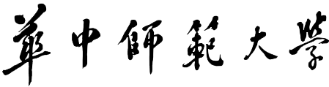
**分类号 密级**

**U D C 编号**



**硕士学位论文**

**基于文本挖掘的细菌-疾病关联知识库构建与预测**

|  |  |
| --- | --- |
| **学位申请人姓名：** | **罗雨晨** |
| **申请学位学生类别：** | **全日制硕士** |
| **申请学位学科专业：** | **计算机应用技术** |
| **指导教师姓名：** | **蒋兴鹏 教授** |

**硕士学位论文**

**基于LLM增强标注的营养-细菌-病毒-疾病知识图谱构建与预测**

**论文作者：**罗雨晨

**指导教师：**蒋兴鹏 教授

**学科专业：**计算机应用技术

**研究方向：**生物信息学

**华中师范大学计算机学院**

**2026年5月**

**Construction and Prediction of Microbes-Diseases Associations Knowledge Base Based on Text Mining**

*A Thesis*

Submitted in Partial Fulfillment of the Requirement

*For the M.S Degree in* *Computer Application Technology*

**By**

**Yuchen Luo**

**Postgraduate Program**

**School of Computer Science**

**Central China Normal University**

Supervisor: Xingpeng Jiang

Academic Title: Professor Signature 

Approved

May, 2026

**华中师范大学学位论文原创性声明和使用授权说明**

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的研究成果。除文中已经标明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

作者签名： 日期：2020 年 6月4 日

学位论文版权使用授权书

学位论文作者完全了解华中师范大学有关保留、使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属华中师范大学。学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许学位论文被查阅和借阅；学校可以公布学位论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存、汇编学位论文。（保密的学位论文在解密后遵守此规定）

保密论文注释：本学位论文属于保密，在 年解密后适用本授权书。

非保密论文注释：本学位论文不属于保密范围，适用本授权书。

作者签名： 导师签名:

日期： 2020 年 6 月 4 日 日期： 2020年 6月 4 日

本人已经认真阅读“CALIS高校学位论文全文数据库发布章程”,同意将本人的学位论文提交“CALIS高校学位论文全文数据库”中全文发布,并可按“章程”中的规定享受相关权益。同意论文提交后滞后：□半年；□一年；□二年发布。

作者签名： 导师签名：

日期： 2020年 6 月 4 日 日期： 2020年6 月 4 日

# 摘 要

在自然界中，微生物无处不在。这些微生物尤其是细菌寄生于人体，通过相互作用调节机体平衡，影响人体的新陈代谢。近几年来，有关微生物的研究数据，随着实验条件与技术水平的迅猛发展而快速扩增，这些数据的积累催生出基因组序列分析、生物网络构建、生物医学文本挖掘等计算方法，用以探究海量的生物医学数据中的潜在知识。本文主要针对细菌-疾病关联提取的问题，通过实体识别、关系抽取、知识库构建、关联预测等方法，从生物医学文本挖掘的角度开展微生物与人体健康的研究。本文的主要工作研究如下：

1. 提出了基于维基百科的细菌与疾病关联挖掘方法。已有研究通过文献检索和提取构建了细菌与疾病关联的数据库，但是没有用到全球最大的公共知识平台维基百科。维基百科里面蕴含了丰富的细菌与疾病实体之间的关联，从而可能成为文献挖掘的较好补充。针对该问题，本文利用文本挖掘工具Kindred，从维基百科正文中获取了细菌与疾病的交互作用。这项工作将已有的细菌与疾病关联数据库扩充了约16%，证明了通过维基百科的文本挖掘方法是文献挖掘方法的有力补充。
2. 开发了一个基于Django的细菌与疾病交互可视化平台。现有的细菌与疾病相关数据库，仅仅考虑了简单的关联信息，功能有限。本文整合了细菌栖息地、宿主等多源属性信息，通过对细菌的属性与疾病关联的相关性分析，可以验证细菌属性与疾病之间的相关关系。该系统平台还实现了关联数据查询、可视化等操作，提供了更加直观便捷的可视化分析工具，也为更深入的关联预测分析提供了可靠依据。
3. 提出了一个基于细菌与疾病知识库的关联预测方法。从文本角度的关联预测分析在通用领域得到了广泛应用，而在生物医学领域尤其是微生物关联的应用尚且较少。本文提出了基于双线性模型、神经网络及翻译模型预测细菌与疾病关联的方法。实验结果表明，翻译表示学习的预测效果优于其他方法，能够快速有效地预测未知的关联，并且能够实现知识库补全的功能。

本文提出的面向细菌与疾病关联知识库研究和可视化平台实现，能够实现细菌与疾病知识挖掘和关系预测的任务，提供了一个直观便捷的可视化分析工具。

**关键词**：生物信息学；文本挖掘；维基百科；细菌与疾病知识库；关联预测

# Abstract

Microbes are everywhere in nature. These microorganisms, especially bacteria, live on the human body, regulate the balance of the body and affect the metabolism through interaction. In recent years, with the rapid development of experimental conditions and technological level, the research data on microorganisms have been rapidly expanded. The accumulation of these data has led to the emergence of genomic sequence analysis, biological network construction, biomedical text mining and other computing methods, which are used to explore the potential knowledge in the massive biomedical data. This study mainly focuses on the extraction of bacteria - disease association from the perspective of biomedical text mining, with the usage of entity recognition, relationship extraction, knowledge base construction, association prediction and other methods to study on microbes and human health. The following shows the main work of this thesis:

(1) To propose a Wikipedia-based method of bacteria and disease association mining. Previous studies have built databases of bacteria and disease through literature search and extraction, except Wikipedia, the world's largest public knowledge platform. Wikipedia contains such an enormous associations between bacteria and disease entities that it might be a good complement to literature mining. To address this issue, this article used Kindred, a text mining tool, to extract the interactions between bacteria and disease from the texts in Wikipedia. This work expands the existing database of bacteria and disease by about 16 percent, proving that text mining via Wikipedia is a powerful complement to literature mining.

(2) To develop a Django-based bacteria and disease interaction visualization platform. Existing databases related to bacteria and disease only consider the simple correlation information, resulting in the function limitation. In this study, the multi-source attribute information such as bacterial habitat and host is integrated, and the correlation between bacterial attribute and disease can be verified by analyzing the correlation between bacterial attribute and disease. This system platform also realizes the query and visualization of correlation data, which provides a more intuitive and convenient tool for visualization and reliable evidence for future correlation prediction analysis.

(3) To propose a correlation prediction method based on knowledge base of bacteria and disease. The association prediction analysis from the perspective of text has been widely used in the general fields. However, the application in the field of biomedicine, especially the microbial association, is still rare. This thesis proposes a method to predict the association between bacteria and disease based on bilinear model, neural network and translation model. The results show that the predictive effect of translation representation learning is better than the other methods, which can predict unknown associations quickly and effectively and realize the function of knowledge base completion.

In this thesis, the research and visualization platform for knowledge base of bacteria and diseases can realize the tasks of knowledge mining and relationship prediction of bacteria and diseases, and provide an intuitive and convenient visual analysis tool.

**Keywords**：Bioinformatic; Text Mining; Wikipedia; Bacteria and Disease Knowledge Base; Link Prediction

# 

**目 录**

[摘 要 I](#_Toc41378007)

[Abstract II](#_Toc41378008)

[第一章 绪论 1](#_Toc41378009)

[1.1 课题研究背景 1](#_Toc41378010)

[1.2 课题研究意义 3](#_Toc41378011)

[1.3 论文的组织结构 4](#_Toc41378012)

[第二章 细菌与疾病知识库研究现状 6](#_Toc41378013)

[2.1 相关概念及研究范围 6](#_Toc41378014)

[2.2 知识库的研究 8](#_Toc41378015)

[2.2.1 基于生物医学文本的知识挖掘 9](#_Toc41378016)

[2.2.2 基于生物医学图谱的知识发现 10](#_Toc41378017)

[2.3 细菌与疾病知识库研究面临的挑战 10](#_Toc41378018)

[2.4 本章小结 11](#_Toc41378019)

[第三章 基于维基百科的细菌与疾病关联挖掘 12](#_Toc41378020)

[3.1 数据提取和预处理 12](#_Toc41378021)

[3.1.1 数据来源 12](#_Toc41378022)

[3.1.2 数据预处理 13](#_Toc41378023)

[3.2 细菌与疾病的交互关系抽取 15](#_Toc41378024)

[3.2.1 实体识别 15](#_Toc41378025)

[3.2.2 关系提取 16](#_Toc41378026)

[3.2.3 对比分析 16](#_Toc41378027)

[3.3 本章小结 18](#_Toc41378028)

[第四章 融合多源信息的细菌与疾病关联可视化平台 19](#_Toc41378029)

[4.1 细菌属性分析 19](#_Toc41378030)

[4.2 系统构建 22](#_Toc41378031)

[4.2.1 数据库设计 22](#_Toc41378032)

[4.2.2 实验环境 24](#_Toc41378033)

[4.2.3 平台构建 25](#_Toc41378034)

[4.3 本章小结 28](#_Toc41378035)

[第五章 基于细菌与疾病知识库的表示学习及关联预测 30](#_Toc41378036)

[5.1 概念与背景 30](#_Toc41378037)

[5.2 实验数据 31](#_Toc41378038)

[5.3 模型设计与开发 31](#_Toc41378039)

[5.3.1 双线性模型 32](#_Toc41378040)

[5.3.2 神经网络模型 35](#_Toc41378041)

[5.3.3 翻译模型 37](#_Toc41378042)

[5.4 模型训练与优化相关工作 38](#_Toc41378043)

[5.4.1 Early Stopping技术 38](#_Toc41378044)

[5.4.2 Dropout技术 39](#_Toc41378045)

[5.5 实验结果与对比分析 39](#_Toc41378046)

[5.5.1 实验评价指标 39](#_Toc41378047)

[5.5.2 参数选择对比 40](#_Toc41378048)

[5.5.3 网络模型对比实验 41](#_Toc41378049)

[5.5.4 模型预测结果 41](#_Toc41378050)

[5.6 本章小结 42](#_Toc41378051)

[第六章 总结与展望 44](#_Toc41378052)

[6.1 总结 44](#_Toc41378053)

[6.2 下一步研究工作 45](#_Toc41378054)

[参考文献 46](#_Toc41378055)

[攻读硕士期间发表的论文和参加的科研项目 53](#_Toc41378056)

[致谢 54](#_Toc41378057)

# 绪论

## 课题研究背景

生物信息学基于庞大而复杂的生物信息数据，覆盖了蛋白质序列分析[[1](#_ENREF_1)]、基因的变异与表达[[2](#_ENREF_2)]，物种代谢网络构建[[3](#_ENREF_3)]，生物医学文本挖掘[[4](#_ENREF_4)]等方方面面。随着生物技术的发展，人们对生物信息的认知越发的深入，因为实验研究所产生的数据量迅速扩增，人们对计算方法、分析手段的要求也在逐步提升。与此同时，随着计算条件与方法的创新发展，越来越多的生物医学研究者擅长借助信息技术手段，快速而高效的展开新的研究，极大的促进了生物医学领域水平的提升。因此，生物医学水平与信息技术水平始终在相互影响，协同并进，共同进步。

在生物医学领域，近几年来，微生物尤其是细菌常常是研究的热门话题，微生物在自然界中的作用也显得越发的重要。微生物在自然界中大量存在，并且在生态系统中扮演了至关重要的角色，生态循环的方方面面均离不开微生物的影响，人体也不例外。人体作为微生物的宿主之一，也拥有成千上万的微生物种群。这些微生物生存在人体的表面，例如皮肤、发囊，更多的，寄生在人体内，例如人体的器官、内脏、肠道等[[5](#_ENREF_5)]。总而言之，微生物与人体新陈代谢息息相关，微生物在人体不仅具有丰富的物种多样性，并且大量的微生物都在影响着生物体的新陈代谢，与我们共同进化与发展[[6](#_ENREF_6)]。然而，微生物总是在宿主内以群居状态寄生，与周围的环境以及宿主机体不断进行着物质交换，因此微生物的生长始终受到周围环境因素的制约，而环境和机体的平衡状态也会受到微生物菌群的影响。已有研究实验表明，随着微生物栖息环境、温度、好氧性等属性因素的改变，微生物的菌群大小、种类也会受到影响[[7](#_ENREF_7)]。

随着社会的快速发展，疾病的尤其是传染病的致死率逐渐攀升。根据世界健康卫生组织（World Health Organization，简称WHO）的数据显示，传染病在世界上的疾病死亡率占四分之一，并且大约二分之一的疾病致死发生在年轻的孩子身上[[8](#_ENREF_8)]。而影响传染病的一个重要因子是人体内的微生物组成，微生物在人类生物体健康中起着至关重要的作用。据研究表明[[9](#_ENREF_9)]，环境因素可以影响微生物区（Microbiota）的组成[[10](#_ENREF_10)]，因此也会导致一些疾病的发生[[11](#_ENREF_11)]。微生物区是指某一种类的微生物赖以生存的微小的环境，例如肠道微生物群，而环境因素通过人体内外的物质交换，能够不断调节这些微生物区的组成与更替，从而影响人体的新陈代谢，导致疾病的发生。例如季节改变导致的温度、湿度的变化对微生物区的组成有较大的影响，吸烟、熬夜等生活习惯也会改变肠道微生物菌群。

生物医学领域的知识纷繁复杂却又彼此紧密相连，近几年来，与细菌和疾病相关的数据库研究迅速增长，如图1.1所示，自1994年以来，生物医学的文献发表数量随着时间的推移呈指数型增长，在2019年，每年的生物医学文献发布数量接近25000篇。生物医学文本是基于实验验证得以发表的文献资料，因此文本信息作为重要的知识成果，将其收集并整理成库对后续的研究分析十分重要。然而文本是一种非结构化数据，研究人员需要通过人工或其他方法去除其中的冗余信息，得到方便存储的结构化知识。仅仅依靠人工的方法从这些非结构化文献中挖掘潜在关联十分困难，因此使得利用计算方法，让机器学会阅读理解文本，并且挖掘潜在关联成为可能。目前已有大量研究通过利用美国国家生物信息中心（National Center for Biotechnology Information，简称NCBI）中提供的生物医学文献[[12](#_ENREF_12)]，将非结构化文本信息提炼成结构化关联，然后整合为知识图谱得以发布。

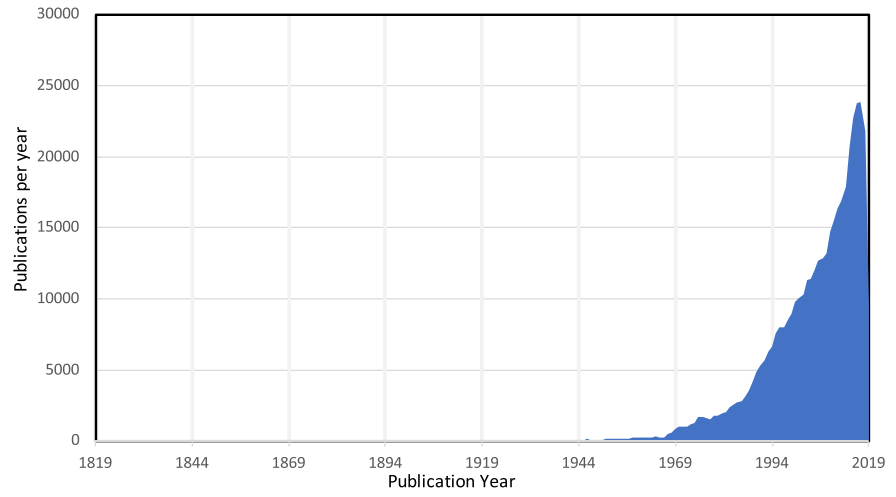


图1.1 细菌与疾病关联文献发表数量与发表年份关系示意图

除此之外，一些公共资源平台例如维基百科、百度百科等也蕴含着丰富的挖掘价值。维基百科是一个由专业的编写人员，集成了大量公共数据和知识的开放资源平台。英文版本的维基百科包括59,000,000篇英文文章，48,000,000个页面。每天都有来自世界各地的成千上百万的访问者共同编辑这些文章，并且创造了成千上百万的新文章，以扩充现有的维基百科知识库。通过超链接来链接这些文章是维基百科的重要特性，来自于世界各地的任何结构化资源都可能被导入维基百科，并且形成紧密的全球性的维基百科知识图谱[[13](#_ENREF_13)]。因此，维基百科这一公共资源平台所蕴含的非结构化的知识对于生物信息研究来说，是一个丰富的数据来源。本文利用了文本挖掘的计算方法，将维基百科中的非结构化知识提取成结构化关联，整合为知识库网络。

通过已有的文献挖掘得到的结构化知识，构建知识库网络，只能蕴含部分已有实验验证的信息，对于后续的生物研究，还需要补充更多的未知关联信息。对于知识库网络，尚且有许多实体间没有建立关联，可以通过网络分析、表示学习等方法，预测出这些未知关联，对已有的知识图谱网络进行补充。对于细菌与疾病知识而言，由于细菌与疾病的分类太多，导致知识库网络更新较慢。利用已有的知识库进行关联预测，不仅能够发现填充未知的细菌与疾病关联，还能为微生物实验验证方向提供有力的证据支撑。

综上所述，微生物与疾病研究在计算科学、生物医学等多个领域早已有较多的研究基础，通过对细菌的作用分析，已经证实了其在人体内的重要作用。本文正是基于这些研究背景，提出了利用计算方法，展开对细菌与疾病关联的知识挖掘与知识发现研究。

## 课题研究意义

一直以来，对于微生物与疾病的研究一直是世界各地实验室关注的焦点。然而，对于微生物与疾病之间的关系的研究[[14](#_ENREF_14)]，目前主要有三种方法：

（1）基于测序的方法。利用基因测序技术与生物实验相结合的方法，研究微生物与机体疾病发生的直接关联性[[15](#_ENREF_15)]；

（2）基于代谢网络的微生物关联分析的方法。通过微生物的代谢物网络分析，找到与疾病的潜在关联[[16](#_ENREF_16)]；

（3）基于文本挖掘的关联预测方法。从海量的生物医学文本中分析句法特征，找到经实验验证的微生物与疾病的关联[[17](#_ENREF_17)]，并通过对现有知识图谱的表示学习，预测更多的未知关联。

