分类号　　密级　　UDC

北京林业大学

专业学位硕士学位论文

基于知识图谱的菜品推荐系统

Dish Recommendation System Based on Knowledge Graph

董洪伟

指　导　教　师　付慧副教授 王立伦高级工程师

学　　院　信息学院

专业学位类型　工程硕士

领　域　名　称　计算机技术

二〇二〇 年　七　月　十二 日

独 创 性 声 明

本人声明所呈交的论文是我个人在导师指导下（或我个人……）进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京林业大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

签　名：_____　日　期：_____

关于论文使用授权的说明

本人完全了解北京林业大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存论文。

(保密的论文在解密后应遵守此规定)

签　名：_____导师签名：_____日　期：_____

摘要

随着我国经济发展和人民生活水平的提高，人们在饮食习惯和生活方式上也发生很大改变，近年来我国癌症患病率呈现明显上升趋势，而导致我国癌症高发的一大重要原因是不合理的饮食。在中国历史悠久的饮食文化中，中国菜有着不可替代的地位，根深蒂固地影响着每一个中国人的日常饮食。本文通过调研现有的与健康饮食相关的软件，如“薄荷 APP”、“美食杰”等，发现其仅给出相关食物的卡路里等营养数值信息以及营养价值等，不能结合用户的身体健康状况、也不能细致的描述饮食状况，因此这些工具很难做出准确而有效的健康饮食推荐。为了满足用户合理膳食、健康饮食需求，本文通过构建中国菜品知识图谱对饮食信息进行更为全面的描述，同时结合协同过滤算法和知识表示学习为用户提供更符合健康饮食要求的推荐结果。本文的主要研究内容如下：

首先构建了菜品知识图谱。对获取的菜品菜谱以及营养学相关文献等数据的知识特征进行分析，实现了领域内实体与关系的划分，并定义多种实体之间的关系；使用 BiLSTM-CRF 对半结构化和非结构化的文本数据进行实体抽取，并对实体部分进行了对齐处理；最后利用图数据库 Neo4j 存储了构建的菜品领域知识图谱。

其次提出一种融合协同过滤和知识表示的推荐算法，既利用了用户行为数据，又包含了菜品本身的相似性信息，采用知识表示学习算法 TransD 将菜品知识图谱中的实体和丰富语义关系精准映射到低维向量空间中，生成基于语义的菜品相似度表达；根据收集到的用户数据构建用户兴趣模型，获取用户行为矩阵，生成基于用户打分的菜品相似度表达，线性融合两种相似度信息获得最终的菜品间相似度表达，并将用户打分矩阵与融合后的相似度表达进行再结合，对用户未打分的菜品进行预测，依据预测评分降序排列并从中选取 Top-N 的菜品作为推荐列表推送给用户。

最后通过结合本文构建的菜品知识图谱和用户对菜品的评分数据，使用本文提出的推荐算法进行实验。结果表明，相较于传统的协同过滤算法，本文结合菜品语义信息的推荐算法在准确度达到了 79.3%、F1 值达到了 76.3%。

关键词：知识图谱，知识表示学习，菜品推荐，健康饮食

Dish Recommendation System Based on Knowledge Graph

Master Candidate: DongHongwei
(Computer Technology)
Directed by Fu Hui and Wang Lilun

ABSTRACT

With the development of China's economy and the improvement of people's living standards, people's eating habits and lifestyle have changed a lot. In recent years, the incidence of cancer in China has shown an obvious upward trend, and one of the important reasons for the high incidence of cancer in China is unreasonable diet. In China's long history of food culture, Chinese food has an irreplaceable position, deeply affecting every Chinese diet. This article investigates the existing software related to healthy diet, such as "Mint APP", "Gourmet Master", etc., and finds that it only gives nutritional value information and nutritional value of calories and other related foods, but cannot be combined with the user's physical health. status, nor describe the dietary status in detail. Therefore, it is difficult for these tools to make accurate and effective recommendations for healthy diet. In order to meet users' reasonable diet and healthy diet needs, this article constructs a Chinese food knowledge graph to describe diet information more comprehensively, and at the same time combines collaborative filtering algorithms and knowledge representation learning to provide users with recommendations that are more in line with healthy diet requirements. The following is the main contents of this paper：

Firstly, the knowledge graph of dishes was constructed. This paper analyzes the knowledge characteristics of the acquired recipes and nutrition related literature date, realizes the division of entities and relationships in the field, and defines the relationships among multiple entities; Using BiLSTM-CRF to extract entities from semi-structured and unstructured text data, and aligns the entities; Finally, the graph database Neo4j is used to store the constructed knowledge graph of the dish domain.

Secondly, a recommendation algorithm combining collaborative filtering and knowledge representation is proposed, which not only uses the user behavior data, but also contains the similarity information of the dishes themselves. By using the knowledge representation learning algorithm TransD, The entities and rich semantic relationships in the knowledge graph of dishes are accurately mapped into the low-dimensional vector space to generate the similarity expression of dishes based on the semantics; Building user interest model based on information collected from users, obtain user behavior matrix, and generate user-based scoring, the similarity expression of dishes based on user scoring is generated, and the final similarity expression between dishes is obtained by linear fusion of two similarity information, and then combine the user scoring matrix and the fused similarity expression to combine the user scoring matrix to predict the unscored dishes of the users. According to the prediction score in descending order and select Top-N dishes from it as a recommendation list and push it to the user.

Finally, by combining the knowledge graph of dishes constructed in this paper and the user's scoring data on the dish, we use the recommendation algorithm proposed in this paper to carry out experiments.

The results show that, compared with the traditional collaborative filtering algorithm, the proposed algorithm combined with the semantic information of dishes achieves an accuracy of 79.3% and an F1 value of 76.3%.

Keywords: knowledge graph, feature representation learning, dish recommendation, healthy eating

目录

1 绪论

1.1 研究背景及意义

人们的日常生活离不开“衣食住行”，其中“食”更是不可或缺的一部分。随着生活水平的提高，人们可以更容易获得想要的食物，然而随之而来的却是饮食不均衡带来的健康问题。在做到健康饮食的同时，兼顾口味爱好更符合人们的饮食期待。近年来，由于人们不科学的饮食习惯，诱发了许多健康问题，如肥胖、高血压、胆固醇等疾病，严重损害人们的身体健康。中国菜是中国文化的重要组成部分之一，又称中华食文化，根深蒂固地影响着每一个中国人的日常饮食，不同地区和民族的菜肴风格又迥然不同，菜品信息的表达需要更加细化并能够自动提取出其中的重要关系。因此本文基于知识图谱构建菜品推荐系统，能够获取菜品数据的更为细致全面的内容表达和重要的关系信息，结合用户的身体特点，为不同人群推荐健康合适的菜肴，以满足人们健康饮食需求。

2019 年 1 月，国家癌症中心发布了最新一期的全国癌症统计数据（郑荣寿等，2019）。报告显示，2015 年全国恶性肿瘤患病人数约有 392.9 万人，相比于 2014 年的 380.4 万增加了 12.5 万，增长率达到 3.2%；也就是说，平均每天有超过 1 万人被确诊为癌症，每分钟大约有 7.5 人被确诊为癌症，而我国癌症高发的主要诱因其中就包括不健康的饮食。随着我国人口老龄化加剧、不良饮食方式及习惯的广泛存在，我国癌症发病率和死亡率还将不断上升，因此合理安排饮食愈发显得重要。

研究表明很多疾病发生都与不合理饮食息息相关，例如营养过剩会导致糖尿病，高血脂等，营养不良则会导致低血糖，抵抗力降低等，不仅影响健康，而且会缩短寿命。为了提高生活质量降低疾病发生率，人们需要合理安排膳食，如此才能帮助人们预防疾病，控制体重，但如何选择合适的饮食成为了一个问题。在大量杂乱的营养信息中，传达给公众的营养信息很少建立在严谨的科学研究之上，人们无法判断出什么是真正适合自己的食物，于是“人们愿意接受纵容自己坏习惯的建议”，情况变得糟糕起来。长久以来，人类对食物的关注点大多停留在食物本身的好与坏上，而忽略了食物对自己是否有益。市场上的健康饮食相关工作的调研，发现其主要是通过给出菜品的卡路里等营养数值信息来指导用户饮食，不考虑用户的身体状况、饮食状况等，很难做出准确而有效的健康饮食推荐。

随着互联网技术的进步及用户需求的不断扩大，知识图谱（Knowledge Graph）在语义网络（Semantic Network）的基础上逐渐发展起来，语义网络借助图结构对信息进行结构化存储。知识图谱保留了语义网络的优点，并在其基础上进行了修改。知识图谱以结构化的形式描述客观世界中实体及其关系，能够将在互联网上的信息以更符合人类的理解和推理的方式表达出来，其表达方式更容易被人接受和使用。基于知

识图谱的发展，知识图谱最先应用在搜索引擎上，使得搜索引擎可以更好的理解用户输入与数据之间的语义关联，为用户提供更加准确的信息，例如在 Google 中搜索“西红柿的学名”，搜索引擎在结果中直接给出了答案，并在右边以卡片的形式返回了西红柿相关的属性和信息，如图 1.1 所示，卡片中显示了西红柿的维基百科介绍、学名、其科目信息等，同时还提供了相关实体的链接便于用户查看。因此基于知识图谱的搜索引擎结合了用户输入与数据之间的语义联系，为用户更快捷的搜索和更有效的过滤信息。

图 1.1 Google 中搜索“西红柿的学名”的搜索结果

Figure.1.1 Search results of "scientific name of Tomato" in Google

除了应用在搜索引擎外，推荐系统也是知识图谱的一大主要应用。本文通过各种与用户相关的信息，如：饮食习惯、疾病等，可以得到精确的用户画像。利用菜品构成及营养学知识创建的菜品知识图谱，通过用户的信息，结合知识库中的知识，推断出用户的体质，利用丰富的菜品知识库既能从饮食上帮助用户补充营养，又能让用户缓解病情或预防疾病。

本文将用户相关的信息与丰富的菜品知识图谱相结合，能够根据用户疾病类型和菜品的营养成分进行推荐。本推荐系统推荐的准确性能够达到 79.3%，在目前已有的推荐系统中位于前列。

1.2 国内外研究现状

1.2.1 知识图谱的研究现状

随着人工智能相关技术的不断进步和发展，知识图谱逐渐成为人工智能的重要分支，知识工程在大数据环境中的成功应用，然而知识图谱作为知识工程研究的分支，并不能作为一个全新的概念，其本质可以看做是语义网络（Semantic Network）的知识库。语义网络借助图结构对信息进行结构化存储，在 20 世纪 50 年代由 M. Ross Quillian 和 Robert F. Simmons 等人提出（Simmons，1965）。知识图谱在语义网络的基础上，保留了语义网络中使用图的结构对信息进行表示的特点，并以结构化的形式对数据中涉及的实体和实体间的关系进行表示（Chen *et al.*，2009）。知识图谱通常以三元组的形式对实体间关系进行表示，可以抽象为$\mathrm{G}=(E,R,S)$，其中$\mathrm{E}=\{e_1,e_2,\cdots,e_{|E|}\}$是代表实体集合，包含实体的个数是$|E|$个；同样地，$\mathrm{R}=\{r_1,r_2,\cdots,r_{|R|}\}$对关系集合进行描述；$\mathrm{S}\subseteq\mathrm{E}\times\mathrm{R}\times\mathrm{E}$则对图谱中所有三元组进行表征。由此，知识图谱以实体为基本单位，通过关系将实体连接以形成知识网络（Maedche *et al.*，2002）。

近年来在构建知识图谱方面踊跃出大量的研究成果。随着人们需求不断增加以及自然语言处理技术的发展，大量知识图谱相关的项目开始开展。从知识图谱构建所依赖的技术上区分，可以分为两类：一类是以 WordNet，HowNet 等知识库为代表的早期以手工构建的本体库；另一类则是以 YAGO，DBPedia 等为代表的借助自动构建技术且面向开放领域的知识图谱。从图谱面向的范围和领域区分，可以分为面向通用领域知识图谱与面向垂直领域知识图谱两类。面向通用领域的知识图谱一般应用于搜索等场景，要求包含知识的全面性，但对知识的准确性的要求相对较低，比较代表性的通用图谱有国内的 Zhishi.me、CN-DBPedia、搜狗知立方和百度知心以及国外的 Freebase、YAGO、DBPedia 等。面向垂直领域的知识图谱主要应用于智能问答、知识推理以及辅助决策等要求知识准确性场景，目前垂直领域的知识图谱主要涉及旅游、交通和医学等方向。

在健康饮食领域，已有一些提供检索和推荐食物的知识图谱。文献（Helmy *et al.*，2015）通过整合食材，宗教、营养、喜好等多个领域的知识来建立数据库，对所有食品进行检索，然后根据用户的身体健康状况向他们推荐适合的食品。此外还有一些国内的相关工作，文献（胡秋明等，2006）构建了一个食物的知识图谱，其中包含了食物成分和合理膳食所需的食谱，以及预防和治疗一些疾病所需要的食材的数据库。

目前，知识图谱的构建主要通过将互联网中形式多样的数据作为数据源，进行知识抽取，经过一系列的整合和对齐后选用合适的存储方式建立知识图谱；根据面向领域以及数据形式的不同会采用不同的构建流程，目前主流的图谱构建流程包括自顶向下（top-down）和自底向上（button-up）两种模式。其中自顶向下的构建方式需要依

赖现有的结构化数据为图谱预先定义本体模式层进行约束，之后再将抽取出的知识加入到知识图谱中；而自底向上的构建模式则是先从半结构化或非结构化数据中利用合适的方法抽取出置信度较高的知识加入知识图谱，再依据建好的知识图谱抽象出顶层的本体模式。目前对于不同的知识图谱构建模式，其共有的核心流程主要可以分为知识抽取、融合以及存储三个步骤；通用的知识图谱构建流程如图 1.2 所示。

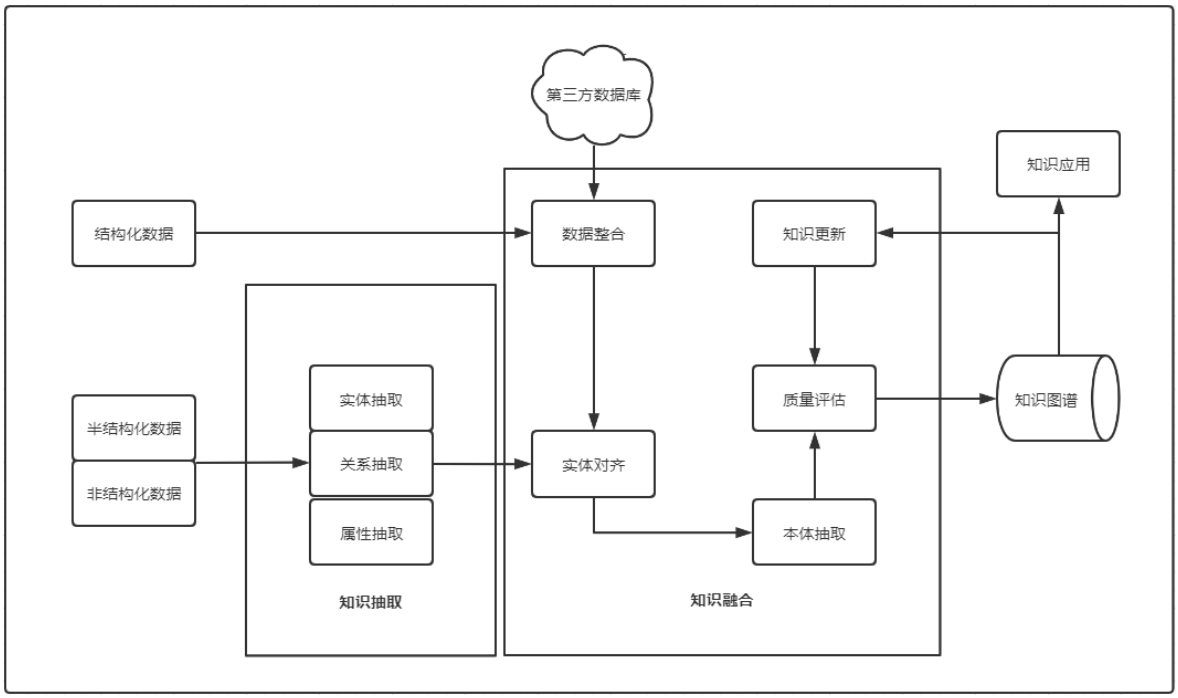

图 1.2 知识图谱构建流程

Figure 1.2 Construction process of knowledge graph

目前在健康饮食领域，多数知识图谱构建工作集中在症状、疾病、药物等医疗主题概念，少有工作将食物与健康进行关联，为日常饮食服务。

1.2.2 推荐系统的研究现状

协同过滤算法作为推荐系统的重要技术，于上世纪九十年代被明尼苏达大学 GroupLens 研究组提出（Resnick *et al*，1994）。推荐系统经过二十多年的长足发展，已经成功的在多个领域展开应用：在线视频、在线音乐、电商以及新闻等各个生活角落。目前国内外的大厂商都在产品上运用了推荐系统，如抖音、斗鱼直播等在视频领域，如网易云、QQ 音乐等音乐领域，如淘宝、亚马逊等电商领域，如头条、网易新闻等新闻领域。

伴随着推荐算法的发展，传统的推荐算法，如协同过滤算法、以及基于内容的推荐算法（黄立威等，2018）由于数据极度稀疏而无法满足人们的需求。随着互联网的发展，越来越多的蕴含用户行为信息的数据产生并被获取，包括图像、文本以及标签等多元异构信息（魏慧娟等，2016）。混合推荐算法通过利用这些辅助信息，可以有

效地缓解传统推荐算法中数据稀疏与冷启动问题。

推荐系统作为一个多学科交叉的领域，包括数据挖掘、自然语言处理、机器学习等学科。同样地，各学科的发展也会给推荐系统带来新的活力。知识图谱中蕴含的丰富实体与实体之间的关系，能够一定程度的解决传统推荐算法存在的难题，知识图谱与推荐系统的结合是必然而有巨大意义的。基于知识图谱的推荐模型大多以现有的推荐模型为基础，通过知识图谱中将项目、用户等实体的结构化知识描述的更加细粒度化，利用知识推理得出更加深层次的信息加入到推荐模型中，丰富推荐结果的多样性。通过引入辅助知识改善早期推荐模型中数据稀疏和冷启动问题。文献（Passant，2010）通过计算知识图谱中实体间关系的语义距离建立音乐推荐模型。文献（Cheekula *et al*，2015）提出了利用 DBpedia 知识图谱中的层次类别信息应用于推荐任务中，通过传播激活算法在知识图谱中寻找推荐实体。文献（张富峥等，2016）通过知识表示学习的方法将知识图谱中的信息加入推荐模型中，提出了协同知识图谱表示学习的推荐模型。文献（吴玺煜等，2018）提出一种基于知识图谱表示学习的协同过滤推荐算法，把协同过滤推荐算法得到的邻近集合替换成知识图谱表示学习得到的语义邻近集合。文献（Wang *et al，2019*）提出了基于知识图谱与注意力机制的新方法，通过 user 和 item 的属性链接成 user-item 实例，从而去除 user-item 二者之间相互独立的假设。文献（Wang *et al，2019*）提出一种用于知识图谱增强推荐的多任务特征学习方法，多任务通过交叉和压缩单元相关联，交叉和压缩单元自动共享潜在特征，并且学习推荐系统中的项目与知识图中的实体之间的高阶交互。文献（Wang *et al，2019*）提出了知识图卷积网络（KGCN），通过在知识图谱上挖掘它们的相关属性来有效地捕捉项目间的相关性。

1.3 论文主要内容及方法

本文首先构建了菜品知识图谱，对获取的菜品菜谱以及营养学相关文献等数据的知识特征进行分析，实现了领域内实体与关系的划分，并定义多种实体之间的关系；使用 BiLSTM-CRF 对半结构化和非结构化的文本数据进行实体抽取，并对实体部分进行了对齐处理；最后利用图数据库 Neo4j 存储了构建的菜品领域知识图谱。

其次提出一种融合协同过滤和知识表示的推荐算法，既利用了用户行为数据，又包含了菜品本身的相似性信息。采用知识表示学习算法 TransD 将菜品知识图谱中的实体和丰富语义关系精准映射到低维向量空间中，生成基于语义的菜品相似度表达；根据收集到的用户数据构建用户兴趣模型，获取用户行为矩阵，生成基于用户打分的菜品相似度表达，线性融合两种相似度信息获得最终的菜品间相似度表达，结合用户打分矩阵，对用户未打分的菜品进行预测，依据预测评分降序排列并从中选取 Top-N 的菜品作为推荐列表推送给用户。

