分类号： 密级：

UDC： 学号：

南 昌 大 学 专 业 学 位 硕 士 研 究 生

学 位 论 文

**面向知识图谱的多源数据图表征**

**及检索优化方法研究**

## Research on Multi-source Data Graph Representation

## and Retrieval Optimization Method Oriented to Knowledge Graph

程婕

培养单位（院、系）：信息工程学院

指导教师姓名、职称：徐子晨 教授

指导教师姓名、职称：卜象平 高级工程师

专业学位种类：工程硕士

专业领域名称：软件工程

论文答辩日期： 2021年6月6日

答辩委员会主席：

评阅人：

年 月 日

|  |
| --- |
| **面向知识图谱的多源数据图表征及检索优化方法研究**  **程婕**  **南昌大学** |

**一、学位论文独创性声明**

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，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，也不包含为获得南昌大学或其他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谢意。

学位论文作者签名（手写）： 签字日期： 年 月 日

**二、学位论文版权使用授权书**

本学位论文作者完全了解南昌大学有关保留、使用学位论文的规定，同意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，允许论文被查阅和借阅。本人授权南昌大学可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编本学位论文。同时授权北京万方数据股份有限公司和中国学术期刊（光盘版）电子杂志社将本学位论文收录到《中国学位论文全文数据库》和《中国优秀博硕士学位论文全文数据库》中全文发表，并通过网络向社会公众提供信息服务，同意按“章程”规定享受相关权益。

学位论文作者签名（手写）： 导师签名（手写）：

签字日期： 年 月 日 签字日期： 年 月 日

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 论文题目 | 面向知识图谱的多源数据图表征及检索优化方法研究 | | | | |
| 姓 名 | 程婕 | 学号 | 411016118347 | 论文级别 | 博士□ 硕士☑ |
| 院/系/所 | 信息工程学院 | | 专业 | 软件工程 | |
| E\_mail | jiechengstudy@outlook.com | | | | |
| 备注： | | | | | |

☑公开 □保密（向校学位办申请获批准为“保密”， 年 月后公开）

**摘 要**

知识图谱生成中所使用的数据存在来源不同、类型多样、利用率低等特征，这些特征易造成多源数据难表征问题和知识图谱检索效率低问题。针对上述问题，迫切需要多源异构数据有效表征及检索优化方法。本文面向多源异构文本数据表征及检索优化方法展开研究，主要工作如下：

第一、本文通过研究多源数据、知识图谱和图检索技术，探究三者间的关联，揭示了知识图谱中的多源数据图表征问题以及图表征下多源数据快速检索问题，并就相关问题的国内外研究现状展开调查与分析。

第二、基于上述问题与内容，本文提出了两种图表征方法，一种为基于马尔科夫决策过程的二分图表征法，该方法基于马尔科夫决策过程和二分图概念提炼并表征多源数据中的数据关系，实现多源数据的融合表征，但该方法存在检索效率低的问题。基于此本文提出了第二种表征方法——基于图相似性度量的二部二分图表征方法，该方法使用最佳决策算法与图相似性度量算法，在第一种方法的基础上，提炼原二分图中的top-k结构形成top-k二分图，并与原二分图形成二部二分图结构。该结构中top-k二分图能够缩小检索范围，提高检索效率，二部二分图结构能实现图的动态演进。

第三、本文分别使用问卷型医疗数据和电子病历型医疗数据对两种方法进行实验验证和性能评估。评估结果表明，与传统向量表征方法相比，本文中的二分图表征方法具有更优的可解释性和表征效果；与二分图表征方法相比，二部二分图表征方法构建时间平均减少65.8%，内存消耗平均减少50.2%，匹配时间平均减少53.4%。二部二分图结构动态演进后，检索准确率平均提高8%。

**关键词：**表征优化；知识图谱；多源数据

# ABSTRACT

The data used in the generation of the knowledge graph has characteristics such as different sources, diverse types, and low utilization. These characteristics are likely to cause the problem of difficult characterization of multi-source data and low efficiency of knowledge graph retrieval. In view of the above problems, there is an urgent need for effective characterization and retrieval optimization methods for multi-source heterogeneous data. This dissertation focuses on the research of multi-source heterogeneous text data representation and retrieval optimization methods. The main work is as follows:

First, this article explores the relationship between multi-source data, knowledge graph and graph retrieval technology, and reveals the problem of multi-source data graph representation in knowledge graph and the problem of fast retrieval of multi-source data under graph representation. Investigate and analyze the current status of domestic and foreign research on the issue.