本文则是采用的第三种方法。近几年来，随着生物医学文献大量涌现，为挖掘相关生物实体的潜在关联提供了必要的条件。利用机器学习或者深度学习的方法，将生物医学文本中隐藏的实体关联关系大规模提取，并以三元组的方式整合为生物医学知识库，为生物实验设计提供了可靠的参考，减少了生物实验中不必要的尝试，在一定程度上节省了实验资源。在生物医学实体数据积累到一定程度，通过计算方法进行预测与推荐，能够为临床问诊提供一定的决策辅助[[18](#_ENREF_18),[19](#_ENREF_19)]。同时，通过进一步了解微生物与疾病之间的关联，也为药物设计[[20](#_ENREF_20)]、靶向治疗方法的研究[[21](#_ENREF_21)]、营养调理方案[[22-24](#_ENREF_22)]等提供了理论依据与数据支撑。

对于微生物和疾病这对特定的实体关系提取，也有一些团队进行了尝试。2016年Ma等人[[12](#_ENREF_12)]人工的阅读了近百篇生物医学文献，从中获取了47种微生物以及37种疾病，并结合相似度算法将不同定植位置的微生物分成了不同的组。然而在实际情况下，微生物与疾病之间的相互作用会受到环境因素或微生物自身属性的影响而发生变化。因此，本文提出了基于整合的细菌与疾病关联知识库，融合了细菌的属性信息，进一步丰富了细菌与疾病交互的知识。

由于通过计算方法从生物医学文本中挖掘到的细菌与疾病关联，都是经过实验验证得到的有效结论，因此通过文本挖掘构建的关联网络可靠性较强，然而尚且还有更多的未知关联蕴含在知识网络结构中。除了构建一个完整的细菌与疾病知识库，本文还利用了双线性模型、深度学习模型以及翻译模型对网络图进行表示学习，通过模型预测网络中更多的未知的关联信息。以此进一步扩充现有的知识库，这些关联也能够在接下来的生物实验中，作为有效的实验依据，降低实验成本，提高实验效率。

综上所述，基于生物医学文本的文本挖掘与文本发现任务，在微生物与疾病关联研究中具有较强的可行性。通过计算方法从维基百科的正文文本中获取到的细菌与疾病交互关联，是经过研究发现的可靠知识，将非结构化文本转化为结构化知识，能够为后续的研究提供更加高效的数据资源；将多个来源的细菌与疾病关联数据以及细菌的属性知识整合，通过标准化手段降低数据的冗余度，能够提供给研究人员更多的便利对这些知识进行查询和分析；整合成的关联网络，利用机器学习和深度学习的方法对现有的知识图谱进行表示学习，对未知的连接进行关联的预测，以获取到更多的知识，为关联发现、疾病治疗等提供可靠依据。

## 论文的组织结构

本文主要研究了有关细菌与疾病关联知识挖掘、知识库的构建以及关联预测等方面的问题，全文主要分为五个部分，结构如下：

第一章，主要介绍了相关细菌与疾病交互研究的背景，研究的意义以及论文的总体组织结构。

第二章，主要介绍了相关细菌与疾病知识库的研究进展，目前国内外的研究现状，其他相关领域的研究方法介绍，以及针对目前的研究内容可能遇到的挑战，接下来则按照图1.2的流程，以此介绍细菌与疾病关联知识库的研究。



图1.2 本文研究流程示意图

第三章，详细介绍细菌与疾病关联挖掘研究过程，主要介绍细菌与疾病的命名实体识别，实体消岐，细菌与疾病的关系提取，以及与其他数据库的对比分析。

第四章，通过整合多个数据源的细菌与疾病的关联，以及细菌的栖息地、宿主等属性关联，构建了基于Django的细菌与疾病的交互可视化平台数据库。

第五章，开展了细菌与疾病知识库的推理与应用研究。首先在已有的细菌与疾病数据库的基础上，通过实体对齐的方法，整合了细菌的属性关联，包括细菌与宿主、细菌与栖息环境的关联，构建了关联预测预料库，在主流模型结构上取得了较好的表现，以此能够预测出知识库中未知的交互关联。

第六章，主要针对两个主要的研究内容以及研究方法的分析，获得了一些结论，并根据结论提出了未来研究的展望。

# 细菌与疾病知识库研究现状

微生物在自然界中广泛存在，而细菌作为微生物中的主要成员之一，在生态系统以及人类的新陈代谢过程中起着至关重要的作用。关于细菌与人类健康状态的研究早已逐步开展，无论是从生物实验、蛋白质序列分析以及基因表达等角度，都已经证明了细菌对人类健康状态有着直接和间接的影响。随着生物医学数据的迅速扩增，利用计算方法在生物医学文本中挖掘非结构化关联的研究也在逐渐推进，本章主要介绍了目前已有的细菌与疾病交互关联挖掘的任务，以及微生物与疾病交互的关联数据库，本章还总结了目前这些研究所面临的诸多挑战。

## 相关概念及研究范围

（1）人类微生物组计划

微生物遍布人体，主要分布在口腔、皮肤黏膜以及胃肠道等人体内外表面[[25](#_ENREF_25)]。2007年首次提出了人类微生物组计划（Human Microbiome Project，简称HMP）[[7](#_ENREF_7)]，与人类基因组计划相类似，即利用高通量技术，全面地刻画人类微生物群落的特征，并且通过跨学科的努力，探究人体微生物与健康或疾病的关联。该计划由美国国家卫生研究院（National Institutes of Health，简称NIH）资助，并且在2013年继续扩展为综合性人类微生物组计划（Integrative Human Microbiome Project，简称iHMP）[[26](#_ENREF_26)]，进一步缩小了研究范围，将人体分为不同的生理部位，以此研究不同部位下微生物组的影响与变化。

除了来自父母遗传的基因组，人类具有的另一个基因组则是通过出生以后的成长阶段逐渐进入人体，因此受到了生活状态、生存方式的影响，而在之前人类对人体中所存在的大量微生物组的组成与影响却不得而知。人体的健康状态平衡、新陈代谢与自然生长需要遗传基因组和微生物组的共同作用。因此，在对人体健康状态、营养物发现、药物与疾病相互作用的研究中，对于这些微生物的作用是不可小觑的。

人类微生物组计划自2007年底开始执行，在第一阶段的研究中，人体肠道微生物的组成是计划的主要研究成果，通过对微生物组的基因测序分析，描绘了健康人体不同身体部位的微生物组群落、微生物通路，并将其与一些代谢表型结合起来。而计划的第二阶段于2019年6月完成[[27](#_ENREF_27)]，并且发布了人体内不同生理部位的微生物与糖尿病[[28](#_ENREF_28),[29](#_ENREF_29)]、早产[[30](#_ENREF_30)]和炎症性肠病[[31](#_ENREF_31)]具有关联性，进一步描绘了肠道微生物与宿主的健康平衡状态随着时间的推移在动态变化。

（2）细菌与疾病交互

数以万亿计的微生物栖息在人体内，通过代谢活动和宿主间的相互作用，影响正常的生理机能。肠道微生物会自我调节以适应人体各个阶段的变化，但是如果微生物失调就会导致人体的心理或生理出现问题。如图2.1所示，肠道微生物在人体内能够导致不同的疾病[[32](#_ENREF_32)]。

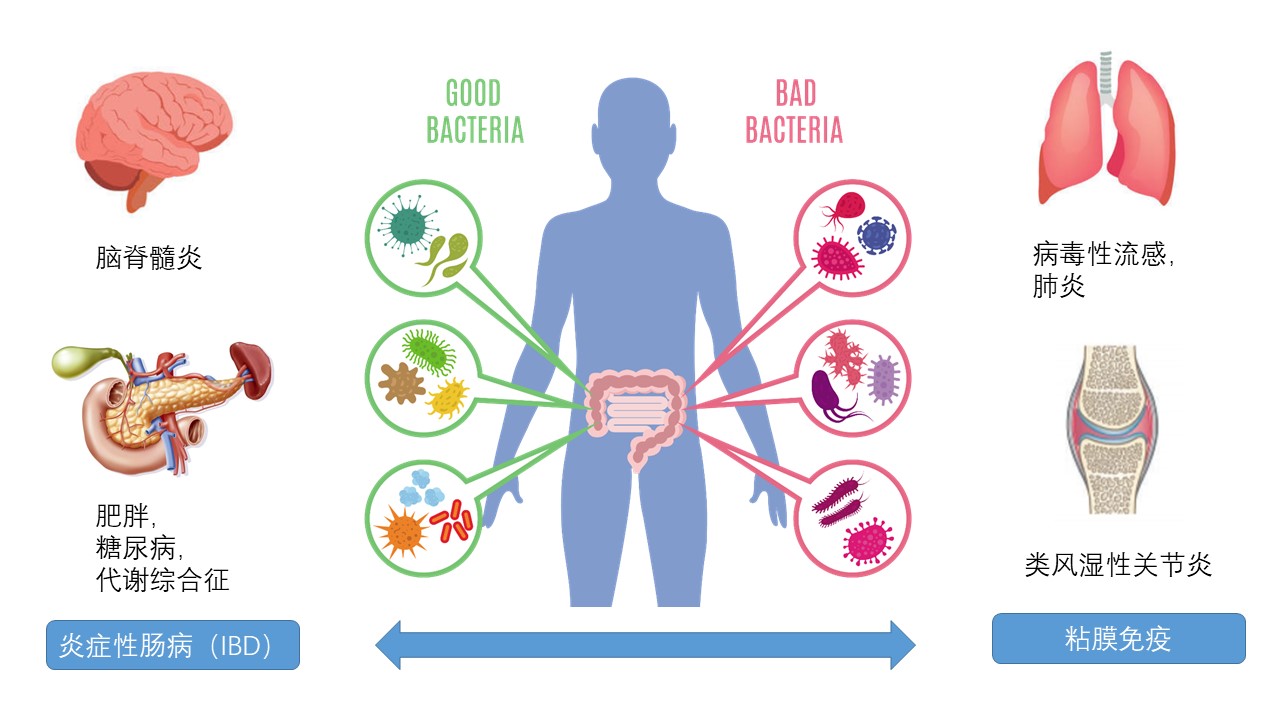


图2.1 微生物可能在人体的不同部位致病示意图

大肠杆菌存在于人的肠道中，并发挥着重要的功能。大肠杆菌[[33](#_ENREF_33)]，也就是聚集在紫色基质上的黄色杆状菌，大多数菌株不仅无害，而且有益。肠道中的大肠杆菌能够制造维生素K和B12，以及击退引起疾病的细菌。大肠杆菌也具有生物特异性，能够在一些动物体内导致腹泻等疾病。又如肺炎链球菌[[34](#_ENREF_34)]可以分裂成子细胞，导致严重的疾病，如脑膜炎和肺炎，但像大肠杆菌一样，一些链球菌是无害的，能够为人体提供必要的代谢交换物质，以维持机体的平衡状态。因此，细菌虽然广泛存在于皮肤、口腔、呼吸道和肠道，但也会由于不同的属性与相互作用，在不同的部位对人体造成一定影响。

人体微生物群落是维持健康和免疫系统的一个组成部分，人体与肠道共生菌之间的交流对免疫反应既有增强作用也有不利影响[[35](#_ENREF_35)]。这些细菌通过自己的栖息与生长，与人体机体之间产生物质交换，并且通过细菌之间以及细菌与环境之间的相互作用，影响人体的机体平衡。若由于某些原因破坏了肠道微生物与宿主之间的平衡，例如过度饮食、天气状态、病毒侵入等原因，则会导致肥胖[[36](#_ENREF_36)]、免疫性关节炎[[37](#_ENREF_37)]、炎症性肠病[[38](#_ENREF_38)]等疾病的诱发。

除了免疫系统疾病以外，近几年来精神类疾病的影响也越发重要。随着社会快节奏发展，人体的机体平衡也受到生活压力与不良生活习惯的双重影响，进而发展为精神类疾病，例如抑郁症、焦躁症等。有文献说明，肠道微生物群与中枢神经系统之间的双向通道，即微生物群-内脏-大脑轴[[39](#_ENREF_39)]，影响着人体及动物中的社会行为的各种复杂方面，通过对肠道微生物群与社会行为之间的关系，能够有助于解释一些精神疾病的起因，例如抑郁症、自闭症等近几年来大量发生的心理缺陷，并且针对这些疾病特征研究新的疗法。

（3）生物医学关联预测

如今生物医学数据量暴增，如何将大规模的生物医学数据格式化、链接化等问题成为扩展下一步研究的关键问题，由此衍生了生物医学知识图谱的概念。知识图谱（Knowledge Graph）是一种以网络形式表示的数据存储格式，最初由通用领域的知识库扩展而来。知识图谱中的实体表示现实条件下的各种角色，在生物医学文本中生物医学的研究对象则均可以作为实体，在知识图谱中实体用点来描述。实体之间通过边来连接，这些边表示实体之间的关系，例如细菌与细菌之间的共生关系，细菌与疾病的致病关系，细菌与病毒之间的寄生关系等，由此看来，知识图谱是基于数据库之上更能适应当下数据结构形式的一种新颖的数据存储模式[[40](#_ENREF_40)]。

关联预测是在知识库或知识图谱的基础上，进一步深入研究所提出的研究方法[[41](#_ENREF_41)]，通过构建实体节点之间的关联网络，将其作为无向图来表示，并且嵌入不同类型的特征[[42](#_ENREF_42)]，预测出潜在的实体关联。例如实验证明了细菌A能够导致B疾病，而研究也发现疾病B与疾病C之间存在相似性，则通过关联预测希望能够发掘细菌A与疾病C之间存在的某种关联，并且将这些未知关联作为进一步的研究方向或实验设计。在当下的类似生物医学领域的研究中，关联预测的方法已经被广泛应用于药物发现、药物与疾病的关联预测等方面。

## 知识库的研究

在生物信息领域，知识库的研究通常是一个重要的研究产出，为的是给研究人员提供更加便捷的查询与存储方式。知识库的构建流程较为复杂，除了必要的数据库系统开发，最重要两个步骤便是知识挖掘与知识发现任务。已有的细菌与疾病知识库数据分别来源于生物医学文献、实验设计资料或临床电子病历等不同方面，随着知识来源的不同，其存储格式与数据标准也不尽相同。有了较为系统的知识库以后，还需要对知识图谱进行图表示学习，在现有的连接上预测更多的未知关联。知识挖掘是在已有的文献中得到经过实验验证的结构化信息，是构建知识库的基础；知识发现是从知识库网络的连接中找到潜在的未知关联，这些关联的有效性还需要进一步验证，是构建知识库的应用。在本文中，对细菌与疾病知识库的研究始终围绕知识挖掘与知识发现两个步骤展开。

### 基于生物医学文本的知识挖掘

一直以来，有关生物医学实体的关联研究是生物信息学等交叉领域研究的重点，随着生物医学领域研究的爆发式增长，生物医学文献数量也在大量累积。在生物医学文本中，文献中蕴含的语义关系是生物医学关系进一步整理与研究的方向，利用计算方法挖掘交互关系是当前针对生物医学实体交互关联研究的重要方法之一。除此以外，基于基因序列分析、生物网络代谢物分析等方法也能够更加直截的对生物实体关联展开研究，这些方法是在未知关联的基础上，通过数据分析手段发现新的关联。而知识挖掘任务是试图通过已有的文献，找到蕴含的已知实体交互关系。

基于生物医学文本的知识挖掘，即利用机器学习或深度学习等计算方法，结合句子的特征，利用实体的共现挖掘微生物与疾病的关联关系。此前，已有大量的实验针对不同的实体关系做出了显著的成果。

在2011年，Sherwin, E等人[[43](#_ENREF_43)]建立了药物与药物相互作用的语料库，并且采用机器学习的方法自动从生物医学文本中获取药物与药物相互作用关系。2017年，Peng等人[[44](#_ENREF_44)]采用机器学习与深度学习模型组合投票的方法，自动检测文本中的化学-蛋白质关系。维基百科作为一个公共资源平台，集成了全球的各大知识库，近些年来，也有一些研究人员开始利用维基百科进行生物医学的研究。Zinovyev等人[[45](#_ENREF_45)]利用维基百科中隐藏的关联网络，研究了蛋白质之间的某些关联。Rollin 等人[[46](#_ENREF_46)]通过维基百科网络分析，研究了世界各个国家每年的传染病的爆发日志，并且提出了可靠的预测分析。

2016年，Ma等人[[47](#_ENREF_47)]人工的阅读了近百篇生物医学文献，从中获取了47种微生物以及37种疾病，并结合相似度算法将不同定殖位置的微生物分成了不同的组。但由于他们团队使用了人工阅读相关文献的方法，获得的数据量太少，并且更新较为缓慢。2018年，Janssens等人[[48](#_ENREF_48)]结合了外部数据，将微生物与疾病实体标准化，将生物医学文本挖掘的结论与临床病例数据进行了统一，为后续的辅助决策与实验设计提供了帮助。还有许多团队在微生物和疾病领域分别进行了命名实体识别等其他任务[[49-51](#_ENREF_49)]。

然而，对于微生物与疾病的关联挖掘上，机器学习、深度学习等计算方法的运用尚且较少，无法适应当下细菌与疾病交互关联研究文献数量迅速扩增的现状。与此同时，由于已发布的细菌与疾病知识库的数据来源不尽相同，导致各数据库中的数据格式与标准未能统一。细菌属性作为影响细菌与疾病交互的关键因素之一，尚且没有数据库包含这一类数据。综上所述，本文提出了基于整合的细菌与疾病交互关联知识库的构建研究。

### 基于生物医学图谱的知识发现

目前世界各地的研究团队针对微生物与疾病的相关研究，发布了相关的数据库，同时，还有许多针对生物医学实体关系抽取的评测任务，包括BioCreative[[52](#_ENREF_52)]，BioNLP[[53-56](#_ENREF_53)]，SemEval[[57](#_ENREF_57)]等，这些涉及到关系抽取的评测任务也很大程度上促进了生物医学实体关系挖掘的进展。对基于生物医学图谱的知识发现研究已经有一些尝试，一些研究团队通过将已有的生物医学实体之间的关联收集起来，并且采用了传统的统计学及机器学习方法，在蛋白质互作用发现，药物发现，化合物作用研究等方面均取得了一些进展。

