产，提出了一种优化潜在表示的新框架 [18]，这个新框架将 So-lar 系统和卢瑟福-波尔模型进行类比交换，如图1-5所示，此新框架建立在三个基本类比之上，不仅提供多关系嵌入方法的统一，而且具有公认的可扩展性。

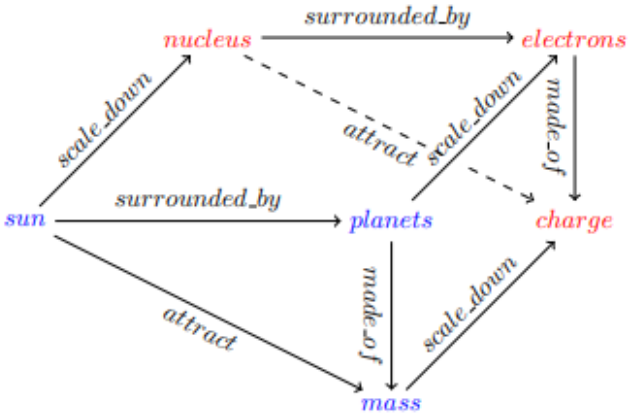

图 1-5 So-lar 系统和卢瑟福-波尔模型的类比交换图 [18]

由于大多数现有的知识图谱都存在不完整性，Vashisihth S, Sanyal S, Nitin V, et al. 通过 Interactive 增加特征交互来改进基于卷积的知识图的嵌入 [19]，根据 ConveE 方法提出了交互 E 增加交互特征，利于链路的预测性能。

知识图谱的嵌入也是预测知识图中缺失链接的强大技术，Zhang Z, Cai J, Zhang Y, et al. 提出了一种新的知识图嵌入模型——层次感知知识图谱嵌入（HAKE）[20]，将实体映射到了极坐标当中，HAKE 能有效地对知识图中的语义层次进行建模，而且在链接预测任务的基准数据集上有着显著优势。

目前的知识图嵌入方法第一非常复杂，需要大量时间进行训练和推理，Ren F, Li J, Zhang H, et al. 提出了一种简单有效的基于萎缩卷积的知识图嵌入方法 [21]。通过使用萎缩卷积增加特征交互，使用残差学习方法解决原始信息遗忘问题和消失/爆炸梯度问题，而且自身结构简单，参数效率高。

在使用神经嵌入方法学习知识库中实体和关系的表示中，Yang B,Yih W,He X,et al. 通过 Frebase 上的 TransE[22] 实现了百分之 73.2 的前十精度，找到了神经嵌入方法学习知识库中实体和关系新的组合公式，还发掘了一种新的方法，该方法将含义和关系表示用于挖掘逻辑规则，他们发现倾向于二元对象的嵌入能够有效的捕捉语义并且关系的构成以矩阵乘法为特征，而基于嵌入的规则提取方法在挖掘涉及组合推理的 Horn 规则时也表现出足够的优越性。

可以看出，国内外针对只是图谱嵌入式的研究已经非常丰富，大部分研究已经证实了嵌入式对于知识图谱构建已经海量数据运用的重要性，但是现有的知识图谱依旧存在着不完整性，知识图谱的嵌入中也缺失相关的链接技术，而目前的知识图谱嵌入方法也普遍复杂，训练与推理需要消耗大量时间。

1.3 研究内容以及章节介绍

本文以领域类知识图谱中的服装名称知识图谱为核心研究问题，通过基于目前主流电商平台的服装商品属性，结合本体概念和图数据库技术，以京东、拼多多等电商网站的服装商品数据作为主要数据来源，通过爬虫技术获取数据，使用 py2neo 等 Python 第三方工具库对服装名称知识图谱的本体模式层进行构建，随后使用语义插件对生成的 RDF 三元组数据完成 RDF 三元组数据到图数据结构的映射，最后将服装商品数据存储到 Neo4j 图数据库中，完成对服装名称知识图谱的构建。

本文主要有一下几个方面的研究工作：

（1）了解并掌握有关网络爬虫，Request、Selenium 等 Python 第三方库的相关知识，学习并采用 Xpath、Css.selector 等特殊元素定位方法解析 HTMl 网页结构，构建出完整的商品服装详情页商品信息爬取方法。

（2）了解知识图谱的现状及其发展趋势, 了解相关知识图谱本体构建的基本原则和方法，了解并掌握 py2neo、Neo4j 图数据库的基本原理及其应用方法，构建出

完整的服装名称知识图谱。

本篇文章主要分为以下四个章节，具体内容如下所示：

第一章是绪论，主要介绍了本文所研究课题服装名称知识图谱构建的研究背景、研究目的和意义，目前知识图谱的发展状况、国内外对于服装知识图谱的研究现状以及本课题的主要研究内容。

第二章是知识图谱的相关理论，主要对知识图谱的基础理论内容进行讲解，其中包括爬虫的基础算法、知识图谱的基础理论和 Neo4j 图数据库的基本理论。

第三章是服装名称知识图谱构建，本章首先介绍了服装名称知识图谱的构建框架设计，然后介绍了服装名称知识图谱本体模式层的设计方法，接着阐述了对于目前主流电商网站平台服装商品数据源的获取以及后续的数据处理，最后介绍了服装名称知识图谱的最终构建和存储过程，并提供了服装名称知识图谱中单商品、多商品以及商品数据查询等结果进行展示。

第四章是总结和展望，主要对本文所研究课题过程中所做的工作进行了总结并且对领域类知识图谱在服装领域上的应用进行了展望。

1.4 本章小结

本章节首先阐述了服装名称知识图谱的研究背景和研究意义，然后描述了目前知识图谱的发展状况以及国内外对于服装名称知识图谱研究的动态状况，提出了采用 crawler、py2neo、Neo4j 等相关技术来构建服装名称知识图谱，接着简单的介绍了本文的研究内容，并对每一章节所讲述的内容进行了归纳总结，以便更好的理解本文架构。

2 相关理论及关键技术

2.1 引言

知识图谱作为一种新兴的知识表示和呈现方式，已经成为了人工智能领域的研究热点之一。通过将实体、属性和关系等元素以图形化的方式进行展示和组织，知识图谱可以帮助我们更好地理解和利用海量的数据，并支持自然语言处理、机器学习、数据挖掘等多个应用场景。在本章中，我们将深入探讨知识图谱相关的理论和技术，包括知识图谱的定义、构建方法、表示方式以及常见的知识图谱推理算法等。此外，本章还会介绍构建知识图谱前知识抽取阶段所用到的爬虫以及构建知识图谱时所用到的 Neo4j 图数据库。

2.2 爬虫

2.2.1 Web 自动化工具

爬虫是一种非常实用的自动化程序，它可以在互联网上抓取数据并进行处理，如图2-1。而 Selenium 则是一种广泛使用的 Web 自动化工具，它可以模拟用户在浏览器中的各种操作，包括页面加载、表单填写和点击按钮等。相比于其他的爬虫工具，Selenium 具有更强的交互性和灵活性。Selenium 提供了一个完整的 Web 自动化解决方案，使得用户可以轻松地模拟出真实用户的操作行为。这样就使得 Selenium 可以处理大多数常见的网站，并且可以很容易地与其他 Python 库集成。由于它可以模拟用户行为，因此可以轻松地获取需要登录或者涉及验证码等复杂操作的数据。此外，Selenium 具有高度可定制性和灵活性。用户可以根据自己的需求对其进行调整，例如选择使用不同的浏览器驱动程序，或者在一次请求中同时打开多个浏览器窗口。这些选项可以让用户更好地控制 Selenium 的运行方式，并根据具体情况进行灵活调整。其中 Selenium WebDriver 是 Selenium 的主要组件，它允许开发人员编写脚本来自动化浏览器操作。WebDriver 支持多种浏览器，包括 Chrome、Firefox、Edge 和 Safari 等。WebDriver 还提供了多种定位元素的方法，如通过 ID、class name、CSS selector 和 XPath 等方式。因此 Selenium 是编写爬虫代码的较好选择之一。

2.2.2 HTML 路径语言

XPath 是一种强大的 HTML 路径语言，可在 HTML 网页结构中选择节点和属性, 如图2-2所示。它基于树结构，并提供了一种简单直观的方式来浏览 HTML 网