Second, based on the above problems and content, this paper proposes two graph representation methods. One is a bipartite graph representation method based on Markov decision process. This method is based on Markov decision process and bipartite graph concept to refine and represent multi-source The data relationship in the data can realize the fusion characterization of multi-source data, but this method has the problem of low retrieval efficiency. Based on this, this paper proposes a second representation method-a two-part bipartite graph representation method based on graph similarity measurement. This method uses the best decision algorithm and graph similarity measurement algorithm. On the basis of the first method, the original method is refined. The top-k structure in the bipartite graph forms a top-k bipartite graph, and forms a two-part bipartite graph structure with the original bipartite graph. In this structure, the top-k bipartite graph can narrow the search scope and improve the retrieval efficiency, and the two-part bipartite graph structure can realize the dynamic evolution of the graph.

Third, this article uses questionnaire medical data and electronic medical record medical data to conduct experimental verification and performance evaluation of the two methods. The evaluation results show that compared with the traditional vector representation method, the bipartite graph representation method in this paper has better interpretability and characterization effect; compared with the bipartite graph representation method, the construction time of the bipartite graph representation method is reduced by 65.8% on average , The memory consumption is reduced by an average of 50.2%, and the matching time is reduced by an average of 53.4%. After the dynamic evolution of the two-part bipartite graph structure, the retrieval accuracy increased by an average of 8%.

**Key Words:** representation optimization; knowledge graph; multi-source data

**目 录**

[第1章 引言 2](#_Toc68374651)

[1.1 研究背景和意义 2](#_Toc68374652)

[1.2 国内外研究现状 2](#_Toc68374653)

[1.2.1 多源数据融合方法研究现状 2](#_Toc68374654)

[1.2.2 多源数据表征方法研究现状 2](#_Toc68374655)

[1.2.3 图检索优化方法研究现状 2](#_Toc68374656)

[1.3 本文主要研究内容 2](#_Toc68374657)

[1.4 本文组织结构 2](#_Toc68374658)

[第2章 基础理论与相关技术背景 2](#_Toc68374659)

[2.1图的定义及相关关系 2](#_Toc68374660)

[2.1.1 图与二分图 2](#_Toc68374661)

[2.1.2 图的表达 2](#_Toc68374662)

[2.1.3 图数据库 2](#_Toc68374663)

[2.2 马尔科夫理论 2](#_Toc68374664)

[2.2.1 马尔科夫过程 2](#_Toc68374665)

[2.2.2 马尔科夫奖励过程 2](#_Toc68374666)

[2.2.3 马尔科夫决策过程 2](#_Toc68374667)

[2.3 相似性度量方法 2](#_Toc68374668)

[2.3.1 语义相似性度量方法 2](#_Toc68374669)

[2.3.2 结构相似性度量方法 2](#_Toc68374670)

[2.4 本章小结 2](#_Toc68374671)

[第3章 基于马尔科夫决策过程的二分图表征方法 2](#_Toc68374672)

[3.1 基于马尔科夫决策过程的图表征 2](#_Toc68374673)

[3.2 基于马尔科夫决策过程的二分图表征 2](#_Toc68374674)

[3.3 面向知识图谱的基于马尔科夫决策过程的二分图表征 2](#_Toc68374675)

[3.4 本章小结 2](#_Toc68374676)

[第4章 基于图相似性度量的二部二分图表征方法 2](#_Toc68374677)