Cameron等人[[58](#_ENREF_58)]提出了一种自动的子图模型，能够根据聚类和构建句法图的方法说明潜在的生物医学实体之间的关联。Malas等人[[59](#_ENREF_59)]则是采用了知识图谱特征的方法，例如大量的媒介关联概念词，上下文语义的类别词等，通过已知的药物和疾病关系来预测新的药物和疾病的诱发关联。Bakalb等人[[60](#_ENREF_60)]使用了简单路径的方法，将生物医学实体关系链接起来，作为逻辑回归模型的特征，以达到药物发现的目的。除了这些方法以外，最近也有一些研究将深度学习的方法应用到药物发现领域[[61-63](#_ENREF_61)]。这些方法着眼于将生物医学数据库中已有的药物与其靶标之间的关联结合起来，却忽略了生物医学文本以及生物属性等其他上下文关系，而这些上下文关系却在真实的关联挖掘中起到重要的作用。

基于细菌与疾病关联知识库的知识发现研究尚且不多，然而随着相关知识库研究的深入，本文也发现了对现有知识库进行关联预测以扩充的重要性[[64](#_ENREF_64)]。因此，本文提出了一种基于整合的细菌与疾病知识库的知识发现任务，采用了双线性模型、深度学习模型以及翻译模型等多模型的综合方法，尝试在已有的知识库中找到更多的未知关联信息，为接下来的实验研究提供有效的依据。

## 细菌与疾病知识库研究面临的挑战

最近几年国内外生物医学研究的不断前进。虽然将计算方法应用于生物医学文本挖掘的研究一直在创新，但是目前有关微生物和疾病知识库的研究尚且面临着诸多困难，而这些困难也促使着细菌与疾病的知识库研究具有更大的挑战性，具体面临的挑战如下：

1. 目前已有的微生物与疾病数据库构建方法仍不能达到快速高效的目的，大多数研究方法仍然采用人工阅读生物医学文本的方式，不符合生物医学数据迅猛增加的态势，因此需要结合计算方法，利用机器学习、深度学习等自动化挖掘工具，达到快速获取，保证持续更新的目的。
2. 由于生物医学文献的独特性，有许多命名标准尚且没有统一。例如在细菌命名中，有许多的缩写、别名、俗称等不同的命名方式；疾病也会由于临床和研究环境下，可能会有不同的命名规范。而在文献的写作过程中，作者往往也会有自己的习惯，在上下文中采用缩写的方式简化冗长的命名，因此对提取数据的标准化是知识库构建过程中必不可少的关键一步。
3. 细菌与疾病的关联研究有时还需要结合上下文的语境，在不同的生存环境下的相同细菌可能导致不同的疾病，而在文本挖掘时仅仅只考虑了细菌与疾病的关联，细菌的属性上下文很容易被忽略。
4. 基于生物医学知识图谱的知识发现任务受到图谱大小的制约，如果图谱大小稀疏并且特征不明显，传统的规则方法很难从中发现新的关联知识，并且传统方法不具有较好的迁移能力。

因此本文提出了基于维基百科挖掘细菌与疾病关联的方法，并结合细菌属性关系，通过实体对齐方法构建了细菌与疾病知识库。由于知识库的数量较少，关系网络较为复杂，本文还提出了基于细菌和疾病知识库关联预测的方法，从已知的细菌关联中发现未知的知识，扩充现有的知识库，为进一步的实验提供可靠的参考依据。

## 本章小结

本章主要对细菌与疾病的知识库研究的相关概念，目前已有类似相关研究的研究方法，以及可能会面临的一些挑战做了详细的介绍。随着人类微生物组计划的提出，微生物研究已经成为当下生物医学领域研究的热点话题，微生物与疾病的因果关系也在不断的被相关研究所论证，因此微生物对于人体生长发育、新陈代谢的影响十分重要。而通用领域内的计算方法的引入也进一步加快了对微生物的研究，计算模型不仅能够缩短研究进程，也更加能够适应当下的研究环境，降低实验的人力与时间成本。目前对于生物医学实体知识库的研究可以分为知识挖掘和知识发现两个进程，可能遇到实体消歧、结合上下文等不同的挑战。综上所述，接下来首先开展的是基于维基百科的细菌与疾病交互关联挖掘的工作。

# 基于维基百科的细菌与疾病关联挖掘

前两章首先介绍了基于整合的细菌与疾病知识库研究的背景，以及研究的意义，说明了有关疾病与细菌知识库的研究在当下生物医学领域的重要性。其次，细菌与疾病知识库的研究主要分为两个部分，分别是知识挖掘与知识发现。本章主要就知识挖掘的工作展开论述，如图1.2所示。本章提出了基于维基百科公共平台的数据挖掘方法，分别进行了数据预处理、命名实体识别、关系提取与实体对齐等工作。

## 数据提取和预处理

本章主要针对公共知识平台维基百科的正文文本进行知识的挖掘。维基百科为研究人员提供了开放的页面数据下载入口，本节主要介绍源数据格式，下载方法以及数据预处理办法。

### 数据来源

细菌的种类繁多，分类树级别复杂。在进行有关细菌的研究之前，首先对细菌的分类树进行了分析。细菌的分类级别由高到低依次为界门纲目科属种，级别越高则研究视野越大，对细菌与疾病的关联越模糊。而在物种级别以下，有亚种和菌株的分类。但由于亚种与菌株的文献较少，且作者在论文中命名尚不规范，因此本节选定了物种级别的细菌与疾病的交互研究。

在维基百科中包括上千篇对特定的细菌物种的索引页面，这些页面对细菌的起源、作用、演化等相关研究进行了介绍。维基百科页面中包含有表格数据和正文数据两个部分，对于细菌而言，表格的结构化数据主要是细菌的分类树信息，对于疾病等交互研究并无直接关联，因此本章主要针对正文的非结构化文本展开研究。维基百科正文页面本身，蕴含着复杂的语义网络，其中有关细菌物种的页面大约6,000多条。本章采用了数据挖掘的办法，在正文描述中，挖掘出对应的细菌与相关疾病之间可能的联系。维基百科的内容可以通过两种方式获得：

（1）维基百科发布的可直接调用API。这种方法适用于针对给定的话题或目标进行研究分析；

（2）下载维基百科官方提供的数据库转储文件Wiki Dumps[[65](#_ENREF_65)]。

由于维基百科的话题覆盖了社会、文化、历史、教育等各个方面的知识，因此第一种方法尚且不能直接用于生物医学领域的文本挖掘研究。我们采用的是第二种方法，即通过转储文件进行数据清洗和预处理，进一步挖掘到细菌与疾病的关联。

### 数据预处理

维基百科上的页面涵盖社会、生活、人物、历史等成千上万个话题，本节要从中找到与细菌相关的页面正文，则需经过一系列的数据预处理步骤。



图3.1 数据预处理流程示意图

上述流程图详细的介绍了本节在数据预处理过程中的具体流程，以及从预处理到命名实体识别与关系抽取过程中引用的外部数据库。如图3.1所示，首先，从维基百科提供的官方下载页面下载了数据库转储文件，其中包含所有维基百科页面当前最新更新版本的条目、模板、图片描述以及基本的元页面[[66](#_ENREF_66)]。由于该数据库转储文件是以XML格式存储的，该格式完整有序但数据量庞大。因此在下载完成后，无需进行解压，而是编写了转储文件处理脚本，将其中的链接信息及正文文本信息整理成约13,000个文本文档，并且按照首字母顺序分块存储。接下来，本节根据Wikidata中的唯一标识找到对应细菌的正文页面文本。

Wikidata是基于维基百科的综合结构化标识，以三元组的方式存储了通用领域的知识及关系。在Wikidata中，前缀Q表示实体，前缀P表示一种关系类型[[67](#_ENREF_67)]。我们利用Wikidata Query Service[[68](#_ENREF_68)]，利用SPARQL查询的方式，得到了细菌436,196个条目，疾病8,483个条目，每一个实体条目都有对应的Wikidata编号[[69](#_ENREF_69)]。

图3.2显示了Wikidata中的唯一标识与维基百科的词条链接正文页面的对应关系。上面的黑色方框内显示的是Wikidata中对应词条“Treponema pallidum”的编码信息，包括不同语言的细菌名称，细菌的物种关系等。Wikidata中关于对象编号分别有两种方式，其中实体条目用“Q”开头的编号表示，关系条目用“P”开头的编号表示。此处用椭圆形标识框凸显的“Q156007”则对应于细菌“Treponema pallidum”的唯一标识。而下面的方框内对应的是维基百科的正文页面关于词条“Treponema pallidum”的介绍内容，包括细菌的介绍、细菌的发现、细菌与疾病以及细菌的图片与分类树等信息。维基百科的正文页面负责将该条目的详细内容展示给用户，包括了条目的详细说明、背景及应用扩展等多角度的话题内容，而Wikidata则通过编码的方式存储正文页面中有限的三元组关系，以构建完整的语义网络。目前Wikidata中尚且没有存储细菌与疾病的互作用关系，而在维基百科的正文文本中，这种关系是蕴含在非结构化的文本语义中的，因此我们可以通过数据挖掘的方法得到这一类互作用关系。



图3.2 Wikidata唯一标识与维基百科正文链接对应示意图

从中提取到关于该细菌的概述“Treponema pallidum is a spirochaete bacterium with various subspecies that cause the diseases syphilis, bejel, and yaws. It is transmitted only amongst humans.”，这句话介绍了该细菌与三种不同疾病的关联，但在Wikidata中没有对应的关系编号描述，本节按照Wikidata的描述方法得到了以下三个三元组，分别描述了细菌“Treponema pallidum”与“syphilis”、“yaws”、“bejel”三种疾病的交互关联：

（1）<Treponema pallidum：Q156007，has effect：P1542，syphilis：Q41083>;

（2）< Treponema pallidum：Q156007，has effect：P1542，yaws:Q76973>;

（3）< Treponema pallidum：Q156007，has effect：P1542，bejel:Q682798>。

由此可见，在维基百科的正文文本中蕴含了许多易于提取的细菌与疾病的交互信息，并且这些信息能够按照Wikidata的编号组合为三元组格式。通过Wikidata的唯一标识搜索，本节在13,000个文本文档中找到了有关细菌物种的6,332篇正文文本，得到了包含标题、正文以及Wikidata编号的文本数据。接下来，我们将使用基于Python的开源Kindred[[70](#_ENREF_70)]工具包对得到的文本数据进行命名实体识别和关系提取等工作。然而，Kindred对文本数据的格式有一定要求，因此本节还将原格式转化成了BioC format格式[[71](#_ENREF_71),[72](#_ENREF_72)]。

## 细菌与疾病的交互关系抽取

接下来，本节将利用Kindred工具对得到的细菌正文文本进行命名实体识别和关系抽取工作，整体的流程如图3.3所示，接下来则按照该流程示意图详细介绍细菌与疾病的交互关系抽取流程。



图3.3 细菌与疾病交互关系抽取流程示意图

### 实体识别

对维基百科的正文文本，经过了去重、格式转换等数据清洗之后，利用Kindred对文本进行命名实体识别。Kindred是斯坦福开发并开源的一个文本挖掘工具包，提供给使用者对生物医学文本识别并提取实体、关系等接口。实体识别则是通过一个精确词典搜索的方案，将文本中的基因、环状RNA或其他的生物医学名词识别出来。

本节通过Wikidata的SPARQL查询得到了疾病及其临床术语别称，通过Kindred分别进行命名实体识别，以及疾病实体标准化。对细菌实体，本节利用了UMLS[[73](#_ENREF_73)]中提供的细菌及其缩写名称，能够有特定性的识别出首字母缩写实体名称，同时完成细菌的实体消歧任务。经过实体识别得到了2000条句子，每个句子中既有细菌实体，又包含疾病实体。

### 关系提取

根据命名实体识别得到的含有目标实体的句子，本节提出了假设：在一个句子s中如果出现了两个目标的实体对<e1,e2>时，便给这个句子做一个关系标记，于是得到一个三元组<e1,r,e2>,这个也可以称为是一个实体对的目标关系。每个句子标识出了细菌的Wikidata编号。通过对维基百科的内容分析，本节发现维基百科对细菌与疾病之间的关系描述比较直白而且客观。

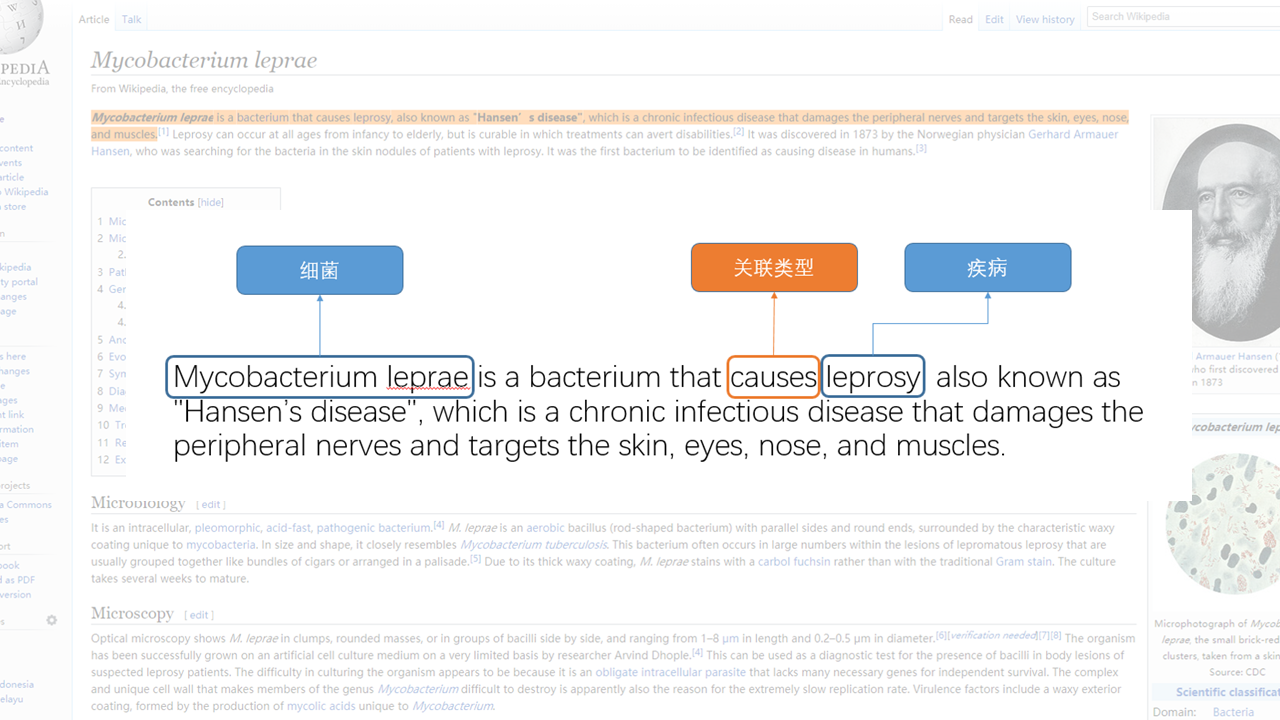


图3.4 维基百科正文关系提取目标句子示意图

例如图3.4中的句子，麻风分枝杆菌（Mycobacterium leprae）是一种能够导致麻风病又名汉森氏病的一种细菌。在维基百科中，类似的句子还有很多，由于维基百科的正文文本的科普性较强，如果某细菌经过证明致病，便会以较为明显的特征词的形式表达。因此本节直接采用了关键词'infect', 'cause', ' pathogen'匹配的方式，对标记后的句子进行过滤，得到了有关细菌与疾病相互作用的目标句子，并利用Kindred对这些句子进行了关系提取。通过对维基百科正文文本的细菌与疾病互作用挖掘，我们得到了132种细菌，152种疾病，新增了208对细菌与疾病之间的关系。在本章的工作中，仅仅研究了物种级别的细菌与疾病之间的相互作用。

### 对比分析

进一步地，本节将获取的微生物与疾病的交互关系与其他数据库已有的数据进行了对比分析。Ma等人[[12](#_ENREF_12)]提出了MicroPattern，通过人工阅读在PubMed上500余篇文献的方法，手工获取到的39种微生物，39种疾病，以及409对微生物与疾病的互作用关系。图3.4显示了维基百科与MicroPattern两种不同数据来源的数据情况对比，分别从细菌种类，疾病种类数以及细菌与疾病关联数三个方面进行比较。

首先，从文献依据角度， MicroPattern中的文献仅仅是最初人工获取的500篇摘要文献，参考意义也十分有限，而维基百科中的细菌条目页面有六千余条，经过筛选的可供文本挖掘的正文文本有1267篇。其次，从细菌与疾病种类角度， MicroPattern细菌与疾病的种类较少，并且存在大量的一种细菌对应多种疾病的情况；维基百科中的细菌与疾病种类则分布较为均匀，且细菌均为物种级别，细菌与疾病的对应情况既包含一对一的对应关联，也包含一对多的对应关系。最后，从细菌与疾病关联数目来看，来自维基百科的关联，虽然在总数上略低于其他数据来源，但新增的208条关联都能找到对应的文献依据，并且与其他两组数据的重复较少，并且只有极少的重复新增。



图3.4 细菌与疾病关联在不同数据库中的数据量大小示意图

此外，我们还加入了NCBI公开数据集的比较，由于公开数据集中的数据格式与来源与我们通过计算方法构建的数据库不同，因此我们仅在细菌与疾病关联的总数上进行了对比。我们从维基百科上又新增加了208条细菌与疾病的相互作用关系，与NCBI公开的数据集中包含了细菌菌株与动植物物种的疾病交互关联870余条和MicroPattern中的细菌与疾病交互409条进行了对比。重复的细菌与疾病的关系只有10条左右。重复较少的原因可能是NCBI属性知识库中主要针对的是细菌的亚种或菌株级别与疾病的作用，并且MicroPattern研究的主要是人体内的某些特定部位的疾病，疾病的种类较少。正是由于细菌与疾病的种类、物种级别的限制，本章得到的细菌与疾病交互才能通过有限的细菌与疾病种类，得到16%的交互新增。因此，从维基百科正文页面也能够通过文本挖掘方法，得到细菌与疾病的交互作用关联，扩充现有的数据库资源。

## 本章小结

维基百科作为一个全世界通用的知识库平台，具有复杂的语义网络信息，蕴含了丰富的可挖掘知识。细菌自身的相互作用纷繁复杂，与化合物、宿主乃至病毒的相互作用更是形成了复杂的关系网络。仅仅通过生物实验的方法去研究这些关系网络是有限的。目前也有越来越多的研究人员将科学计算、数据挖掘的方法应用到微生物关联的分析中。