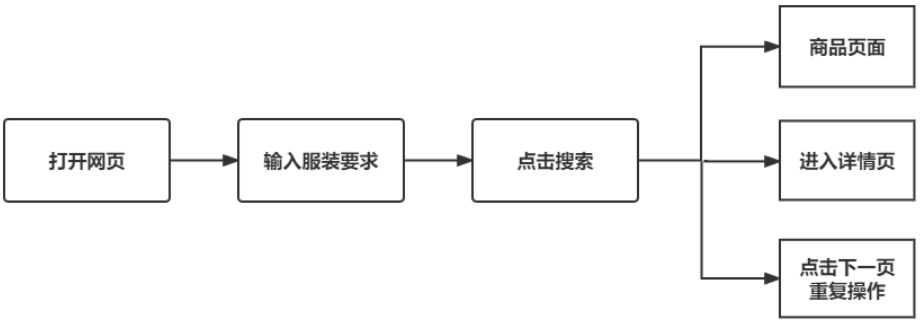

图 2-1 Web 自动爬取流程

页结构，允许用户指定路径来定位特定节点或节点组。这种功能使 XPath 成为数据提取和操作的必备工具，特别是在需要处理大量 HTML 数据的应用程序中。XPath 的一个关键优点是它与广泛的编程语言兼容，包括流行的平台，如 JavaScript、Java 和 Python。这种多功能性使它成为开发人员和数据科学家的流行选择，他们需要一种灵活高效的工具来处理 HTML 网页数据。除了选择节点和属性的能力外，XPath 还包括丰富的函数和运算符集合，可实现 HTML 网页数据的高级操作。例如，XPath 可以用于对数字数据进行计算、执行字符串操作，甚至操作日期和时间等。总的来说，XPath 是一种非常有用和广泛采用的技术，其简单性和灵活性使它成为爬虫 HTML 网页路径选择的理想工具。

图 2-2 XPath 路径定位特殊元素

CSS（层叠样式表）是一种用于描述网页内容视觉外观的强大而多功能的语言。它为 Web 开发提供了一种精确和高效的方法，用于定义 HTML 文档和其他基于 Web 的内容的演示样式，如布局、排版、颜色和视觉效果。CSS 的一个重要优点是其能够将结构和呈现分离开来，通过将内容与视觉风格分离，网站可以创建

更灵活、响应式的布局，这也使得可以在不修改底层 HTML 代码的情况下改变网站的外观和感觉。CSS 可以将 HTML 中的元素进行排版、布局和装饰等操作，使得网页更加美观、易读和易用。然而，由于 CSS 增加了网页的复杂度，给爬虫的数据获取和解析带来了一定的挑战，于是在爬虫中使用 CSS 可以帮助我们快速精准地获取 Web 页面的数据，使用 CSS 的语法来定义样式，基于选择器、属性和值。选择器用于针对特定的 HTML 元素进行定位，而属性定义了这些元素的视觉属性，例如字体大小、背景颜色或填充。值指定每个属性的具体设置，例如颜色值或测量单位，通过特殊元素定位的方法便可以获取想要的 HTML 网页数据。如图2-3所示。

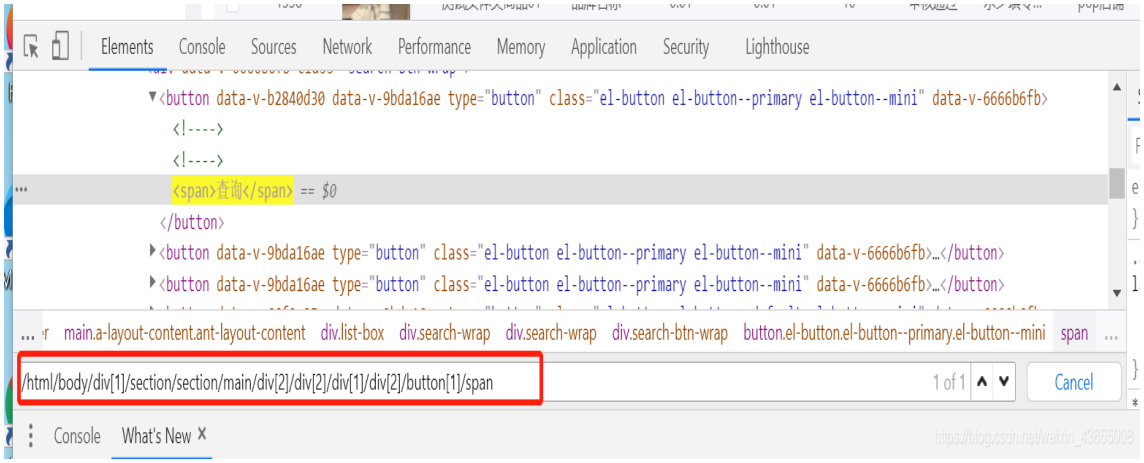

图 2-3 CSS 路径定位特殊元素

2.2.3 哈希算法

哈希算法可以用于文本去重，其基本思路是将文本内容转换为一定长度的哈希值，再根据哈希值判断两个文本是否相同。具体步骤如下：首先对每个文本进行哈希运算，得到唯一的哈希值。将哈希值存储到一个哈希表中，如果该哈希值已经存在于哈希表中，则说明该文本已经出现过，直接忽略即可。如果该哈希值不存在于哈希表中，则将该哈希值添加到哈希表中，并且保存该文本以备后续比较。MD5 就是一种常用的哈希函数，可以将任意长度的消息映射为一个 128 位的数字指纹，使用 MD5 函数的编程实现可以参考各种编程语言提供的相关库或者函数。在 Python 中可以使用 hashlib 模块中的 md5() 函数对字符串进行加密已达到对文本数据进行二进制转换的目的，最后通过比对二进制数码的相似度达到去重的目的。

2.3 知识图谱

2.3.1 知识图谱构建流程

知识图谱是一种将现实世界中的事物通过它们之间的关系连接起来形成网状结构的技术。随着信息技术的发展，Google 于 2012 年首次应用这一概念 [23] 使得我们可以更加方便地获取、传递和共享知识。知识图谱作为一种新兴的信息组织方式，通过建立实体间的关系，将大量的结构化和非结构化数据转化为可理解、可查询的形式。同时，搜索引擎的出现也使得我们能够快速地获得我们所需的信息，提高了效率和用户体验。网络信息技术的不断进步，加上人工智能技术的不断革新，让我们在处理大规模数据时，能够更好地挖掘相互之间的关系，进行知识查询和文本分析。开放链接数据则提供了更加丰富、多样的数据资源，让不同领域的数据产生联系和相互影响。最终，所有这些技术都是为了更好地描述和探索客观存在的事物，帮助我们更好地理解世界。

知识图谱采用图结构来描述现实世界中的客观事物及其关系。它由三元组数据：实体，实体的属性，实体与实体、实体与属性的关系组成，实体与实体之间或者实体与属性通过相互之间的关系进行连接。知识图谱的知识表示形式有以下两种，即：实体——关系——实体（例如：苹果和香蕉是水果、苹果是苹果树的果实）；实体——关系——属性（例如：苹果是红色的，衣服的价格是 89 元），其中实体是知识图谱的核心概念，是对同类事物所做的抽象描述；关系则指实体之间的所存在的联系，每个实体可能与其他实体存在一种或多种关系；而属性则用于描述实体所拥有的特性、特征或意义。如图2-4所示。