最后通过结合本文构建的菜品知识图谱和用户对菜品的评分数据，使用本文提出

的推荐算法进行实验。结果表明，相较于传统的协同过滤算法，本文结合菜品语义信息的推荐算法在准确度达到了 79.3%、F1 值达到了 76.3%。

其工作流程如图 1.3 所示：

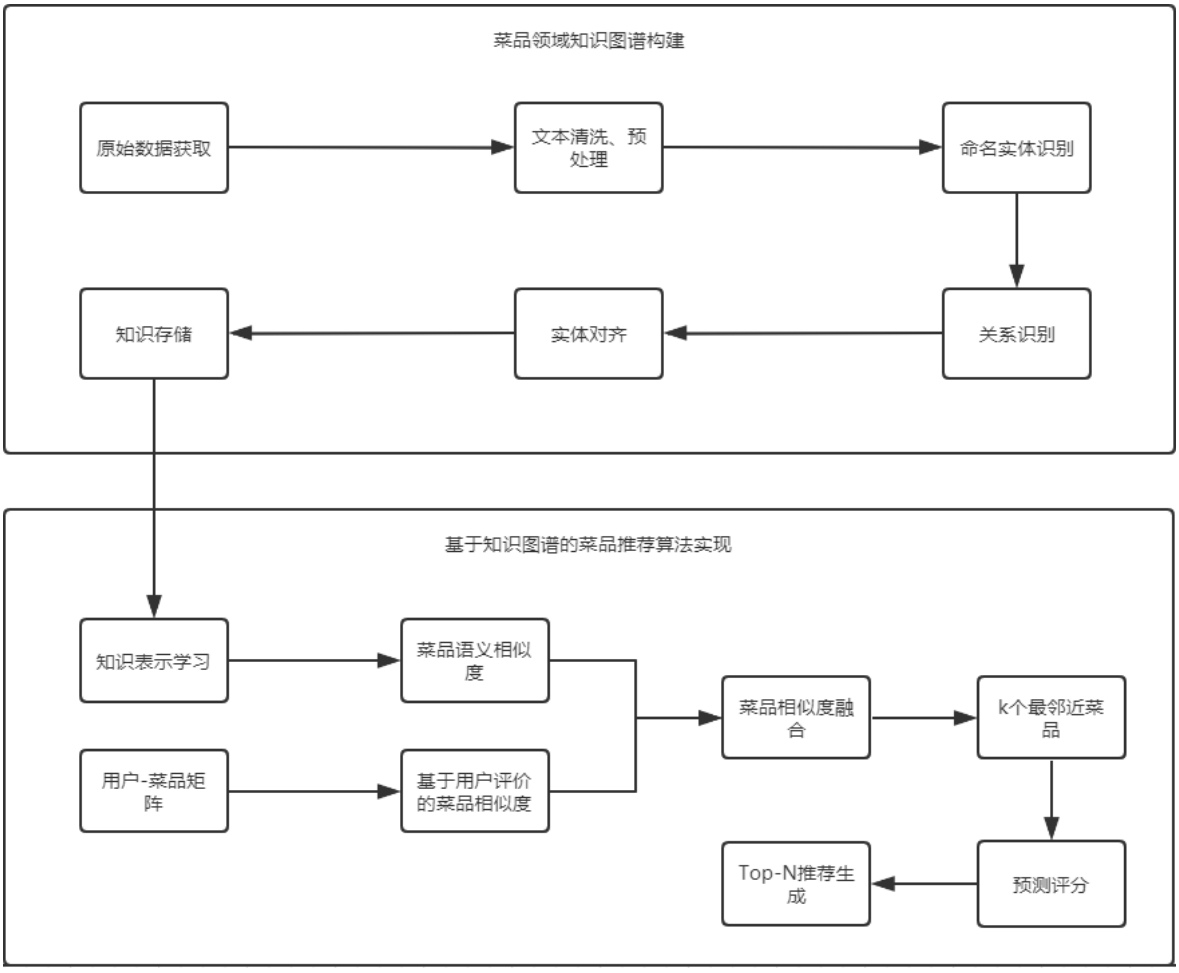

图 1.3 技术路线

Figure.1.3 Technical route

本文所做研究和工作具体的技术路线如图 1.3 所示。首先针对获取到的菜品数据进行文本清洗以及语义角色标注的处理，通过实体及关系的抽取，对菜品的知识图谱进行构建。然后由此进行知识表示学习得到菜品与关系实体向量并计算菜品语义相似度，通过基于用户与菜品的评分信息构建评分矩阵。最后基于上述实现，将知识图谱菜品语义相似度和基于用户的菜品相似度融合，按照降序排列用户未打过分的菜品得分，实现基于知识图谱的菜品推荐系统。

1.4 论文组织结构及安排

本文主要是由以下六部分构成：

第一章、详细介绍研究的背景和意义，简述知识图谱和推荐算法领域的研究现状，针对知识图谱目前很少应用在饮食领域以及推荐算法数据稀疏等问题，提出本文的研究目标，并对本文的主要研究内容进行了简要的概述。

第二章、对知识图谱构建的相关理论和方法进行介绍，包括知识图谱的起源、发展和分类；知识图谱的构建方法以及知识图谱的存储形式；同时介绍了协同过滤算法、基于内容的推荐算法和基于知识图谱的推荐算法。

第三章、菜品领域知识图谱的构建方法，包括数据预处理，根据数据特点定义实体以及实体间的关系，同时详细介绍了实体识别和关系抽取的过程，最后使用图数据库 Neo4j 存储构建的菜品知识图谱。

第四章、提取知识图谱的菜品语义表示。利用知识表示学习技术，将知识图谱中实体和关系转变为向量表达。利用 TransD 对知识图谱中实体与关系进行学习，将实体与实体之间的关系抽象成头向量表达和关系向量表达，用实体链接和三元组分类检测向量提取准确性，提取出菜品的语义向量。

第五章、将知识图谱方法与协同过滤算法相结合，提出 TransD-CF 推荐算法。融合基于知识图谱的菜品间相似度信息和基于用户打分的菜品相似度信息，然后依据融合的菜品相似度来预测用户未打分菜品的得分并降序排列进行菜品推荐。通过实验证明，与传统的协同过滤算法相比较，本文提出的算法推荐准确率提高了 3.2%。

第六章、总结全文，分析了本文的主要贡献和不足之处，并对基于知识图谱的菜品推荐系统的进一步完善进行展望。

2 相关理论及方法简介

以知识图谱作为基础的菜品推荐系统涵盖知识图谱的构建和菜品推荐算法两大主要部分。本章重点介绍与以上两部分有密切关联的理论及方法。

2.1 知识图谱

知识图谱在构建之中会牵扯到很多种专业技术，包括但不局限于知识逻辑化建模、知识抽取、知识描述、知识概念层和数据层融合、智能系统下知识推理、质量评估等，最基础也是最重要的任务是知识抽取，针对不同领域使用适应其数据特点的方法和技术，将实体、属性及关系等组成知识图谱所需的主要元素进行抽取，抽取过程对于最终知识图谱的优劣结果尤为重要；同时选择合适的工具对抽取出的知识进行存储也是知识图谱构建的关键任务。本节不仅包含原始数据转变成而形成知识图谱过程中技术介绍，同时会介绍目前知识表示学习的现状以及知识存储所需工具。

2.1.1 知识抽取

知识抽取是知识逻辑建模时获取知识和能否构成知识库所必不可缺过程，在这个过程中的知识材料都是未经处理的，是通过知识提取从蕴含着信息的载体中将知识经过鉴别、理解、筛选、总结出来，并将这些知识存储成知识库。通常用户评价知识图谱的质量高低会从以下两方面进行：其一，知识图谱的宽度限度中是否包含用户所迫切需求的知识服务；其二，知识图谱能否提供给用户无需再处理的精准信息也是评价知识图谱优劣的重要指标。因此，要想持续性为用户提供优良的知识服务，新知识的实时更新与扩充、知识的精度提升对一个知识图谱至关重要，这也对知识抽取提出挑战。

知识抽取常常在相异的信息载体，相异的数据结构间进行，所获得的数据信息来源包含数据库、表格、列表和纯文本等。通过从信息源抽取到的实体、关系、属性去构建 SPO 关系或者多元关系。知识抽取核心要素有子信息元素实体抽取、解决两者以上的多实体语义链接的关系抽取和为实体构造属性的属性抽取。

2.1.1.1 实体抽取

实体抽取的任务是利用原始数据对其涉及的命名实体进行识别和抽取，有时也被叫作命名实体识别（Named Entity Recognition，NER）（滕青青等，2010）。命名实体识别第一次被作为明确的研究对象是在 1995 年举办的第六届 MUC（Message Understanding Conferences）会议上（龚启文等，2019）。随着技术的发展和推进，命名实体识别技术由早期的基于词典和规则的策略过渡到基于传统机器学习，之后深度学习的进步和广泛应用，开始进入深度学习的时代。

在早期基于规则和词典的方法中，比较具有代表性的是 Collins 等于 1999 年提出的 DLCoTrain 方法，学者们借助机器自动发现和规则生成技术，利用人工构建的词典

和规则进行实体的识别、抽取和分类；与之类似的还有使用 Bootstraping 进行规则自动生成的方法。随着机器学习在自然语言处理领域的运用，命名实体识别也开始进入基于统计机器学习的时期。该时期所提出的方法，本质都是将 NER 任务转化为分类任务。

近年来随着深度学习技术的成熟以及词向量技术在自然语言处理中的运用，学者们开始将深度学习技术运用于命名实体识别，其中循环神经网络（Recurrent Neural Network，RNN）及其变种模型双向长短时记忆-条件随机场模型（BiLSTM-CRF），常被使用且取得较好的效果。本文基于 BiLSTM-CRF 模型进行实体抽取，下面将对 BiLSTM-CRF 模型进行详细介绍。

1.双向长短时记忆-条件随机场模型

在自然语言处理领域，通常需要对句子中的语义信息进行理解和分析，但句中的各个成分并不是单独存在的，而是具有很强的语义关联性，传统的统计机器学习方法以及全连接神经网络模型，很难在学习中保留句子前后文的相关信息进行推断和预测；相比之下循环神经网络拥有较强的推理能力，该模型为递归式神经网络，以映射后的序列数据作为输入，模型的循环单元以链的形式链接，递归则在序列数据的传播方向进行（杨丽等）。随着对 RNN 的应用以及研究的深入，Bengio 等人发现由于模型架构的特性，在输入序列长度增加的情况下，传统的 RNN 模型存在着梯度消失和梯度爆炸的现象（Bengio *et al.*，1994）。为了解决这种由于长距离依赖而产生的问题，诞生了一种循环神经网络的改进模型，即长短期记忆网络（Long Short-Term Memory，LSTM）。目前基于 LSTM 的深度学习模型已广泛应用于命名实体识别、语音合成与处理等多个自然语言处理领域。

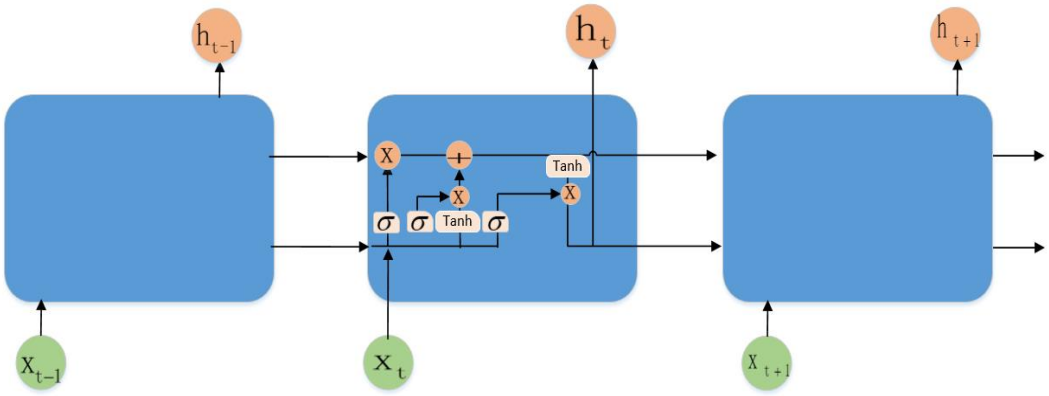

图 2.1 LSTM 结构图

Figure 2.1 Structure of LSTM

相较于原始的 RNN 模型，LSTM 拥有输入门、遗忘门以及输出门三个门结构，它们可以增加或消除信息传输到神经元（细胞）状态的能力，使得模型可以记住需要长期保存的特征，并将无需保存的特征信息遗忘，从而解决长距离依赖的问题。在 LSTM 模型的每个循环单元中，输入序列先后通过遗忘门和和输出门，以决定在后续的传播过程中遗忘的信息和保存的信息；最后通过输出门，输出门决定输出的值。LSTM 模型的结构如图 2.1 所示，可以将单次传播流程分为三步：

第一步遗忘：这一步的作用为根据上一层的输出的信息计算出遗忘后所需保留的信息，在遗忘门将上一层的输出 $h_{(t-1)}$以及本层的输入 x_t 一起作为输入，首先经过 sigmoid 激活函数得到 f_t；随后根据前一个状态 $C_{(t-1)}$得到要保留的信息 $S_{(t-1)}$，其中 $C_{(t-1)}$代表了本层细胞被遗忘的概率，其取值在[0,1]之间，0 代表全部遗忘，1 代表全部保留；

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1,}\ x_t\right] + b_f\right) \tag{2-1}$$

$$S_{t-1} = f_t * C_{t-1} \tag{2-2}$$

第二步：更新状态矩阵。首先根据上一层的输出 $h_{(t-1)}$以及本层的输入 x_t 计算出输入门的输出值 i_t，随后 tanh 层同样使用上一层的输出 $h_{(t-1)}$以及本层的输入 x_t 产生一个新的当前状态 $\tilde{C}_t$，此状态会决定本层加入多少新信息；最后将经过遗忘门后保留的部分 $S_{(t-1)}$与新添加信息的权重求和，生成新的状态矩阵 C_t；

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1,}\ x_t\right] + b_i\right) \tag{2-3}$$

$$\tilde{C}_t = \tanh\left(W_C \cdot \left[h_{t-1,}\ x_t\right] + b_C\right) \tag{2-4}$$

$$C_t = S_{t-1} + i_t * \tilde{C}_t \tag{2-5}$$

第三步：计算输出 h_t。首先根据上一层的输出 $h_{(t-1)}$以及本层的输入 x_t 使用 sigmoid 激活函数计算出 O_t；本层输出则为输入们产生的新状态 C_t 经 tanh 作用后与 O_t 的乘积。

$$O_t = \sigma\left(W_o \cdot \left[h_{t-1,}\ x_t\right] + b_o\right) \tag{2-6}$$

$$h_t = O_t * \tanh(C_t) \tag{2-7}$$

目前在命名实体识别（NER）任务中常用的深度学习模型架构为 BiLSTM-CRF，即在双向长短时记忆网络（BiLSTM）模型后叠加随机条件场（CRF）模型，将 NER 转换为序列标注任务（李纲等，2020）。相较于 LSTM 模型，双向长短期记忆网络模型可以同步保留输入序列的前序和后续特征，更加适用于需要考虑上下文的自然语言处理任务（LeCun *et al.*，2015）。在命名实体识别中，BiLSTM 模型的输入是经过 Embedding 层映射后的向量，输出的则是句中所有单词针对每一个类别标签的预测分

数，在不添加 CRF 层的情况下，可以选择得分最高的标签作为单词的标签，BiLSTM 的结构如图 2-2 所示。

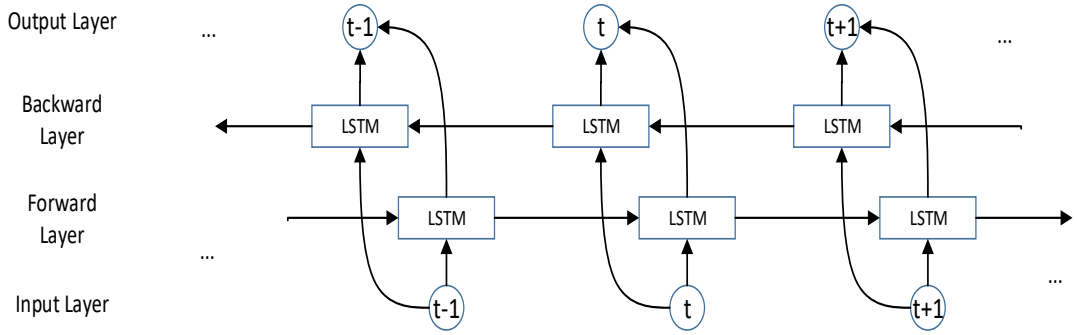

图 2.2 BiLSTM 模型结构

Figure 2.2 Structure of BiLSTM

然而单纯使用 BiLSTM 模型进行标签预测无法学习到句中标签的特征，例如某些标签的转移在命名实体识别中并不合法。CRF 层的作用就是在训练过程中自动学习标注数据中标签的特征和规则，进一步得到最终实体标签的得分，这样整个模型既保留了输入序列的前后序特征，也学习到了标签前后的特征。在 BiLSTM-CRF 模型中训练数据经 Embedding 层向量化后首先输入到 BiLSTM 层，得到当前句子的分数矩阵，而后将分数矩阵输入到 CRF 层。

CRF 层的原理如下：

（1） 对于标注含有 m 个标签（本文中$m = 7$）的数据，分数矩阵$S = n * m$，n 为句子分词后词向量的数量，则$S[i][j]$表示句中第 j 个单词其标签为 i 的得分；

（2） 将分数矩阵 S 输入到 CRF 层，得到到状态转移概率矩阵 T；

（3） 对于转移概率矩阵 T，$T[i][j]$表示由标签由 i 转移到 j 的概率，据此得到对于输入的分数矩阵 S 其最终的预测标签序列 y 的最终得分，其中 y 的计算公式如(2-8)所示。

$$\mathrm{Score}(M,y) = \sum_{i=0}^{n} T_{y_i,\ y_{i+1}} + \sum_{i=1}^{n} S_{i,\ y_i} \tag{2-8}$$

最终模型通过 softmax 层后得到预测的标签序列；

相对英文文本而言，汉语文本不存在利用空格符切分汉语文本一说。因此如果按照英文文本的分词方式来对汉语文本中命名实体进行分词操作，将会导致菜品实体边界认知出现许多无法预知的问题难点。而菜品实体边界因分词带来的误差将会延伸至实体模型的名称确立环节，最终导致菜品识别精确度和准确率下降或无法识别。所以在以下的研究中通过字级别而非字符级别的 BiLSTM-CRF 模型方式，对具有一定结构性的半结构化文本及无规则的数据结构非结构文本完成命名实体识别，任何实体都拥有与之相映射的字符标记。本研究以文献（Dongle *et al*，2016）作为参考，按照文

献中实体识别模型的方法进行名称确立，并完成相关的字符表达，使之能够成为实体识别最终输入。文章中所用的模型框架于图 2-3。

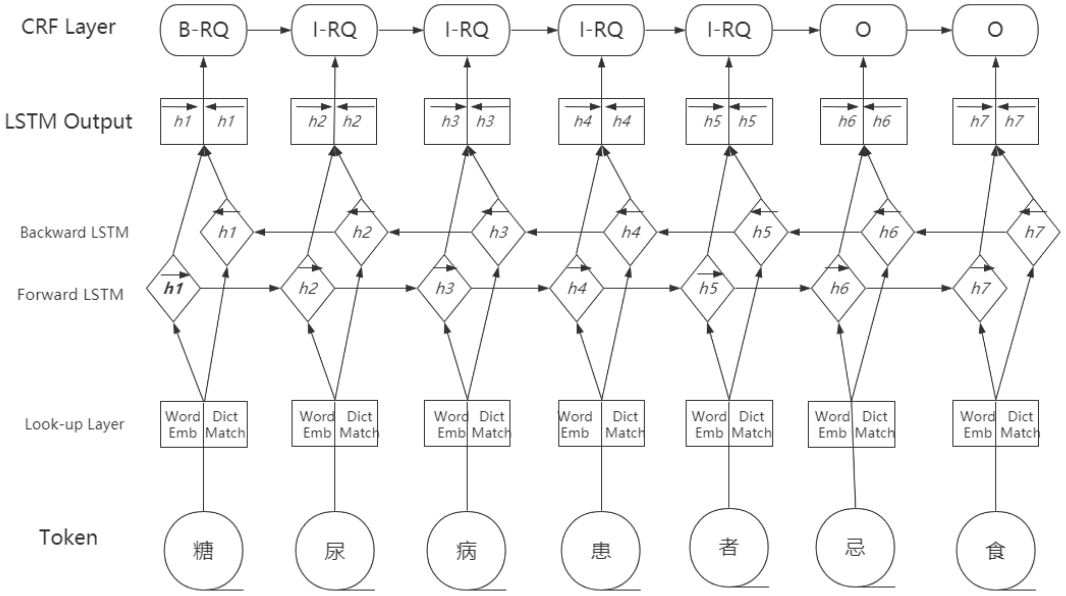

图 2.3 BiLSTM-CRF 模型结构

Figure 2.3 Structure of BiLSTM-CRF

2.1.1.2 关系抽取

关系抽取最终要达到的目标为根据数据信息得到实体与实体存在的某种联系，从而将离散的实体联系起来获得原始数据完整的语义信息（刘方驰等，2013）。研究前期所提出的关系抽取方法多是针对特定领域语料的关系进行抽取，需要自行预先对语料中关系进行定义（杨棣威，2019）。针对特定领域关系的抽取，最早采用的是基于模式匹配方法，该类方法通常首先对文本进行分割，然后利用分词、词性标注以及句法及语义分析的相关技术结合制定的模板规则进行规则的抽取，此类方法依赖于自然语言处理工具，对结构性强的语料较为适用，拥有较高的精度，但泛化能力不高（Zhou and Min，2007）。随着机器学习的应用，上述缺点得到了有效解决，根据标注数据的不同，可以分为有监督关系抽取、半监督关系抽取以及无监督关系抽取三种（奚斌，2008）。针对特定领域，关系抽取技术的具体分类如图 2.4 所示。