本章提出了基于维基百科的细菌与疾病的关联挖掘的方法，从维基百科的非结构化正文页面文本中获得了细菌与疾病的关联。与其他的数据来源相比，本章方法得到的数据在细菌的种类，疾病的种类等上都有提升，在细菌与疾病的交互关联数目上，得到了16%的新增，每一条关联都有文献来源与之对应，扩充了现有的数据库资源。并且通过生物医学文本的实体关联挖掘工具Kindred自动化提取，能够实现细菌疾病关联的大规模更新，进一步丰富交互网络。

然而在细菌的交互研究中，细菌的属性也被一些研究证明对细菌与其他实体的交互存在一定的影响。因此，在接下来的研究中，本章将结合细菌的属性特征做进一步的分析，将细菌的栖息环境、细菌宿主等属性信息整合在细菌与疾病交互知识库中。并且通过整合的细菌关联数据，设计开发了基于整合的细菌与疾病关联可视化平台，为进一步的研究提供了数据支撑。

# 融合多源信息的细菌与疾病关联可视化平台

第三章通过采用文本挖掘的命名实体识别、关系抽取技术，从维基百科的正文文本中获得了细菌与疾病之间的关联。进一步的研究发现，细菌的属性特征与细菌的致病特征息息相关，因此本章将维基百科中挖掘到的细菌与疾病关系与细菌的属性数据融合，如图1.2所示，构建了基于Django的细菌与疾病关联可视化平台。

## 细菌属性分析

细菌的关系作用复杂，已有研究[[74](#_ENREF_74)]证明，细菌的属性特征对细菌的相互作用有一定的影响，因此在构建细菌的疾病知识库前，本章将细菌的属性特征与细菌疾病的相互作用整合在一起。我们从NCBI网站上获得了微生物的属性存储文件，共有23条属性类型，9000多个微生物物种，其中包含有细菌、真菌、古细菌等多个类型的微生物物种[[75](#_ENREF_75)]。

本节对其中的细菌进行了筛选，去除了缺失值较多的属性条目，并且去掉了冗余的微生物组计划研究项目号等暂时不需要的ID特征。最终得到了7765个细菌物种，11种属性类型：序列特征（sequence），基因组大小（genome size），GC容量（作为细菌分类树划分依据GC content），革兰氏属性（gram strain），细菌形状特征（shape），菌群大小范围（arrangement），菌群栖息环境（habitat），细菌的运动状态（motility），细菌的喜盐性（salinity），细菌的需氧性（oxygen requirement），生长环境温度（temperature range）。在对细菌的属性特征进一步分析之前，本节首先对该11种属性特征进行了特征工程处理。图4.1总结了本节针对每一种不同类型的特征所进行的处理方法。

首先按照属性特征的数据类型将11种属性分别划分为三类，分别是数值类型，字符串类型与布尔类型的属性。首先，数值类型的特征分别是：基因组大小，GC含量，序列特征。数值类型的特征缺失值较少，对于少量缺失，本节进行了去重处理。其次，字符串类型的特征分别是：生长环境温度，菌群大小范围，细菌的运动状态，细菌形状特征，细菌的需氧性，细菌的喜盐性，菌群栖息环境。对于生长环境温度，菌群大小范围，由于均为区间型数据，因此本节采用了取中位数的办法，将数据替换为区间中位数；对于其他五种字符串类型的数据，由于均为字符串类型的分类数据，且分类较为稀疏，因此本节采用数值转换的方法，将其替换为数值类型的多分类。最后，布尔类型的特征是：革兰氏属性。主要是阴性和阳性两种类别，且缺失值较少，因此采用了二分类数值替换的方法，将其转化为数值类型的二分类特征。至此本节完成了11种属性特征的预处理过程。



图4.1 细菌属性的预处理流程示意图

经过一系列的细菌属性预处理，本节将预处理后的细菌属性数据进行了绘图与属性分析。为了验证细菌的生长环境与机体平衡状态的关联，本节选择了人体与牛、植物、小鼠、猪、鱼等不同动植物体与细菌的交互进行了对比。根据细菌在不同机体或同一机体的不同部位的致病情况，将细菌的菌群栖息环境（habitat）与细菌的宿主（host）属性结合并绘图后分析，如图4.2所示，细菌在人体的不同部位能够导致不同的疾病，也能在不同生物体内导致疾病。

细菌在人体或其他生物体内，可以导致不同类型的疾病。例如，沙眼衣原体（Chlamydia trachomatis）可以在人体导致淋病（Gonorrhea）和失明（Blindness）。同时，同一种微生物在不同的宿主体内，也能够导致不同的疾病。而沙眼衣原体不仅可以在人体内致病，也可以在老鼠体内导致肺炎（Pneumonitis）。除此以外，巴斯德菌（Pasteurella multocida）在人体可以导致萎缩性鼻炎（Atrophic rhinitis），也可以在羊体内导致出血性败血症（Pasteurellosis）。因此，对细菌的致病研究不仅是人体机体健康状态研究的关键，细菌也对其他各种动植物体有重要作用。

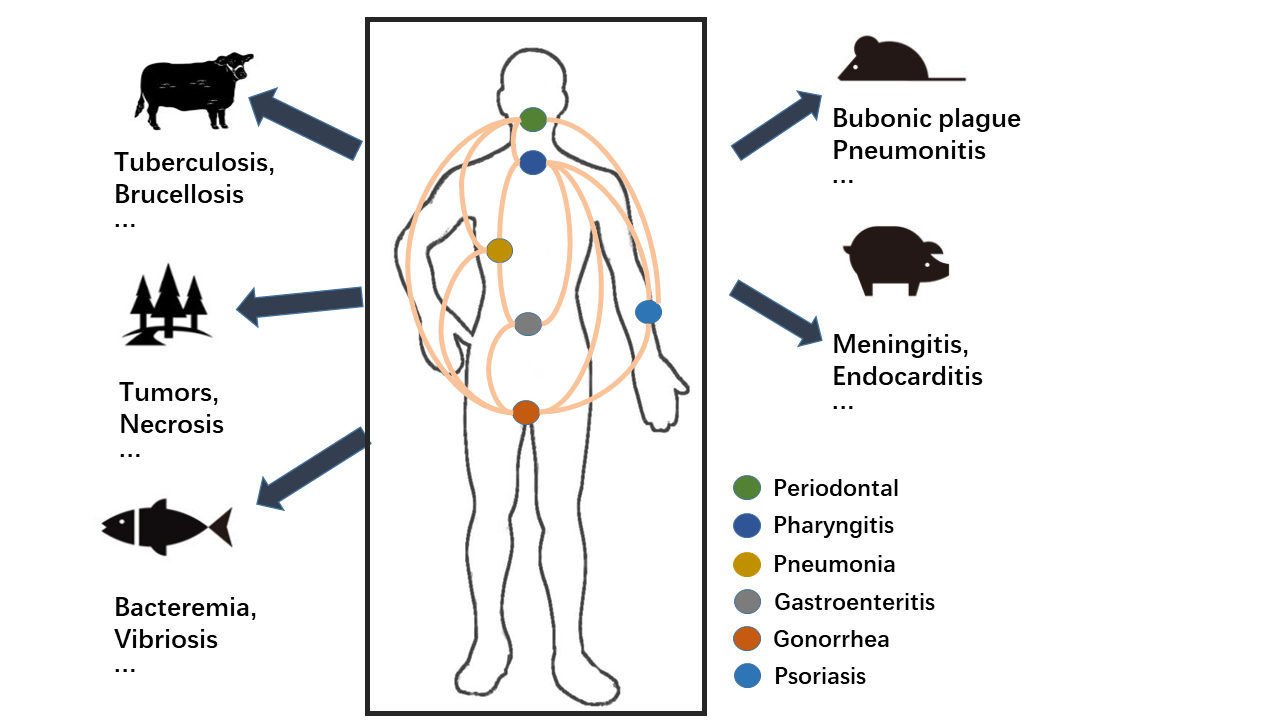


图4.2 细菌可能在人体的不同部位，或者在不同的生物体内致病示意图

接下来，本节将对细菌的属性关联数据进行可视化分析。本节选用了革兰氏属性（gram strain）、细菌的喜盐性（salinity）、细菌的需氧性（oxygen requirement）、生长环境温度（temperature range）、菌群栖息环境（habitat）五种属性对细菌的致病性进行了聚类分析，图4.3表明了本节针对这五种属性进行的属性分析过程。

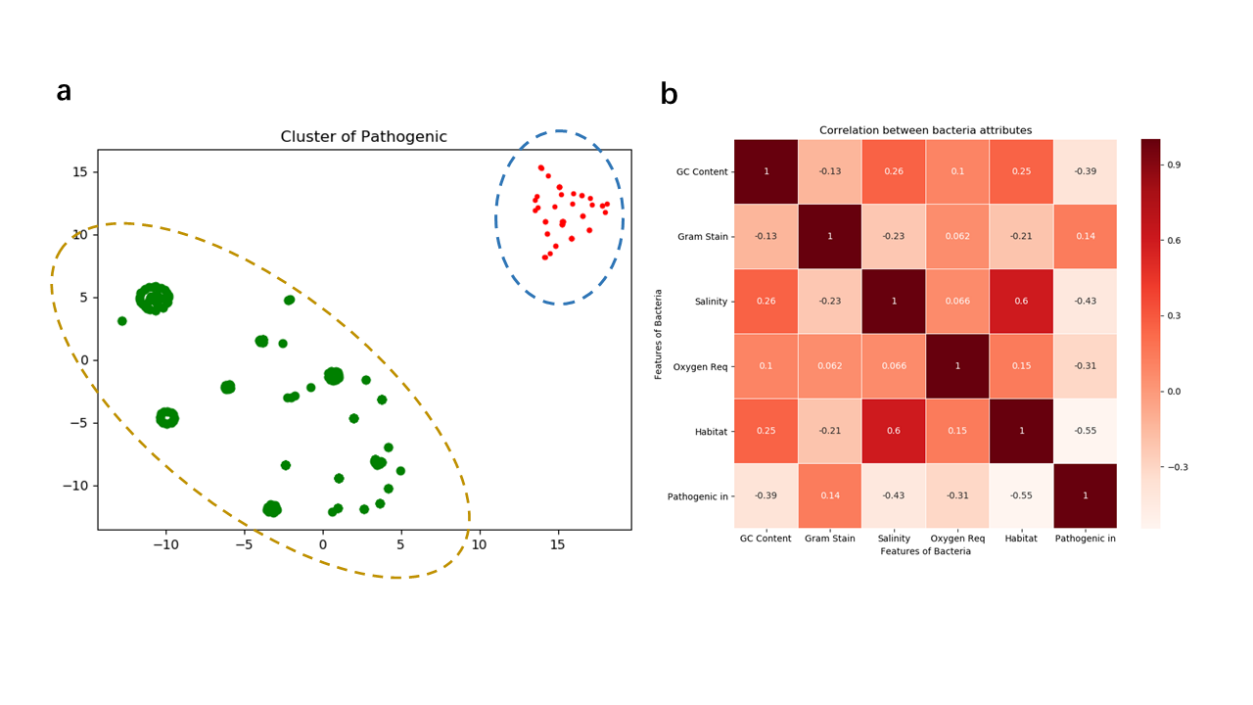


图4.3 细菌属性分析示意图

首先，如图4.3中的a图所示，通过五种属性对细菌的致病性进行了二分类，其中每一个点表示一条具有属性特征的细菌样本，绿色的表示该细菌致病，红色的表明该细菌不是致病细菌。结果表明上述五种细菌属性确实与细菌的致病特征有关联，并且五种属性对细菌的致病特征有较强的关联性。

除此之外，本节还研究了细菌属性两两之间的关联，如图4.3中的b图所示。此时选用了GC容量（GC content），细菌的喜盐性（salinity），细菌的需氧性（oxygen requirement），革兰氏属性（gram strain），菌群栖息环境（habitat）以及致病标签六种属性，每一个矩形框内数字的正负表明了分别表示了正负的相关关系，数字的大小和颜色的深浅表示属性两两之间相关关系的强弱。可以看出，细菌的喜盐性（salinity）和菌群栖息环境（habitat）之间有较强的正相关关系，细菌的喜盐性（salinity）与致病性，菌群栖息环境（habitat）与致病性之间也有一定的负相关关系。

综上所述，细菌的属性特征两两之间彼此影响，共同作用。在不同的环境条件下，对机体的平衡状态造成一定影响。通过对细菌的属性分析，进一步认识到了细菌属性对细菌的相关作用研究的重要性。因此，在接下来的研究中，本节将整合细菌的属性数据，并且构建一个基于整合的细菌与疾病交互可视化平台。

## 系统构建

数据来源的多样性造成了数据的异构特性，将不同来源下的数据进行整合与归一化，并发布为基于Web的知识库网站，是一种新的数据可视化分析方案[[76](#_ENREF_76)]；同时，构建的知识库也能够为下一步细菌与疾病关联预测提供数据支撑。接下来，本节将从数据库设计、网站构建与效果展示等几个方面介绍该细菌与疾病知识库系统的构建流程。

### 数据库设计

细菌本身的属性关联复杂，而细菌与疾病的关联交互类型单一，因此在构建知识库的数据库结构时，本节从细菌的属性中，提取了两种属性关系，与细菌与疾病的交互关系一同作为知识库系统的关联类型。首先，对系统的数据来源进行了整合，如图4.4所示，从公共知识平台上的三种实体类型：细菌的属性，细菌以及疾病。通过上文对细菌的属性分析，得到了两种更容易表述为实体类型的宿主与栖息地实体作为细菌的交互对象，并且抽取出细菌与宿主的寄生关系，细菌与栖息地的栖息关系，组成了该细菌与疾病交互知识库系统的三种关系类型。

此外，本节整合了维基百科等多个数据源的细菌与疾病之间的交互关联。我们将NCBI中涉及的细菌物种级别的致病交互关联与MicroPattern中的部分细菌与疾病交互关联与维基百科中的数据进行了整合。在知识库系统的开发过程中，本章将MicroPattern与维基百科中的细菌与疾病关联分别对应的PubMed文献编号与Wikidata编号进行了一一对应，以保证数据来源的可靠性。

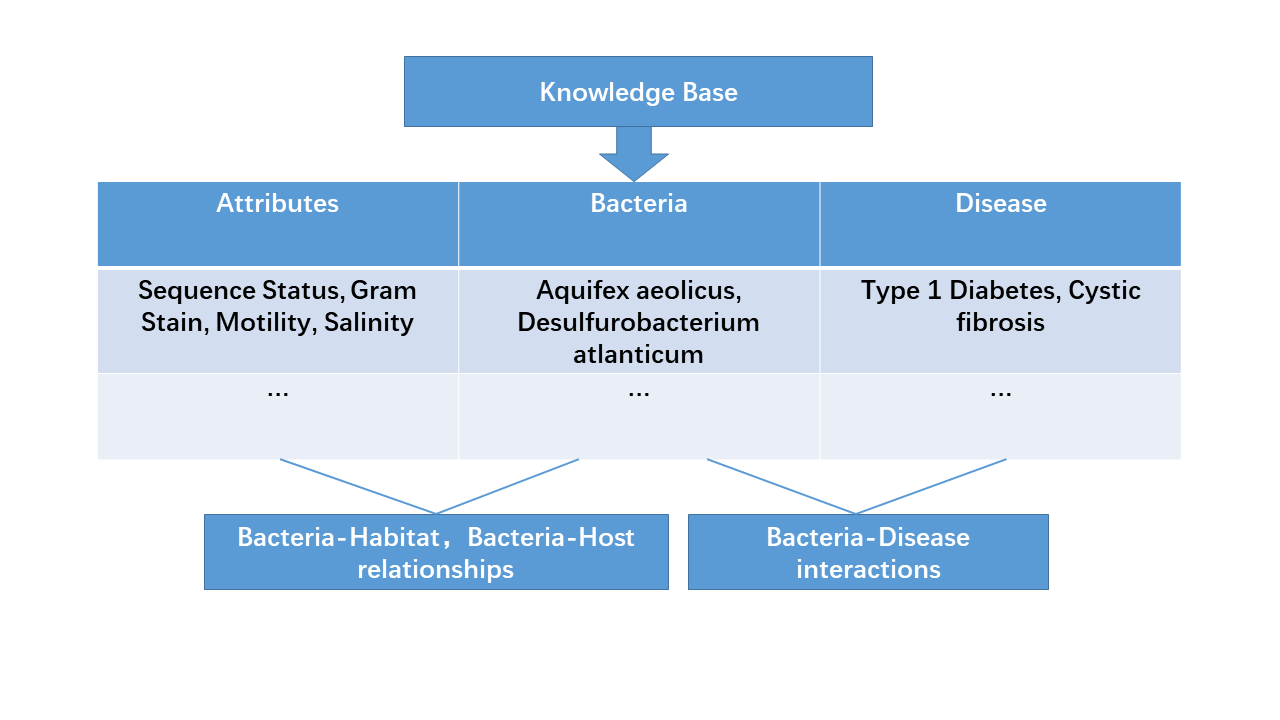


图4.4 知识库的数据类型组成示意图

随后，本节对系统的数据库进行了设计。本节旨在构建关系型数据库，并且选用了MYSQL作为数据库，数据库表设计如表4.1所示。对细菌、疾病、宿主以及栖息地及其属性构建了四张实体属性表；而围绕细菌实体展开的关联交互表，则由于主要是一对一，一对多的关联，因此本节设计了一个关系表，其中用关系类型标识细菌与其他实体的关系。

表4.1 系统的主要数据库表及其描述

|  |  |  |
| --- | --- | --- |
| **编号** | **表名** | **具体描述** |
| 1 | Bacteria information | 细菌名称，NCBI taxonomy 编号以及具体的细菌属性描述 |
| 2 | Disease information | 疾病名称，PMID 或者相应的Wikidata 编号 |
| 3 | Host information | 细菌的宿主名称及其属性描述 |
| 4 | Habitat information | 细菌的栖息地名称及其属性描述 |
| 5 | Relation | 多种关系类型，例如细菌与疾病的交互，细菌与宿主之间的寄生关系，以及细菌与栖息地的栖息关系 |