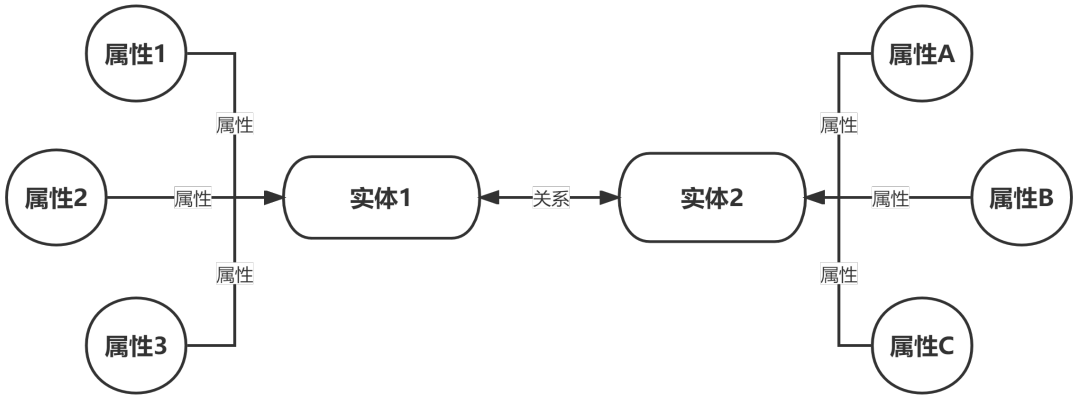

图 2-4 实体关系

知识图谱采用两个层次结构，即数据层和本体模式层，来实现知识的存储和描述。数据层是从各种数据源中提取的半结构化数据，通过三元组来描述现实世

界中存在的事实，并使用 Neo4 作为主流图数据库进行存储。同时，本体模式层作为知识图谱的核心框架，对数据层进行抽象和概括总结，以描述和限制实体、关系和属性。在本体模式层中，知识被统一标准化，并按照特定规则进行定义和制约，从而保证知识的一致性和完整性。

知识图谱构建方法通常包括自顶向下和自底向上两种。在自顶向下的构建方法中，需要首先依据应用领域内权威专家的指导创建本体模式层，然后结合本体模式层对数据进行实例化与对应，从而构建出知识图谱，这种方法一般以结构化数据为基础 [24]。例如，Freebase 采用的就是自顶向下的构建方式，其主要数据源来自 Wikipedia[25]，适用于构建垂直领域类知识图谱。但垂直领域知识图谱构建全程需要领域专家参与，因此成本较高。自底向上的构建方法适用于构建通用类型知识图谱，它的过程是首先提取相关知识数据，然后从中选取置信度高的知识添加至知识图谱中。接着，根据知识库中的知识总结抽象实体概念，并构建本体模式层。由于通用类型知识图谱包含的数据种类繁多、数据来源广泛，因此构建本体模式层是必要的。相较于自顶向下的方法，自底向上构建方法对专家参与的要求较低，被广泛应用于工业界和学术界。知识图谱的构建流程可以大致分为六个部分：知识抽取、知识融合、本体模式层构建、知识推理、质量评估以及知识存储 [26]。这些步骤均需要仔细处理，以保障知识图谱质量和准确性。构建知识图谱的流程如图2-5所示。

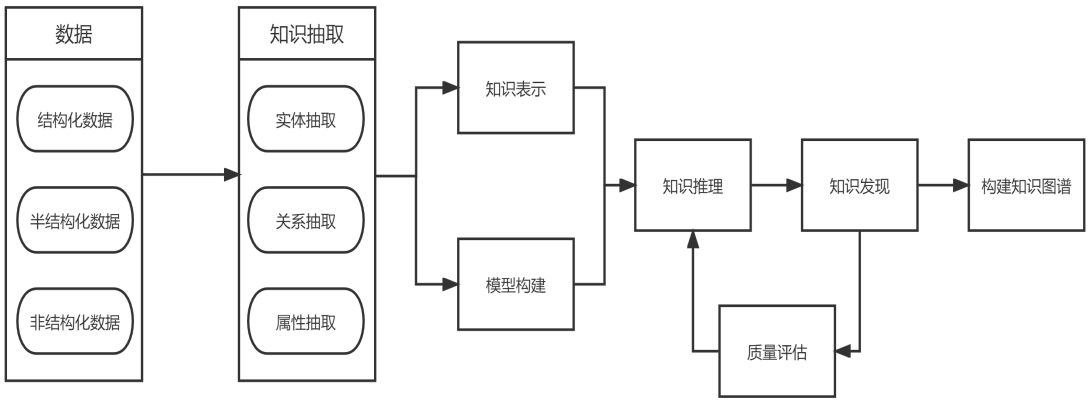

图 2-5 知识图谱构建流程

2.3.2 知识抽取

知识抽取是指从非结构化或半结构化文本数据中提取出有用的信息和知识，并将其转化为结构化数据的过程。这些结构化数据可以被用于构建知识图谱、搜索引擎、数据挖掘等应用。知识抽取通常包括三个主要步骤：命名实体识别、关

系抽取和事件抽取。其中，命名实体识别是指从文本中识别出具有特定意义的实体，如人名、地名、组织机构名等；关系抽取是指从文本中识别出实体之间存在的语义关系，如“X 是 Y 的创始人”、“X 属于 Y”等；事件抽取则是指从文本中识别出实体之间发生的事件，如“X 与 Y 签署了一项协议”。自然语言处理技术在知识抽取中扮演着重要的角色，如词性标注、句法分析、语义角色标注等。同时，机器学习算法也可以被用于训练模型，从而提高知识抽取的准确率。知识抽取相关技术如图2-6所示。

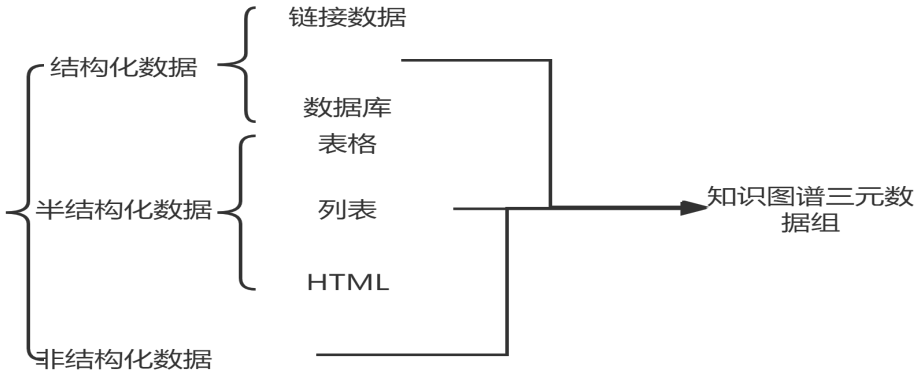

图 2-6 知识抽取相关技术

2.3.3 知识融合

知识融合指的是将来自不同数据源的同一知识的多种表达形式进行统一表示，以实现对实体的更全面描述。不同知识库对同一实体的描述方向可能存在差异，有些更注重某些单一属性或特性，而有些则更关注实体之间的关系描述。为应对这种情况，知识融合可通过实体对齐等方法将来自多个不同知识库中相关知识完全汇总到同一个知识库中，从而实现对实体信息的更全面描述。例如，在西游记中，齐天大圣、美猴王和孙悟空都是相同实体在不同数据库中的不同表示形式。因此，知识融合可以补充并整合不同来源的同一实体知识，形成更加全面和准确的实体描述。知识融合是非常重要的任务，它可以提高知识图谱中实体描述的全面性和准确性，同时也可增强知识图谱的可用性和应用效果。

2.3.4 本体模式层构建

知识图谱由本体模式层和数据层两部分组成，其中本体模式层是知识图谱的核心，用于具体描述知识图谱中实体、属性和关系等的抽象概念。在建立本体模式层时，通常需要参考领域规范或标准，并采用一些基本的参考模型和规范。这

些标准不仅包含对数据的要求，也将其纳入到本体模式层中，以形成初始的基础本体模式层。在不断获取新的数据并更新知识图谱的过程中，本体模式层也需要根据新数据的特点进行适时调整和更新。同时，在行业相关的知识图谱中，可以复用现有的本体概念，以提高本体模式层的效率和精确度。因此，本体模式层是知识图谱中非常重要的一部分，它为知识图谱提供了基础框架和参考标准，保证了知识图谱中实体、属性和关系等的一致性和可拓展性。同时，本体模式层也为知识图谱的应用提供了坚实的基础。

2.3.5 质量评估

构建知识图谱需要进行质量评估，以保证所获取的知识符合知识图谱的要求，并将满足要求的知识添加到知识图谱中，不同类型的知识图谱会采用不同的质量评估方式。例如，在通用类型知识图谱中，常采用基于众包法的方式获取知识。多人参与标注某个知识点时，如果该知识点有唯一答案，则可以使用投票方法，选择得票数最高的标注结果作为该知识点的答案。而对于垂直领域中的专业知识数据，需要借助领域权威专家的指导来获得符合领域要求的数据。例如，在医学领域的知识图谱构建中，需要依赖医学专家的经验和知识进行数据标注和质量评估，确保知识图谱中的医疗数据准确可靠。总之，质量评估在知识图谱构建中具有重要作用，它确保了知识图谱中知识的准确性和完整性，并为知识检索、推荐等应用提供了可靠的基础。