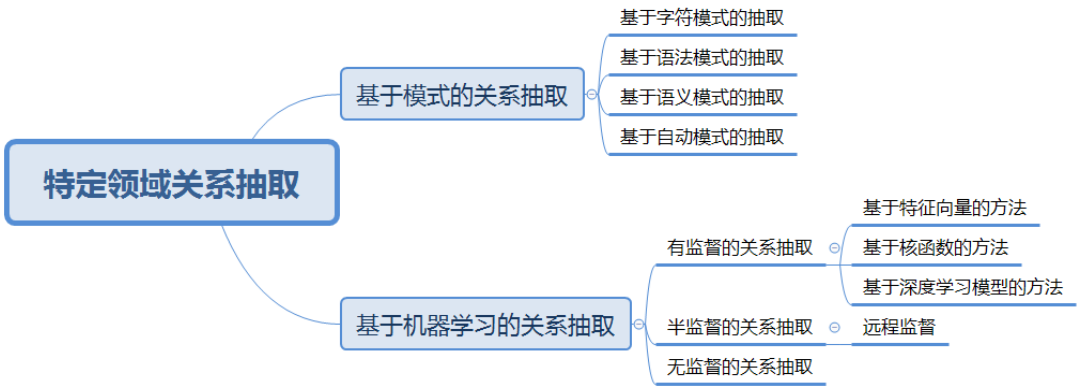

图 2.4 特定领域关系抽取方法分类

Figure.2.4 The classification of domain-specific relational extraction method

有监督的关系抽取使用人工标注好的数据利用分类的思想对文本中的关系进行识别和抽取。基于特征向量（feature-based）和基于核函数（kernel-based）的两类方法构成了有监督的关系抽取任务中主要技术类别（Bengio *et al.*，1994）。基于特征向量的方法依据语料的句法信息和语义信息进行特征向量的选取和构造，进而对关系识别模型进行训练（Von *et al.*，1999），比较有代表性的模型有 Kambhalta 提出的基于最大熵分类器的关系抽取模型等（Strehl and Ghosh，2003）。为了解决和避免基于特征向量的关系抽取方法需要依靠经验选取和建立特征向量的弊端，Zelenko 等在 2003 年提出了基于核函数的关系抽取方法，可以有效利用句中的长距离特征（Zelenko *et al.*，2003）。随后学者们在基于核函数的方法上做出许多贡献，例如 Culotta 等人使用依存树核对新闻内容进行关系抽取；Bunescu 提出了基于最短依存树核的抽取模型；Zhang 等人提出基于解析树的卷积核模型；但总体上由于其约束较强容易出现召回率低等问题（Goldreich *et al.*，1991）。

半监督的关系抽取主要利用预先设定的关系类型以及人工筛选出的实体对作为种子不断迭代学习以获得关系集的一中方式，降低了对人工标注数据的依赖。目前半监督的关系抽取最常用的是由 Brin 提出的基于 Bootstrapping 的方法（Huang *et al.*，2011）；在此基础上 Agichtein 等进行了改进，利用向量相似度增大样本标注量提出了 Snowball 方法；2008 年国内学者陈锦秀等人利用图策略构建了基于图的半监督抽取模型（陈锦秀等，2008）。总体上半监督的关系抽取降低了对人工标注数据的依赖，但初始种子集的质量决定了最后关系抽取的效果，并在迭代学习过程中会产生语义漂移现象，导致回召率普遍偏高。

无监督的关系抽取利用预先抽取的实体关系对语料进行聚类，根据聚类后的结果，对语料中的关系进行标注，是一种自底向上的关系抽取方式。1994 年，Hasegawa 等通过对重复出现阈值进行设定以识别潜在语义关系，首次将无监督方式应用于关系抽取（Hasegawa *et al.*，1994）。无监督关系抽取的方式增强了领域的普适性，相比于半

监督的方法进一步降低了对人工标注数据的依赖，但由于其关系的标注较为宽泛且对于出现频率不高的实体和关系难以捕捉，导致其最终的准确率和回召率较低。

近年来随着深度学习模型的发展和应用，学者们开始运用深度学习的方法对关系进行抽取；作为有监督的抽取方式之一，其通常需要利用词向量技术对训练语料中的字、词进行向量化表示和映射，并利用深度神经网络模型训练数据，并根据当前任务的类型使用不同的函数输出并对结果进行预测。

2.1.2 知识存储

在实际运用中，针对不同的数据类型，需要选择不同且合适的存储工具进行存储。对于知识图谱的存储也类似，知识图谱由利用大量不同结构类型数据抽取出的结构化知识组成，其中针对数据来源为结构化及半结构化数据的规模较小的知识图谱，常见的做法是选择传统的关系型数据库对融合后的知识进行存储；但对于面向开放领域及互联网开放文本构建的知识图谱而言，出于对关系推理及对复杂关系进行查询的需要，常使用图数据库（Graph Database）作为存储知识图谱的工具。相较于关系型数据库而言，图数据库借助图结构对存在关联的数据进行表达和存储，并对符合图语义的信息有较高的查询效率，可以有效推动语义 Web 分析、自然语言处理等领域的发展。

目前研究中和市场上常见的图数据库中，以 Neo4j 最具代表性。其作为一款图数据库，对于以三元组对数据进行表达，且数据间关联性较强的知识图谱而言，具有较高的性能；Neo4j 本质上是一个具备完全的事务特性的 Java 持久化引擎，使用符合 SQL 标准的查询语言 Cypher（CQL）进行查询，并支持数据索引；同时提供了丰富的 API 以供多种编程语言访问调用。本文构建的菜品知识图谱选择使用 Neo4j 进行存储，用于后续的展示与知识表示学习。

2.1.3 知识图谱表示学习

知识图谱表示学习是面向知识图谱中实体与实体间关系的表示学习，把实体和关系向低维向量空间投射，且不丢失实体间联系，方便高效率的计算实体之间的复杂语义关系，又称为知识表示学习、知识图谱特征学习。对于知识图谱的构建、推理以及应用有着重大作用。目前，知识表示学习已经在很多应用中取得了闪耀成果，尤其是将知识图谱与推荐系统相结合，弥补推荐系统数据稀疏与冷启动问题，有效的提高推荐准确率。

知识图谱表示学习的经典代表模型有距离模型、能量模型、双线性模型、矩阵分解模型和翻译模型等（刘知远等，2016）。距离模型 SE（Bordes *et al*，2011）通过对头、尾实体进行映射来计算两投映向量在空间中的距离；能量模型 SME（Bordes *et al*，2014）通过多个矩阵映射实体和关系，进而发现它们之间的潜在联系；双线性模型 LFM（Jenatton *et al*，2012）通过将关系进行双线性转换得到实体与关系的二阶关联，

矩阵分解模型的代表方法是 RESACL 模型（Nickel *et al*，2012），其核心思想是将三元组对应的张量值分解成实体与关系的表示；而翻译模型由于其参数少，计算复杂度低而高效，得到了广泛的使用，其发展历程简单介绍如下：

Mikolov 等人提出 word2vec（Mikolov *et al*，2013），作者发现了平移不变现象。受平移不变现象的启发，翻译模型 TransE（Bordes，2013）将三元组 S（h,r,t）中的关系 r 看作是某种平移向量，因此头实体 h 经过平移后得到尾实体 t。模型如图 2.5 所示；

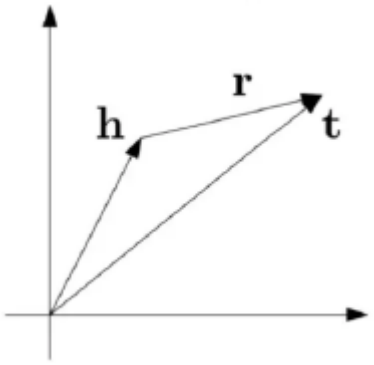

图 2.5 TransE 模型

Figure 2.5 TransE model

翻译模型 TransE 由于其参数少、计算复杂度低而高效，得到了广泛关注。然而由于 TransE 模型在处理一对一关系时的良好表现并没有在处理一对多、多对一或者多对多复杂关系的时表现出来。模型 TransH 为了解决 TransE 在处理复杂关系时出现的问题，提出实体不再单一只有一种表示，而是不同关系下有不同的表示。模型 TransR 提出实体和关系应彼此远离处于不同的语义空间，一个实体是多种属性的综合体，然而不同关系关注实体的不同属性，应该将实体和关系嵌入到不同的空间中，在对应的关系空间中实现向量化表示。

TransD 模型在研究学习 TransR 模型的同时，发现 TransR 模型尚有不足之处：TransR 模型重在将实体的不同关系投影到不同的空间，将每一关系有且仅代表一种语义关系，然而有时一关系可能代表不同的语义。而且 TransR 模型的参数在关系投映矩阵的引入后增加剧烈，从而导致计算开销变大。

针对 TransR 模型的不足之处，TransD 模型对 TransR 模型进行了改进，认为不同的实体应映射到不同的语义空间中，且通过将矩阵运算转变成向量运算，大大减少了计算量。TransD 的模型如图 2.6 所示；

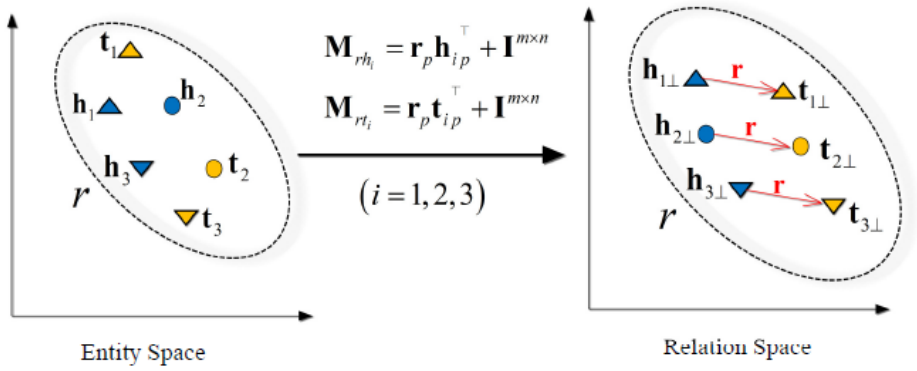

图 2.6 TransD 模型

Figure 2.6 TransD model

令 h_p、r_p、t_p 为 h，r，t 的映射向量，根据映射向量得到头、尾实体到关系空间的头、尾投影矩阵，$\mathbf{M}_{rh}$、$\mathbf{M}_{rt}$ 分别是实体 h,t 的映射矩阵，$\mathbf{h}_{ip}, \mathbf{t}_{ip}(i=1,2,3)$ 及关系 r_p 为投影向量，$\mathbf{h}_{\perp}$、$\mathbf{t}_{\perp}$ 分别为头尾实体的投影向量。

$$\begin{aligned} \mathbf{M}_{rh} &= \mathbf{r}_p \mathbf{h}_p^{\mathrm{T}} + \mathbf{I}^{m\times n} \\ \mathbf{M}_{rt} &= \mathbf{r}_p \mathbf{t}_p^{\mathrm{T}} + \mathbf{I}^{m\times n} \\ \mathbf{h}_{\perp} &= \mathbf{M}_{rh}\mathbf{h} \\ \mathbf{t}_{\perp} &= \mathbf{M}_{rt}\mathbf{t} \end{aligned} \tag{2-9}$$

其损失函数如下：

$$f_r(\mathbf{h},\mathbf{t}) = -\left\|\mathbf{h}_{\perp} + \mathbf{r} - \mathbf{t}_{\perp}\right\|_2^2 \tag{2-10}$$

文献（Dettmers *et al*，2018）设计一种参数高效的、计算快速的2D卷积神经网络用来做知识图谱的表示学习。文献（Sun *et al*，*2019*）将每个关系定义为在复矢量空间中从源实体到目标实体的旋转，从而能够建模和推断各种关系。

本文采用TransD模型，将菜品知识图谱中的实体与关系投影到低维的向量空间。

2.2 推荐算法

随着互联网信息数量的成倍上升，在浩如烟海的数据中找寻真实且自己需要的信息将会变得愈发困难。面对数据爆炸问题，如何寻找用户需要的资源成为一大难题。专家和学者纷纷提出自己的想法，然而大多都无功而返，直到推荐算法的出现，这一

难题才真正面临着解决。推荐算法通过各种推荐技术向用户推荐他们真正需要的信息。近年来，推荐算法得到了长足的发展。

2.2.1 协同过滤算法

协同过滤算法是又被叫作基于用户行为的推荐算法，是推荐领域中使用最广泛的算法，其算法核心思想是仅通过用户过往交互信息挖掘出用户潜在喜欢的信息或商品，不需要预先获得用户或物品的特征值。协同过滤算法大致可分为基于用户(User-based)的协同过滤算法和基于物品(Item-based)的协同过滤算法（BREESE *et al*，2013）。

2.2.1.1 基于用户的协同过滤算法

基于用户的协同过滤算法在 1992 年被提出，其基本思想是据给用户推荐和他兴趣相似的用户感兴趣的物品（王国霞等，2012）。算法计算步骤主要有两部分构成：（1）寻找兴趣与丙相似的用户集，（2）将用户集中用户感兴趣的而丙没有过交互信息的是物品推荐给丙。其工作原理如图 2.7 所示。

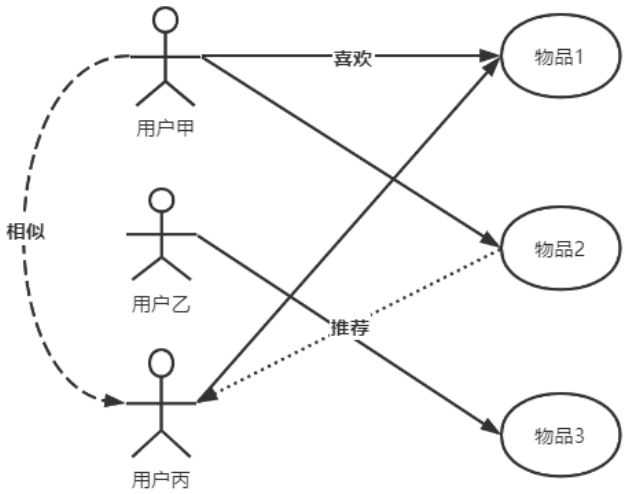

图 2.7 基于用户的协同过滤

Figure 2.7 User-based collaborative filtering

基于用户的推荐算法可以反映用户所在的一个群体中大家都喜欢的商品，足够社交却个性化不足。因为用户的兴趣不可能一直不发生变化，当用户兴趣发生改变时，此时的基于用户的推荐算法就面临着数据缺失问题。

2.2.1.2 基于物品(Item-based)的协同过滤算法

基于物品的协同过滤算法是目前使用最广泛的算法，很是受到工业界的追捧以及学术界的深入研究。其基本思想是为用户推荐他们之前喜欢的物品的相似物品。算法的计算步骤主要有两部分组成：（1）计算物品相似的物品。需要注意的是，这里的相似并非计算物品之间的相似度，而是依据“喜欢 A 物品的人大多也喜欢物品 B，因此物品 A 和 B 相似”这种假设。如图 2.5 所示，甲和乙都喜欢物品 1 和 2，而丙喜欢物

品 1，假设喜欢 1 的都喜欢 2，所以把物品 2 推荐给丙。（2）计算物品评分。其工作原理如图 2.8 所示：

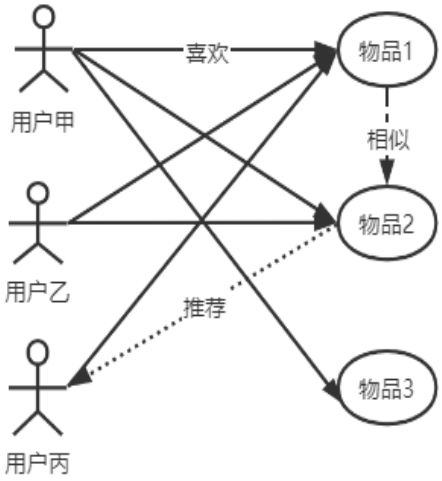

图 2.8 基于物品的协同过滤

Figure 2.8 Item-based collaborative filtering

Item-CF 推荐结果更个性化且具有可解释性，能够反应用户的兴趣爱好，被广泛的应用到电商系统中。Item-CF 当面对大量物品时会出现计算效率缓慢，需要一定的缓冲时间，同时 Item-CF 也存在新用户冷启动的问题。

2.2.2.3 相似度算法

当推荐算法为用户推荐物品时，首先要获得用户的相关信息，从而计算与用户相似度高的用户或物品推荐给用户，因此在推荐算法中计算相似度是一个重要步骤。常见的相似度算法有皮尔逊相似系数、余弦相似度和欧式（Euclidean）距离。

皮尔逊相关系数是由协方差除以两个变量的标准差而得到的。当两个变量的方差都不为 0 时，相关系数才有意义，其取值范围为[-1,1]。

在推荐算法中，皮尔逊相似性度量法主要应用在计算物品与物品之间或者用户与用户之间的相似性度量，其公式如 2-11 所示：

$$\mathrm{sim}(i,j)=\frac{\sum_{u\in U}\left(R_{u,i}-\overline{R}_i\right)\left(R_{u,j}-\overline{R}_j\right)}{\sqrt{\sum_{u\in U}\left(R_{u,i}-\overline{R}_i\right)^2}\sqrt{\sum_{u\in U}\left(R_{u,j}-\overline{R}_j\right)^2}} \quad (2\text{-}11)$$

$\overline{R}_i$ 表示对物品 i，j 都评价过的用户集对项目 i 的评分均值，$\overline{R}_j$ 表示均对物品 i，j 评价过的用户集对项目 j 的评分均值。

余弦（Cosine）相似度又叫作余弦距离，其用向量空间中两向量夹角的余弦值衡

量两向量的相似程度。向量的夹角越小，则对应样本的相似度就越高；向量的夹角越大，则相对应样本的相似度就越低。这叫做“余弦相似性”。余弦相似度的计算如公式 2-13：

$$\mathrm{sim}_{\mathrm{cosine}}(X,Y)=\cos\theta=\frac{X\mathrm{g}Y}{\|X\|\ \|Y\|}=\frac{\sum_{i=1}^{n}\left(x_i\times y_i\right)}{\sqrt{\sum_{i=1}^{n}x_i^2\times\sum_{i=1}^{n}y_i^2}} \tag{2-12}$$

欧式（Euclidean）距离可以用来计算两个用户之间的距离，假设用户 A=（U_1,V_1），用户 B= （U_2，V_2），那么它们之间的欧氏距离如公式 2-13 所示；

$$sim(A,B)=\sqrt{(V_1-V_2)^2+(U_1-U_2)^2} \tag{2-13}$$

2.2.2 基于内容的推荐算法

基于内容的推荐算法开始是应用在信息检索当中，其主要利用用户相关信息（profile）及用户操作行为来构建推荐算法模型，为用户提供推荐服务。用户相关信息是指人口统计学信息(如年龄、性别、偏好、地域、收入等等)。用户操作行为可以是评论、收藏、点赞、浏览、点击、购买等。基于内容的推荐算法依赖于用户自身的行为为用户提供推荐，不涉及到其他用户的行为。

基于内容的推荐算法基本思想是为用户推荐与他感兴趣内容相似的物品，比如用户喜欢喜剧电影，直接给他推荐电影《憨豆先生》 。这个推荐过程既考虑了用户的兴趣，也结合考虑了电影的内容。因此不需要用户提供行为数据就可为用户推荐感兴趣的内容，可以很好地解决新用户冷启动的问题。