此外，本节还根据数据库表绘制了数据库关系E-R图，如图4.5所示。



图4.5 数据库实体关系E-R图

在数据库中的实体关系主要为三种：细菌与宿主的寄生关系，细菌与疾病的致病关系以及细菌与栖息地的栖息关系。且三种关系均为多对多的关联，即每种实体都有多种对应关系。细菌信息表中除了细菌的基本信息以外，还包含上一节整合的细菌属性信息；每一个实体都包含基本的系统ID，能够更好的对实体进行编码与调用；每一个实体信息还包含创建时间和修改时间，该属性用来记录数据的生成与修改状态，能够方便数据库管理。

完成系统的数据库表设计，并且将基于整合的细菌与疾病关联数据导入数据库中以后，我们便着手开始了系统的构建编码工作。对于系统的设计功能，主要设计了属性可视化展示、细菌与疾病交互可视化展示以及其他简单的搜索查询功能。

### 实验环境

该细菌与疾病交互平台系统是在Python的开源框架Django环境下进行开发的[[77](#_ENREF_77)]。Django是基于Python的Web框架，主要用来构建软件层，并且协调体系结构的组件和系统之间的信息工作流。最后，我们采用了HTML, CSS, JavaScript, JQuery进行前端页面开发[[78](#_ENREF_78)]。图4.6展示了Django的前后端交互，首先浏览器通过HTTP请求将用户操作发送给系统前端的url，然后系统通过Views模块与Models模块调用后端数据库；Models通过SQL语句对数据库进行数据提取，并且返回给Models。Django框架的前后端交互符合MVP模式（Model-View-Presenter）。



图4.6 系统前后端交互示意图

本系统还使用了AJAX[[79](#_ENREF_79)]将数据库的数据传到前端并且显示。AJAX是一种Web开发技术，用于构建异步Web应用程序和前端函数，以便为用户可视化Web服务器上的搜索结果。本节使用了Chart.js[[80](#_ENREF_80)]用于可视化。Chart.js是一个开源库，提供了多种不同的功能函数用于绘图。

### 平台构建

通过将外部数据库中的数据与基于维基百科挖掘到的细菌与疾病的关联数据做了整合，开发了这个基于整合的细菌与疾病关联数据库DMDA（Database for Microbes and Diseases Associations）。本节主要介绍了可视化平台的功能设计，图4.7展示了该数据库平台的功能设计图。平台的功能主要分为细菌属性可视化、细菌与疾病交互检索可视化以及开源下载等功能。



图4.7 DMDA平台总体功能结构图

本细菌与疾病关联的可视化系统功能主要服务于研究人员关于细菌相关研究的查询分析操作，因此平台将细菌属性可视化与细菌与疾病关联可视化进行了分块设计。细菌属性可视化是对细菌的属性特征详细分析，并通过异步更新对数据库中的知识进行动态更新展示。细菌与疾病交互可视化是针对系统中的三种关联：细菌的致病关系，细菌的栖息关系以及细菌的寄生关系，分别进行查询可视化，采用的是Echarts的方式，能够更加直观的表示与细菌有关的交互。

该平台整合了细菌的属性知识，能够为用户提供基于细菌属性的可视化分析的方案。下面主要展示该平台的细菌属性可视化界面以及细菌与疾病交互查询可视化界面。图4.8展示了本节可视化界面的部分具体内容，中间黑色方框内展示的是平台的可视化操作页面，其周围箭头指向的分别是细菌的属性可视化部分放大的图像。用户点击进入可视化操作（VISUALIZATION）选项卡，可以看到利用Echarts.js和AJAX技术对后台的微生物属性数据进行的动态数据可视化展示。该页面通过动态刷新技术，将数据库中的细菌属性通过可视化柱形图、饼图等多种方式绘制，更加直观的展示出细菌的属性分布与特征，为研究人员对细菌的属性的进一步分析提供了参考。

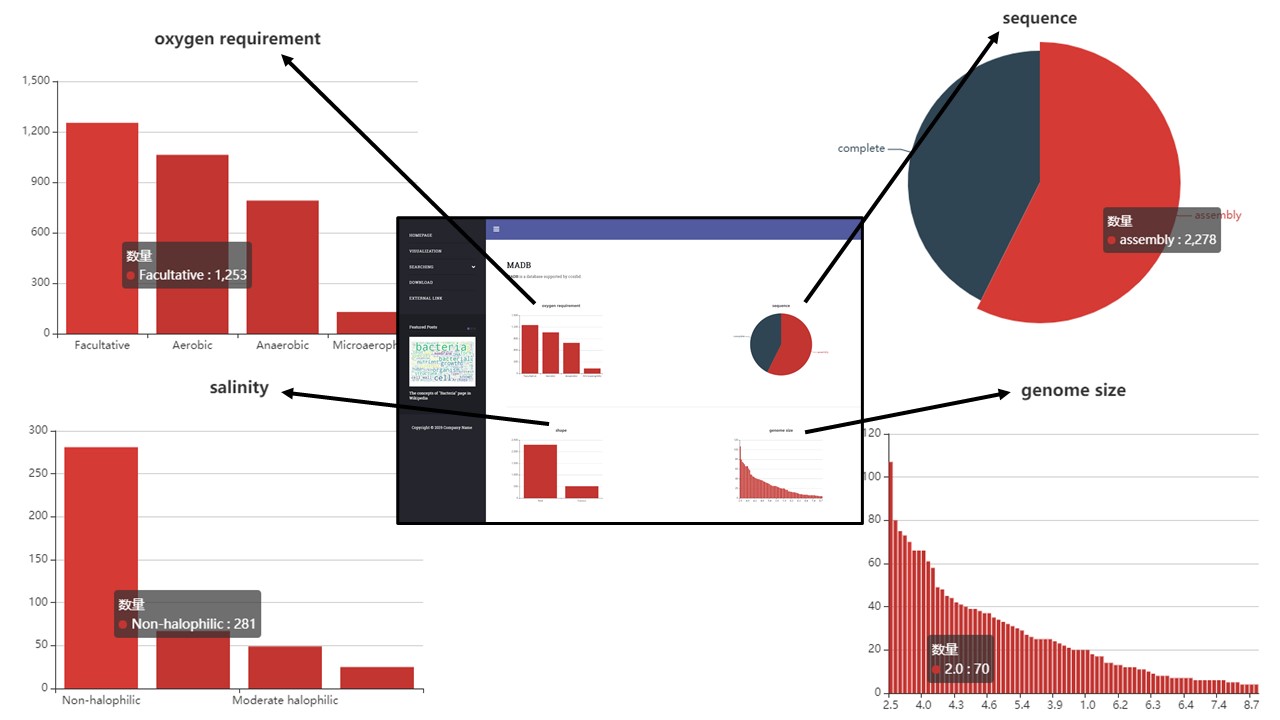


图4.8 DMDA细菌属性可视化页面示意图

本节可视化界面给出了8种具体的细菌属性类型，分别是：序列特征（sequence），基因组大小（genome size），GC容量（作为细菌分类树划分依据GC content），细菌形状特征（shape），菌群大小范围（arrangement），细菌的喜盐性（salinity），细菌的需氧性（oxygen requirement），生长环境温度（temperature range），并且针对每种属性特征，分别采取了不同的图像呈现方式。

此外，DMDA平台还设计了关于细菌条目和疾病条目的详细信息数据列表查询功能，在细菌与疾病条目查询（DATALISTS）模块，能够方便直接获取细菌与疾病的相关信息。本模块通过表格展示了数据库中全部的细菌属性信息，用户可以通过细菌名或疾病名称，建立索引，查询到细菌或疾病的具体条目，同时每个条目同时链接了Wikidata或PubMed的唯一标识，以供查询者能够快速链接到源页面，对该条目进行详细的研究参考。

在细菌与疾病交互网络分析（NETWORK）模块，本节还给用户提供了查询可视化的界面。图4.9展示了本节利用该模块实现整个查询可视化功能的示意图。本节将细菌与疾病的交互关系，细菌与栖息地的栖息关系以及细菌与宿主的寄生关系整合到了这个模块。

以查询细菌'Actinomyces graevenitzii C83'的交互关系为例：用户通过细菌名称条目进行查询，第一步进入细菌与疾病交互网络分析（NETWORK）模块；第二步输入实体名称，例如输入的'Actinomyces graevenitzii C83'；第三步输入实体对象的类别，如细菌'Bacteria'；第四步则自动显示该实体的关联对象结果，查询到与该细菌相关的疾病'Actinomycosis'，栖息地类型为'Host-associated'，宿主生物为'Human'。

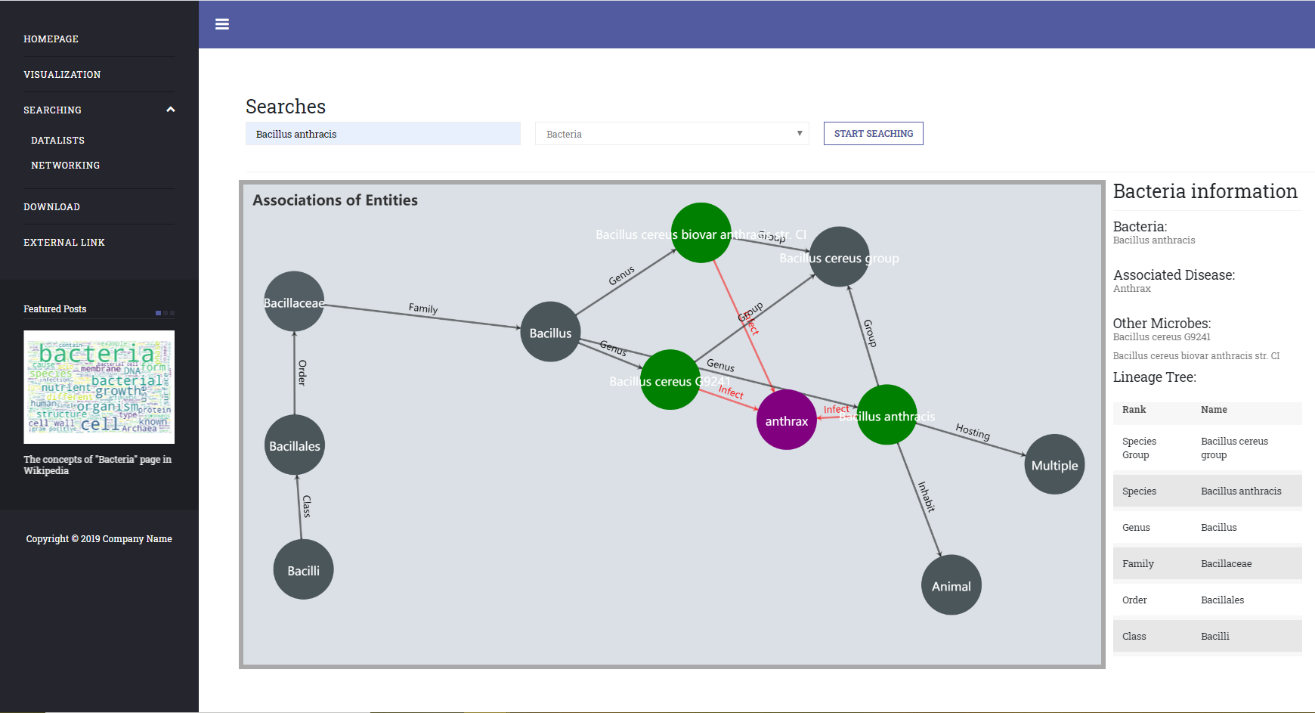


图4.9 DMDA细菌与疾病关联查询可视化页面示意图

本节将这些三元组关系进行了可视化的展示，每一个圆表示一个实体类型，带箭头的有向边表示一种关系类型，其中紫色的圆表示细菌实体，绿色的圆代表疾病实体，灰色的圆代表其他实体。平台的右边展示了图中的细菌基本信息，包括细菌的名称，细菌关联的疾病名称，与疾病关联的其他细菌以及细菌的物种分类树等。

此外，该可视化知识平台还提供了开源的数据下载（DOWNLOAD）功能（https://github.com/ccszbd/DMDA），数据来源的外部网站链接（维基百科，PubMed以及NCBI等），用户可以通过下载这些数据资源做进一步的研究分析。

## 本章小结

本章通过基于维基百科的细菌与疾病的关联挖掘的方法挖掘到细菌与疾病的关联，并且整合了NCBI中细菌的属性知识，将细菌的致病性与细菌自身的属性进行了综合的分析。本章发现了细菌的属性可能存在着两两的相互影响，并且细菌的属性之间可能共同影响到细菌的致病作用。这种快速提取关联性的方法，也可以为生物研究者在进一步研究细菌属性的情况下提供快速决断的支撑方案。

此外，本章还结合了一系列的数据资源提出了基于Web平台的可视化的分析方案，将细菌属性、细菌与疾病的相互作用关联可视化的呈现在本章的系统中。与此前其他研究人员构建的细菌与疾病数据库相比，本数据库平台能够提供在线的交互查询可视化，使用者不仅可以从知识平台上开源下载本章的数据进行下一步的研究，还可以通过平台，在线的进行用户个性化的知识分析。此外，本章还整合了细菌的属性知识，为细菌属性与致病特征的关联研究提供了可靠的依据。专业的生物研究人员也可以将平台提供的知识作为他们生物医学实验的先验性参考，有利于缩短生物实验的时间周期。

然而，基于平台得到的细菌交互关联是经过生物医学实验验证得到的，具有一定的可靠性与延展性。通过计算方法从这些复杂的交互网络中，找到潜在的实体关联，将成为数据库补全或进一步实验验证的可靠依据。因此，在后面的研究中，本章将对以上得到的细菌关联数据做进一步的关联预测，旨在通过计算方法预测实体之间可能的关系连接。

# 基于细菌与疾病知识库的表示学习及关联预测

前两章通过数据挖掘的研究方法，从维基百科的正文页面文本中获得了细菌与疾病的交互关联；基于这些关联与从NCBI中得到的细菌属性关联。本章还构建了基于整合的细菌与疾病交互可视化平台。接下来本章将对已经构建的知识库开展推理与应用研究，如图1.2所示。细菌与其他实体的交互关联复杂，构成了维度较高的交互网络。因此，本章旨在通过表示学习等方法，预测出细菌的交互网络中潜在的未知连接，以获得新的知识。

## 概念与背景

知识图谱是一种以图的形式存储领域知识的大规模语义网络，当下维基百科、谷歌搜索还有以中文为核心的百度搜索均有自己的知识图谱网络。在一个知识图谱网络中，通常用一个节点来标识一种事实实体，可能是人物、国家、动植物等；两个节点之间的连边来表示实体之间的关系，例如是（is\_a）、位于（located\_in）、属于（part\_of）等。人们通常将现实中的事实事件用这种三元组的形式表示，即为<实体头，关系，实体尾>。例如，<大肠杆菌，寄生于，人体>就是一个三元组。

然而，通用领域的知识图谱尚且存在着三元组缺失的问题，特定领域的知识图谱更是如此。对于细菌与疾病的知识库，从生物医学文本中挖掘潜在的关联依然会受到现有的研究的实验进度限制，导致研究的细菌关联有较多的缺失；进一步通过实验方法继续开展生物研究更是耗费人力物力。因此本章设计了细菌与疾病知识库补全任务，通过机器学习模型，在已有的知识网络中自动寻找到缺失的实体边。由于此时并不会通过模型得到新的实体或关系类型，只是在现有的实体之间找到可能的边类型，因此也被成为关联预测的任务。

现有的表示学习模型主要分为三类：

1. 基于距离的翻译模型；（2）双线性模型；（3）神经网络模型。

本章将基于三种不同类型的表示学习模型，分别对细菌与疾病关联DMDA数据展开关联预测任务。首先对细菌与疾病关联知识库中的数据进行了预处理，得到了规范的三元组关联数据；接下来针对不同模型进行了调参调优，比较了各个模型在不同参数下的表现；最后选择了在DMDA数据集下表现最优的模型与参数，并且进行了预测结果的分析。

## 实验数据

本节通过前两章的数据挖掘方法从维基百科正文页面得到了细菌与疾病的交互关联，并且将这些关联关系与NCBI中的细菌属性整合到一起，得到了基于整合的细菌与疾病关联知识库。从细菌与疾病的知识库中，本节得到了四种关系类型，分别是致病关系、寄生关系、栖息关系以及细菌的共生关系。此外，本节从NCBI taxonomy[[75](#_ENREF_75)]中，利用爬虫技术得到了43万条细菌的分类树上下级关系。总数据集在各分类关系下的个数如下表5.1所示，其中实体一表示三元组中的实体头部，实体二表示三元组中的实体尾部，关系类型则是实体之间的连边。

表5.1 数据集中包含各实体与关系数目

|  |  |  |  |
| --- | --- | --- | --- |
| **实体一** | **实体二** | **关系类型** | **关系数** |
| 细菌物种 | 疾病 | Pathogenic to | 2594 |
| 细菌物种 | 细菌宿主（动、植物或微生物） | Host in | 1743 |
| 细菌物种 | 细菌共生群落 | Group in | 7586 |
| 细菌物种 | 栖息地（生态或食物环境） | Locate in | 1279 |
| 细菌物种 | 细菌属 | Genus is | 2196 |
| 细菌属 | 细菌科 | Family is | 2171 |
| 细菌科 | 细菌目 | Order is | 2171 |
| 细菌目 | 细菌纲 | Class is | 2166 |

进一步地，通过整合上表中的数据，本节得到了包含有21,905条关系边的数据集，其中实体种类有9,832个，关系类型有7种。本节按照8:1:1的比例对数据集进行了划分。按照该比例划分数据集的原因在于，本节的细菌与疾病关联三元组DMDA较稀疏，为了更好的学习到网络图的表示，本节将尽可能多的提供训练集大小，同时保证验证集的参数调整。如下表5.2所示，本节分别列出了训练集、测试集和验证集种所包含的实体一、实体二、关系类型个数以及包含的关系边条数。

表5.2 划分后的数据集中各实体与关系数目

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **实体1数** | **实体2数** | **关系类型数** | **关系数** | **所占比例** |
| **训练集** | 7,865 | 7,749 | 8 | 17,524 | 8 |
| **测试集** | 991 | 976 | 8 | 2,191 | 1 |
| **验证集** | 983 | 965 | 8 | 2,191 | 1 |