2.3.6 知识推理

知识推理是指基于已有的本体模式层和数据层，利用叙述逻辑系统软件对实体、关系及图谱结构的信息进行组合和分析，从而获得新的知识，增强知识图谱的完整性。叙述逻辑是一种简单的语言表达形式，由基础元素、定义、相互联系和个体四个基础组成部分构成。例如，当一个实体包含在另一个实体中，而后者又被包含在第三个实体中时，第一个实体也被认为被包含在第三个实体中。通过推理所得到的新知识可以更新和完善知识图谱，提高其涵盖范围和精确性。因此，知识推理是一种有效的方法，能够大幅度拓展知识图谱的内容，并使其更加完备和准确。

2.4 Neo4j 图数据库

随着现代社会的快速发展，各行业之间形成了一个庞大且复杂的相互关联网络。在这种情况下，新兴技术如 5G、人工智能和大数据开始广泛应用于许多领域，并且带来了极高的商业价值。因此，需要将应用方向转向图数据，一种通过相关

关系将实体进行关联的数据模式，以建立跨越多个领域的模型和分析数据。

在这种情况下图数据库应运而生。目前，市场上有很多类型的图数据库，例如Neo4j、Arango DB 和 Fauna。选择使用 Neo4j 图数据库来存储服装知识图谱，用户可以通过图形可视化方式查询或添加服装信息要素。Neo4j 图数据库支持多种数据导入方式，包括关系型数据库导入、CSV 文件自动导入和 Cypher 语言导入，这使得抽取的三元组能够以简洁明了的方式展现节点之间的相互关系和实体属性。

与传统的关系型数据库相比，图数据库更能直接、真实地表达数据之间的相关性。它能够高效地查询关联数据，易于数据建模，并提供更加自然和直接的形式来存储知识图谱。因此，使用图数据库来存储知识图谱是一种更加自然和高效的做法，可以帮助我们更好地理解数据之间的关系。

2.5 本章小结

本章首先对知识图谱做了一个简要的阐述，然后重点介绍了知识图谱构建的基本流程，知识图谱的构建流程可以大致分为六个部分：知识抽取、知识融合、本体模式层构建、知识推理、质量评估以及知识存储，从这六方面入手进行详细介绍。最后介绍了在构建知识图谱中所需要想要用到两项关键技术：爬虫技术和Neo4j 图数据库。

3 服装名称知识图谱构建

3.1 引言

本章的核心在于探讨如何设计和构建一个服装名称知识图谱，包括了服装名称知识图谱基础理念的总结，详细介绍了服装名称知识图谱的构建框架设计，构建本体模式层以及获取服装数据的方法和技术，最终总结了服装名称知识图谱的存储方式并对不同类型、不同数量商品的知识图谱进行了结果展示。

3.2 服装名称知识图谱构建框架设计

构建服装行业的本体模式层，首先需要对服装行业的知识结构、知识分类等进行足够的认识、学习和分析。这个过程需要结合领域权威专家的理论指导，以确保本体模式层的完整性和正确性。在本体模式层的构建过程中，可以采用现有的本体库或者自定义本体库。如果采用现有的本体库，则需要根据行业特点进行本体库的筛选和适配。如果需要自定义本体库，则需要对行业的概念体系和分类进行深入研究，以确保本体库的准确性和完整性。同时，在本体模式层的构建过程中，还需要考虑到数据层的构建。数据层是指具体的实例数据，包括产品信息、企业信息、消费者信息等。在构建数据层时，需要根据本体模式层的体系结构进行分类和整理，并将其与本体模式层进行关联。最终，通过本体模式层和数据层的结合，可以构建出一个完整的服装行业知识图谱，以支持相关领域的应用和研究。服装名称知识图谱的构建流程如图3-1所示

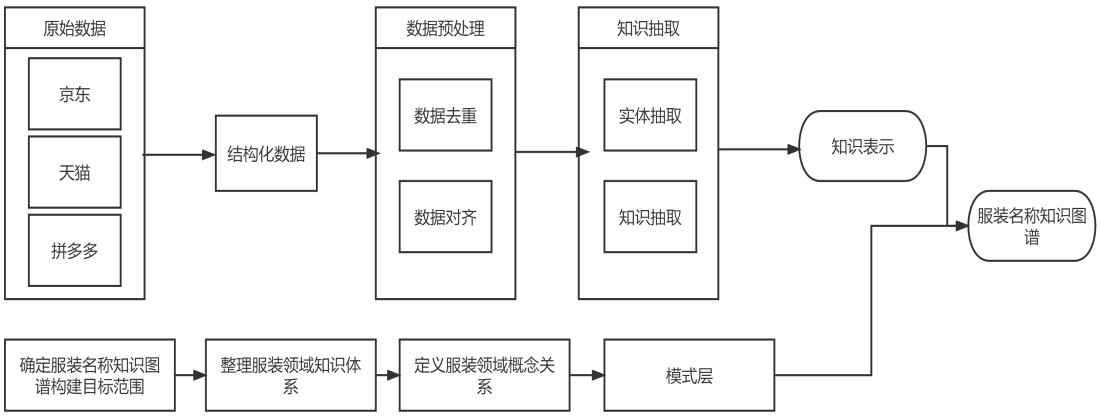

图 3-1 服装名称知识图谱构建流程

根据图 3.1 所示，构建服装名称知识图谱的过程可以分为两个主要步骤：服装

本体模式层的构建和服装数据的获取。在进行模式层构建时，首先需要明确服装知识图谱的应用目标和范围，并对服装领域中的相关概念、关系和属性等信息进行整理和分析。这些信息将被存储到后续用于构建服装名称知识图谱本体模式层的数据库中。在数据层的获取上，主要是根据已经构建好的服装名称知识本体模式层，从目前主流的电商网站获取相对应的服装商品数据，包括实体、关系和属性等信息，然后将这些商品数据、商品信息存入数据库中。最终，将数据层和模式层进行映射存储到 Neo4j 图数据库中用于完成服装名称知识图谱的构建。

3.3 服装知识图谱模式层设计

在知识表示和语义网络中，我们可以使用本体模式层来建立概念。每个概念都有一组相关属性，这些属性描述了该概念的特征和特性。概念之间存在不同的关系，如包含、相似、对立等。此外，概念与属性之间也存在着关联和联系。在服装领域，我们可以将不同的服装品类和类型作为概念进行建模，并且添加各种属性来描述这些服装的特征和特性，例如颜色、尺寸、材质等。此外，我们还可以利用概念之间的关系来建立更高级别的概念，例如“夏季服装”、“正式服装”等。通过这样的方式，我们可以构建一个完整的知识结构，以帮助人们更好地理解和应用服装领域的知识。总的来说，在知识表示和语义网络中，我们可以利用本体模式层、概念、概念属性、概念之间关系、概念与属性之间关系等关键词来构建一个完整的知识体系。在服装领域，我们可以利用这些工具和方法来帮助人们更好地理解和应用相关的知识。

3.3.1 模式层建模方法介绍

经典的本体模式层构建方法有七步法 [27]、循环获取法 [28] 等。本体模式层构建的七步法是一种用于创建本体模式的方法，它包括以下七个步骤：1、确定范围：明确你要构建本体模式的范围和目标，以便你能够专注于重点领域。2、收集信息：收集与本体模式相关的信息，包括相关文献、专家意见、现有本体和其他知识资源。3、确定类别和属性：根据收集到的信息，确定需要构建的类别和属性，并定义它们之间的关系。4、创建本体模式：使用本体编辑器创建本体模式，包括类别、属性、关系和实例。5、审查和测试：审查和测试本体模式，以确保其符合预期并满足需求。6、集成和扩展：将本体模式集成到你的应用程序中，并在需要时扩展它以适应新的需求。7、维护和更新：定期检查和更新本体模式，以反映变化的业务需求和知识。在使用本体模式层构建的七步法时，重要的是要遵循每个步骤，并仔细记录进展。这将帮助确保能够有效地构建本体模式，并最终达到目标。