[4.1 有限马尔科夫决策过程二分图相似性度量 2](#_Toc68374678)

[4.1.1 基于马尔科夫决策过程二分图的相似性 2](#_Toc68374679)

[4.1.2 二分图状态相似性度量 2](#_Toc68374680)

[4.1.3 二分图动作相似性度量 2](#_Toc68374681)

[4.1.4 有限马尔科夫决策过程二分图相似性度量方法 2](#_Toc68374682)

[4.2 基于最佳决策及相似性度量的top-k二分图表征 2](#_Toc68374683)

[4.3 基于图相似性度量的二部二分图表征及其动态演进 2](#_Toc68374684)

[4.3.1 基于最佳决策和图相似性度量的二部二分图表征 2](#_Toc68374685)

[4.3.2 二部二分图的动态演进 2](#_Toc68374686)

[4.4 本章小结 2](#_Toc68374687)

[第5章 实验设计与分析 2](#_Toc68374688)

[5.1 二分图表征方法实验分析与应用 2](#_Toc68374689)

[5.1.1 实验数据描述 2](#_Toc68374690)

[5.1.2 实验过程描述 2](#_Toc68374691)

[5.1.3 实验结果分析 2](#_Toc68374692)

[5.1.4 可解释性评估 2](#_Toc68374693)

[5.2 二部二分图表征方法实验分析与应用 2](#_Toc68374694)

[5.2.1 实验数据描述 2](#_Toc68374695)

[5.2.2 实验过程描述 2](#_Toc68374696)

[5.2.3 实验结果分析 2](#_Toc68374697)

[5.3 本章小结 2](#_Toc68374698)

[第6章 总结与展望 2](#_Toc68374699)

[致 谢 2](#_Toc68374700)

[参考文献 2](#_Toc68374701)

[攻读学位期间的研究成果 2](#_Toc68374702)

**第1章 引言**

**1.1 研究背景和意义**

在这个时代下，数据内容呈现出爆炸式的增长态势。此时，如何更好的利用数据以及数据中所传递出来的知识成为了热点研究问题[1]，而为了更好的开展相应研究，知识图谱技术应运而生。Google于2012年提出知识图谱( Knowledge Graph )的概念[2]，它是一种用于增强搜索引擎功能的知识库，其本质上是一个以图的形式揭示实体之间关系的语义网络，该网络可以提取数据中的实体，并将实体与实体之间的相互关系形式化的描述并展示出来[3]。通过与其他技术的结合（如推理技术等[4]），知识图谱可以给一些研究（如人工智能系统等）提供先验知识，从而让这些研究具有解决复杂问题的能力[4][5]。也正因为这个原因，知识图谱在智能搜索[6][7]、个性化推荐[8][9]等工作中有重要应用价值。

然而，由于数据生产与采集方式各有不同，数据结构（如结构化数据和非结构化数据等）和存储方式（如文件、关系型数据库、非关系型数据库等）各有差异，知识图谱生成过程中所使用的数据具有数据来源多样、数据类型不同，即多源异构的特点[10]，而如何利用这些数据并实现数据融合是知识图谱构建过程中的重要一环[11]。实际上，目前很多行业和领域的数据处理工作中都面临着多源数据融合问题，例如智慧城市[12][13][14]、智慧医疗[15][16]等等。但由于多源数据往往具有高维、异构的特点，所以数据处理和融合工作具有较大难度。因此，对多源数据融合方法的研究有着极为重要的理论价值和实践意义。