## 模型设计与开发

现有的知识库通常被表示为网络的形式存储，研究想要通过现有实体间的连接预测到新的关联，采用人工的方法推理需要耗费大量人力物力，并且还会受到数据维数高，数据稀疏等因素的约束。现有以机器学习、深度学习模型为代表的表示学习模型，能够首先将实体与关系的映射表示为低维空间中的语义关系，能够有效的缓解数据稀疏的问题。然后再通过一个打分模型为三元组的预测打分，以此给出任意三元组的现实概率。

### 双线性模型

双线性模型通过基于一个评分函数，以匹配实体和关系的潜在语义，学习他们的向量空间表示，使得正样本的得分高于负样本，以预测未知三元组的概率。以下将详细介绍几种典型的双线性模型。

#### RESCAL

RESCAL[[81](#_ENREF_81)]是2011年被提出来的双线性模型。该模型将张量引入三元组的表示中，主要有三个方面的好处：

（1）可以用一个张量表示多重的二元关系；

（2）对于不必要的信息无需进行多余的处理，例如知识库或其他从数据中推断出的变量；

（3）通过因式分解的方法，能够使高维且稀疏的关系表示得到很好的结果。

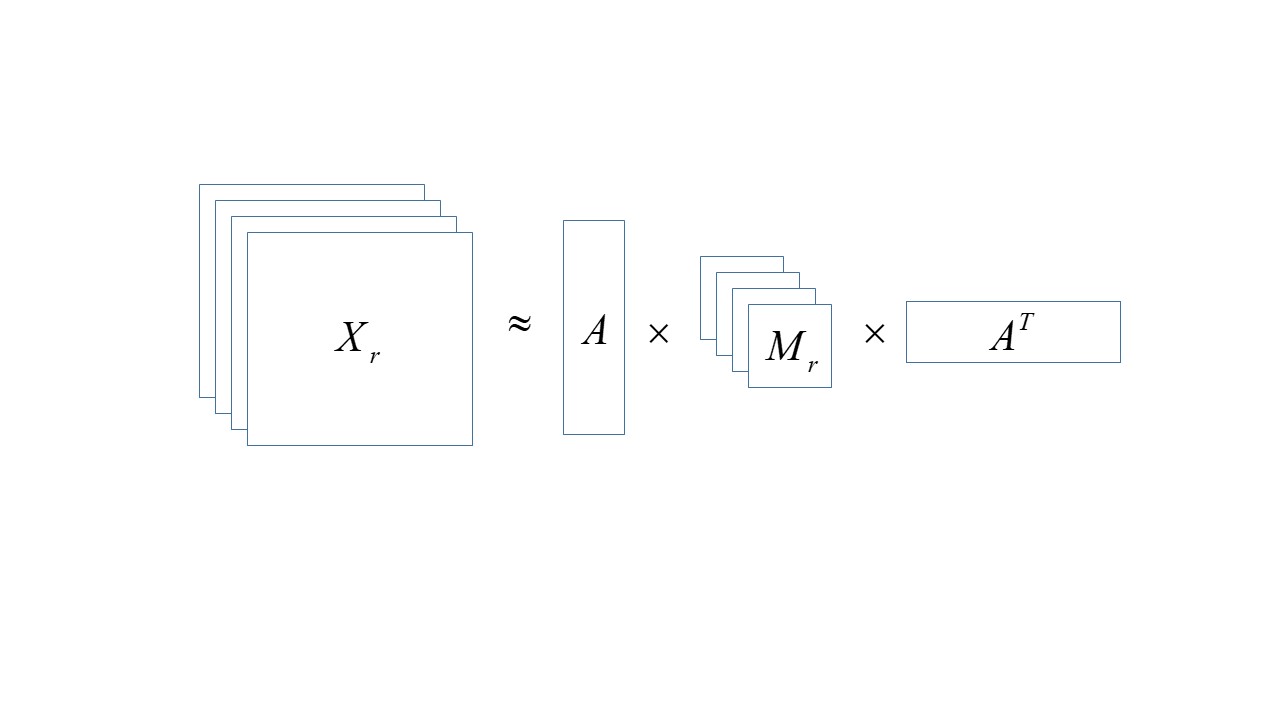


图5.1 RESCAL模型的张量分解示意图

他将每一种关系都表示为一个满秩矩阵，每个实体名称则用一个词向量表示，利用图5.1所示的张量因式分解，定义一个张量 ，学习二元关系数据的固有表示，以此预测未知的二元关系。于是用来给未知二元关系打分的评分函数可以表示为：

 (公式5.1)

然而，由于满秩矩阵容易导致过拟合，因此后面的研究对于矩阵又做出了其他假设。

#### DistMult

DistMult是最简单的一个乘法模型[[82](#_ENREF_82)]，在2015年被提出，它将每一个实体和每一种关系都表示为低维的嵌入，由一个普通的神经网络框架来进行多类型关系的表示学习。首先给出关于知识库的一个关系三元组表示，其中表示关系的头部实体，表示关系的尾部实体，表示实体间的关系类型。DistMult模型希望通过神经网络的结构学习到事实三元组关系的表示，以此能够在验证三元组上给出更高的分数或者较低的权值。

神经网络的第一层是将输入的实体对表示为低维的向量。每一个输入的实体都对应一个高维的向量，要么是一个独热（one-hot）向量，或者是词向量。我们用和分别表示输入的实体和；用表示第一层的投影矩阵，最后学习到的实体表示和可以表示为：

 (公式5.2)

其中可以是一个线性或者非线性函数，而为参数矩阵，可以使用预训练的向量或随机初始化得到。神经网络的第二层则将这两个向量组合成一个标量，然后通过带有关系参数的打分函数进行比较。我们用一个基本的线性变换和一个双线性变化来统一打分函数，于是和可以表示为：

 (公式5.3)

其中和为关系参数。此处，还可以考虑通过一种简单的限制，用表示一个对角矩阵，则打分函数可以写为：

 (公式5.4)

以此减少关系参数的个数。神经网络的结构以及组件完成，然后通过最小化间隔损失函数，能够使正样本的得分高于负样本，其中负样本需要满足，使用表示三元组的得分函数，于是训练时的损失可以表示为：

 (公式5.5)

DistMult将实体与关系转化为向量表示，嵌入为对角矩阵，但仅仅遵循了实体间的对称关系，忽略了实体的语义关系。

#### ComplEx

2016年Trouillon 等人提出了ComplEx模型[[83](#_ENREF_83)]，作为DistMult模型的延申，引入了虚部嵌入来改善模型的非对称关系与逆关系。

定义一个三元组，对于一个实体，令表示实体的实部嵌入，表示实体的虚部嵌入；对实体间的关系，令表示关系的实部嵌入，表示关系的虚部嵌入。于是在ComplEx模型中三元组的映射函数可以表示为：

 (公式5.6)

此时。于是很容易得到ComplEx模型的映射函数可以简写为。ComplEx模型证明了在实向量的上加入虚向量的约束，能够一定程度上降低训练时间，并且也能够达到较好的训练结果。此外，还有与ComplEx类似的HolE等模型，能够更好的解决环状的关系，其嵌入方式与ComplEx等同。

#### ANALOGY

在以上几种框架的基础上，基于知识库逻辑推理的ANALOGY框架应运而生[[84](#_ENREF_84)]。逻辑推理在知识库中的应用已有很多，而ANALOGY提供的是一个基于类比结构潜在语义嵌入，类似数学公式的推理，将实体和关系转化为一个知识图谱。例如，用分别表示四种不同的实体，用分别表示两种不同的关系类型，假设在模型中这两种关系为线性的。对多关系网络嵌入，假设实体与关系的三元组满足：

 (公式5.7)

则这类关系图可以通过图嵌入可视化，如图5.2的平行四边形所示。每一个节点都表示为一个实体，每一条边为一个线性关系连接两个实体。

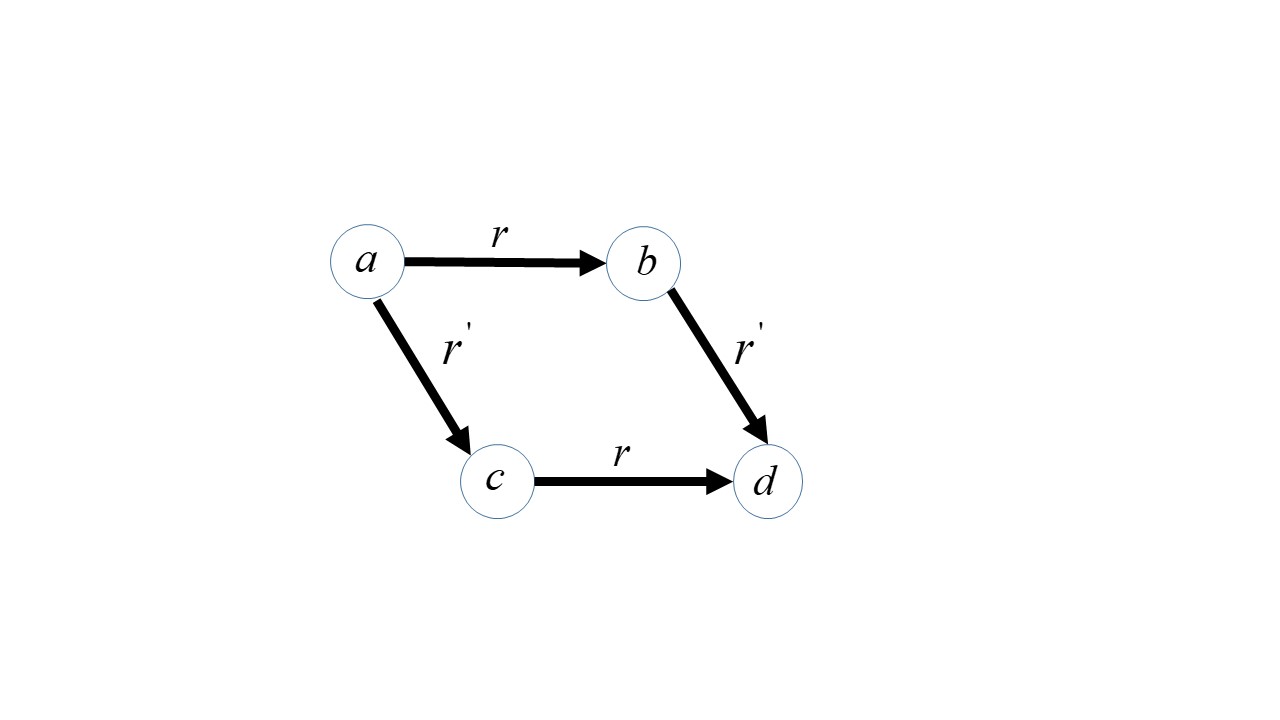


图5.2 语义信息逻辑分析得到的平行四边形示意图，每一条边表示一个线性映射

本章在后续的实验中均验证了这四种模型的有效性，并且证明了双线性模型能够很好的表示实体与关系，然而却对真实关系的语义信息考虑不全面，并且对于更加复杂的关系类型无法很好的学习。

### 神经网络模型

神经网络模型在近几年来受到了研究的广泛关注。现在已有双线性模型的方法能够从现存的数据库中预测出未知的关系，随着深度学习在自然语言处理任务中的广泛运用，于是神经网络也被引入知识库的关联预测任务中来。

2013年，Socher等人提出了NTN[[85](#_ENREF_85)]模型，使用了全连接神经网络结构，将实体表示为词向量，关系则由一个神经张量网络中的参数表示，由此预测新的三元组的概率。2018年，Nguyen[[86](#_ENREF_86)]等人使用了卷积神经网络来定义评分函数，他们将三元组表示为一个三维矩阵，将这个三维矩阵作为卷积层的输入，得到一个特征映射以表示这个输入三元组。给真实的三元组赋予一定权重以预测新的三元组是否有效。最近，随着图卷积神经网络研究的深入，在关联预测的任务中也尝试引入图卷积神经网络以提升最后的预测效果。

#### 图卷积神经网络

图卷积神经网络（Graph Convolutional Neural Network，简称GCN）[[87](#_ENREF_87)]是面向图结构表示学习的卷积神经网络。最初由文本的卷积神经网络（Convolutional Neural Network，下称CNN）[[88](#_ENREF_88)]演变而来。CNN主要处理图像、文本中，图形坐标或文本的欧氏空间距离，能够被转化为矩阵，有效地提取到特征。随着数据维度的扩增，出现了许多不符合欧氏距离的数据，例如社交网络、生物医学交互网络、化学与分子结构网络等等，这些数据彼此交错连接，形成了由节点和连边组成的拓扑图，这些数据无法在空间上保持卷积核的平移不变性，因此也就无法利用传统的CNN等方法做卷积核的运算。

DeepWalk[[89](#_ENREF_89)]的提出，开启了图论学习在深度学习中的进程。最早提出的图神经网络（Graph Neural Network，简称GNN）[[90](#_ENREF_90)]主要解决的还是分子结构的分类问题，尽管许多图像或文本数据也可以转换为拓扑图的结构。GNN主要基于不动点理论，也就是说全局更新函数无论迭代多少次，其变量都会收敛到一个固定的点。换句话说，GNN通过这种方式能够传递图中的信息，并且进行预测。正是因为这一特性，导致了GNN也存在局限性，例如不能很好的学习到边的某些特性。

GCN的出现更好的解决了这一问题，它重新定义了一个基于图结构的卷积操作。如图5.3所示。模型的输入为一张由节点和连边组成的图结构，第一个卷积核的输入则是一张完整的图。第一层卷积层学习节点的信息和特征，然后每个节点的另据节点进行一次卷积，用卷积的结果来更新该节点。接着经过一个激活函数例如ReLU后得到第二层卷积层的输入，不断重复上述过程，最后得到图中节点的信息表示。GCN通过多层的卷积，每一层的参数均不相同，能够更好的学习到节点的表示，从节点的信息生成整张图的表示，以此做接下来的预测。

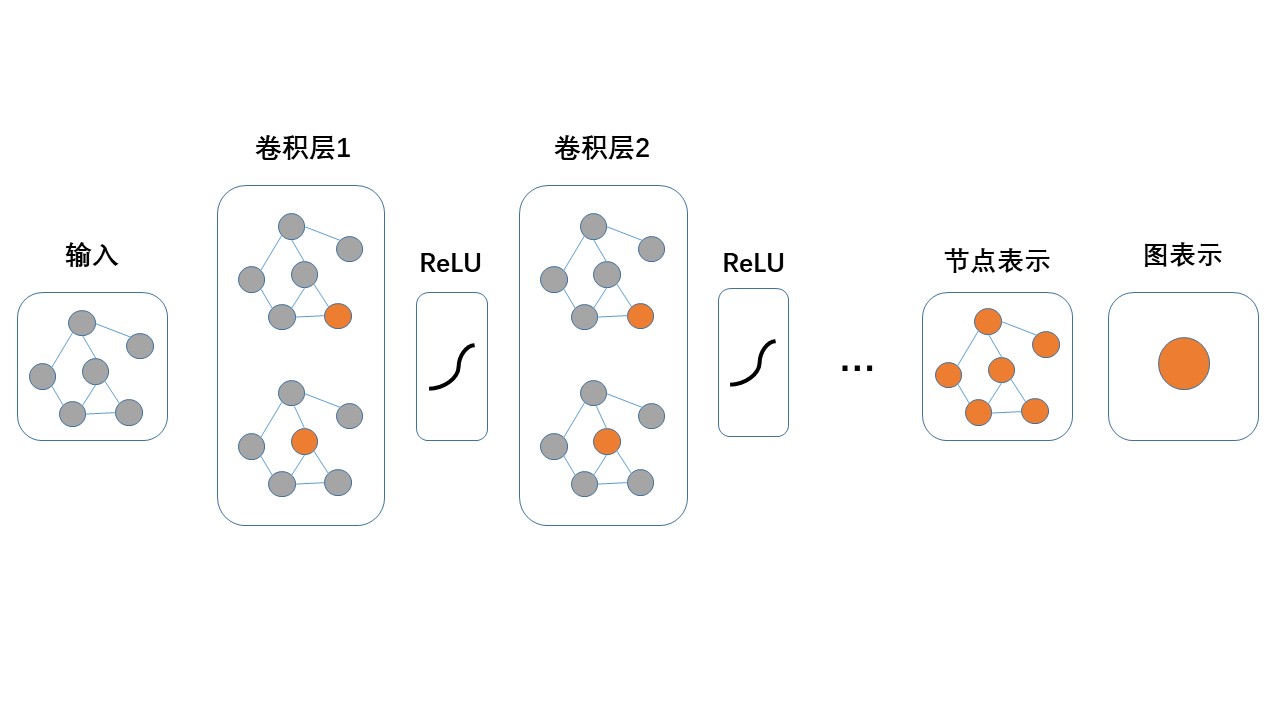


图5.3 GCN的模型结构示意图

在上述模型结构的驱动下，本节引入了一种图结构的表示方法：，其中表示一张输入的图结构，用表示图中的节点（实体），图中带标签的边（关系）表示为，这里的表示一个三元组，用表示一个关系类型。于是定义了一个简单的传播模型，用于计算带标签的有向边（关系）和节点（实体）的前向更新：



 (公式5.8)

其中表示节点在神经网络的第层的隐藏状态，表示该层的维数；这里的表示节点在关系下的邻居节点，这里的表示用来归一化的常数，可以实现学习到或者定义。这种方法通过一个归一化和累积学习邻居节点的特征，能够不断更新节点的表示，同时，通过引入特定于关系的转化，将边的类型和方向也表示为特征，能够更好的学习到图的表示。

图卷积网络的模型结构虽然在设计上能够很好的学习到节点的表示，但也可能存在当节点数到达一定数量级时，GCN学习到的特征可能会太过平滑，每个节点的表示都类似，导致模型的泛化性能变差。

### 翻译模型

为了解决传统方法参数较多，训练过程复杂，并且泛化性能较弱等缺点，人们引入了翻译模型对图进行表示学习。受到词向量[[91](#_ENREF_91)]的启发，研究发现通过将不同空间下的实体表示映射到同一个向量空间下，利用平移不变性，能够将知识图谱中三元组的关系向量近似地翻译为头实体到尾实体的映射。下面介绍的是翻译模型中最为经典的TransE模型。