由于基于内容的推荐算法在给用户推荐物品时，该物品往往已经与用户有着很高的相似度，推荐的物品都是与用户较为相关且熟悉的，用户很难在推荐结果中获得惊喜。

2.2.3 基于知识图谱的推荐算法

随着互联网的发展，越来越多的蕴含用户行为信息的数据产生并被获取，包括音频、视频以及标签等多元异构信息，混合推荐算法通过使用这些辅助信息，可以有效地缓解传统推荐算法中数据稀疏与冷启动问题。知识图谱中蕴含的丰富实体与实体之间的关系，能够一定程度的解决传统协同过滤算法存在的问题。

基于知识图谱的推荐模型大多以现有的推荐模型为基础，通过知识图谱中将项目、用户等实体的结构化知识描述的更加细粒度化，利用知识推理得出更加深层次的信息加入到推荐模型中，使得推荐结果更具有多样性和解释性。根据推荐过程中知识图谱表示形式，本文将基于知识图谱的推荐方法大致分为三类：基于本体的推荐方法、基于开放链接数据(Linked Open Data，LOD)的推荐方法以及基于知识图谱特征学习

（Knowledge Graph Embedding）的推荐方法（常亮等，2019）。

（1）基于本体的推荐算法

基于本体的推荐算法本质上还是依据传统算法的核心思想，但是与传统算法不同的是，在某些具体细节上的实现，还是利用了本体技术的优势所在，并通过这些优势对推荐算法进行一定程度的改进，进而尝试解决传统推荐算法中存在的问题。Moreno等设计并实现了基于本体的知识图谱和协同过滤的推荐算法混合（Moreno *et al*，2013）。把旅游景点和周边信息加到图谱中，，然后使用基于内容的推荐算法和协同过滤算法对信息进行过滤，使用 Pearson 进行相似度计量，最终实现了较好的推荐效果。其优点在于：可以清晰准确的描述概念之间的上下级关系，把数据之间的关联性更进一步牢牢增强，同时也能够更加精细化的分析用户的喜好。缺点在于：在构建本体时前期投入费时费力，后期可扩展性不强，且因为某些领域的领域性太强，需要领域专家通过手工来设计构建本体文件，开销太大。

（2）基于开放连接数据（LOD）的推荐算法

同基于本体的推荐算法一样，基于 LOD 的推荐算法本质上也是利用传统推荐算法的核心思想，只不过其可以将丰富的语义关系融入到现有方法中，从而对现有方法打来提升。基于 LOD 推荐算法重点计算用户的偏好以及物品之间的小相似度，通过利用 LOD 中的丰富语义，可以很好地刻画用户画像或者物品画像，精细化的对用户或物品间进行相似度衡量。Oramas 等（Oramas *et al*，2017）通过建立知识图谱来完成声音和音乐的推荐，工作详细讲述了如何创建知识图谱并将之与混合推荐算法相结合完成声音和音乐的推荐任务。基于 LOD 的推荐算法借助数据之间的关联，能够形成很强的逻辑能力并且可以具备一定程度的推理能力，从而发现一些不易发现的语义关联，从而提高推荐效果的准确率与可解释性。然而由于过度依赖外部数据库，导致外部数据的质量很大程度上决定了其推荐效果，同时由于外部知识库的数量庞大，在推荐过程中会存在计算开销过大的问题。

（3）基于知识图谱表示学习（Knowledge Represent Learning，KRL）的推荐算法

现阶段，基于知识图谱表示学习的推荐系统也是大多与传统的推荐算法相结合。通过 KRL 技术可以将知识图谱中的实体和关系嵌入到低维稠密的向量空间中，然后通过计算实体向量在低维的向量空间中的距离衡量实体之间的联系，大大降低了计算开销。

目前，将知识图谱表示学习应用到推荐系统中常用两种方式——依次学习和联合学习。

①依次学习。如图 2.9 所示，为了将知识图谱中的特征引入，首先需要提取知识图谱中特征得到实体向量和关系向量，并将其输入到推荐系统中。紧接着推荐系统对知识图谱表示学习得到的实体向量和关系向量与推荐算法原有的输入信息一并处理，

得到关于用户以及物品的特征向量。

依次学习的优点在于知识表示学习模块和推荐系统模块相互独立，通过知识表示学习方法得到关于实体和关系的向量后可直接用于推荐系统的训练，节约大量用于训练的时间，尤其当知识图谱规模较大时，依次学习的优点就更加明显。同时，依次学习的优点注定它的缺点存在：由于知识表示学习模块和推荐系统模块的相互独立，无法做到端到端的训练，在缺乏推荐系统模块的监督下，通过知识表示学习得到的特征向量是否适合推荐任务还需要进一步的实验来进行验证。

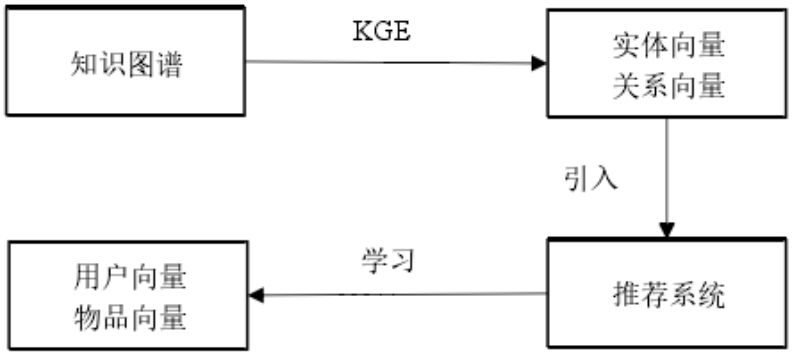

图 2.9 依次学习

Figure 2.9 Learning in turn

②联合学习。如图 2.10 所示，不同于依次学习，联合学习采用端到端的训练方式，将知识图谱和推荐算法一同训练，开始阶段对知识图谱的实体和关系向量随机复制，训练过程中将原始数据、知识图谱的向量特征和推荐模型与推荐系统相结合，并不断地调整实体和关系的特征值。联合学习的代表模型有 RippleNet（王宏伟等，2018），在 RippleNet 中推荐系统的输入为知识图谱的原始信息，并不对知识图谱进行表示学习，而是充分利用了知识图谱的图结构，使得数据在图结构中传播就像水波纹一样层层荡开，而周围的数据也都受到影响。美中不足的是，RippleNet 推荐系统仅利用了知识图谱中的信息，没有使用推荐系统中基本的用户物品评分信息等，推荐结果很难符合用户的喜爱。。

联合学习的优点在于能够将知识图谱和推荐系统结合进行端到端的学习，根据学习状况进行调整得到更好的推荐结果。但联合学习使得训练开销变大，尤其是在面对一些大型知识图谱时，这个问题更加的凸现出来，另外联合学习将两个模块一同训练，最终的目标函数如何结合以及训练参数的调整都需要实验才能确定。

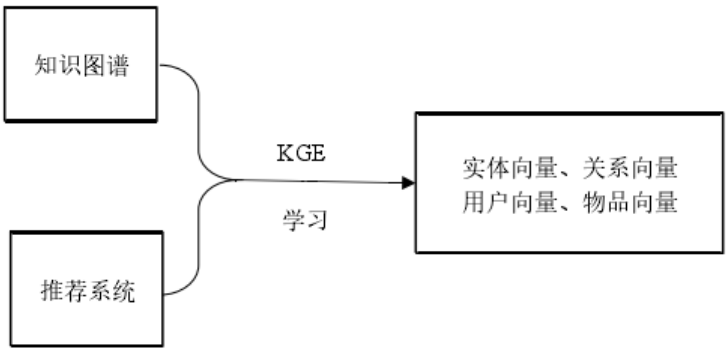

图 2.10 联合学习

Figure 2.10 Joint learning

综上所述，基于知识图谱的推荐方法大致经历了三个阶段：基于本体的推荐、基于 LOD 的推荐以及基于知识图谱表示学习的推荐，知识图谱与推荐系统相结合的方法也越来成熟。本文拟采用依次学习的方法，将菜品知识图谱中的语义信息通过知识表示学习应用到推荐系统中，与用户信息相结合，弥补传统推荐算法存在的冷启动以及数据稀疏问题。

2.3 本章小结

本章主要对本文研究过程中涉及的相关理论知识和技术做了总结和概括。首先对知识图谱构建过程中涉及的知识抽取技术、知识表示学习技术以及知识存储工具进行了总结概括，之后对推荐算法中的常用算法进行了分析和介绍。本章介绍的方法与技术，是后续工作展开的坚实基础。

3 菜品领域知识图谱构建

为了使推荐系统的信息来源更加丰富，不单单只考虑用户与物品的交互行为数据，而是把菜品本身的信息也加入到推荐系统，从而使得菜品推荐结果更具有可解释性和可信度，本章以菜品知识为例提出一种构建领域知识图谱的方法，以便为后续的菜品研究奠定基础。

3.1 总体框架

针对菜品领域文本，总结和分析已有健康饮食领域知识图谱（Zhao *et a*l，2014），本文的菜品知识图谱构建的总体框架如图 3.1 所示。

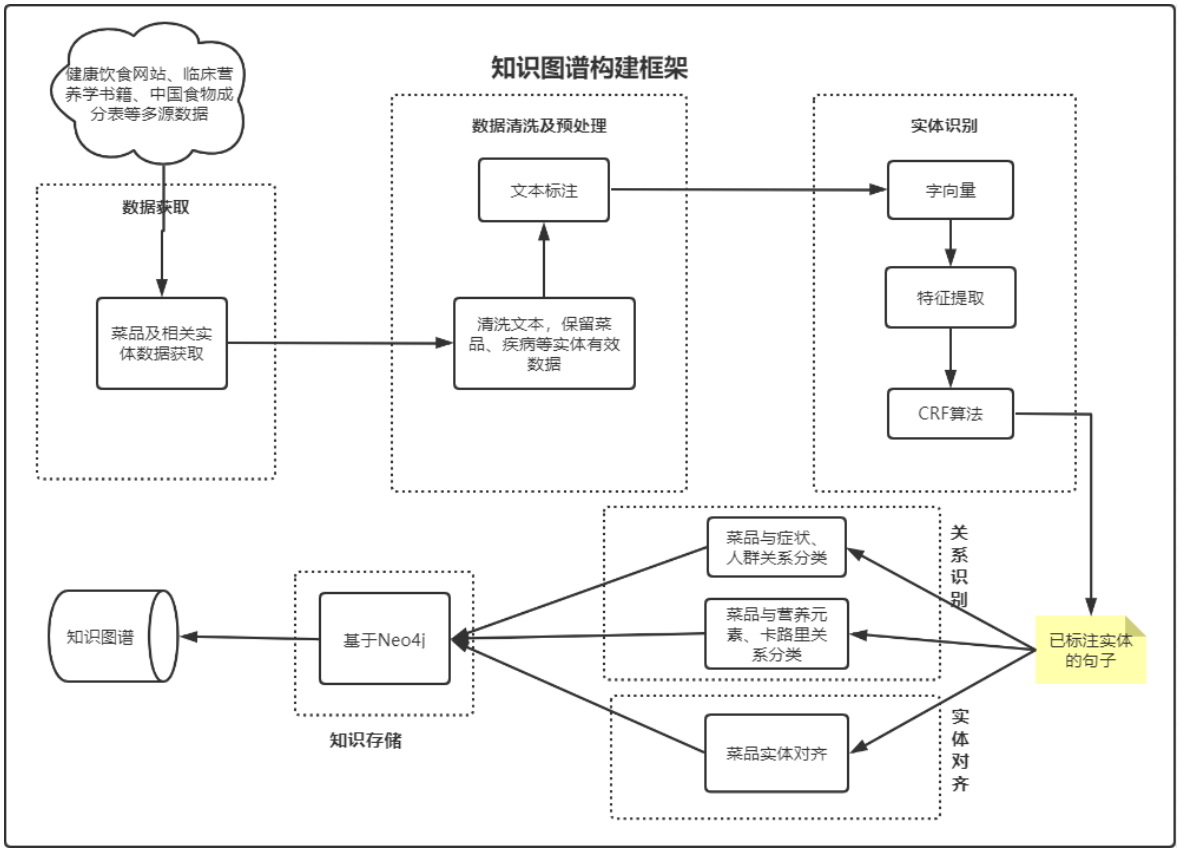

图 3.1 知识图谱构建总体架构

Figure 3.1 The overall architecture of knowledge graph construction

整个构建流程包括了六个模块：原始数据获取、数据清洗与预处理、实体识别、关系识别、实体对齐以及知识图谱存储。针对各模块的具体作用及流程描述如下：

（1） 原始数据获取：针对三个健康饮食相关网站（美食杰，薄荷网，豆果网）使用 Scrapy 爬虫获取菜品的名称、简介、功效等半结构化和非结构化文本；从《常见疾病饮食忌讳》（陈思，2012）、《临床营养学》（蔡威，2012）和中国食物成分

表（杨月欣，2005），从中可以获取半结构化和结构化数据。

（2） 数据清洗与预处理：保留数据中的有效信息，按照文本结构使用正则匹配提取菜品及其相关基本属性，同时将不同类型的数据进行标注以便进一步使用。

（3） 实体识别：抽取文本数据中的实体，如：疾病、菜品、营养元素等。本文通过使用 BiLSTM-CRF 算法进行实体识别并设计对比实验，通过对多种算法的准确率、召回率进行评价来验证实体识别的效率。

（4） 关系识别：在菜品与疾病、人群实体的关系分类中，通过计算菜品文本投映到多维空间向量，利用机器学习算法将菜品实体间的关系分为禁忌食用和适宜食用两种关系。最后通过多次实验评估本文的关系识别准确度；

（5） 实体对齐：利用菜品的相关属性信息判断不同数据来源的同一菜品实体是否对齐。本文利用余弦相似度计算菜品之间的相似性得到特征向量，再利用机器学习算法将菜品间的实体关系分为“相同”以及“不同”两种，实际上将实体对齐转化成二分类问题。

（6） 知识存储：该模块将上述步骤抽取出的实体及关系使用 Neo4j 图数据库进行存储，完成菜品知识图谱的构建。

3.2 数据获取及预处理

本文数据来自多个数据源，其中包括三个食谱相关的垂直网站（美食杰，薄荷网，豆果网），选择使用 Python 的分布式爬虫框架 Scrapy 对其菜品构成及营养学文本进行爬取。Scrapy 是一个适用于提取网页中结构性数据的应用程序框架，由引擎、调度器、下载器以及管道等不同模块组成，可以对网页数据进行分布式获取，从中可以获得半结构化和非结构化数据；还有《常见疾病饮食忌讳》、《临床营养学》和中国食物成分表，从中可以获取半结构化和结构化数据。经提取后本文获得菜品相关的语句 7182 条，以及从《临床营养学》和《饮食忌讳》中摘取的语句 4260 条，通过正则匹配等方法去除无效文本以及人工筛选后，共获得可用文本 10374 条语句作为实验语料。

由于数据多源，通过使用正则表达式制定不同的规则从文本中提取相关实体及其属性，但是非结构化的自然语言数据中隐藏着许多相关实体，对该菜品的描述非常重要，却没有被充分挖掘，因此本文采用基于 BiLSTM-CRF 命名识别方法对非结构化数据进行实体识别。本文利用清华大学提出的开源软件 THULAC 对文本数据进行分词处理（孙茂松等，2016），实现文本数据中的词与实体相对应。同时人工标注部分词性以提高菜品领域实体识别的准确率。

（1）命名实体分类

在数据清洗过程中，根据数据的特点，本文对构建的知识图谱定义了一些重要的

实体分类：食材、菜品、营养元素、卡路里、人群、疾病共六类，与食物和健康饮食主题密切相关。如表 3.1 所示分别对每个实体类别进行定义：

表 3.1 相关实体定义及举例

Table 3.1 Definition and examples of entity type related entities

实体	定义	举例
食材	菜品的组成成分；	排骨、牛腩和土豆等；
菜品	各类品种的菜肴；	西红柿炒鸡蛋、红烧排骨等；
营养元素	人体必需的微量物质；	碳水化合物、碘、钙等；
卡路里	一种热量单位；	热量来自于碳水化合物、脂肪、蛋白质。 例如：碳水化合物产生热能= 4 大卡/克；
人群	不同人按照依据某种共有特征分为一组，比如按照年龄、性别、所患疾病等特征；	老年人、男性、痛风患者等；
疾病	身体常见的不适病症；	感冒，高血压等；

本文对菜品做出了分类。根据菜品的食材、菜名等把菜品分为 22 种口味（例如：酸辣、清淡、香咸等），这些作为菜品实体的属性，以便结合人群及其身体状况进行推荐。同时通过查阅中国食物成分表，对菜品所含卡路里以及营养元素等信息进行了等级划分，共分为五个等级，其中第 1、2 等级代表含量较低，第 3、4 等级代表含量较高，第 5 等级代表含量很高，将其作为卡路里和营养元素的相关属性存储到知识图谱当中，以便于更精确化的描述菜品的具体信息，对于刻画菜品之间相似度极有帮助，进而有助于给用户推荐健康菜品。

同时，也给出了菜品属性定义，如表 3.2 所示：

表 3.2 菜品及其属性定义

Table 3.2 Definition of dishes and their attributes

属性值	描述	举例
id	菜品在图数据库中的位置	鱼香肉丝@id:349
名称	菜品的在图数据库中的名称	糖醋鱼@name:糖醋鱼
口味	菜品的味道	辣子鸡@taste:麻辣
烹饪方式	菜品的具体做法	酸辣土豆丝@method:“1. 将土豆削皮，切细丝…炒熟即可。”

（2）相关实体间的语义关系

① 禁忌食用

包括:“人群-禁忌食用-菜品”、“人群-禁忌食用-食材”、“疾病--禁忌食用-菜品”、“疾病--禁忌食用-食材”。北齐 颜之推 《颜氏家训·养生》:“若其爱养神明，调护气息，慎节气息，均适寒暄，禁忌食饮，吾无间然。”饮食因人而异，不同的人可能因为身体状况存在某些食物过敏亦或影响身体状况，出现不可预知的症状。确定禁忌食用关系，为将来的菜品推荐生成菜品提供安全性保障。

a 人群同菜品（食材）的禁忌食用关系

例：在营养学书籍中关于绿豆的描述:“老人、病后体虚者不宜食用绿豆。”

根据句子中关于“不宜食用”的食用关系描述，得到三元组关系如下所示：

（老人，禁忌食用，绿豆）

（病后体虚者，禁忌食用，绿豆）

b 疾病同菜品（食材）的禁忌食用关系

例:《临床营养学》中，关于冠心病有如下描述:“冠心病忌食红烧肉。”

根据句子中关于“忌食”的食用关系描述，得到三元组关系如下所示：

（冠心病，禁忌食用，红烧肉）

② 适宜食用

包括:“人群--适宜食用-菜品”、“人群-适宜食用-食材”、“疾病--适宜食用-菜品”、“疾病--适宜食用-食材”，指当人体食用某种食物有益于或对健康有益。适宜食用关系的设立是为让菜品推荐系统在生成菜品时，能够考虑更多的影响因素，得到更加健康的饮食推荐。

a 疾病与菜品（食材）的适宜食用关系

例如:《临床营养学》中关于胃病的营养治疗有如下描述:“胃病忌吃“生冷硬”，可以多吃小米粥。”

根据句子中关于“可以多吃”的食用关系描述，得到三元组关系如下所示：

（胃病，适宜食用，小米粥）

b 人群与菜品（食材）的适宜食用关系

例如：在营养学书籍中，对高血压患者饮食有如下描述:“燕麦具有降胆固醇和降血脂作用…适合高血压和糖尿病人对食疗的需要。”

根据句子中关于“适合”的食用关系描述，得到三元组关系如下所示：

（高血压患者，适宜食用，燕麦）

（糖尿病患者，适宜食用，燕麦）

③ 营养构成

例如:在食谱文本中，对清蒸鲈鱼有如下描述:“清蒸鲈鱼富含蛋白质、脂肪、钙、维生素等成分,具有很高的营养价值。”

根据句子中关于“富含”的营养构成关系描述，得到三元组关系如下所示：

（清蒸鲈鱼，营养构成，蛋白质）

（清蒸鲈鱼，营养构成，脂肪）

（清蒸鲈鱼，营养构成，维生素）

④ 组成成分

例如：在食谱文本中，对鱼香肉丝的用料有如下描述：“鱼香肉丝的用料有猪肉、胡萝卜、黑木耳、葱末、姜末等。”

根据句子中关于“用料”的组成关系描述，得到三元组关系如下所示：

（鱼香肉丝，组成成分，猪肉）

（鱼香肉丝，组成成分，黑木耳）

（鱼香肉丝，组成成分，胡萝卜）

⑤ 卡路里含量

卡路里含量关系是指根据菜品组成成分及口味等信息判定其含有的卡路里含量。在构建菜品知识图谱时，本文把卡路里进行了等级划分并存入到知识图谱当中，以便于更等级化、精确化的描述菜品的具体信息，对于刻画菜品之间相似度极有帮助，进而有助于算法推荐健康菜品给用户。