在服装名称知识图谱的本体模式层构建的过程中，本文在“七步法”的构建基础上进行了改进来达到构建服装名称知识图谱本体模式层的目的，具体的流程图如图3-2所示。

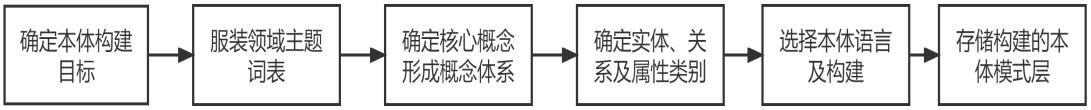

图 3-2 本体模式层构建“六步法”构建流程图

3.3.2 服装商品本体范围确定

在构建本体模式层时，首要步骤是确定服装本体的应用领域。本文的重点在于探讨如何构建适用于电商网站上的服装商品本体，因此，本体知识的范围和领域主要关注于电商网站上的服装商品。

3.3.3 服装商品核心类及属性定义

目前，尚未出现具有公开特定本体模式的服装商品领域描述，但已有相对完整的元素描述。为构建服装主题词表，参考了专业书籍《服装形象设计》[29]，并结合京东商城网站和国家标准 GB/T15557-2008(服装术语) 进行补充。根据该词表与本文所要构建的本体模式层自身需求确定了最终核心概念，进而确定了七种核心类别——商品、店铺、品牌、价格、材质、面料和类型。在确认这些核心类别后，对它们的名称、类别之间的关系以及属性进行了标准化。如图3-3所示。

核心类名	核心类描述	实体示例
商品	服装商品名称	红蜻蜓短袖t恤女夏季新款韩版洋气时尚百搭显瘦女装夏装打底圆领上衣衫
店铺	销售店铺	恒源祥服饰官方旗舰店
品牌	服装品牌	恒源祥
价格	销售价格	298.00
材质	面料使用的材质	羊毛
面料	服装使用的面料	纯羊毛
类型	服装类型	针织衫

图 3-3 本体模式层核心类

为了创建服装领域本体类，我们需要定义该类特有的属性和描述信息。这些属性可以分为两种类型：对象属性和数据属性。对象属性用于表示不同本体类之间的关联关系，其定义域和值域都是本体类；而数据属性则用于表示本体类固有的属性特征，其定义域是具体的本体类，而值域则是该本体类的固有数据类型。在使用 Python 工具库 Py2neo 时，可以通过 Node Graph 功能来构建服装商品的对象属性。这些属性根据不同类别之间的关系进行定义。例如，关系：商品–面料的定义域为商品，值域为面料，表示这个商品由某种面料制成。样例如图3-4展示了对象属性的设置。同时，可以添加数据属性以描述不同服装的属性特征。例如，如图3-5显示了基于获取到的男装服装数据所设置的不同数据属性。

为了构建本体模型的架构并定义相关的类和属性，我们需要规定一些规则。这些规则将用于从 xls 文件中导入有关商品、店铺、品牌、价格、材质、面料和类型的实例数据，以形成领域内知识组织架构。因此，我们将建立一个完整的知识管理系统，以支持对这些领域内实体的有效管理和查询。通过这种方式，我们可以实现更加高效和准确的知识表达，提高知识的可重用性和可扩展性，从而为实现人工智能技术在该领域的应用奠定基础。

实体A	实体B	实体关系
商品	店铺	商品___店铺
商品	品牌	商品___品牌
商品	价格	商品___价格
商品	材质	商品___材质
商品	面料	商品___面料
商品	类型	商品___类型
商品	商品	商品___商品

图 3-4 实体关系

性别	类型	属性
男士	T 恤	适用人群、流行元素、图案、风格、领型
	夹克	适用季节、领型、口袋设计、衣门襟、流行元素、下摆设计、适用人群、风格、图案
	卫衣	风格、适用季节、领型、适用人群、流行元素、图案
	衬衫	适用季节、领型、适用人群、流行元素、风格、图案、适用场景
	针织衫	衣门襟、领型、适用人群、穿着方式、风格、图案
	风衣	衣门襟、适用季节、适用人群、衣长、风格、领型、款式、口袋设计、下摆设计
	休闲裤	适用季节、裤门襟、腰型、裤型、适用人群、流行元素、裤长、风格、图案、弹力

图 3-5 服装属性设置（男装为例）

3.4 服装信息抽取

3.4.1 数据获取

为了实现对服装商品数据的结构化采集，本文使用了基于 Selenium 框架的爬虫技术。具体而言，我们主要针对互联网上大型电商网站（例如京东、天猫等）展开爬取工作。首先，我们通过关键字查询获取商品列表页，并从中提取与关键字相关的商品信息。随后，我们自动访问每个商品对应的详情页，以获取其详细描述信息（如图3-6所示）。通过以上方式，我们成功地收集到了大量有关服装商品的数据，以供进一步分析研究使用。

图 3-6 京东服装商品详情页

为了爬取网站上相应数据信息，我们采用 Python 语言，并结合了 Python 的第三方工具库 Request、BeautifulSoup 和 Selenium 来实现自动爬取数据的功能。

Request 库是一种基于 Urllib 创建的简易 Http 库，它可以方便地发起 Http 请求。而 Beautiful Soup 则是一个第三方 Python 库，它可以从半结构化数据如 HTML 或 XML 文件中提取所需要的数据。我们还使用 Selenium 工具库来模仿网上现实用户在电商平台上浏览商品的操作，以此来找到想要的的服装商品信息，获取并记录所找到的每个商品的详情页的 Url。接着，我们对这些网页的 HTML 格式进行解析，使用基于 XPath 和 CSS 的路径选择方法定位特殊元素来匹配网页标签，获取我们所需要的数据，并将其存储为 CSV 文件。整个流程可见图3-7。

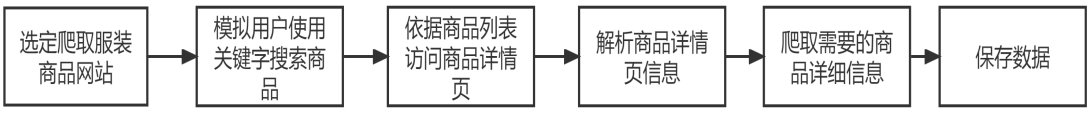

图 3-7 数据获取流程

3.4.2 数据清洗

数据清洗是确保数据准确性的重要步骤，它通过检查和校验数据来解决重复、错乱等问题。首先，需要对爬取的数据进行分析，并定义错误类型以及相应的清理和转换规则。接着，搜索错误记录并修改与之关联的具体错误，最终将经过处理后的数据回流以替换原始数据，从而提高数据质量。本文针对商品、店铺、品牌、价格、材质、面料和类型这七类实体数据进行了数据清洗。通过哈希算法对重复数据进行去重对齐处理，同时进行数据对齐和错误数据纠正。例如，将“罗蒙女装 T 恤”和“罗蒙女款半长袖 T 恤”视为同一实体进行统一处理，使其成为“罗蒙 T 恤”。最终，我们成功完成了数据清洗，输入结果如图3-8所示。