在多源数据融合工作中，如何在多源数据中求同存异、如何快速找寻数据之间的关系是关键一步[3]。由于图在关系表征上的优越性[17]以及图结构与知识图谱之间的天然契合[18][19]，以图为基础的相关技术给多源数据融合工作，尤其是面向知识图谱的多源数据融合工作提供了新的思路。然而，以图为基础的关系表征方法和融合技术同样带来了一系列思考与问题，例如，如何将多源的大规模数据快速转换成对应实体、如何快速提取实体间的关系进行关系表征以及如何将关系表征快速嵌入到知识图谱的语义网络中完成多源数据的图表征工作等[20][21][22]。除此之外，传统图结构以及知识图谱具有静态、显性的特点，但实际数据处理工作中往往面对过程性和决策性数据，这些数据除了具备多源、异构、高维度等特点之外，还具有一定的动态性、隐含性与主观性，因此，图和知识图谱的动态演进同样是相关领域的研究热点[23][24]。在此基础上，图与知识图谱所面临的另一个问题是如何进行快速检索。这是因为知识图谱的主要工作是通过检索为下游应用提供建议或决策，而基于大规模数据构建的知识图谱是一个结构复杂且维度极高的语义网络，在该网络开展检索工作往往面临着开销大、效率低等问题。

针对上述情况，本文主要研究两个问题，分别是面向知识图谱的多源数据图表征问题以及图表征下多源数据的快速图检索问题。围绕这两个问题，本课题将利用问卷型医疗数据和电子病历型医疗数据，重点研究这两类特定数据在特征级多源数据融合下的图表征与快速图检索技术，充分发挥知识图谱在实体关系构建与查询方面的潜在优势的同时，实现面向知识图谱的多源数据图表征与检索优化工作。

**1.2 国内外研究现状**

对于多源数据融合表征及检索优化问题，国内外许多研究者进行了深入地研究。这些研究根据方向、内容和方法大致可以划分为三个部分，分别是多源数据融合部分、多源数据表征部分以及图检索优化部分。本节将分别对这三个部分及其涉及的方法进行介绍。

**1.2.1 多源数据融合方法研究现状**

多源数据融合最初被称为多传感器信息融合，随着近年来的发展逐渐引申为将来自不同来源的数据进行整合并链接在一起，目的是为了得到更加统一且准确的数据，从而增加后续的决策过程[25]。目前，国内外对多源数据融合工作提出了三个针对不同阶段的融合方法，分别是针对前期融合的数据级融合方法、针对中期融合的特征级融合方法以及针对后期融合的决策级融合方法[26]。三种多源数据融合方法的实现难度依次递增，其中，数据级和特征级的融合方法是使用较多的方法。而又因为数据集融合方法对多源数据间的关联性考虑较少，不能很好的利用数据之间的关系，因此，相比之下，特征级融合方法是目前使用最多的融合方法。

（1）数据级融合方法

数据级融合方法是直接对原始数据进行处理，然后在数据挖掘任务的不同阶段使用不同的数据完成相应的融合分析，这种方法是典型的松耦合融合方式，即数据和数据之间的融合关系较为宽松，这种融合方法使用较为简单并保留了大部分数据的原始信息，因此信息损失量相对较小。但这类方法对单源数据的要求较高并且计算量大，若单源数据的处理情况不理想，就会影响到后期多源数据的利用，所以容错能力相对较差。

这类方法中比较常见的融合技术有加权平均法[27]、卡尔曼滤波法[28]、蒙特卡洛法[29]等。例如，Garcia等人[27]采用像素加权平均策略对深度传感器数据进行融合，该方法基于联合双边采样将低分辨率图像和高分辨率图像融合在一起，从而提高传感器的分辨率。但该方法也具有这类方法的弊端，即如果单源数据处理不理想，那么加权平均之后的效果未必会有直接分析某一单源数据的效果好。

（2）特征级融合方法

中期融合的特征级融合也称为基于数据关系的融合或模型级融合，该方法是从不同数据源的数据中学习数据的含义与数据的关系，然后将多源数据关系融合成新的数据关系，通过对新的数据关系的特征提取来表征实体之间的关联，进而完成多源数据融合工作[12]。例如Zhang等人[30]将数据分别送入ResNet、VGG16和VGG19三个模型，将三个模型的产出融合成新的数据关系与特征并将其输入神经网络模型进行进一步的分析，这种做法就是非常典型的中期融合方法。