例如：

（香菇炒牛肉，卡路里含量，4）

（手撕包菜，卡路里含量，2）

⑥ 适量进食

例如：在营养学中，对糖尿病人群有如下描述：“糖尿病人群为了减少脂肪，可以选择含蛋白质的食物适量代餐。”

根据句子中关于“适量代餐”的适量进食关系描述，得到三元组关系如下所示：

（糖尿病人群，适量进食，蛋白质）

实体关系如图 3.2 所示：

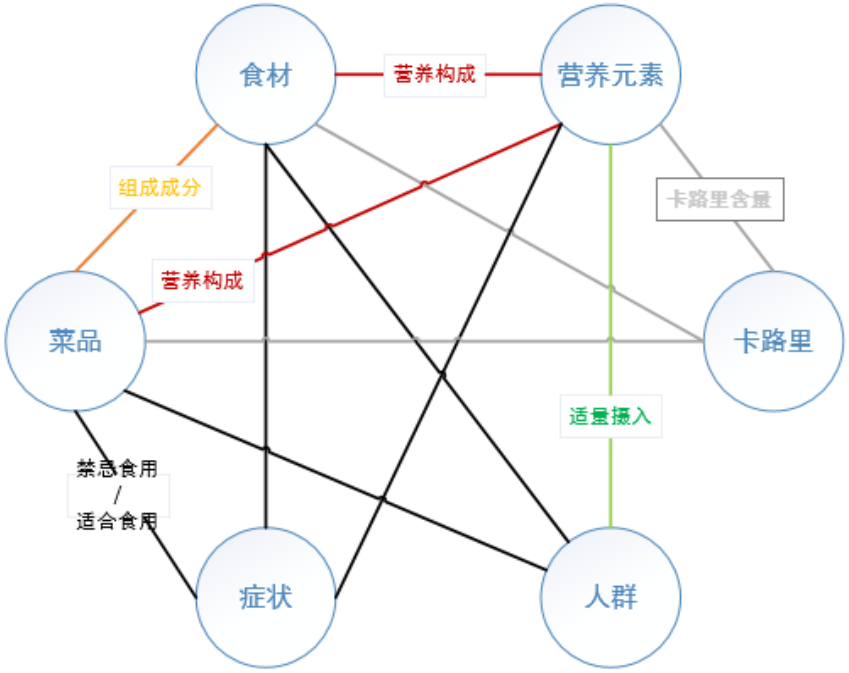

图 3.2 菜品知识图谱的实体关系

Figure 3.2 Entity relationship of dishes knowledge graph

其中“橙色线条”代表“组成成分”关系；“红色线条”代表“营养构成”关系；“黑色线条”代表“禁忌食用/适合食用”关系；“蓝色”代表“易患”关系；“灰色”代表“卡路里含量”关系；“绿色”代表“适量摄入”关系。

3.3 实体抽取

3.3.1 训练数据集标注

在 NLP 领域，通常将一个句子视为一个序列，句中的词或单个字符作为一个元素，序列标注的任务是对每一个元素选择一个预定义的标签进行标注。

目前通常将命名实体识别任务转化为序列标注，常用的序列标注方法为 BIO 序列标注法；该方法将句子中的每一个元素标注为“B-EntityTag”、“I-EntityTag”或“O”；其中“B-EntityTag” 表示此元素在句中处于 EntityTag 所指类型实体的开头位置，“I-EntityTag”则表示该元素处于 EntityTag 所指类型实体中间位置；“O”表示该元素在句中不属于任何实体成分。

本文为了提高菜品领域实体识别的准确率，将 3.2 节中的数据进行序列标注，以生成训练数据集。实体共分为 6 类，分别为食材类（INGREDIENT）、菜品类（DISH）、疾病类（DISEASE）、卡路里（CALORIE）、人群（CROWD）以及营养元素（NUTRIENT），根据 BIO 标注的规则，共有“B- INGREDIENT”、“I- INGREDIENT”、“B- DISH”、“I- DISH”以及“O”等 6 类 13 种标签，标注后的示例如图 3.3 所示。

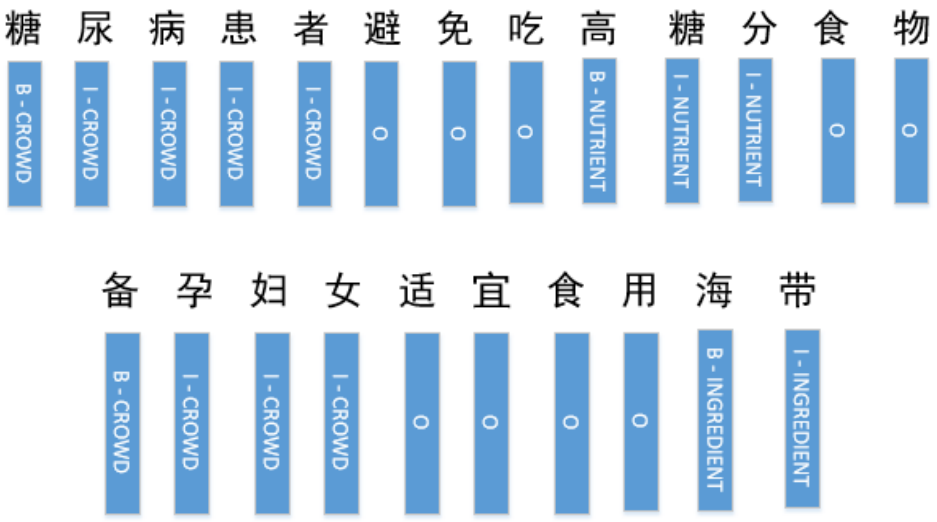

图 3.3 序列标注结果

Figure 3.3 Sequence labeling results

3.3.2 实验与分析

本文使用 2.1.1 节中介绍的 BiLSTM-CRF 进行命名实体识别的模型训练，共选取了 10374 条包含菜品饮食相关的语句，经过 BIO 标注后，实验时以 6:2:2 的常用比例划分训练集、测试集以及验证集；数据经由模型的嵌入层（Embedding）向量化后作为模型的输入，以供其进行学习并对实体标签进行预测。

本文模型的搭建环境基于 Python3.6 以及 Tensorflow1.2.0 实现，训练过程中对模型的参数设定及解释如表 3.4 所示。

表 3.4 模型训练参数

Table 3.4 Model training parameters

参数名	参数设置	参数详解
lstm_dim	100	前后 LSTM 模型隐藏层数
clip	5	梯度阈值
batch_size	64	批处理大小
dropout	0.5	Tensorflow 防治过拟合机制
lr	0.001	初始学习速率
optimizer	AdamOptimizer	Tensorflow 优化器
epoch	50	

本文通过设计对比实验以对菜品命名实体识别进行评价，对比实验为基于 CRF 的命名实体识别和基于 LSTM 模型的命名实体识别。其中 CRF 模型采用的是 CRF++0.58，LSTM 命名实体抽取模型基于 TensorFlow 1.12.0 实现。

针对本文训练的命名实体识别模型，分别采用准确率（Precision）、召回率（Recall）以及 F1 值（F1-Measure）三个指标来进行评价，其计算公式分别为（3-1）、（3-2）和（3-3）：

$$\text{Precision} = \frac{N_r}{N_{all}} \times 100\% \tag{3-1}$$

$$\text{Recall} = \frac{N_r}{N_{predict}} \times 100\% \tag{3-2}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \times 100\% \tag{3-3}$$

其中N_r表示模型预测正确的结果数量，N_{all}表示模型进行预测的样本总数，$N_{predict}$则表示所有预测结果为正的样本总数。

经过 25 轮的迭代训练后，实验结果如表 3.5 所示：

表 3.5 模型的综合效果

Table 3.5 comprehensive effect of the model

Model	P(%)	R(%)	F1(%)
CRF	85.16	78.12	81.48
LSTM	89.8	83.06	86.29
BiLSTM-CRF	90.8	84.9	87.75

模型的综合准确率、召回率和$F1$值分别达到了 90.8%、84.9%和 87.75%，相比 CRF 模型在 F1 值上高了 6.2%，相比 LSTM 模型在 F1 值上提高了 1.5%，在三个模型中效果最佳。该结果证明了本文使用菜品文本数据作为训练数据，搭建模型进行训练的方法取得了较好的效果，同时模型对菜品领域的实体有较好的识别能力。具体针对本文六类实体的训练 F1 值如表 3.6 所示：

表 3.6 模型在六类实体的 F1 值

Table 3.6 F1 value of the model in six types of entities

实体类别	CRF	LSTM	BiLSTM-CRF
菜品	84.26	86.73	91.35
人群	75.36	81.6	80.3
疾病	80.33	87.26	85.11
食材	76.82	83.21	89.89
营养元素	87.59	89.78	89.51
卡路里	84.54	89.18	90.36

该结果证明，BiLSTM-CRF 模型在菜品、食材以及卡路里等实体上取得更好的实验效果，LSTM 在识别人群实体、疾病实体、营养元素实体三类上效果最优。

3.4 关系抽取

上一节中，本文介绍了菜品相关实体识别的工作，关系抽取是构建知识图谱的另一个重要步骤，本节将对各类实体间的关系抽取方法进行详细介绍。同时，由于本文菜品实体对齐工作是通过判断两两菜品是否相同或不同来达到实体对齐效果，因此，可将实体对齐工作当成二分类任务来处理：根据实体间的属性来抽取特征，进而判断菜品是否相同。

3.4.1 实体关系抽取

在菜品与疾病、人群实体的关系分类中，通过计算菜品文本投映到多维空间向量，利用机器学习算法将菜品实体间的关系分为禁忌食用和适宜食用两种关系。最后通过多次实验评估本文的关系识别准确度。

针对菜品与疾病、人群间的关系，本文选择使用曹明宇等人提出的基于 TF-IDF 算法和词向量的相似度计算方法（曹明宇等，2019），并将得到的 TF_IDF 向量输入到机器学习模型中进行分类提取。将菜品关系识别算法具体描述如下：

（1）分词处理 3.2 节中获取的文本数据，并分别计算分词后每个单词w在文本中出现的频率，即词频TF_w，其计算公式如(3-4)所示；

$$TF_w = \frac{单词w在文本中出现的次数}{文本分词后的单词总数} \tag{3-4}$$

（2）分别对单词 w 的逆向文件频率IDF_w进行计算，逆向文件频率反映了单词在文本中的区分能力，该值的大小与其区分能力成正比；IDF_w的计算公式如(3-5)所示；

$$IDF_w = \log\left(\frac{文本分词总数}{包含单词w的文本+1}\right) \tag{3-5}$$

（3）结合词频以及逆向文件频率计算单词 w 在分词文本中的权重TF_IDF_w，其计算公式如(3-6)所示；

$$\mathrm{TF}_{IDF_w} = TF_w * IDF_w \tag{3-6}$$

（4）本文采用支持向量机（SVM）（Joachims，2005）、朴素贝叶斯（NB）算法（Dai *et al*，2007）、最近邻（KNN）算法和 LSTM（Hochreiter *et al*，1997）进行特征提取。NB 算法、SVM 算法和 KNN 算法的输入为步骤（3）得到的IDF_w向量，LSTM 网络的输入为词汇向量随机初始化入。上述三种算法输出“菜品-禁忌食用-人群/疾病”和“菜品-适宜食用-人群/疾病”的概率，二者概率进行比较，概率大的即为从句子提取出的关系。

本文使用 3.2 节获取的文本数据进行菜品与疾病/人群关系抽取，共选取 2500 条文本，以 6:2:2 的常用比例划分为训练集、测试集以及验证集。对训练集文本数据进

行标注，以提高关系抽取模型的准确率。

本文模型的搭建环境基于 Python3.6 以及 Tensorflow1.2.0 实现。通过 Scikit-learn(sklearn)库实现支持向量机（SVM 算法）以及贝叶斯（NB）算法，Scikit-learn(sklearn)库是基于 Python 的函数库，对常见的机器学习算法进行了封装处理，可以直接调用函数处理数据。

LSTM 网络输入是步骤（1）中经过分词处理后，采用 word2vec 预训练数据初始化，输出为菜品与疾病/人群之间的概率。训练过程中对模型的参数设定及解释如表 3.7 所示。

表 3.7 LSTM 模型训练参数

Table 3.7 training parameters of LSTM model

参数名称	参数值	参数详解
batch size	128	批处理大小
lstm_dim	200	前后 LSTM 模型隐藏层数
epochs	10	训练次数
dropout	0.5	Tensorflow 防治过拟合机制 dropout 函数参数
lr	0.001	初始学习速率
clip	5	梯度阈值
activation function	Softmax	激活函数

针对本文的菜品实体抽取任务，采用准确率（Precision）、召回率（Recall）以及F_1值三个指标进行评价，实验结果如表 3.8 所示。

表 3.8 菜品实体与疾病、人群实体之间的关系分类效果

Table 3.8 Classification effect of relationship between dish entity, disease and crowd entity

任务	分类器	P	R	F1
（菜品，禁忌食用，疾病/人群）	SVM	0.98	0.92	0.96
	NB	0.87	0.92	0.89
	KNN	0.94	0.94	0.94
	LSTM	**0.99**	**0.94**	**0.97**
（菜品，适宜食用，疾病/人群）	SVM	0.95	**0.93**	0.94
	NB	0.67	0.84	0.75
	KNN	0.92	0.91	0.91
	LSTM	**0.99**	0.92	**0.95**

可以看出，LSTM 网络在菜品与疾病/人群关系抽取任务中综合表现最好，在二分类中F_1值分别达到了 97%和 95%。该结果证明了本文对文本数据进行标注作为训练

数据，搭建模型进行训练的方法取得了较好的效果，同时模型对菜品领域关系抽取有较好的识别能力，为构建菜品知识图谱提供了有力的支撑。

3.4.2 菜品实体对齐

本质上，菜品实体对齐就是根据菜品属性对两两菜品进行匹配，看是否为同一菜品。通过利用菜品的属性，如口味、别名等构建属性特征，然后利用机器学习算法计算特征从而判断两菜品是否相同，输出菜品间的关系。

考虑到菜品的属性，本文主要从以下几个角度对菜品的属性进行特征学习：

（1）菜品的不同名称

本文获取的菜品数据来自不同的网站，不同数据源对同一菜品实体的称呼可能有所不同，然而利用这些不同名称，可以帮助判断两两菜品是否相同，因为同一菜品实体即使名称有所不同，但其存在着比较接近的食材名称，例如：“西红柿炒鸡蛋”和“番茄炒蛋”，二者虽然名称不同，但是都含有食材：鸡蛋。因此根据名称属性，可以帮助判断菜品之间是否存在相同或不同关系。

（2）菜品口味

不同菜品的口味差别可能很大，然而同一菜品一般来说具有相同的口味，比如：“蚂蚁上树”和“肉末粉条”，从名称来看二者毫无关联，但是从口味来看它们都具有“微辣”这一相同口味。当判断不同菜品之间是否相关时，口味也是重要属性。

（3）菜品与其他实体间的关系也可作为菜品的其他属性作为补充，例如：菜品与食材之间的组成关系、菜品与营养元素之间的关系、菜品与疾病之间的关系，都可当做菜品的属性进行特征提取，从而判断两两菜品是否存在相同。

通过学习菜品属性的特征，为每个菜品生成对应的特征向量，利用机器学习相关算法，包括支持向量机（SVM）、朴素贝叶斯（NB）算法和 K 最邻近（KNN）算法训练分类器模型，通过实验对比三个模型效率并选择其中效果最好的作为本文菜品实体对齐模型。

实体对齐的实验数据为人工标注的两两菜品对共 300 对，由于数据规模较小，为了充分利用数据，采用 K 折交叉验证法构建训练集和验证集，其中 K 值取 5。实验如表 3.9 所示：

表 3.9 菜品实体对齐

Table 3.9 Dish entity alignment

任务	机器学习算法	P	R	F1
（菜品，相同，菜品）	SVM	**0.92**	0.94	**0.93**
	NB	0.83	**0.97**	0.89
	KNN	0.74	0.68	0.71
（菜品，不同，菜品）	SVM	**0.95**	**0.93**	**0.94**
	NB	0.93	0.89	0.91
	KNN	0.87	0.91	0.89

从实验结果来看，算法支持向量机 SVM 取得了较好的准确率，在两个菜品实体对齐任务上的综合成绩也是最好，分别达到了 93%以及 94%。本文最终选择 SVM 算法进行菜品实体对齐。

最终，经过菜品实体识别和关系抽取，本文从菜品相关数据源中共获得领域内实体 5284 个，其中包括食材（INGREDIENT）类实体 3617 个，菜品（DISH）类实体 1283 个，疾病（DISEASE）类实体 144 个，人群（CROWD）类实体 194 个，营养元素（NUTRIENT）类实体 35 个，卡路里（CALORIE）类实体 11。

抽取到关系共 6 种，包括 hasIngredient、hasCalorie、hasNutrilon、SuitDish、UnSuiDisht、hasRightAmount，分别对应实体之间的关系：菜品-组成成分-食材、菜品/食材-卡路里含量-等级、菜品/食材-营养元素-等级、菜品/食材-适合食用-人群/症状、菜品/食材-禁忌食用-人群/症状以及人群-适量摄入-营养元素。

通过本章菜品领域知识图谱构建，共得到三元组 20261 条。详细结果如表 3.10 所示:

表 3.10 实体与关系获取结果统计

Table 3.10 experimental results of entity and relationship acquisition

	数量	详细信息
实体	5284	1283 个菜品实体，3617 个食材实体，144 个疾病实体等共 5284 个。
关系	6	食材-组成成分-菜品，菜品-禁忌/适合-疾病等共 6 种
三元组	20261	例如：西红柿炒蛋-组成成分-西红柿

3.5 基于图数据库 NEO4J 的知识存储

3.5.1 关系型数据库

关系型数据库在知识图谱应用的一些场景下，其设计和使用会带来诸多不便。在实际应用场景中，信息的更新迭代速度飞速加快，不断地有新的实体和关系加入，如果选择使用关系型数据库存储数据时，每当增加新数据，尤其是大量的数据添加进来，会对数据库的结构和效率造成影响，带来很大的工作量，而图数据库可以在不影响数据库结构的前提下完成数据添加，其只是插入节点或者边就能够很好的完成工作，在面对频繁的插入和删除实体与关系任务上表现出良好的性能。

同时，在查询速度上图数据库也有着巨大的优势。除了最简单的查询，Neo4j 在其他复杂查询的性能表现上都明显表现更好。在查询深度为 3 时的查询速度比关系型数据库快 4 倍。在查询深度为 4 时的查询速度，则要比关系型数据库好五个数量级。深度为 5 时，图数据库的速度甚至要比关系型数据库要快万倍。

因此，本文对基于关系数据库的知识存储方式进行改进，利用图数据库来存储知识图谱。

3.5.2 基于图数据库 Neo4j 的知识存储

在实际运用中，针对不同的数据类型，需要选择不同且合适的存储工具进行存储。对于知识图谱的存储也类似，知识图谱由利用大量不同结构类型数据抽取出的结构化知识组成，其中针对数据来源为结构化及半结构化数据的规模较小的知识图谱，常见的做法是选择传统的关系型数据库对融合后的知识进行存储；但对于面向开放领域及互联网开放文本构建的知识图谱而言，出于对关系推理及对复杂关系进行查询的需要，常使用图数据库（Graph Database）作为存储知识图谱的工具。相较于关系型数据库而言，图数据库借助图结构对存在关联的数据进行表达和存储，并对符合图语义的信息有较高的查询效率，可以有效推动语义 Web 分析、自然语言处理等领域的发展。

目前研究中和市场上常见的图数据库中，以 Neo4j 最具代表性。其作为一款图数据库，对于以三元组对数据进行表达，且数据间关联性较强的知识图谱而言，具有较高的性能；Neo4j 本质上是一个具备完全的事务特性的 Java 持久化引擎，使用符合 SQL 标准的查询语言 Cypher（CQL）进行查询，并支持数据索引；同时提供了丰富的 API 以供多种编程语言访问调用。

本文使用图数据库 Neo4j 作为菜品领域知识图谱的存储工具，将 3.4 节获得的实体及其对应关系使用 Neo4j 的标准化查询 Cypher（CQL）导入到数据库中。初步构建的知识图谱如图 3.4 所示。