#### TransE

通过传统的图表示学习可以看出，层次关系在知识库网络的关系中非常普遍，考虑在树中反复学习节点与邻居节点，很容易导致学习到的节点彼此接近，最后模型的泛化性能自然会较弱。2013年Bordes[[92](#_ENREF_92)]等人提出了TransE模型。考虑到将多维关系网络中的实体和关系映射到低维的向量空间，TransE首先提出了一个假设。在实体嵌入的空间中，假设一个三元组，那么尾实体的在低维空间中的向量应该近似等于头实体加上关系的向量。由于该模型之学习了每个实体及其关系的向量，因此参数相对较少，训练更加简单。

首先，设置一个三元组集，实体集，以及所有的关系集。其中包含所有真实的三元组，和满足，满足，模型旨在学习这些实体和关系的向量表示。用表示训练集的负样本，则可以在正样本集中随机替换头实体或者尾实体得到，表示为：

 (公式5.9)

在训练过程中，通过不断的调整，使得三元组不断满足，或者无限远离。定义映射函数来对实体和关系建模，并且将他们映射到相同的向量空间中。为了使模型更好的学习到向量的表示，定义了损失函数，通过最小化间隔损失得到更好的训练模型。



 (公式5.10)

运算符表示的正数部分，而表示的间隔参数。该损失函数倾向于使得正样本三元组会比随机生成的负样本的取得更低的值，达到预期的训练标准。

## 模型训练与优化相关工作

构建好网络模型以后，构建了训练集进行训练，在训练过程中，通过随机替换正样本中的头节点或者尾节点生成负样本。在训练过程中，我们通过验证集选择合适的超参数，采用了早停（Early Stopping）技术，能够有效减少训练时间，有效缓解过拟合的问题。同时，为了降低模型的复杂程度，我们也利用Dropout方法，在训练过程中随机抛弃一部分的节点，优化训练结果。模型的训练和优化工作，都是为了使得最终的模型能够在有效的训练时间内达到更好的泛化性能。

### Early Stopping技术

本章设置了训练集和验证集，训练集用来模型学习向量表示，而验证集则用来验证模型的泛化能力。随着训练迭代次数的增加，模型的泛化能力会逐渐提升。

理想状态下，模型的训练误差以及在验证集上的表现会随着迭代次数的增加而趋于稳定，但真实状态下，模型的验证误差很有可能会出现反弹，模型在验证集的表现逐渐变差，意味着模型的训练效果太好，而泛化性能变差，模型状态变得过拟合。

为了防止模型过拟合的现象发生，本节采用了早停法，使模型在验证集上表现变差时及时终止训练。本节设置了每迭代1000次，将模型应用于验证集上去评估模型的泛化能力，如果此时的验证误差低于之前的验证误差，则说明模型还没有达到最优的表现，保存好此时的模型以后，继续迭代训练；若此时的验证误差与上一次的验证误差出现反弹，结束训练；若到达最大迭代次数，模型还没有达到最优的表现，则说明可能时设置的迭代次数较小或者学习率太大等原因。

早停法除了预防模型过拟合，也能降低模型的复杂程度，提高训练效率。因此在翻译模型、双线性模型的训练过程中均加入了早停策略，以优化模型结构，得到更优的训练效果。

### Dropout技术

在神经网络模型的训练过程中，模型的复杂程度与模型最终的泛化性能表现息息相关。模型的复杂程度较高时，容易出现过拟合的现象，因此可以通过简化模型结构来防止模型的过拟合，我们采用的是随机丢弃神经元的Dropout方法。

Dropout方法即在训练的迭代过程中随机丢弃一些神经元，导致每次迭代训练的神经元结构各不相同。经过多次迭代之后，能够得到不同的子网络；同时，由于迭代过程中，训练的参数始终没有发生改变，不同模型之间的参数能够实现共享，从而也能保证模型的正则化，不会因为模型数目的增多导致训练时效太长和效率降低等问题。这种随机的神经元失活方法能够有效防止某些节点特征可能出现的偶然结果现象，增强模型的鲁棒性。在训练完成以后，通过多个不同的网络训练得到的子模型再做模型的集成，对于固定的验证集与测试集做测试，能够得到更稳定的模型表现。

在训练过程中，Dropout的参数设定也需要斟酌，本节使用dropout率为0.5或0.8，然后观察验证集上的模型表现来选择合适的参数。除此以外，还有L1、L2正则化方式也能有效降低网络的复杂程度，使得模型达到更好的泛化效果。

## 实验结果与对比分析

本章确定并设计了模型的结构，设置好训练时可以有效防止模型过拟合的方法，接下来便开始模型的训练与对比过程。首先将介绍基于知识库网络的表示学习任务的评价指标，该评价方式的优劣能够更加直观的反映模型的真实表现。然后本章对图卷积神经网络进行了模型的参数选择对比。最后本章将双线性模型，神经网络模型以及翻译模型等多个模型结构的表现进行了综合的对比，并且将在测试集上表现最好的模型进行了案例分析，得到了模型最后的预测效果展示。

### 实验评价指标

在实验过程中，对于测试集中的每个三元组，通常将头实体或者尾实体随机替换，创建一组新的候选三元组，然后通过模型预测给每个候选三元组打分，并且按照分数降序排序。本节将排序的结果分为“原始”（Raw）和“过滤”（Filtered）两种情况，由于替换实体时很可能出现现有的真实三元组，导致模型给出的分数较高，干扰最终的结果。因此在“过滤”的情况下，会将这些真实三元组去掉，因此最后的结果会普遍高于“原始”情况。

我们使用了平均倒数排名（Mean reciprocal rank，下称MRR）[[93](#_ENREF_93)]和命中N（Hits at N，下称H@N）作为评价指标。MRR是指正确答案在总排序下的倒数均值，如公式（11）表示：

 (公式5.11)

其中表示第i个候选三元组打分排序的正确答案的排名。MRR是一个对推荐系统算法进行评价的指标，他把标准答案在结果序列中的排名位置的倒数求平均，然后作为推荐的准确度。而H@N则表示正确答案在结果序列中排名的位置，最后求排名在序列的百分之N个数的平均，N的取值通常为1、3、10。则最后的结果更高的MRR或者H@N则表示模型的性能更优。

### 参数选择对比

本节使用了各个模型训练了所有的基准方法，并且根据。对于RESCAL模型，将正则化参数这是为0，并且选择了维度的取值范围为，使得模型在验证集上的预测平均损失最低。对ComplEx、ANALOGY、Dismult模型，本节选择了学习率的取值范围为，维度的取值范围为。对于T ransE，本节选择的学习率的取值范围为，间隔的取值范围为，维度的取值范围为，根据模型在验证集上的不同表现，最终确定了的参数设置。

对于GCN图神经网络模型，本节还增加了Dropout技术。本节设置了神经网络层数的的取值范围为，神经网络的嵌入维数的取值范围为，学习率的取值范围为，Dropout率的取值范围为。根据模型在验证集上的表现，最终确定采用单层的200维的网络嵌入，的参数设置。

### 网络模型对比实验

经过对不同模型参数的反复调节，本节固定了特定的参数，并且按照8:1:1的比例划分了训练集、验证集和测试集。本节将训练集对不同的模型分别进行了训练，并且在测试集上测试了模型的预测效果，如表5.3所示。

表5.3 不同模型在DMDA数据集下的最优结果对比

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **模型** | **DMDA** | | | | | |
| **MRR** | | | **Hits@N** | | |
| **Raw** | **Filtered** | **1** | | **3** | **10** |
| **GCN** | 32.2% | 41.2% | 39.9% | | 41.4% | 44% |
| **ComplEx** | 37.4% | 44.5% | 40.3% | | 43.3% | 57.8% |
| **Distmult** | 38% | 45.3% | 40.4% | | 47.3% | 58.3% |
| **ANALOGY** | 42.8% | 49.4% | 44.6% | | 51.1% | 58.9% |
| **RESCAL** | 38.1% | 44.7% | 40.3% | | 45% | 58.1% |
| **TransE** | 43.5% | 49.9% | 45.5% | | 51.7% | 59.6% |

上述模型均经过了参数调优，模型优化等过程，均达到了较好的实验结果。实验看到，目前TransE模型达到了更好的性能，相比于图卷积神经网络，TransE模型的MRR提升了约10%，H@1提升了约7%，H@3提升了约10%，H@10提升了约15%。与双线性模型中表现最优的ANALOGY相比，其MRR也提升了约1%。因此可以得出结论，对于细菌与疾病关联网络DMDA，翻译模型TransE能够更好的学习到图的结构表示，达到更优的预测效果。

实验结果表明对于细菌与疾病知识库网络的表示学习与关联预测任务，TransE模型的效果较好，最显著的特点是该模型具有较少的参数，计算复杂度低，能够更好的建立实体和关系之间的语义关联。

### 模型预测结果

本节将训练效果较优的TransE模型应用于DMDA的测试集，通过已知的头实体和实体关联来预测尾实体，并且给出了如表5.4所示的关联预测示例。在关联预测时，首先标注一个完整的真实三元组，包含一个头实体，一种关系类型标签以及一个尾实体。训练好的模型会根据这个头实体与关系类型得到预测序列，该预测序列则是模型认为的尾实体最优解序列，序列按照降序排列，位于前十位的则是模型认为最有可能是尾实体的实体名称。在表5.4中，列出了5种不同关系类型下的实体预测示例 ，按照预测结果的次序排列 ，对正确的实体名称进行了加粗表示。

表5.4 TransE模型在DMDA数据集下关联预测结果示例

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **头实体** | **Actinomyces odontolyticus F0309** | **Bilophila wadsworthia** | **Escherichia coli O157:H7** | **Paenibacillus larvae subsp. larvae B-3650** | |
| **关系**  **类型** | **host\_in** | **Pathogenic\_to** | **locate\_in** | **group\_in** | |
| **1** | **Human** | Psoriasis | mice | Alphaproteobacteria |
| **2** | Cattle | Crohn's Disease | cheese from dairy B | Actinobacteria |
| **3** | Animal | Behcet's Disease | patients | Gammaproteobacteria |
| **4** | Fish | Pneumonia | Euprymna scolopes | **Firmicutes** |
| **5** | Rodent | **Cirrhosis** | Borrelia burgdorferi CA-11.2A | Other Bacteria |
| **6** | Horses | Tuberculosis | **Lettuce** | Betaproteobacteria |
| **7** | Thyroid carcinoma | Autism | Snowella | Chloroflexi |
| **8** | Micrococcus | Aggressive periodontitis | Campylobacter hyointestinalis subsp. hyointestinalis LMG 9260 | Cyanobacteria |
| **9** | Porcine | Irritable Bowel Syndrome | chicken nugget processing plant | S. equorum Mu2 |
| **10** | Lactobacillus | Lyme Disease | susceptible individuals | Thermotogae |

通过表5.4可以看出，即使正确的实体并不总是在第一个的位置，然而通过模型预测出来的实体序列都是满足常理的。例如Bilophila wadsworthia菌致病关系的实体应该是某种疾病名称，则排名在Cirrhosis正确实体之前的都是某种疾病的名称，说明这些实体在向量空间上与Cirrhosis十分接近。

## 本章小结

本章首先将DMDA的数据进一步整合，得到了细菌与疾病关联预测的数据集。在图表示学习任务中，常用的算法模型主要分为双线性模型、深度学习神经网络模型与翻译模型。本章分别针对不同的类型介绍了几种经典模型结构，并且将这些模型对DMDA数据训练。在训练的过程中，对参数相似的模型选择了类似的参数选择方式，保证了模型在同一数据集和同一参数环境下的学习；对图神经网络模型，本章采用了Dropout技术，在训练过程中适当减少神经元个数，有效提高了训练效率，防止过拟合现象。经过模型的调参与调优，本章确定了模型的参数，并且进行了训练。最终得到了训练好的模型，并且在测试集上进行了对比分析。

分析结果显示，TransE翻译模型在测试集上的表现最优，双线性模型的表现也尚佳。原因是TransE翻译模型能够更好的学习到节点的表示，而双线性模型却忽略了DMDA数据集中隐含的上下文语义关系，因此表现欠佳。神经网络模型在该数据集上的表现稍弱，其原因可能是由于DMDA数据集规模较小，并且数据分布较为稀疏等。也有研究表明[[87](#_ENREF_87)]，神经网络模型在大规模，密集分布的数据集上表现较优，因此该模型受到了数据集本身属性的限制。最后本章将TransE模型对测试集预测出来的结果进行了展示。发现在预测的结果序列中，无论正确的尾实体是否在序列的第一位，其序列中的实体都是十分类似且符合常识的，因此模型在数据集上的关联预测过程是有效的。

通过此次关联预测任务发现，现有的模型应用于本章的细菌与疾病关联数据集是有效的，经过训练以后，能够达到一定的知识库补全的目的，对于未知的三元组关联，能够预测出有效的尾实体。而在细菌与知识库关联预测中，尚有更多的语义关联和领域知识，而通用领域的算法模型没有考虑到，因此接下来的工作可能针对这一方面进一步的改进。

# 总结与展望

## 总结

随着全球研究人员对生物信息、健康医疗领域的密切关注，微生物与人类疾病的关联挖掘也成为了研究的热点话题。肠道微生物在人体内栖息、繁衍，时时刻刻影响着人类的机体平衡与新陈代谢。已有的人类疾病种类繁多，随着生活条件、环境状态的改变，机体的平衡也在不断倾斜。人体及自然界中的微生物不断进行物质交换，微生物种群间形成共生、竞争、拮抗、捕食等交互，同时也将交互产物提供给机体所需，或是影响机体代谢。计算模型与科学技术的大发展，推动着生物信息学领域的快速前进，序列分析、模型预测、数据存储等依赖着高性能的计算工具，也激发了研究人员在这些领域内的不断创新。

生物医学文本数量扩增，以文本为载体的生物医学数据日益增多，如何更好的利用、存储这些文本资源，成为计算科学与生物信息科学的一个重大任务。计算人员通过将通用领域的自然语言处理技术、文本挖掘方法应用于生物医学领域，将机器学习、深度学习模型用于学习海量的生物医学文本。通用知识平台也蕴含了丰富的文本信息，通过统计学方法中复杂的映射公式，将文本处理为机器可识别的数学矩阵、向量模型，而语言学思想，使文本的上下文语义、词义也转化为文本特征，作为重要的文本挖掘依据。让机器理解文本语义，是文本挖掘的关键一步，为提取结构化知识打下基础。文本是一种非结构化的知识，将其中蕴含的结构化知识提取出来，能够节省存储空间，并且能够为进一步数据分析提供数据来源。

生物医学领域的数据库成百上千，数据标准却良莠不齐，本文旨在微生物与疾病交互领域，发布一个基于整合的细菌与疾病交互知识库，不仅囊括基于文本挖掘得到的细菌与疾病交互关联，也整合了细菌的属性特征。数据库资源不仅能够将现有的生物医学数据系统地整合在一起，也能够通过一些在线分析、可视化展示等功能，为研究人员提供方便快捷的线上数据分析与处理平台。

通过分子序列结构、代谢网络分析，能够利用统计学方法对现有的细菌与疾病的关联做进一步的关联预测。受到社交网络、搜索引擎、化学分子结构的启发，以及图数据库存储的广泛使用，人们开始研究对知识库网络的表示学习与关联预测。本文也针对细菌与疾病知识库网络提出了基于知识图谱的表示学习与细菌与疾病关联预测。通过将双线性模型、深度学习图神经网络模型以及翻译模型等应用于细菌与疾病交互网络，将学习到的节点表示归纳为图的表示，预测出未知的细菌与其他实体的交互关联，补充现有的知识库网络，并且也为进一步的实验研究设计提供参考依据。

本文对细菌与疾病知识库的研究工作，从文本挖掘、知识库整合与构建以及未知关联的知识发现三个方向依次进行了研究。通过研究过程的深入，该细菌与疾病知识库能够不断更新，为研究者提供开源的知识查询与表示平台，也能够为微生物实验设计提供参考，提高生物医学研究效率。

## 下一步研究工作

尽管本文按照知识挖掘、知识库构建以及知识发现三个流程开展了基于细菌与疾病知识库的研究，受到计算条件、硬件设施的限制，也依然出现了一些可改进之处，也为下一步的研究方向提供了一些思考：

1. 考虑集成学习的方法，开发细菌的属性挖掘工具，实现细菌与疾病知识库的细菌属性知识与细菌与疾病关联知识的同步更新；
2. 充分利用生物医学文本资源，除了在维基百科等公共平台，在PubMed等生物医学数据库中，也可以通过文本挖掘方法，提取更多的细菌与疾病关联交互；
3. 完善细菌与疾病知识库功能，按照细菌种群、细菌属性等分类方法，结合细菌的致病类型做更细致的可视化分析方案；
4. 进一步结合文本的语义特征、词特征，丰富知识库网络，加入“Attention”机制，使得模型更好的学习知识图谱表示。

综上所述，基于细菌与疾病的知识库网络研究还有许多的研究方向，然而想要进一步了解微生物对人体的作用，仅仅开展细菌与疾病的研究远远不够。细菌与细菌之间的两两交互、细菌对抗生素的作用、细菌与病毒的作用等都将成为细菌在人体内作用的关键一环。有关细菌的研究道阻且长，但对于细菌对人体的作用研究，进一步对人体的机体健康研究始终是本文的研究目标与方向。

# 参考文献

[1] Hoffmann R , Krallinger M , Andres E , et al. Text Mining for Metabolic Pathways, Signaling Cascades, and Protein Networks[J]. ence s STKE, 2005, 2005(283):pe21.

[2] Flemming H C , Wingender J , Szewzyk U , et al. Biofilms: an emergent form of bacterial life[J]. Nature Reviews Microbiology, 2016, 14(9):563-575.

[3] Mengqian Y , Ying Z , Qingwei C , et al. Metabonomics Research Progress on Liver Diseases[J]. Canadian Journal of Gastroenterology & Hepatology, 2017, 2017:1-10.

[4] Allahyari M , Pouriyeh S , Assefi M , et al. Text Summarization Techniques: A Brief Survey[J]. International Journal of Advanced Computer ence & Applications, 2017, 8(10):397-405.

[5] Sender R, Fuchs S, Milo R. Revised Estimates for the Number of Human and Bacteria Cells in the Body[J]. PLoS Biology, 2016, 14(8): e1002533.

[6] Mclean M H , Dieguez D , Miller L M , et al. Does the microbiota play a role in the pathogenesis of autoimmune diseases?[J]. Gut, 2015, 64(2):332-341.