商品	价格	店铺	详情页	品牌	详情	材质	面料	类型
黛维亚（DE	149	黛维亚（DE	https://it	品牌：黛	商品名称：黛维亚（DEVEIA）纯棉白色短袖衬衫女2023春季新款			
真维斯女装	39.9	真维斯官方	https://it	品牌：真	商品名称：真维斯女装 2023夏季新款 休闲冰丝阔腿裤 JRP B款			
浪莎冰丝运	59.9	浪莎女装京	https://it	品牌：浪	商品名称：浪莎LSSQ-A059-876商品编号：100010925389商品毛重			
QGF短袖t恤	48	QGF女装旗	https://it	品牌：QGF	商品名称：QGF短袖t恤女美式复古oversize潮牌夏季新款宽松显			
红蜻蜓 短	89	红蜻蜓女装	https://it	品牌：红	商品名称：红蜻蜓 短袖t恤女早春夏季新款T恤女装夏装冰丝打底			
雅黛丝薇	168	雅黛丝薇服	https://it	品牌：雅	商品名称：雅黛丝薇 春装2023年女装新款春秋法式小香风连衣裙			
ZGWT轻奢高	318	ZGWT旗舰店	https://it	品牌：ZGW	商品名称：ZGWT轻奢高档品牌白色短袖T恤女宽松夏季时尚小个子			
宏资琦新款	136	宏姿琦服饰	https://it	品牌：宏	商品名称：宏资琦新款连衣裙2023夏季女装法式气质收腰显瘦吊			
爱心东东妙	149	众为女装专	https://it	品牌：妙	商品名称：妙漫物语（Miaomanwuyu）女装2023年新款连衣裙子夏			
俞兆林女装	48	俞兆林女服	https://it	品牌：俞	商品名称：俞兆林T恤商品编号：100010966680商品毛重：200.0			
红蜻蜓 纯	89	红蜻蜓服饰	https://it	品牌：红	商品名称：红蜻蜓 纯棉短袖t恤女宽松正肩夏季女装2023年新款			
森马短袖t	49.99	森马京东自	https://it	品牌：森	商品名称：森马109321100020商品编号：100011540565商品毛重			
莉社（LISH	99	莉社官方旗	https://it	品牌：莉	商品名称：莉社（LISHE）冰丝阔腿裤女高腰垂感2023夏季新款宽			
娇茹妮100%	39	娇茹妮服饰	https://it	品牌：娇	商品名称：娇茹妮100%纯棉美式国潮宽松短袖T恤女ins2023年超			

图 3-8 女装服装数据示例

3.5 服装知识图谱的构建与存储

本文使用 Neo4j 图数据库作为存储工具，采用 Java 语言实现。该数据库基于相互连接的图来存储数据，与传统关系数据库相比，具有更强的分析处理能力和清晰的可视化效果。此外，采用基于 Python 语言 py2neo 工具库中 NodeMatcher 的检索方式，查询效率远胜于传统的基于 OWL 文档的检索和存储方式。因此，在服装知识图谱的存储上，本文将主要采用 Neo4j 图数据库。

3.5.1 映射匹配分析与本体知识存储

本文利用 Python 的 Py2neo 工具库构建服装领域的本体模式层，并通过实例化数据生成 RDF 格式数据，以 < 主语，谓语，宾语 > 的三元组形式存储。为了将本体模式层和数据层与本体概念和实例之间进行映射，我们使用 Neo4j 图数据库来存储 RDF 格式数据。在 Neo4j 图数据库中，主要包括节点、关系、属性、标签和路径等五类要素。节点和关系都可以拥有多种属性元素，这些属性元素可以通过键值对表示并独立存储。我们将服装 RDF 格式数据中的关系、实体对象以及实体对象的特征属性与 Neo4j 图数据库中的节点、关系和特征属性进行匹配映射来实现存储。节点与本体模式层中的类名和实例名进行匹配映射，例如“商品”与“女装格子衫”、“面料”与“棉”、“材质”与“纯棉”等；关系则与本体模式层中的对象属性匹配，如“商品–价格”、“商品–材质”等；特征属性则与本体中的数据属性相匹配，例如“商品”本体类的数据属性一般包含“品牌”、“面料”、“材质”等，“类型”属性包括“款式”、“风格”等。通过以上匹配映射，成功实现了将服装知识图谱存储到 Neo4j 图数据库中。

将 RDF 格式数据存入 Neo4j 图数据库后，我们可以对通过 Python 编译环境下的 Vscode 对各实体类型和关系进行不同的调整，例如调整节点的数量和想要构建的关系或者调整实体节点显示的属性等等，以提高可视化显著性。

3.5.2 结果展示和查询

利用 Neo4j 图数据库完成服装知识图谱的知识存储后, 可以使用 Neo4j 图数据库提供 Cypher 语言或者 Python 的第三方工具库 Py2neo 对数据库中的数据完成增（CREATE）、删（DELETE）、改（SET）、查（MATCH）以及推理，轻松实现对服装商品、品牌、类型材质、面料等服装实体、服装实体属性以及节点之间的关联关系的检索和遍历。生成的知识图谱用不同的颜色区分不同实体类型的节点，节点之间的连线代表实体之间的关系。

3.5.2.1 服装名称知识图谱商品属性关系展示

1）单件服装商品属性

在单件服装商品属性图中核心类“商品”将被作为实体 A，其他核心类例如“价格、材质、面料”等其他 6 个核心类将被作为实体间的关系，而他们所对应的属性将被作为其他的实体，通过与实体 A 的关系连接起来。

如图3-9所示是女装的单件商品属性知识图谱

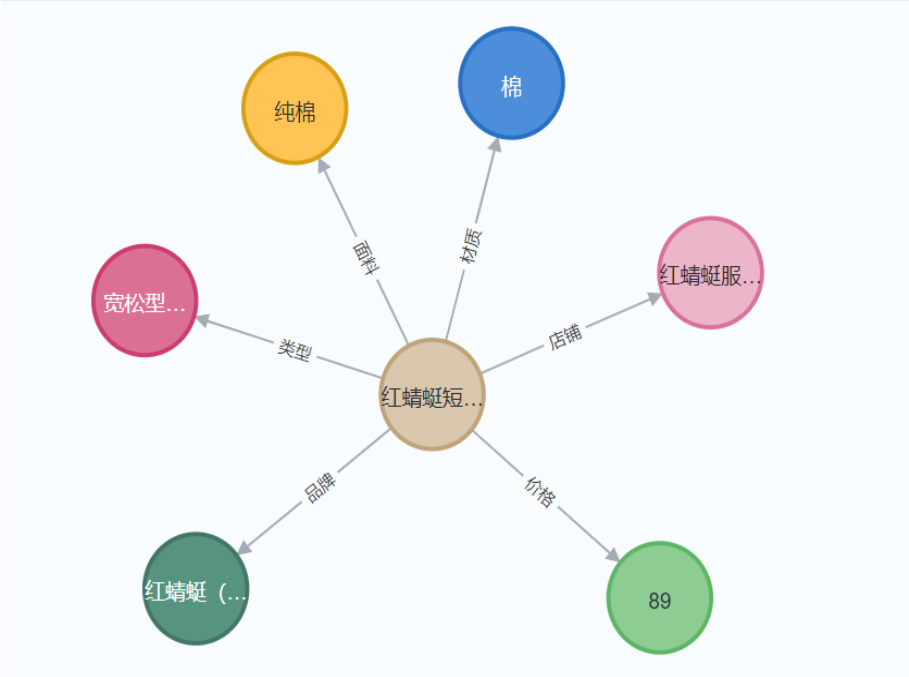

图 3-9 红蜻蜓短袖 t 恤女装夏装打底圆领上衣衫（女装）

2）多件服装商品属性

在单件服装商品属性图的基础上寻找多个单件服装商品属性图的共同实体或者共同关系进行连接从而形成多件服装商品属性的知识图谱。

如图3-10所示为价格都是 89 元的红蜻蜓品牌女装服装

3）结果查询

通过运用 Python 语言 py2neo 工具库中 NodeMatcher 查询语句对构建的服装知识图谱进行查询，例如查询与“红蜻蜓品牌女装服装”相关的服装内容。查询结果如图3-11所示，其中包含所有属于“红蜻蜓品牌女装服装”类型的服装实体节点。通过单击任意一个实体节点，可以呈现该服装的详细信息，如图3-12。此外，双击该节点后可以展示路径为 1 的所有实体节点，使得 Neo4j 图数据库的操作更加简便。这样，我们可以充分利用 Neo4j 图数据库强大的功能来对知识图谱进行查询和分析。