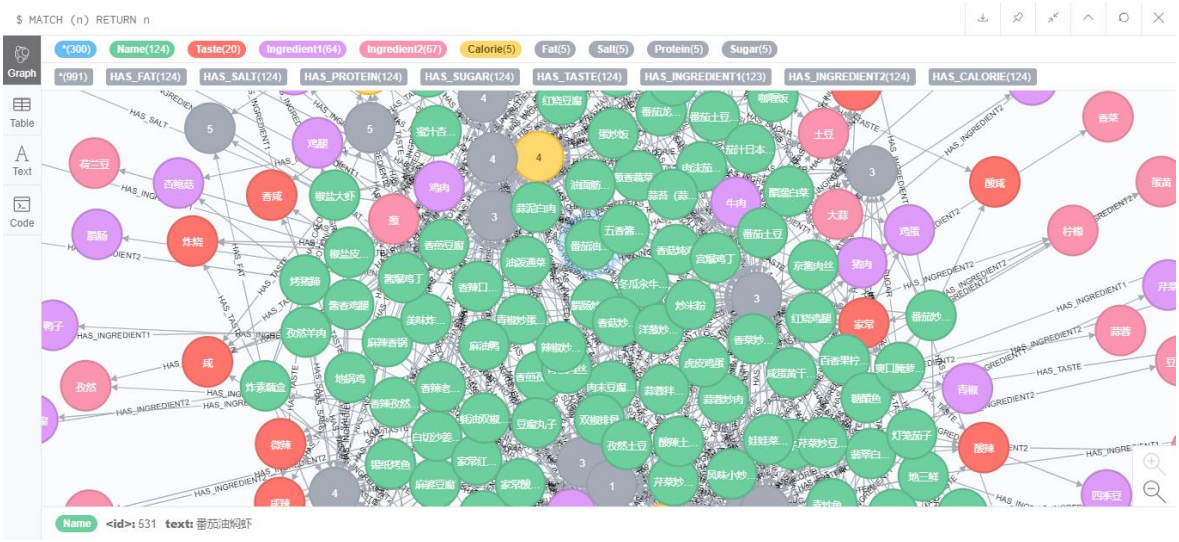

图 3.4 初步构建的知识图谱

Figure 3.4 The knowledge graph of preliminary construction

如图 3.4 所示，初步构建的知识图谱。其中绿色代表菜品实体，灰色代表卡路里含量，紫色代表食材。

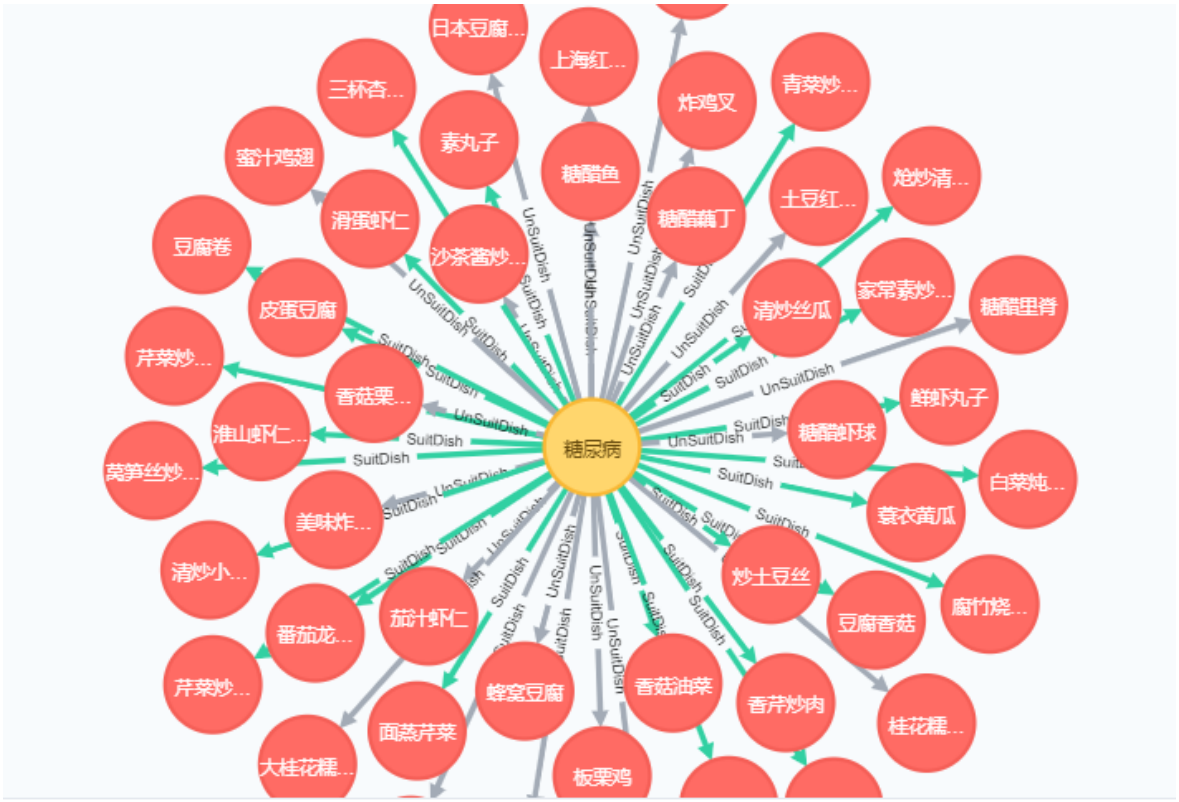

图 3.5 糖尿病与菜品关系可视化

Figure 3.5 Visualization of the relationship between diabetes and dishes

如图 3.5 所示，糖尿病与菜品关系可视化。其中，黄色节点表示疾病名称，红

色节点表示菜品名称，灰色连线代表禁忌食用的菜品，绿色连线代表适合食用的菜品。

3.6 本章小结

本章从数据获取及清洗到最终知识图谱的构建和存储，详细介绍了本文构建菜品领域知识图谱的过程。首先对菜品领域相关的文本进行获取；针对数据特点进行清洗并针对清洗结果定义了知识图谱的实体及实体之间的关系；之后使用 BiLSTM-CRF 模型进行实体识别，并且该模型在评估中取得了较好的效果，解决了目前菜品领域命名实体识别准确率低的问题；使用机器学习的方法对实体关系进行抽取和对齐，并给出实验结果。最后本文使用图数据库 Neo4j 存储知识，将菜品领域知识图谱可视化，为后续的推荐工作打下坚实基础。

4 基于知识图谱的菜品语义表示

在第三章本文构建了菜品领域知识图谱，本章将介绍利用菜品知识图谱向量化表示的方法，通过把知识图谱向量化后与用户行为向量进行匹配来进行推荐。首先介绍了传统的向量化表示方法缺点：容易将两个可能存在关联的实体当作完全不同的实体导致对实体语义信息挖掘深度不够。然后介绍基于知识图谱表示学习模型 TransD 对实体与关系进行向量化表示的方法，其通过对知识图谱中的实体与关系投影到低维空间进行向量化表示。知识图谱表示学习将实体与关系向量化表示概括起来有两方面优点：一方面可以降低知识图谱的高维度和结构复杂的特点，使得知识图谱更加灵活，应用更加广泛；另一方面可以降低由于知识图谱高复杂度和不连续性带来的计算方面的开销。

4.1 传统的向量化表示

机器学习算法取得如此巨大的成功，很大一部分原因取决于向量化表示，向量化表示使得算法在构建分类器是更容易抽取到有用的信息，同时可以很大程度上降低计算开销。本质上，向量化表示就是将对象，包括文本、图像等任何可以描述的对象表示成一系列不间断的实值嵌入到低维向量空间中，其中一系列的实值代表向量化对象的有用信息，比如特征值、属性值或者其他可以表示该对象的信息。

传统的向量化表示以空间上的相似度表示语义上的相似度，将对象间的相似度计算变成计算向量在空间上的相似度，常见的相似度方法是余弦相似度、皮尔逊相似度以及欧氏距离。

随着向量化表示的不断深入研究，学者发现向量化表示也有着其难以掩饰的缺点。向量化对象之间往往存在着一定的联系，比如当向量化对象是不同菜品时，向量化表示会将每一个菜品认为是独立存在的实体，它们之间不存在任何联系。这固然是向量化表示的优点：模型简单而又效果明显，但也因此带来了问题。向量化表示方法无法将表示的对象之间的存在关系保留下来，导致刻画对象的向量化表示有失准确，进而影响对象之间相似度的计算。另外，向量空间模型基于关键字的处理方式，依据的是词频，计算两个文档的相似度依赖两个文档的共同拥有的字数，因而无法辨别自然语言的模糊性，更谈不上理解自然语言。

4.2 知识图谱表示学习模型 TransD

知识图谱表示学习的核心思想是将知识图谱中边和节点表示的实体与关系映射到一定维度且连续的向量空间中，表示为低维度稠密向量，能够达到降低知识图谱的高维度和结构复杂的特点，使得知识图谱更加灵活且不丢失原有的语义关联，应用更加广泛，例如：进行实体链接与预测、关系推理等。

知识图谱表示学习的经典代表模型有距离模型、能量模型、双线性模型、矩阵分解模型和翻译模型等（刘知远等，2016）。距离模型 SE（Bordes *et al*，2011）通过对头、尾实体进行映射来计算两投映向量在空间中的距离；能量模型 SME（Bordes *et al*，2014）通过多个矩阵映射实体和关系，进而发现它们之间的潜在联系；双线性模型 LFM（Jenatton *et al*，2012）通过将关系进行双线性转换得到实体与关系的二阶关联，矩阵分解模型的代表方法是 RESACL 模型（Nickel *et al*，2012），其核心思想是将三元组对应的张量值分解成实体与关系的表示；而翻译模型由于其参数少，计算复杂度低而高效，得到了广泛的使用，其发展历程简单介绍如下：

Mikolov 等人提出 word2vec（Mikolov *et al*，2013），作者发现了平移不变现象。受平移不变现象的启发，翻译模型 TransE（Bordes，2013）将三元组 S（h,r,t）中的关系 r 看作是某种平移向量，因此头实体 h 经过平移后得到尾实体 t。模型如图 4.1 所示

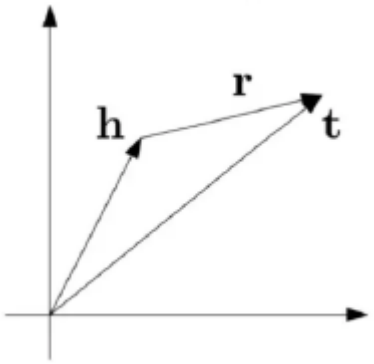

图 4.1 TransE 模型

Figure 4.1 TransE model

翻译模型 TransE 由于其参数少、计算复杂度低而高效，得到了广泛关注。然而由于 TransE 模型在处理一对一关系时的良好表现并没有在处理一对多、多对一或者多对多复杂关系的时表现出来。模型 TransH 为了解决 TransE 在处理复杂关系时出现的问题，提出实体不再单一只有一种表示，而是不同关系下有不同的表示。模型 TransR 提出实体和关系应彼此远离处于不同的语义空间，一个实体是多种属性的综合体，然而不同关系关注实体的不同属性，应该将实体和关系嵌入到不同的空间中，在对应的关系空间中实现向量化表示。

TransD 模型在研究学习 TransR 模型的同时，发现 TransR 模型尚有不足之处：TransR 模型重在将实体的不同关系投影到不同的空间，将每一关系有且仅代表一种语义关系，然而有时一关系可能代表不同的语义。而且 TransR 模型的参数在关系投映矩阵的引入后增加剧烈，从而导致计算开销变大。

针对 TransR 模型的不足之处，TransD 模型对 TransR 模型进行了改进，认为不同的实体应映射到不同的语义空间中，且通过将矩阵运算转变成向量运算，大大减少了计算量。TransD 的模型如图 4.2 所示；

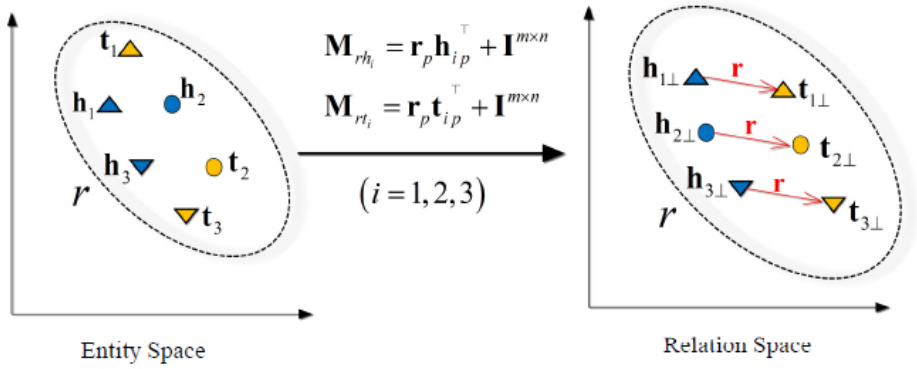

图 4.2 TransD 模型

Figure 4.2 TransD model

令 h_p、r_p、t_p 为 h，r，t 的映射向量，根据映射向量得到头、尾实体到关系空间的头、尾投影矩阵，$\mathbf{M}_{rh}$、$\mathbf{M}_{rt}$ 分别是实体 h,t 的映射矩阵，$\mathbf{h}_{ip}, \mathbf{t}_{ip}(i=1,2,3)$ 及关系 r_p 为投影向量，$\mathbf{h}_{\perp}$、$\mathbf{t}_{\perp}$ 分别为头尾实体的投影向量。

$$\begin{aligned} \mathbf{M}_{rh} &= \mathbf{r}_p \mathbf{h}_p^{\mathrm{T}} + \mathbf{I}^{m\times n} \\ \mathbf{M}_{rt} &= \mathbf{r}_p \mathbf{t}_p^{\mathrm{T}} + \mathbf{I}^{m\times n} \\ \mathbf{h}_{\perp} &= \mathbf{M}_{rh}\mathbf{h} \\ \mathbf{t}_{\perp} &= \mathbf{M}_{rt}\mathbf{t} \end{aligned} \tag{4.1}$$

其损失函数如下：

$$f_r(\mathbf{h},\mathbf{t}) = -\left\|\mathbf{h}_{\perp} + \mathbf{r} - \mathbf{t}_{\perp}\right\|_2^2 \tag{4-2}$$

4.3 TransD 模型训练过程

4.3.1. 负三元组构建方法

TransD 模型训练过程在于对目标函数的不断优化，优化训练需要正确的三元组指导目标函数迭代更新，同时也需要错误的三元组来训练目标函数，从而使得正、负三元组的距离不断拉开，达到模型训练效果。翻译模型 TransE 使用如公式 4-3 所示的负三元组构建方法：

$$S'_{(h,r,t)}=\left\{\left(h',r,t\right)|h'\in E\right\}\cup\left\{\left(h,r,t'\right)|t'\in E\right\} \quad (4\text{-}3)$$

负三元组构建公式中 E 为实体集合，h' 和 t' 是实体集合中随机抽取的一个实体，$\left(h',r,t\right)$ 为被替换头实体而得到的负三元组，$\left(h,r,t'\right)$ 为被替换尾实体的负三元组，$S'_{(h,r,t)}$ 代表构建的负三元组集合。这种随机抽取头、尾实体来替换构成负三元组的方法存在不足之处。

知识图谱中的实体关系错综复杂，当使用上述随机抽取头、尾实体构建负三元组的方法时，会增加很多错误的负三元组。比如：头实体 h 是一对多的关系，$\left(h,r,t\right)$ 和 $\left(h,r,t'\right)$ 都是正确的三元组，通过随机替换尾实体将 $\left(h,r,t\right)$ 替换成 $\left(h,r,t'\right)$ 时，$\left(h,r,t'\right)$ 会被标记为负三元组，实际上 $\left(h,r,t'\right)$ 也是正三元组。虽然 TransE 模型对此情况进行了约束，但实际效果并不理想。

为了解决 TransE 算法出现的错误负三元组问题，TransH 算法提出了一种新的负三元组构建方法，主要针对当面对一对多或多对一的情况，随机抽取策略无法起到作用。采用 bern 表示新的伯努利算法，unif 表示随机抽取算法，新的负三元组构建方法主要思想是：尽可能的对于 1 对多/多对一的三元组，给予更多的头/尾实体被抽取的机会。论文采用伯努利分布来分配头、尾实体的替换，尽可能减少错误的负三元组产生。通过计算每个头实体对应的平均尾实体数量 D_{tph}，每个尾实体对应的平均头实体数量 D_{hpt}，使用公式 4-4 计算概率值 P：

$$\mathrm{P}=\frac{\mathrm{D}_{tph}}{\mathrm{D}_{tph}+\mathrm{D}_{hpt}} \quad (4\text{-}4)$$

概率值 P 是指替换头实体的概率为 P，那么替换尾实体的概率即为 $1\text{-}P$。根据关系的头尾实体连接数量来设置不同的替换头实体或尾实体的概率，具体如下表 4.1 所示：

表 4.1 负三元组抽样策略

Table 4.1 negative triple sampling strategy

关系类型	D_{tph} 与 D_{hpt}	抽样策略
一对多	$D_{tph}\geq 1.5$ 且 $D_{hpt}<1.5$	较多的替换头实体
多对一	$D_{tph}<1.5$ 且 $D_{hpt}\geq 1.5$	较多的替换尾实体
多对多	$D_{tph}\geq 1.5$ 且 $D_{hpt}\geq 1.5$	考虑连接该关系的头实体和尾实体的数量

4.3.2. 目标函数

给定三元组(h,r,t)，h为头实体，r为关系，t为尾实体，将实体与关系投影到实体空间与关系空间中，得到三元组的实体向量和空间向量满足$|h+r|\approx t$。即：如果三元组是正确的，那么尾实体t与头实体h和关系r的加和相接近。反之，如果是错误的三元组，则尾实体t与头实体h和关系r的加和相远离。正是基于此思想，TransD 算法的目标函数如公式 4-5 所示：

$$\mathcal{L}=\sum_{(h,r,t)\in\Delta}\ \sum_{(h',r',t')\in\Delta'}\left[f_r(h,t)+\gamma-f_r\cdot(h',t')\right]_+ \tag{4-5}$$

其中，$[*]_+$表示$\max(0,*)$，f_r表示一种衡量三元组能量的距离函数，Δ是正确的三元组的集合，Δ'是正确的三元组的头、尾实体被替换组成的负三元组集合，Υ是正负三元组距离值（margin），常设为 1。

4.3.3. 模型训练算法

本文采用随机梯度下降算法（SGD）作为模型的训练算法，模型的优化过程是为了最小化目标函数L，也就是使$f_r(h,t)$最小，$f_r(h',t')$最大。TransE 算法主要可分为如下三步：

（1）随机初始化头、尾实体e和关系向量r并对随机初始化的向量做归一化处理。具体处理方法如公式 4-6；

$$e=e/\|\text{uniform}\left(-\frac{6}{\sqrt{d}},\frac{6}{\sqrt{d}}\right)\|;\quad r=r/\|\text{uniform}\left(-\frac{6}{\sqrt{d}},\frac{6}{\sqrt{d}}\right)\| \tag{4-6}$$

（2）对正三元组抽样并对抽样的三元组进行实体替换，形成负三元组。

（3）优化目标函数L，得到实体和关系的向量表示。

当目标函数L不断优化达到最小值，即$f_r(h,t)$最小，$f_r(h',t')$最大时，训练结束。通过将实体与关系投影到特定的实体空间和关系空间，使得尾实体向量与关系向量的加和无限等于头实体向量，视为二者存在语义上的联系，应用在后续的推荐工作中，计算菜品的语义相似度。

4.4 实验与分析

4.4.1. 实验数据集

本文在第三章构建了菜品领域知识图谱，用于作为本次实验的实验数据。菜品领域知识图谱的规模如表 3.10 所示，包含实体 5284 个、关系 6 中、三元组 20261 个。以 8:1:1 的比例划分训练集，测试集和验证集，实验数据详细情况见表 4.2 所示；

表 4.2 知识图谱包含的实体类型，关系类型及其数量

Table 4.2 Entity type, relationship type and quantity of knowledge graph

	数量			详细信息
	Train	Valid	Test	
实体	4228	528	528	1283 个菜品实体，3617 个食材实体，144 个疾病实体等
关系	6	6	6	食材-组成成分-菜品，菜品-禁忌/适合-疾病等共 6 种
三元组	16209	2026	2026	例如：西红柿炒蛋-组成成分-西红柿

4.4.2. 实验设置

（1）评价指标

针对本文训练的知识表示模型 TransD，分别采用实体链接预测和三元组分类两个任务来对模型进行评价。

① 实体链接预测

该任务的主要思想是对不完整的三元组（h,r)/(r，t)，预测头实体/尾实体。

评价指标：

a ：MeanRank：预测序列中平均到第多少个才能匹配到正确的缺失实体。

b ：Hits@10:前十个预测中正确的缺失实体存在的概率

显然，MeanRank 越小且 Hits@10 越高则表示模型训练效果好。此外，知识图谱复杂的结构决定存在大量的一对多或多对一的情况，从而出现类似 TransE 模型中随机抽取实体构建负三元组的情况：当我们预测序列排序时，被认为错误但是本就已存在于知识图谱的三元组会对正确的缺失实体排名造成影响。因此，可以移除测试集中此类三元组，称未移除的数据“Raw”，移除后的数据为“Filt”.