[7] Turnbaugh P J , Ley R E , Hamady M , et al. The Human Microbiome Project[J]. Nature, 2007, 449(7164):804-810.

[8] Dekel S, Ein-Dor T, Berman Z, et al. Delivery mode is associated with maternal mental health following childbirth[J], Archives of women's mental health, 2019, 22(6): 817-824.

[9] Moschen A R , Wieser V , Tilg H . Dietary Factors: Major Regulators of the Gut's Microbiota[J]. Gut and liver, 2012, 6(4):411-416.

[10] Baumler A J, Sperandio V. Interactions between the microbiota and pathogenic bacteria in the gut[J]. Nature, 2016, 535(7610): 85-93.

[11] Scher J U, Ubeda C, Artacho A, et al. Decreased bacterial diversity characterizes the altered gut microbiota in patients with psoriatic arthritis, resembling dysbiosis in inflammatory bowel disease[J]. Arthritis Rheumatol, 2015, 67(1): 128-39.

[12] Ma W, Huang C, Zhou Y, et al. MicroPattern: a web-based tool for microbe set enrichment analysis and disease similarity calculation based on a list of microbes[J]. Sci Rep, 2017, 7: 40200.

[13] Brandes U, Kenis P, Lerner J, et al. Network analysis of collaboration structure in Wikipedia[C]. Proceedings of the 18th international conference on World wide web, 2009: 731-740.

[14] Li C, Lim K M, Chng K R, et al. Predicting microbial interactions through computational approaches[J]. Methods, 2016, 102: 12-9.

[15] Ghandikota S, Hershey G K K, Mersha T B. GENEASE: real time bioinformatics tool for multi-omics and disease ontology exploration, analysis and visualization[J]. Bioinformatics, 2018, 34(18): 3160-3168.

[16] Cheng L, Yang H, Zhao H, et al. MetSigDis: a manually curated resource for the metabolic signatures of diseases[J]. Brief Bioinform, 2019, 20(1): 203-209.

[17] Badal V D, Wright D, Katsis Y, et al. Challenges in the construction of knowledge bases for human microbiome-disease associations[J]. Microbiome, 2019, 7(1): 129.

[18] Lee M, Kim J W, Jang B. DOVE: An Infectious Disease Outbreak Statistics Visualization System[J]. IEEE Access, 2018, 6: 47206-47216.

[19] Theys K, Lemey P, Vandamme A M, et al. Advances in Visualization Tools for Phylogenomic and Phylodynamic Studies of Viral Diseases[J]. Front Public Health, 2019, 7: 208.

[20] Paun A , Yau C , Danska J S . The Influence of the Microbiome on Type 1 Diabetes[J]. Journal of Immunology, 2017, 198(2):590-595.

[21] Gil-Cruz C, Perez-Shibayama C, De Martin A, et al. Microbiota-derived peptide mimics drive lethal inflammatory cardiomyopathy[J], Science, 2019, 366: 881-886.

[22] Vázquez-Baeza Y, Callewaert C, Debelius J, et al. Impacts of the human gut microbiome on therapeutics[J], Annual Review of Pharmacology & Toxicology, 2018, 58: 253-270.

[23] Subramanian S, Huq S, Yatsunenko T, et al. Persistent gut microbiota immaturity in malnourished Bangladeshi children[J], Nature, 2014, 510(7505): 417-421.

[24] Noronha A, Modamio J, Jarosz Y, et al. The Virtual Metabolic Human database: integrating human and gut microbiome metabolism with nutrition and disease[J], Nuclc Acids Research 2019, 47(D1): D614-D624.

[25] Sender R, Fuchs S, Milo R. Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans[J]. Cell, 2016, 164(3): 337-40.

[26] Atul J Butte. The Integrative Human Microbiome Project: Dynamic Analysis of Microbiome-Host Omics Profiles during Periods of Human Health and Disease[J]. Cell Host & Microbe, 2014, 16(3):276-289.

[27] Proctor L M, Creasy H H, Fettweis J M, et al. The integrative human microbiome project[J], Nature, 2019, 569(7758): 641-648.

[28] Zhao L , Zhang F , Ding X , et al. Gut bacteria selectively promoted by dietary fibers alleviate type 2 diabetes[J]. ence, 2018, 359(6380):1151-1156.

[29] Paun A , Yau C , Danska J S . The Influence of the Microbiome on Type 1 Diabetes[J]. Journal of Immunology, 2017, 198(2):590-595.

[30] Walker, Allan W. The importance of appropriate initial bacterial colonization of the intestine in newborn, child and adult health[J]. Pediatric Research, 2017, 82(3): 387-395.

[31] Vrakas S , Mountzouris K C , Michalopoulos G , et al. Intestinal Bacteria Composition and Translocation of Bacteria in Inflammatory Bowel Disease[J]. Plos One, 2017, 12(1):e0170034.

[32] Sartor R B, Wu G D. Roles for Intestinal Bacteria, Viruses, and Fungi in Pathogenesis of Inflammatory Bowel Diseases and Therapeutic Approaches[J]. Gastroenterology, 2017, 152(2): 327-339 e4.

[33] Saeedi P, Yazdanparast M, Behzadi E, et al. A review on strategies for decreasing E. coli O157: H7 risk in animals[J], Microbial Pathogenesis, 2017, 103: 186-195.

[34] Engholm D H, Kilian M, Goodsell D S, et al. A visual review of the human pathogen Streptococcus pneumoniae[J], Fems Microbiology Reviews, 2017, 41(6): 854-879.

[35] Neish A S. Microbes in gastrointestinal health and disease[J]. Gastroenterology, 2009, 136(1): 65-80.

[36] Compare D, Rocco A, Sanduzzi Zamparelli M, et al. The Gut Bacteria-Driven Obesity Development[J]. Digestive Diseases, 2016, 34(3):221-229.

[37] Cheng Z, Meade J, Mankia K, et al. Periodontal disease and periodontal bacteria as triggers for rheumatoid arthritis[J], Best Practice & Research Clinical Rheumatology, 2017, 31(1): 19-30.

[38] Lopez J, Grinspan A. Fecal microbiota transplantation for inflammatory bowel disease[J], Gastroenterology hepatology, 2016, 12(6): 374.

[39] Grenham S, Clarke G, Cryan J F, et al. Brain-gut-microbe communication in health and disease[J]. Front Physiol, 2011, 2: 94.

[40] Dhariwal A, Chong J, Habib S, et al. MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data[J]. Nucleic Acids Res, 2017, 45(W1): W180-W188.

[41] 刘知远, 孙茂松, 林衍凯, 等. 知识表示学习研究进展[J], 2016, 53(2): 247-261.

[42] Trouillon T, Welbl J, Riedel S, et al. Complex embeddings for simple link prediction[C]. international conference on machine learning, 2016: 2071-2080.

[43] Segurabedmar I, Martinez P, De Pablosanchez C, et al. Using a shallow linguistic kernel for drug-drug interaction extraction[J]. Journal of Biomedical Informatics, 2011, 44(5): 789-804.

[44] Peng Y, Rios A, Kavuluru R, et al. Chemical-protein relation extraction with ensembles of SVM, CNN, and RNN models.[J]. arXiv: Computation and Language, 2018.

[45] Zinovyev A, Czerwinska U, Cantini L, et al. Collective intelligence defines biological functions in Wikipedia as communities in the hidden protein connection network[J]. PLOS Computational Biology, 2020, 16(2).

[46] Rollin G, Lages J, Shepelyansky D L. World Influence of Infectious Diseases From Wikipedia Network Analysis[J]. IEEE Access, 2019, 7: 26073-26087.

[47] Ma W, Zhang L, Zeng P, et al. An analysis of human microbe–disease associations[J]. Briefings in Bioinformatics, 2017, 18(1): 85-97.

[48] Janssens Y, Nielandt J, Bronselaer A, et al. Disbiome database: linking the microbiome to disease.[J]. BMC Microbiology, 2018, 18(1): 50-50.

[49] Li X, Fu C, Zhong R, et al. A hybrid deep learning framework for bacterial named entity recognition with domain features[J], BMC Bioinformatics, 2019, 20(16): 1-9.

[50] Wright D. NormCo: Deep disease normalization for biomedical knowledge base construction[D]. UC San Diego, 2019.

[51] Kwon S, Yoon S. End-to-End Representation Learning for Chemical-Chemical Interaction Prediction[J]. IEEE/ACM Trans Comput Biol Bioinform, 2019, 16(5): 1436-1447.

[52] Warikoo N, Chang Y, Hsu W, et al. LPTK: a linguistic pattern-aware dependency tree kernel approach for the BioCreative VI CHEMPROT task[J]. Database, 2018, 2018(2018).

[53] Nédellec C, Bossy R, Kim J-D, et al. Overview of BioNLP shared task 2013[C]. Proceedings of the BioNLP shared task 2013 workshop, 2013: 1-7.

[54] Bossy R, Golik W, Ratkovic Z, et al. Bionlp shared task 2013–an overview of the bacteria biotope task[C]. Proceedings of the BioNLP Shared Task 2013 Workshop, 2013: 161-169.

[55] Bossy R, Deléger L, Chaix E, et al. Bacteria Biotope at BioNLP Open Shared Tasks 2019[C]. Proceedings of The 5th Workshop on BioNLP Open Shared Tasks, 2019: 121-131.

[56] Mehryary F, Hakala K, Kaewphan S, et al. End-to-end system for bacteria habitat extraction[C]. BioNLP 2017, 2017: 80-90.

[57] May J, Priyadarshi J. Semeval-2017 task 9: Abstract meaning representation parsing and generation[C]. Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), 2017: 536-545.

[58] Cameron D, Kavuluru R, Rindflesch T C, et al. Context-driven automatic subgraph creation for literature-based discovery[J]. Journal of Biomedical Informatics, 2015: 141-157.

[59] Malas T B, Vlietstra W J, Kudrin R, et al. Drug prioritization using the semantic properties of a knowledge graph[J]. Scientific Reports, 2019, 9(1).

[60] Bakal G, Talari P, Kakani E, et al. Exploiting semantic patterns over biomedical knowledge graphs for predicting treatment and causative relations[J]. Journal of Biomedical Informatics, 2018: 189-199.

[61] Baskin I I, Winkler D A, Tetko I V, et al. A renaissance of neural networks in drug discovery.[J]. Expert Opinion on Drug Discovery, 2016, 11(8): 785-795.

[62] Chen H, Engkvist O, Wang Y, et al. The rise of deep learning in drug discovery.[J]. Drug Discovery Today, 2018, 23(6): 1241-1250.

[63] Lounkine E, Keiser M J, Whitebread S, et al. Large-scale prediction and testing of drug activity on side-effect targets[J]. Nature, 2012, 486(7403): 361-367.

[64] Nourani E, Reshadat V. Association extraction from biomedical literature based on representation and transfer learning[J]. Journal of Theoretical Biology, 2020, 488: 110112.

[65] Fairchild G, De Silva L, Del Valle S Y, et al. Eliciting disease data from Wikipedia articles[C]. Ninth International AAAI Conference on Web and Social Media, 2015.

[66] Zesch T, Müller C, Gurevych I. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary[C]. LREC, 2008: 1646-1652.

[67] Sorokin D, Gurevych I. Context-aware representations for knowledge base relation extraction[C]. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 1784-1789.

[68] Erxleben F, Günther M, Krötzsch M, et al. Introducing Wikidata to the linked data web[C]. International Semantic Web Conference, 2014: 50-65.

[69] Vrandecic D, Krotzsch M. Wikidata: a free collaborative knowledgebase[J]. Communications of The ACM, 2014, 57(10): 78-85.

[70] Lever J, Jones S J. Painless relation extraction with kindred[C]. BioNLP 2017, 2017: 176-183.

[71] Lever J, Zhao E Y, Grewal J K, et al. CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer[J]. Nature Methods, 2019, 16(6): 505-507.

[72] Comeau D C, Islamaj Doğan R, Ciccarese P, et al. BioC: a minimalist approach to interoperability for biomedical text processing[J], Database, 2013, 2013.

[73] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology[J]. Nucleic Acids Research, 2004, 32(90001): 267-270.

[74] Ortiz-Alvarez R, Fierer N, De Los Rios A, et al. Consistent changes in the taxonomic structure and functional attributes of bacterial communities during primary succession[J]. ISME J, 2018, 12(7): 1658-1667.

[75] Federhen S. The NCBI Taxonomy database.[J]. Nucleic Acids Research, 2012: 136-143.

[76] Chen H, Zhang Z, Jiang S, et al. New insights on human essential genes based on integrated analysis and the construction of the HEGIAP web-based platform.[J]. Briefings in Bioinformatics, 2019.

[77] Burch C. Django, a web framework using Python: tutorial presentation[J]. Journal of Computing Sciences in Colleges, 2010, 25(5): 154-155.

[78] Pilgrim M. HTML5: up and running: dive into the future of web development[M]. O'Reilly Media, Inc, 2010.

[79] Garrett J J. Ajax: A new approach to web applications[J], 2005.

[80] Pokorný P, Stokláska K. Chart visualization of large data amount[C]. Computer Science On-line Conference, 2017: 460-468.

[81] Nickel M, Tresp V, Kriegel H-P. A three-way model for collective learning on multi-relational data[C]. Icml, 2011: 809-816.

[82] Yang B, Yih W, He X, et al. Embedding Entities and Relations for Learning and Inference in Knowledge Bases[C]. international conference on learning representations, 2015.

[83] Hayashi K, Shimbo M. On the Equivalence of Holographic and Complex Embeddings for Link Prediction[C]. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2017: 554-559.

[84] Liu H, Wu Y, Yang Y. Analogical inference for multi-relational embeddings[C]. Proceedings of the 34th International Conference on Machine Learning-Volume 70, 2017: 2168-2178.

[85] Socher R, Chen D, Manning C D, et al. Reasoning with neural tensor networks for knowledge base completion[C]. Advances in neural information processing systems, 2013: 926-934.

[86] Nguyen D Q, Nguyen T D, Nguyen D Q, et al. A Novel Embedding Model for Knowledge Base Completion Based on Convolutional Neural Network[J]. north american chapter of the association for computational linguistics, 2018: 327-333.

[87] Schlichtkrull M, Kipf T N, Bloem P, et al. Modeling relational data with graph convolutional networks[C]. European Semantic Web Conference, 2018: 593-607.

[88] Kalchbrenner N, Grefenstette E, Blunsom P, et al. A Convolutional Neural Network for Modelling Sentences[C]. meeting of the association for computational linguistics, 2014: 655-665.

[89] Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations[C]. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014: 701-710.

[90] Scarselli F , Gori M , Tsoi A C , et al. The graph neural network model[J]. IEEE Transactions on Neural Networks, 2009, 20(1):61-80.

[91] Rong X . word2vec Parameter Learning Explained[J]. Computer ence, 2014.

[92] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data[C]. Advances in neural information processing systems, 2013: 2787-2795.

[93] Chapelle O, Metlzer D, Zhang Y, et al. Expected reciprocal rank for graded relevance[C]. Proceedings of the 18th ACM conference on Information and knowledge management, 2009: 621-630.

# 攻读硕士期间发表的论文和参加的科研项目

**公开发表的学术论文：**

1. Li X, Fu C（付诚诚）, Zhong R, et al. A hybrid deep learning framework for bacterial named entity recognition with domain features[J]. BMC bioinformatics, 2019, 20(16): 1-9.
2. Li X, Fu C(付诚诚), Zhong R, et al. Bacterial Named Entity Recognition Based on Language Model[C]. 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2019: 2715-2721.
3. Zhong R, Li X, Sun X, Fu C(付诚诚), He T, & Jiang X (2019). Microbial Interaction Extraction based on Feature Vector and SVM [C], Asia Pacific Bioinformatics Conference (APBC). Poster paper.
4. Zhong R, Li X, Sun X, Fu C(付诚诚), He T, & Jiang X (2019). Microbial Interaction Extraction from Biomedical Literature using Max-Bi-LSTM [C]. 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Short paper.

**攻读硕士学位期间参加的科研项目：**

1. 国家自然科学基金重点项目—高通量微生物组学数据模式提取和分析（项目编号：61532008，时间：2016年1月—2020年12月）
2. 国家重点研发计划—精准医学研究重点专项（项目编号：2017YFC0909502， 时间：2017年7月—2019年12月）
3. 国家自然科学基金面上项目—基于多组学数据的微生物高阶相互作用预测关键技术研究（项目编号：61872157，时间：2019年1月—2022年12月）

# 致谢

两年的硕士生涯即将告一段落，随着论文致谢部分的开始，我将对这两年的华师光阴做一段小结。

首先需要感谢我的硕士生导师蒋兴鹏教授。在这短暂而充实的硕士研究生生涯中，蒋老师一直对我十分照顾。从最初引导我迈入生物信息学领域的大门，到教导我们如何阅读英文文献，到后来一直鼓励我不断探索新的研究内容，蒋老师作为一个年轻的硕士生导师，亦师亦友，见证了我从一个科研小白，成长为一个热爱科研，忠于探索的硕士研究生。这篇毕业论文的完成，也离不开蒋老师的鼓励与教导。此外，课题组的沈显君教授，赵卫中教授以及院长何婷婷老师，也一直伴随着我的成长，无论是从科研还是生活上，都对我给予了许多帮助。我十分庆幸，这两年的学习经历虽然坎坷而辛苦，但有这些老师的帮助与鼓励，我才能够乘风破浪，收获成长。

其次，陪我日夜兼程的也少不了课题组的小伙伴们。感谢贾易、孙晗、郑佳、李旭升、钟然、祝青、刘冲等师兄师姐对我科研与生活的关照，感谢孙霞、王宇洁、杨玉娟、韦智鑫、霍斑、牟汉林等同门与我的相互帮助与关心，感谢蒋肖斌师妹对我的论文修改提供的帮助。因为同门之间的团结友爱，才使得实验组能够一直积极向上，共同进步。

最后，也是最需要感谢的是家人的关心与陪伴，一个和谐幸福的家庭是我学习与生活中坚强的后盾。两年的光阴很短，但正是这些重要的人和事，陪伴我度过了这段重要的时光。接下来，我将继续整理行装，开启我人生的下一段重要的旅程。我愿将这些人曾给予我的帮助，化作前行的动力，始终在学术与科研的道路上，探索新知识的方向。