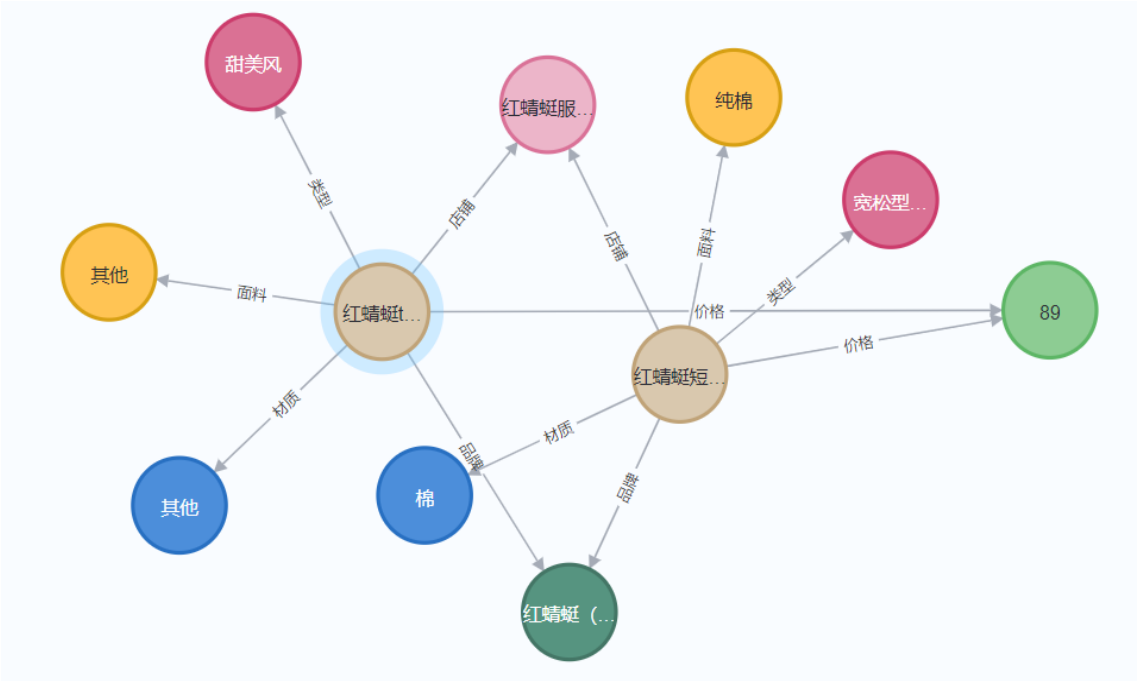

图 3-10 红蜻蜓品牌女装服装

综上所述，本章重点介绍了使用工具 Py2neo 和 Neo4j 图数据库将本体模式层和数据层相结合构建服装名称知识图谱。所构建的知识图谱应用于服装研究，提高了服装数据的有效利用效率，并从商品、品牌、店铺、面料、价格、材质和类型等多个角度进行全方位的探索和多维度分析。利用 Neo4j 图数据库的灵活性、简单可靠性、可扩展性和高可用性特点，通过 Cypher 语言实现对服装数据的有效而准确的搜索和推理，同时避免了使用关系型数据库时低效率和缺陷问题。此方法的查询效率高于基于本体的查询，推理准确性也得到了改善。因此，本研究探讨了将 Neo4j 图数据库应用于构建服装知识图谱的好处，为服装领域的进一步发展和研究提供了更多选择。

3.6 本章小结

本章介绍了一种基于电商网站数据的服装知识图谱构建方法，包括数据爬取、信息抽取和知识存储等多个环节。在知识存储阶段，本文则采用 Neo4j 图数据库存储服装知识图谱。通过对爬取数据中的产地、品牌、店铺、服装、材质、面料、类型等实体进行抽取，并在本体模式层构建之上，使用图数据库对服装知识进行存储，最终完成了服装知识图谱的构建。利用该知识图谱，可以方便地对各种服装数据进行查询检索和推理，以便更好地利用服装资源。

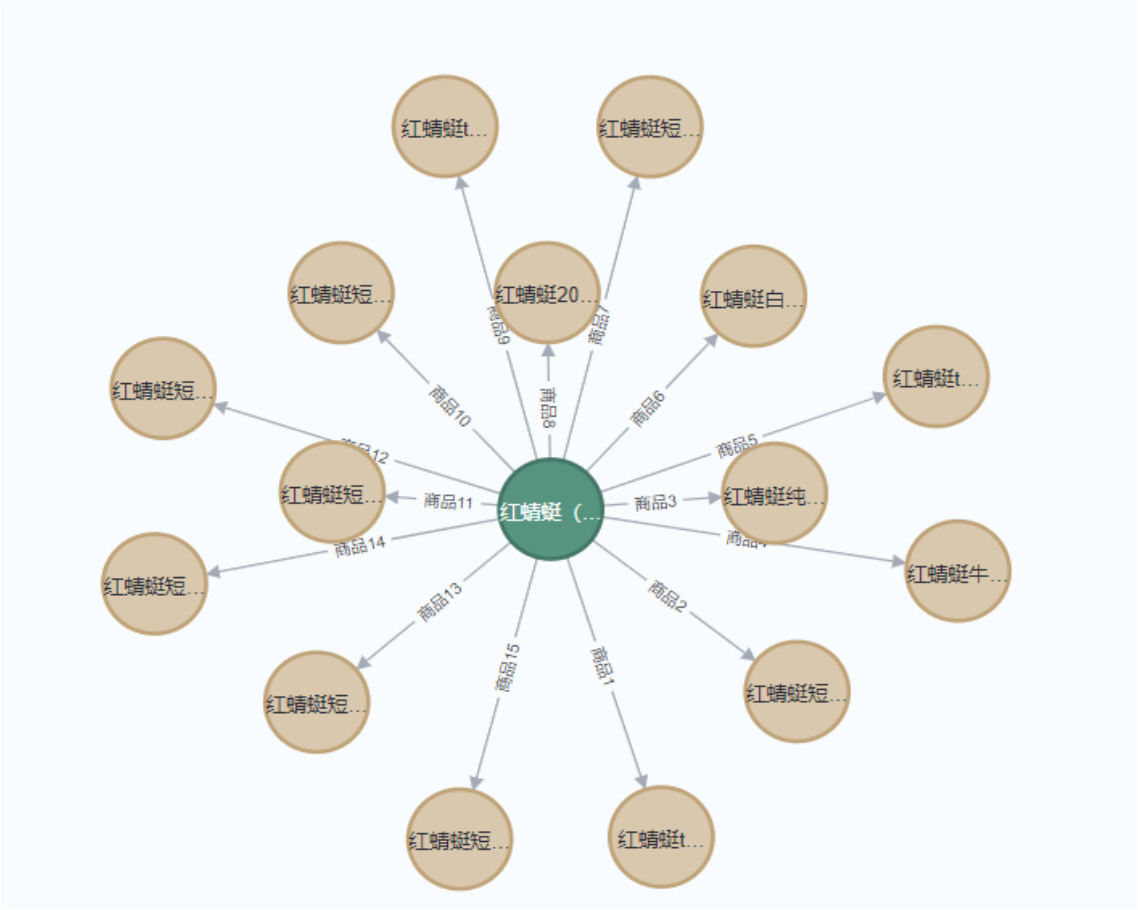

图 3-11 红蜻蜓品牌女装服装

n
(:商品 {name: "红蜻蜓短袖t恤女夏季新款韩版洋气时尚百搭显瘦女装夏装打底圆领上衣衫"})
(:商品 {name: "红蜻蜓t恤女短袖2023年夏季新款女装修身打底格子百搭女士上衣服衫"})
(:商品 {name: "红蜻蜓短袖t恤女早春夏季新款t恤女装夏装冰丝打底时尚百搭衫上衣ins潮"})
(:商品 {name: "红蜻蜓纯棉短袖t恤女宽松正肩夏季女装2023年新款韩版半袖体恤夏装上衣"})

图 3-12 红蜻蜓品牌女装服装

4 总结与展望

4.1 总结

随着人工智能技术的不断发展，知识图谱也越来越成熟，并且在人机交互、智能安防、情绪分析、娱乐活动、智能医疗和在线教育等众多领域得到了广泛应用。然而，由于大量需求的出现，人们对知识图谱的准确性要求越来越高。因此，该技术成为知识图谱领域中一个重要的研究课题。为了解决属性、实体、数据等条件带来的实体关系以及数据准确性参差，许多研究者通过构建领域类知识图谱提出了许多提高知识图谱数据关系精度的方法。尽管这些新方法效果不错，但领域类知识图谱应用过少与发展时间较短导致各类领域的知识图谱无法很好地被应用和发挥其价值。因此，如何快速准确的实时更新知识图谱的数据库以及调整实体关系并且对不同用户提供更加个性化的数据关系仍然是当前研究的热点。本文针对服装领域的服装名称知识图谱的构建进行了以下研究工作：