② 三元组分类

该任务的核心思想是判定三元组是否真实存在。

评价指标：通常使用准确率（Accuracy）、召回率（Recall）以及 F_1 值。

（2）超参数的设置

本文的超参数设置如表 4.3 所示，其中向量维度 d 的取值范围为{50，75，100，125，150}，间隔 ϒ 的取值范围为{0.1，0.5，1.0}，权重 C 的取值范围为{0.0625，0.25，0.5，0.1}，学习率 lr 的取值范围为{0.001，0.005，0.01}，最大迭代次数设为 100。

表 4.3 超参数设置

Table 4.3 Hyper parameter setting

参数	范围
向量维度d	{50, 75, 100, 125, 150}
间隔ϒ	{0.1, 0.5, 1.0}
学习率lr	{0.001, 0.005, 0.01}
权重C	{0.0625, 0.25, 0.5, 1}
最大迭代次数	100

通过调整参数对模型进行多次试验，分析实验结果，从而得出最优的超参数配置。

① 原始抽样算法unif 最优参数配置为：d= 50，ϒ= 1.0，lr= 0.01，C= 0.25；

② 伯努利抽样算法bern最优参数配置为：d= 75，ϒ= 1.0，lr= 0.005，C= 0.5。

4.4.3.实验结果与分析

本章总共设计了三个对比试验，如表 4.4 所示。

（1）不同模型的对比实验

通过将 TransD 模型训练结果同 TransE、TransH、TransR 模型进行对比，可以看出 TransD 模型取得了较好的成绩。

（2）不同负三元组构建抽样算法的对比实验

将随机抽样的 TransD（unif）同伯努利抽样的 TransD（bern）进行对比，可以看出基于伯努利抽样的 TransD（bern）模型总体上要高于 TransD（unif）模型。

（3）有无移除被认为错误的三元组的对比实验

通过对比四个模型在两种数据上的表现，可以在发现“Filt”数据上的表现明显优于在“Raw”数据集上的表现。

表 4.4 实体链接预测实验结果

Table 4.4 Link prediction test results

模型	MeanRank		Hits@10	
	Raw	Filt	Raw	Filt
TransE	54	48	73.2	83.1
TransR	43	40	75.4	83.9
TransH	**41**	39	74.5	85.3
TransD(unif)	44	42	**75.7**	86.9
TransD(bern)	39	**36**	73.1	**87.3**

表 4.5 三元组分类实验结果

Table 4.5 Experimental results of triple classification

模型	准确率(%)
TransE	73.4
TransR	77.5
TransH	84.5
TransD(unif)	84.9
TransD(bern)	**86.2**

根据表 4.4，4.5 的实验结果表明，相比较与 TransE 等翻译模型，TransD 模型在实体链接预测及三元组分类任务上取得了最好的成绩，证明 TransD 模型适合作为知识图谱表示学习的模型对实体与关系进行向量化表示。

4.5 本章小结

本章详细介绍了将第三章构建的菜品领域知识图谱进行向量化表示的过程。首先指出传统向量化表示方法的不足之处，然后介绍知识图谱表示学习的相关模型并着重介绍了一系列的翻译模型，之后使用知识图谱技术的知识表示模型 TransD 对实体与关系进行分布式表示，通过训练模型，得到实体在向量空间中的精确映射，一方面可以降低知识图谱的高维度和结构复杂的特点，使得知识图谱更加灵活，应用更加广泛；另一方面可以降低由于知识图谱高复杂度和不连续性带来的计算方面的开销，为推荐系统提供较准确的菜品语义信息，同时也为后续的推荐工作打下基础。

5 基于知识图谱的菜品推荐算法

传统协同过滤算法主要根据用户的历史行为数据，包括用户搜索、浏览、点击、收藏、评论等，对用户的偏好进行收集，刻画用户像来给用户推荐可能喜欢的物品，而用户的历史行为数据很容易因为现实生活中的某些因素缺失，当缺失用户-物品交互数据时，传统的协同过滤算法推荐效果也大打折扣。

因此，传统协同过滤算法面临两大难点：一方面，当推荐算法应用到具体的商业中时，往往会发现数据稀疏问题愈发地突出。由于用户能够接触的物品有限，导致用户有过历史行为数据的物品数占比极小。另一方面，每当有新用户加入或者新物品加入时，由于缺少新用户或物品历史数据，难以这些物品和用户进行推荐。

协同过滤算法利用用户的历史行为数据，包括用户隐性反馈数据和显性反馈数据（吴玺煜等，2018），给用户推荐合适的物品。其中隐性反馈数据如购买记录、点击记录等，显性反馈数据如评分信息、留言等。协同过滤算法利用这些用户与物品之间有过关联交互的信息，进而挖掘用户潜在的喜好物品来对用户进行推荐。但是这些用户数据主要考虑的是用户与物品的交互数据，尚未把物品内在的信息挖掘考虑。

5.1 TransD-CF 推荐算法框架

针对协同过滤算法仅考虑用户与物品交互的历史行为数据，忽略了被推荐的物品潜在的语义关联现象，本文提出一种融合协同过滤和知识表示的算法，既利用了用户行为数据，又包含了物品本身的相似性信息，可以有效的缓解协同过滤算法存在的数据稀疏以及冷启动问题，并期望通过这种方式提高推荐系统的性能。

TransD-CF 推荐算法的具体思想为：首先将本文第三章构建的菜品领域知识谱中的菜品实体投影到特定的低维度稠密且连续的向量空间，根据向量空间中的实体向量计算菜品实体在空间上的距离来表示菜品在语义角度上的相似度，再计算协同过滤算法中基于用户历史行为数据的菜品相似度，然后把知识图谱中的菜品实体相似度同协同过滤算法的菜品相似度相结合，得到融合的菜品相似度。将融合得到的菜品相似度代替原协同过滤算法中的用户-菜品评分矩阵，结合菜品本身的语义关联信息为用户推荐合适的菜品。算法流程如图 5.1 所示。

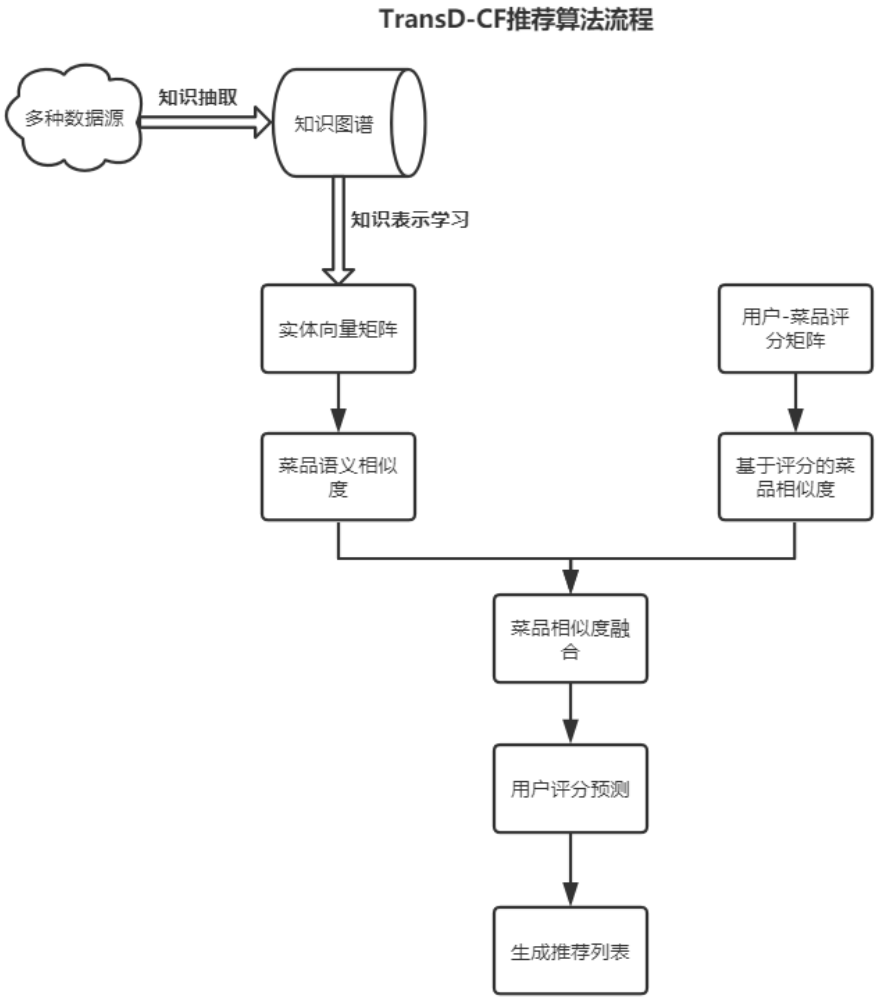

图 5.1 TransD-CF 推荐算法流程

Figure 5.1 TransD-CF recommended algorithm flow

5.2 相似度计算

由第四章的知识图谱表示学习算法 TransD 习得菜品语义相似度 sim_{graph}，在通过利用协同过滤算法中用户与菜品的交互数据计算用户-菜品相似度 sim_{CF}。将上述两相似度乘以相关的系数并做加法结合，得到最终的既包含菜品语义相似度又包含用户评分的菜品相似度。

5.2.1 基于知识图谱的菜品相似度

基于第三章构建的菜品领域知识图谱，本文在第四章用知识图谱表示学习算法 TransD 将菜品实体与关系实体映射到低维稠密向量空间，本节将通过计算菜品实体向量在低维空间上的距离远近来衡量菜品实体在语义上的相似度，利用空间上的距离刻画菜品语义上相似度，弥补了协同过滤算法为有效考虑推荐物品的语义关联。

首先将菜品实体向量和关系向量嵌入到d维的向量空间，菜品I_i表示为d维的向量：

$$Ii = (E_{1i}, E_{2i}, \dots, E_{di})^T \quad (5\text{-}1)$$

如公式5-1所示，E_{di}表示菜品I_i的嵌入向量在第d维上的实值。

其次根据菜品I_i、I_j在d维向量空间的实体向量计算其距离，由于在使用TransD模型进行知识表示学习训练时，为了使正负三元组在空间上能够被分隔开，使用欧几里德距离来计算目标函数。为了保持一致性，本文在计算菜品向量见得距离也采用欧几里德距离，公式如5-2所示：

$$d(Ii, Ij) = \sqrt{\sum_{k=1}^{d}(E_{ki} - E_{kj})} \quad (5\text{-}2)$$

其中$d(Ii, Ij)$表示菜品I_i和菜品I_j的距离，$d(Ii, Ij)$越小，菜品语义相似度越高。为了规范计算，将菜品I_i和菜品I_j的距离转换至(0,1]之间，转换如公式5-3所示。

$$\mathrm{sim}_{\mathrm{graph}}(i, j) = \frac{1}{1 + d\left(I_i, I_j\right)} \quad (5\text{-}3)$$

$\mathrm{sim}_{\mathrm{graph}}(i, j)$值越大说明菜品$i$和菜品$j$语义越相似。

5.2.2 用户-菜品评分的菜品相似度

用户的行为数据能反映出用户对菜品的适宜程度，根据用户-菜品的评分信息可以刻画用户的画像，从而为用户推荐其可能喜欢的菜品。本文通过用户-菜品的评分信息将评分数据向量化。倘若推荐算法中有用户集$U = (U1, U2, \dots, Um)$和 n 个菜品$I = (I1, I2, \dots, In)$，那么用户 菜品评分信息可以表示为矩阵$R_{m\times n}$：

$$\mathbf{R}_{m\times n} = \begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1j} & \cdots & R_{1n} \\ R_{21} & R_{22} & \cdots & R_{2j} & \cdots & R_{2n} \\ \vdots & \vdots & & \vdots & & \vdots \\ R_{i1} & R_{i2} & \cdots & R_{ij} & \cdots & R_{in} \\ \vdots & \vdots & & \vdots & & \vdots \\ R_{m1} & R_{m2} & \cdots & R_{mj} & \cdots & R_{mn} \end{bmatrix} \quad (5\text{-}4)$$

如公式5-4所示，R_{kj}为用户U_k对菜品j的评分（$1 \le k \le m, 1 \le j \le n$），评分高低反应了用户对该菜品的适宜程度。如果两个菜品是相似的，那用户对其评分也应当是相似的。基于项目协同过滤的推荐算法就是利用所有用户对菜品的评价向量来衡量菜品的相似，从而实现推荐。所有用户对菜品j的评分向量I_i表示如公式(1)所示：

$$I_i = (S_{1j}, S_{2j}, \ldots S_{kj}, \ldots, S_{mj})^T \tag{5-5}$$

在公式 5-5 中 S_{kj}为用户k对菜品j的评分（$1 \leqslant k \leqslant m$），$m$为用户数。

衡量菜品相似度本文使用的是最常用应用最为广泛的相似度计算方法有余弦相似度，其计算如公式 5-6 所示。

$$\mathrm{sim}_{CF}(i, j) = \cos\left(S_i, S_j\right) = \frac{S_i \cdot S_j}{\left\|S_j\right\| \cdot \left\|S_j\right\|} = \frac{\sum_{m=1}^{k} \mathrm{S}_{ki} \cdot \mathrm{S}_{kj}}{\sqrt{\sum_{k=1}^{m} \mathrm{S}_{ki}^2} \cdot \sqrt{\sum_{k=1}^{m} \mathrm{S}_{kj}^2}} \tag{5-6}$$

在公式 5-6 中sim_{CF}(i,j)为菜品i和菜品j的余弦相似度，S_i和S_j分别为用户对菜品i和j的评分向量，S_{ki}为用户k对菜品 i的评分。sim_{CF}(i,j)值的大小与菜品 i 和菜品 j 的相似度呈正比。

5.2.3 相似度融合

由第四章的知识图谱表示学习算法 TransD 习得菜品语义相似度$\mathrm{sim}_{graph}(i, j)$，通过利用协同过滤算法中用户与菜品的交互数据计算用户-菜品相似度$\mathrm{sim}_{CF}(i, j)$。将上述两相似度乘以相关的系数并做加法结合，得到最终的既包含菜品语义相似度又包含用户评分的菜品相似度。具体计算方法如公式 5-7 所示：

$$\mathrm{sim}(i, j) = \alpha \cdot \mathrm{sim}_{\mathrm{graph}}(i, j) + (1-\alpha)\mathrm{sim}_{CF}(i, j) \tag{5-7}$$

系数α为融合比例，表示融合后的菜品相似度中$\mathrm{sim}_{\mathrm{graph}}$和$\mathrm{sim}_{CF}$所占的比例，其值范围为[0,1]。最终使用矩阵的形式将融合完成的菜品相似度$\mathrm{sim}(i, j)$来表示。

5.3 评分预测

评分预测是通过预测用户对未评分菜品的评分，并根据评分的高低进行 Top-N 排序，将前 N 个菜品推荐给用户。根据上节得到的融合菜品相似度矩阵后，用 p_{ui} 表示预测用户u对菜品i的评分，其计算公式如公式 5-8 所示；

$$p_{ui} = \frac{\sum_{j \in N(u) \cap \mathrm{S}(i,k)} \left(\mathrm{sim(i,j)} \times R_{uj}\right)}{\sum_{j \in N(u) \cap \mathrm{S}(i,k)} \mathrm{sim(i,j)}} \tag{5-8}$$

其中$\mathrm{sim}(i, j)$为菜品 i 和菜品 j 的相似度，R_{uj}为用户 u 对菜品 j 的已经打过评分，$N(u)$为用户u评分过的菜品的集合，集合$S(i,k)$为菜品 i 最相似的 k 个菜品集。

评分预测的思想是首先找到与菜品i相似的k个菜品集合$S(i,k)$，然后根据用户评过分的菜品集合$N(u)$来与$S(i,k)$求交集，得到与菜品 i 相似且用户没有评过分的备选

菜品集合，再为备选菜品集合中的每个菜品生成相对应的评分，该评分表示该菜品的相似度在所有菜品相似度中所占的比例，评分越高的菜品代表与用户用过评分的菜品相似度越高。最后将菜品的预测评分结果按照降序排列，相似度越高的菜品优先推荐。

5.4 实验与分析

5.4.1 实验数据

知识图谱本文使用第三章构建的菜品领域知识图谱，共 1283 个菜品实体，3617 食材实体，144 个疾病实体等。

为了使得饮食推荐系统能够得出专业的菜品推荐方案，本实验使用的用户数据集是来自于医院患者的饮食营养调查报告，结合营养医生的诊治方案。初始数据一共 2523 份，后期经专业营养师整理得到有效报告 920 份。将得到的数据进行汇总处理。利用这些评分数据，构建了基于 TransD-CF 的菜品推荐模型。营养建议报告涉及的具体内容如下：

（1） 基础情况：年龄，性别，身高，体重，腰围，近 1-3 个月体重变化等；

（2） 病史：是否患有高血压，是否患有糖尿病，是否患有肥胖症等；

（3） 饮食已知：糖、盐、脂肪、蛋白质等菜品所含的营养元素对该病的影响，该病适合吃的食材和菜品以及该病禁忌吃的食材和菜品；

通过营养医生的诊断，对上述营养报告适合与不适合该患者的菜品打分。菜品对患者的适合程度通过其对菜品的打分评价来衡量，等级分为 1-5，等级越高说明该菜品越适合患者食用。通过观察分析，本文把评价等级为 4 和 5 的认为是患者适合的菜品，1-3 分的为患者不适合食用的菜品。

由于用户与菜品的评分数据规模较小，为了充分利用数据，实验采用 K 折交叉验证法划分训练集和验证集，其中 K 值取 5。

5.4.2 评价指标

针对本文训练的 TransD-CF 推荐算法给出的 Top-N 推荐列表做评估，分别采用准确率（Precision）、召回率（Recall）以及$F1$值（F1-Measure）来对推荐系统的预测准确率进行评价。定义如公式 5-9、5-10、5-11 所示；

$$\text{Precision} = \frac{TP}{TP+FP} \tag{5-9}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{5-10}$$

$$\text{F1} = \frac{2*Precision*Recall}{Precision+Recall} \tag{5-11}$$

其中公式中出现的符号见表 5.1 所示，其中 TP 表示为系统推荐并且用户认为适合，FP 表示为系统推荐但是用户感觉不适合。

表 5.1 评价指标相关符号

Table 5.1 Relevant symbols of evaluation indexes

推荐算法	适合	不适合
推荐	TP	FP
未推荐	FN	TN

5.4.3 实验结果及分析

（1）不同邻近个数 k 值对比实验

选取不同邻近个数 k 值是影响算法性能的一个重要参数，本文实验分别选取 10、20、30、40、50 作为邻近数进行对比试验，知识图谱表示学习嵌入维度设定为 75，融合权重α设定为 0.7，实验结果如图 5.2 所示：

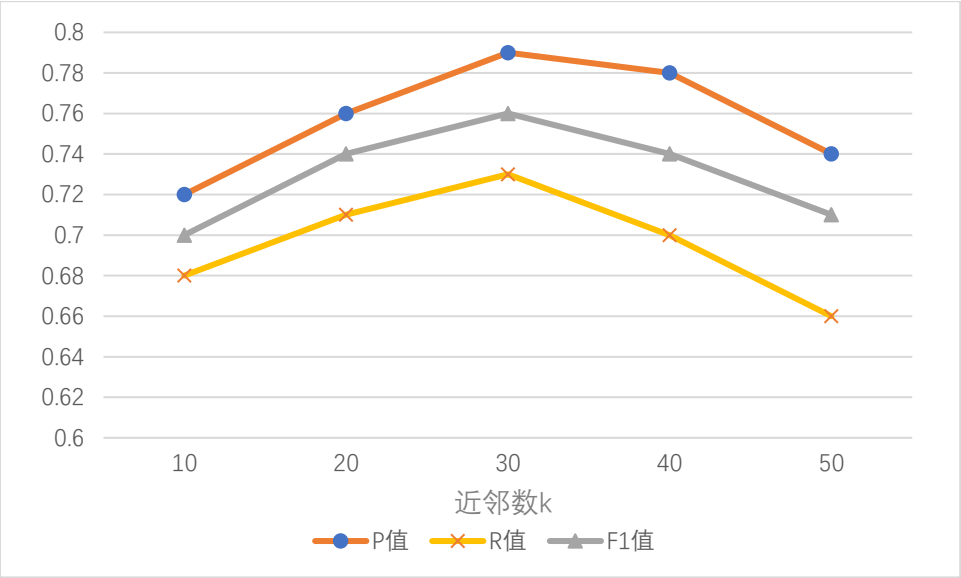

图 5.2 不同近邻个数 k 下，Precision 值、Recall 值以及 F1 值的曲线图

Figure 5.2 Curve of precision value, recall value and F1 value under different k dimensions

根据多次实验结果对比发现，随着近邻数 k 增大，P 值、R 值、以及 F1 值都呈现出上升趋势，说明近邻数 k 增加能够提高推荐精度。当近邻数 k 达到 30 时推荐效果最佳，但随着近邻数 k 进一步增加，其推荐精度反而下降，表明当近邻数过大时对推荐精度作用有限甚至是反作用，同时还增加计算量。因此当近邻数 k 取 30 时，算法的性能表现取得最优。

（2）不同融合权重因子对比实验。

融合权重因子α控制了基于知识图谱的菜品语义相似度和基于协同过滤的用户与菜品评分矩阵的菜品相似度在最终融合相似度中的占比，是本文算法的一个关键因子。α取值范围为[0，1]，实验从α取 0.1 开始，每次递增 0.1，到 1 结束，一共 10 次。

知识图谱表示学习嵌入维度d设定为 75，菜品相似度最大邻近个数k设定为 30，实验参数配置分为表示学习参数和菜品近邻数设置，知识图谱表示学习参数为第四章节实验得出的最佳配置参数：lr=0.005，γ=1，d=75，C=0.5。图 5.3 给出菜品近邻数k为 30 时，TransD-CF 推荐算法在不同融合比例下，Precision 值、Recall 值以及$F1$值的曲线图。通过多次对比实验结果，发现融合比例α为 0.7，即基于菜品知识图谱的语义相似度占融合菜品相似度的 70%时，算法的性能表现最佳。

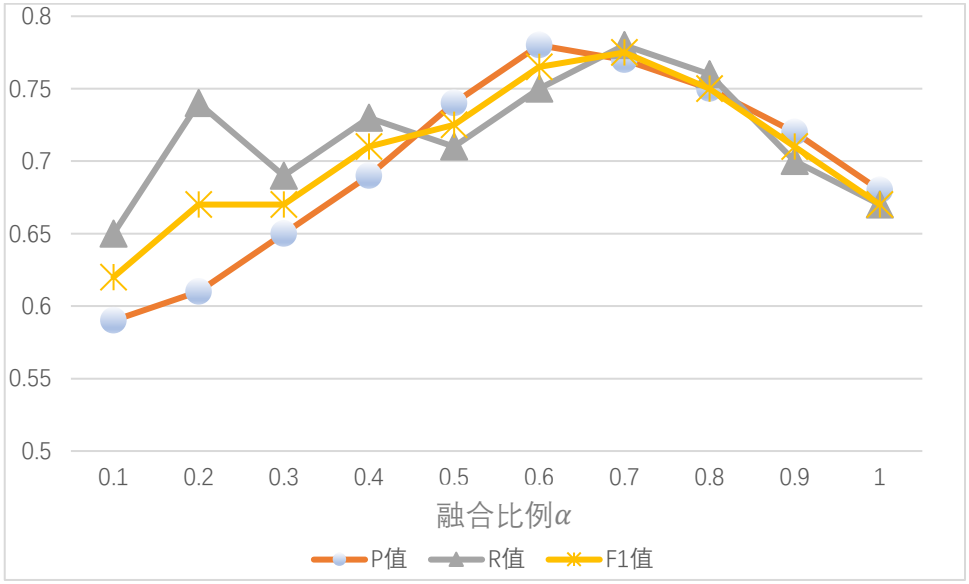

图 5.3 在不同融合比例下，Precision 值、Recall 值以及 F1 值的曲线图

Figure 5.3 Curve of precision value, recall value and F1value under different fusion proportions

（3）算法性能比较。

本文通过与基于物品的协同过滤算法(Item-CF)、基于余弦相似度的算法以及基于调整的余弦相似度算法进行比较证明本文提出的基于知识图谱的推荐算法 TransD-CF 的准确性。实验参数分别为：知识图谱表示学习嵌入维度 d 为 75，菜品相似度最大近邻数 k 为 30，融合比例因子α取值为 0.7。实验结果如图 5.4 所示。

通过对比实验结果表明，本文提出的算法 TransD-CF 在一定程度上提高了推荐结果的准确性，与传统的协同过滤算法相比准确率提高了 3.2%，达到 79.3%，F1 值提高了 1.3%，达到 76.3%。实验证明了本文提出的 TransD-CF 推荐算法的有效性。

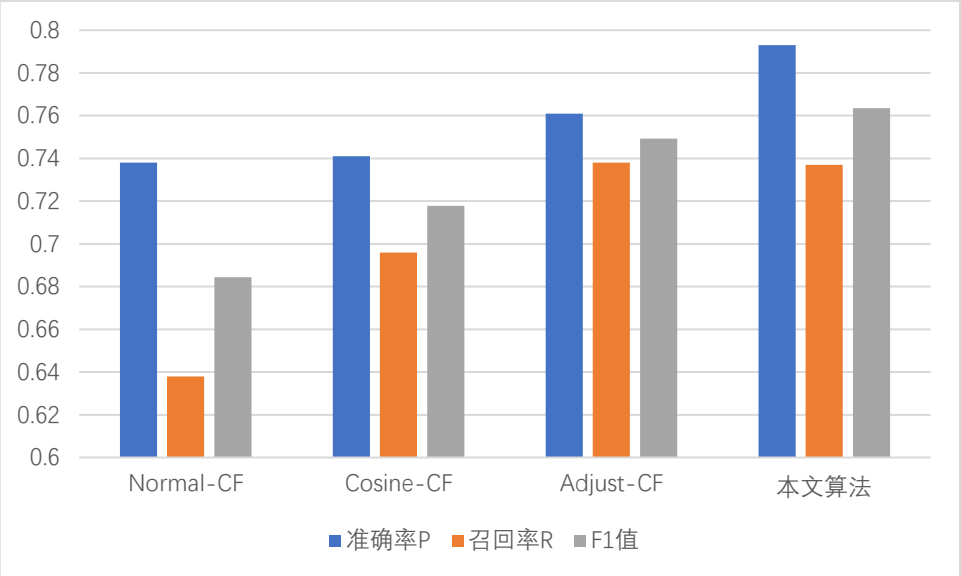

图 5.4 不同推荐算法实验对比结果

Figure 5.4 Experimental comparison results of different recommended algorithms

5.5 本章小结

本章提出了基于知识图谱的菜品推荐算法 TransD-CF，并详细介绍了算法的流程与具体实现。首先介绍了协同过滤算法因数据稀疏问题导致推荐算法效果不佳，同时协同过滤算法未考虑推荐的物品间的语义关联，然后提出本文的推荐算法 TransD-CF，一种融合协同过滤和知识表示的算法，既利用了用户行为数据，又包含了物品本身的语义相似性信息，可以有效的缓解协同过滤算法存在的数据稀疏以及冷启动问题，并对算法的流程与具体实现做了详细的介绍。最后通过实验证明本文提出的 TransD-CF 推荐算法的准确性。

6 全文总结与展望

6.1 全文总结

随着我国经济发展和人民生活水平的提高，人们在饮食习惯和生活方式上也发生很大改变，近年来我国癌症患病率呈现明显上升趋势，而导致我国癌症高发的一大重要原因是不合理的饮食。在中国历史悠久的饮食文化中，中国菜有着不可替代的地位，根深蒂固地影响着每一个中国人的日常饮食。本文通过调研现有的与健康饮食相关的软件，如“薄荷 APP”、“美食杰”等，发现其仅给出相关食物的卡路里等营养数值信息以及营养价值等，不能结合用户的身体健康状况、也不能细致的描述饮食状况，因此这些工具很难做出准确而有效的健康饮食推荐。为了满足用户合理膳食、健康饮食需求，本文通过构建中国菜品知识图谱对饮食信息进行更为全面的描述，同时结合协同过滤算法和知识表示学习为用户提供更符合健康饮食要求的推荐结果。本文的主要研究内容如下：

首先构建了菜品知识图谱。对获取的菜品菜谱以及营养学相关文献等数据的知识特征进行分析，实现了领域内实体与关系的划分，并定义多种实体之间的关系；使用 BiLSTM-CRF 对半结构化和非结构化的文本数据进行实体抽取，并对实体部分进行了对齐处理；最后利用图数据库 Neo4j 存储了构建的菜品领域知识图谱。

其次提出一种融合协同过滤和知识表示的推荐算法，既利用了用户行为数据，又包含了菜品本身的相似性信息，采用知识表示学习算法 TransD 将菜品知识图谱中的实体和丰富语义关系精准映射到低维向量空间中，生成基于语义的菜品相似度表达；根据收集到的用户数据构建用户兴趣模型，获取用户行为矩阵，生成基于用户打分的菜品相似度表达，线性融合两种相似度信息获得最终的菜品间相似度表达，并将用户打分矩阵与融合后的相似度表达进行再结合，对用户未打分的菜品进行预测，依据预测评分降序排列并从中选取 Top-N 的菜品作为推荐列表推送给用户。

最后通过结合本文构建的菜品知识图谱和用户对菜品的评分数据，使用本文提出的推荐算法进行实验。结果表明，相较于传统的协同过滤算法，本文结合菜品语义信息的推荐算法在准确度达到了 79.3%、F1 值达到了 76.3%。

6.2 后续工作展望

本文工作待优化部分在于知识图谱构建部分的数据有待扩展，需进一步提升知识图谱的规模来提高推荐结果的准确性，并且可以尝试移植到移动端，为用户提供更便捷、随身的服务。

同时，本文设计的算法主要面向用户针对简单的推荐，今后考虑结合知识推理及图模型，设计实现面向复杂关系的推荐算法。

参考文献

[1] 蔡威. 临床营养学[M]. 复旦大学出版社, 2012.

[2] 曹倩, 赵一鸣. 知识图谱的技术实现流程及相关应用[J]. 情报理论与实践, 2015,38(12):127-132.

[3] 常亮, 张伟涛, 古天龙, 孙文平, 宾辰忠. 知识图谱的推荐系统综述[J]. 智能系统学报, 2019,14(02):207-216.

[4] 陈思. 常见疾病饮食禁忌[J]. 养生月刊, 2012,33(05):455-457.

[5] 陈思佳. 实体关系抽取技术研究[D]. 北京邮电大学, 2014.

[6] 陈振宏, 兰艳艳, 郭嘉丰, 程学旗. 基于差异合并的分布式随机梯度下降算法[J]. 计算机学报, 2015,38(10):2054-2063.

[7] 陈宗言, 颜俊. 基于稀疏数据预处理的协同过滤推荐算法[J]. 计算机技术与发展, 2016,26(07):59-64.

[8] 代丽, 樊粤湘. 个性化推荐系统综述[J]. 计算机时代, 2019(06):9-11.

[9] 黄恒琪, 于娟, 廖晓, 席运江. 知识图谱研究综述[J]. 计算机系统应用, 2019,28(06):1-12.

[10] 黄立威, 江碧涛, 吕守业, 刘艳博, 李德毅. 基于深度学习的推荐系统研究综述[J]. 计算机学报, 2018,41(07):1619-1647.

[11] 蒋勋, 徐绪堪. 面向知识服务的知识库逻辑结构模型[J]. 图书与情报, 2013(06):23-31.

[12] 李纲, 潘荣清, 毛进, 操玉杰. 整合BiLSTM-CRF网络和词典资源的中文电子病历实体识别[J]. 现代情报, 2020,40(04):3-12.

[13] 刘方驰, 钟志农, 雷霖, 吴烨. 基于机器学习的实体关系抽取方法[J]. 兵工自动化, 2013,32(09):57-62.

[14] 刘知远, 孙茂松, 林衍凯, 谢若冰. 知识表示学习研究进展[J]. 计算机研究与发展, 2016,53(02):247-261.

[15] 钱丽萍. 浅谈饮食与糖尿病[J]. 大家健康(学术版), 2015,9(02):283.

[16] 孙镇, 王惠临. 命名实体识别研究进展综述[J]. 现代图书情报技术, 2010(06):42-47.

[17] 覃玉冰, 邓春林, 杨柳. 基于皮尔逊相关系数的网络舆情评估指标体系构建研究[J]. 情报探索, 2018(10):15-19.

[18] 滕青青, 吉久明, 郑荣廷, 李楠. 基于文献的中文命名实体识别算法适用性分析研究[J]. 情报杂志, 2010,29(09):157-161.

[19] 魏慧娟, 戴牡红, 宁勇余. 基于最近邻居聚类的协同过滤推荐算法[J]. 中国科学技术大学学报, 2016,46(09):736-742.

[20] 吴玺煜, 陈启买, 刘海, 贺超波. 基于知识图谱表示学习的协同过滤推荐算法[J]. 计算机工程, 2018a,44(02):226-232.

[21] 杨月欣. 中国食物成分表[M]. 北京医科大学出版社, 2005.

[22] 姚贤明, 甘健侯, 徐坚. 面向中文开放领域的多元实体关系抽取研究[J]. 智能系统学报, 2019,14(03):597-604.

[23] 张振亚, 王进, 程红梅, 王煦法. 基于余弦相似度的文本空间索引方法研究[J]. 计算机科学, 2005(09):160-163.

[24] 郑荣寿, 孙可欣, 张思维, 曾红梅, 邹小农, 陈茹, 顾秀瑛, 魏文强, 赫捷. 2015年中国恶性肿瘤流行情况分析[J]. 中华肿瘤杂志, 2019(01):19-28.

[25] BENGIO Y. Learning Long-term Dependencies With Gradient Descent is Difficult[J]. IEEE Trans Neural Netw, 1994,5.

[26] Bordes A, Glorot X, Weston J, et al. A semantic matching energy function for learning with multi-relational data Application to word-sense disambiguation[J]. Machine Learning, 2014,94(2):233-259.

[27] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data, 2013[C].

[28] Bordes A, Weston J, Collobert R, et al. Learning structured embeddings of knowledge bases, 2011a[C].

[29] Bordes A, Weston J, Collobert R, et al. Learning Structured Embeddings of Knowledge Bases: Aaai Conference on Artificial Intelligence, 2011b[C].

[30] Cheekula S K, Kapanipathi P, Doran D, et al. Entity recommendations using hierarchical knowledge bases[J]. 2015.

[31] Chen H, Li D, Wu Z, et al. Semantic web for integrated network analysis in biomedicine[J]. Briefings in Bioinformatics, 2009(2):2.

[32] Dai W, Xue G, Yang Q, et al. Transferring naive bayes classifiers for text classification, 2007[C].

[33] Dettmers T, Minervini P, Stenetorp P, et al. Convolutional 2d knowledge graph embeddings[C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[34] Dong C, Zhang J, Zong C, et al. Character-based LSTM-CRF with radical-level features for Chinese named entity recognition[M]//Natural Language Understanding and Intelligent Applications. Springer, 2016:239-250.

[35] Ertu U G Rul D C C E. FoodWiki: a mobile app examines side effects of food additives via semantic web[J]. Journal of medical systems, 2016,40(2):41.

[36] Helmy T, Al-Nazer A, Al-Bukhitan S, et al. Health, food and user's profile ontologies for personalized information retrieval[J]. Procedia Computer Science, 2015,52:1071-1076.

[37] Hochreiter S, Schmidhuber J U R. Long short-term memory[J]. Neural computation, 1997,9(8):1735-1780.

[38] Hu C, Cai W C, Huang L J, et al. A nutrition analysis system based on recipe ontology[J].

University of Taipei Medical, 2006.

[39] Ji G, He S, Xu L, et al. Knowledge Graph Embedding via Dynamic Mapping Matrix, 2015[C].

[40] Jouili S, Vansteenberghe V. An Empirical Comparison of Graph Databases: International Conference on Social Computing, 2013[C].

[41] Karatzoglou A, Hidasi B A Z. Deep learning for recommender systems, 2017[C].

[42] Lecun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015,521(7553):436.

[43] Lin Y, Liu Z, Sun M, et al. Learning entity and relation embeddings for knowledge graph completion, 2015[C].

[44] Lu C, Laublet P, Stankovic M. Travel Attractions Recommendation with Knowledge Graphs, 2016[C].

[45] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013a.

[46] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer Science, 2013b.

[47] Miller G A. Nouns in WordNet: a lexical inheritance system[J]. International journal of Lexicography, 1990,3(4):245-264.

[48] Mohagaonkar S, Rawlani A, Srivastava P, et al. HerbNet: Intelligent Knowledge Discovery in MySQL Database for Acute Ailments[J]. International Journal of Information Systems \& Management Science, 2018,1(2).

[49] Moreno A, Vails A, Isern D, et al. SigTur/E-Destination: Ontology-based personalized recommendation of Tourism and Leisure Activities[J]. Engineering Applications of Artificial Intelligence, 2013,26(1):633-651.

[50] Oramas S, Ostuni V C, Noia T D, et al. Sound and Music Recommendation with Knowledge Graphs[J]. Acm Transactions on Intelligent Systems & Technology, 2016,8(2):1-21.

[51] Passant A. dbrec—music recommendations using DBpedia, 2010[C]. Springer.

[52] RESNICK P. GroupLens : An open architecture for collaborative filtering of netnews[J]. Proc Cscw, 1994.

[53] Simmons, R. F. Answering English questions by computer: a survey[J]. Communications of the Acm, 1965,8(1):53-70.

[54] Sun Z , Deng Z H , Nie J Y , et al. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space[J]. 2019.

[55] Ting Y, Zhao Q, Chen R. Dietary recommendation based on recipe ontology, 2014[C]. IEEE.

[56] Wang Z, Zhang J, Feng J, et al. Knowledge graph embedding by translating on hyperplanes, 2014[C].

[57] Wang X,He X, Cao Y, et al. Kgat: Knowledge graph attention network for

recommendation[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019: 950-958.

[58] Wang H, Zhang F, Zhao M, et al. Multi-task feature learning for knowledge graph enhanced recommendation[C]//The World Wide Web Conference. 2019: 2000-2010.

[59] Wang H, Zhao M, Xie X, et al. Knowledge graph convolutional networks for recommender systems[C]//The world wide web conference. 2019: 3307-3313.

[60] Zhang F, Yuan N J, Lian D, et al. Collaborative knowledge base embedding for recommender systems, 2016[C].

个人简介

董洪伟，1995 年 6 月生，安徽阜阳人，2014～2018 年在武汉纺织大学数学与计算机学院软件工程专业学习，获工学学士学位；2018～2020 就读北京林业大学信息学院计算机技术专业型硕士，在读期间，顺利通过研究生各阶段的培养计划环节要求，积极参加科研课题项目，研究方向为人工智能，并在该方向发表申请软件著作权一项。

第一导师简介

付慧，女，1978.11。籍贯吉林农安县，副教授，CCF 会员。目前主要研究航拍图像拼接，图像识别和林业信息系统。发表十多篇学术论文，曾在《计算机研究与发展》,《计算机辅助设计与图形学学报》、《中南林业大学学报》、《北京林业大学学报》等刊物上发表论文。

第二导师简介

王立伦，男，河北人，2017 年至今在北京立防科技有限公司技术部任职高级工程师、技术总监，主要从事移动端和 IoT 设备上的软件架构研发工作。参与研发基于 O2C 的无人机平台，致力于将深度学习应用到无人机领域。

致谢

回望这两年，特别是在完成毕业设计以及毕业论文的过程，幸得各位的关心与帮助，使我倍感温暖的同时充满动力。在此，衷心感谢所有关心和帮助我的人。

首先，要感谢我的导师付慧老师。“古之学者必有师，师者，所以传道受业解惑也。”从毕业设计的开题到毕业论文的撰写，付慧老师给予了悉心的教导，在指引道路的同时帮助我解决了很多难题，让我受益匪浅。“饮其流者怀其源，学其成者念吾师。”老师对学术的热情与严谨、对专业知识的探索以及独特的见解无不深刻影响着我，督促着我不断前行，再接再厉。

其次，要感谢寝室的舍友，研究生涯难免枯燥与彷徨，而你们给我带来了欢声笑语，也带来了积极的学习态度。难以忘记，实验碰到瓶颈时你们认真的聆听和宝贵的意见、身体不适时你们关心的神情与温暖的举动、户外运动时你们的陪伴与追赶。感谢实验室的老师与同学，实验室虽然不大却充满温暖，老师认真且负责，为我提供了一个安静、整洁的实验环境；同学热情且友善，刻苦的科研态度让我成为更好的自己。行文至此，五味杂陈，祝愿诸君前程似锦。

同时要感谢我的家人们，你们的无私付出换来我求学时平和的生活，你们的默默支持与鼓励，陪伴我在学业上的道路上一路前行。

最后，感谢在百忙之中评阅论文和参加答辩的各位专家老师，感谢您们的辛勤劳动，也感谢您们对我的论文提出的宝贵意见。