（1）根据知识图谱的发展趋势与本次课题的研究要求，了解了有关知识图谱的基础知识，本体模式层构建的概况以及经典的本体构建方法，之后提出了本文所采用的方法。

（2）完成了网络爬虫的算法编写。基于 Request、Selenium 等 Python 第三方库，结合 XPath、CSS 等路径选择方法，面向不同电商网站的多种不同类型服装数据通过特殊元素定位等方法完成了数据的获取。

（3）对知识图谱的构建方法进行了分析与研究，完成了服装名称知识图谱的构建。采用 Python 的工具库 py2neo，结合本体概念和图数据库技术，面向多种不同类型服装数据通过知识抽取、知识存储、实体链接等方法完成了服装知识图谱构建。

4.2 展望

本论文对服装名称知识图谱构建的研究进行了取得了一定的成果并且能够进行可视喉展示，但在日后的工作中还存在不少需要改进的地方，具体如下：

（1）本论文再进行服装名称知识图谱构建时所采取的数据库尽管符合要求，但还是存在一些数据量不够、代表性数据不足、数据不够全面，或者服装属性类别分类不均匀、不合理和实体关系不明确的情况，导致出现了某几种品牌的服装没有在图谱数据库中出现，知识图谱无法搜索到的问题，在之后的工作中需要注意爬取数据时电商网站的可代表性以及所要爬取品牌在服装品牌上的知名度与涉

猎范围，减少由于爬取到服装属性对图谱呈现结果造成的影响。

（2）本次论文所构建的服装名称知识图谱只适用于静态数据的检索，对于实时更新的服装商品数据来讲，暂时无法实现，后期将会尝试对于已有的服装商品数据进行处理，使得用户在使用服装名称知识图谱时能够得到更加个性化的数据。

（3）目前在知识图谱补全方面的研究大多集中于知识图谱的嵌入式应用，知识图谱嵌入技术的优势在于，它可以通过学习知识图谱中的语义信息来解决知识图谱中存在的缺失问题，并可以提高各种机器学习模型的性能，是以后服装名称知识图谱继续优化的研究重点。

致谢

在短暂而充实的四年大学生涯中，我深刻地感受到了人生的无限可能和珍贵。尽管经历了疫情的考验，但我依然收获颇丰。这一切离不开杭州电子科技大学自动化学院老师们的倾囊相授，让我打开了自动化领域的大门，真正领略了自动化为社会和家庭带来的变革。我的专业是自动化，作为 2019 级的一员，我在同学们的相互竞争和互相帮助鼓励中感受到了宛如家庭般的温暖和自豪。最重要的是，我要感谢我的导师韩志敏老师，她为我提供了巨大的帮助和指导，在研究课题的道路上给予了我很多启示和支持。同时，也要感谢我的室友们和亲人们的陪伴和支持，还有杭州电子科技大学自动化学院各位老师的辛勤付出和谆谆教诲。最后，再次感谢所有关心、鼓励和支持我的同学和朋友们，是你们让我在这三年中不断成长和进步。

参考文献

[1] 中华人民共和国商务部电子商务和信息化司. 中国电子商务报告 (2021)[M]. 北京 : 中国商务出版社, 2022 : 2 – 7.

[2] 智研咨询. 《2020-2026 年中国服装纺织电子商务行业市场需求潜力及战略咨询研究报告》[M]. 北京 : 中国产业信息网, 2020 : 546 – 579.

[3] Dong X, Gabrilovich E, Heitz G. Knowledge vault:a web-scale approach to probabilistic knowledge fusion[M]. New York : ACM, 2014.

[4] Lehmann J, Isele R, Jakob M. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia[J]. Semantic Web, 2015, 6(2) : 167 – 195.

[5] Jin H, Li C, Zhang J. XLORE2:large-scale cross-lingual knowledge graph construction and application[J]. Data Intelligence, 2019, 1(1) : 77 – 98.

[6] Xing N, Sun X, Wang H. Zhishi.me[M]. Berlin : Springer Berlin Heidelberg, 2011.

[7] Granberg J, Minock M. A natural language interface over the musicbrainz database[C] // Proceedings of the QALD-1 workshop,1st Workshop on Question Answering over Linked Data, Heraklion, Greece,30 May 2011. 2011 : 38 – 43.

[8] Weible C. The Internet movie database: a reference guide to Hollywood and beyond[J]. Internet reference services quarterly, 2011, 6(2) : 47 – 50.

[9] Wick M. Boutreux C.GeoNames[J]. GeoNames Geographical Database, 2011 : 3 – 8.

[10] Derose P, Shen W, Fei C. DBLife: A Community Information Management Platform for the Database Research Community (Demo)[C] // Proceedings of the third Biennial Conference on Innovative Data Systems Research, Asilomar,USA,Online Proceedings. 2007 : 169 – 172.

[11] 余晓鹏. 服装知识图谱构建及嵌入研究 [D]. 武汉 : 武汉纺织大学, 2022.

[12] 李娜. 基于知识图谱的问答系统的研究 [D]. 四川 : 电子科技大学, 2021.

[13] 李铁飞, 生龙, 吴迪. BERT-TECNN 模型的文本分类方法研究 [J]. 计算机工程与应用, 2021, 57(18): 186–193.

[14] Hochreiter S, Schmidhuber J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8): 1735–1780.

[15] Weible C. The Internet movie database: a reference guide to Hollywood and beyond[J]. Internet reference services quarterly, 2001, 6(2): 47–50.

[16] 俞逸洋. 基于服装服饰知识图谱的研究与应用 [D]. 武汉: 武汉纺织大学, 2022.

[17] 潘王蕾, 何瑛. 基于个性化推荐的服装知识图谱构建 [J]. 服装学报, 2022: 2–5.

[18] Liu，h, Wu Y, Yang Y. Analogical inference for multi-relational embeddings[C] // Proceedings of the 2017 International Conference on Machine Learning. New York: ACM. 2017: 2168–2178.

[19] Vashisihth S, Sanyal S, Nitin V. Interacte:Improving convolution-based knowledge graph embeddings by increasing feature interactions[C] // Proceedings of the 2020 AAAI Conference on Artificial Intelligence.New York:AAAI: Vol 34. 2020: 3009–3016.

[20] Zhang Z, Cai J, Zhang Y. Learning hierarchy-aware knowledge graph embeddings for link prediction[C] // Proceedings of the 2020 AAAI Conference on Artificial Intelligence.New York:AAAI: Vol 34. 2020: 3065–3072.

[21] Ren F, Li J, Zhang H. Knowledge Graph Embedding with Atrous Convolution and Residual Learning[C] // Proceedings of the 2020 International Conference on Computational Linguistics.New York:ACM. 2020: 1532–1543.

[22] Yang B, Yih W, He X. Embedding entities and relations for learning and inference in knowledge bases[J]. arXiv preprint arXiv:1412.6575, 2014: 1–5.

[23] 申豪杰. 基于知识图谱的电影知识问答系统研究与实现 [D]. 重庆: 重庆师范大学, 2019.

[24] 李轩. 基于知识图谱的教育领域知识问答系统的研究与应用 [D]. 长春: 吉林大学, 2019.

[25] 徐增林, 盛泳潘, 贺丽荣. 知识图谱技术综述 [J]. 电子科技大学学报, 2016, 45(4): 589–606.

[26] 魏力. 基于知识图谱的风控模型的研究与实现 [D]. 南京: 南京航空航天大学, 2019.

[27] Noy N, Mcguinness D. Ontology development 101: A guide to creating your first ontology[J]. And Stanford Medical Informatics, 2001, 17(2): 137–141.

[28] Kietz J, Volz R, Maedche A. Extracting a domain-specific ontology from a corporate intranet[C] // Proceedings of the 2nd workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning-Volume. 2000: 167–175.

[29] 倪浑. 服装形象设计 [M]. 北京: 中国纺织出版社, 2012: 2–4.

附录

卓逸挺，男，2001 年 2 月生。目前就读于杭州电子科技大学自动化学院。

代码：

1. xxxxx. Stability analysis for a class of switched systems under perturbations with applications to consensus[J]. IET Control Theory & Applications, 2017, 11(9):1341-1350. (SCI)