特征级融合中常见的融合方法包括模糊推理法[31][32]、神经网络法[33]和产生式规则法等。其中，模糊推理法一般基于模糊推理规则对多源数据的身份信息进行模糊融合，以得出更可靠的目标识别[31]。相比于其他方法,该方法更适合于不精确或不确定的多源数据（如感官数据）。例如，Sagardip等人[32]基于模糊推理和模糊聚类开发了一种多传感器数据融合技术，该技术利用获得的聚类信息在模糊推理的基础上实现了对感官数据的预测。神经网络法在数据融合上较为灵活，它既可以从模型输入的角度完成多源数据的融合，也可以依据模型的输出，通过多个模型的组合拼接来完成多源数据的融合分析。例如，Fraser等人[33]提出了一种基于神经网络的数据融合方法，该方法设计了一种粒状神经网络（Granular Neural Network）,该网络可以处理数据库中的多源粒度数据，然后基于不同种类的粒度数据做出决策，除此之外，该网络还可以输入输出数据之间的内部粒度关系，从而做出预测。这种方法就是从模型输入就已经开始数据融合工作，而Zhang等人[30]的方法则是依据模型的输出，将某些模型的输出当作模型的输入来完成数据融合工作。

由于特征级融合方法在使用过程中不用每次对原始数据进行处理，所提取出来的关系和特征在一定场景下能够重复使用，所以可以有效减少数据分析过程中的计算量。但不足之处在于，该类方法的精度损失问题明显大于数据级融合方法。

（3）决策级融合方法

后期融合又称为基于语义的融合或决策级融合，它是在理解多源数据含义和不同数据特征关系的前提下，在数据融合过程中利用人类思考问题的方式抽象不同数据的语义来完成跨源数据融合工作。这种融合方式更加智能，并且具备较好的容错能力和抗干扰能力。缺点是相比于前两种方法，这种方法会带来较大的精度损失。

决策级融合中常见的融合方法包括Dempster-Shafer法（D-S法）[34][35]和贝叶斯推理法[36][37]等。其中，Dempster-Shafer法基于人对客观世界的认识，对论证的合成提供了系统性的合成公式，可以在数据源提供的信息不连续、不完整且没有单个数据源可以产生压倒性的绝对可能性来识别最可能的事件时，实现多源数据的融合。例如，Nour-Eddin等人[34]使用D-S法将来自感应环路传感器和收费站的行驶时间数据合并，以产生一个改进的行驶时间估计，从而提高对驾驶员和交通管理人员行驶时间的估计。Khan等人[35]基于D-S法提出了一种具有新颖熵函数的8步算法，并将8步算法应用于时域以实现时域数据的顺序组合。和其他方法相比，该方法显示出更好的抗干扰能力。但这类方法的不足之处在于论证之间必须相互独立，且论证合成规则的合理性与有效性存在较大的争议。贝叶斯推理法的基本思想是将多源数据融合工作分成多个等级进行，给定前一个等级数据的似然估计后，依据后一个等级的数据对前一个等级的似然估计进行更新，从而实现多源数据的融合[36]。例如Victor等人[37]基于贝叶斯推理法提出了一种两阶段数据融合的状态监测方法，该方法允许将来自不同状态监视系统的数据进行组合，并分别在本地和全局两个范围内分别使用贝叶斯推理法对多个监视系统的数据进行融合，从而减少因机器故障而导致的系统指示错误。这类方法的不足之处在于先验知识条件概率的获取相对较难，且该方法要求所使用的概率之间相互独立，这一要求在实际应用中较难实现。

**1.2.2 多源数据表征方法研究现状**

表1.1列举了部分特征级多源数据表征方法及其拓展。其中，Wangner[38]和Kirtchenko[39]等利用情感词典来提取多源文本数据离散特征，并将他们融合输入模型训练情感分类器，饶元淇等人[40]面向多特征向量的轨迹数据提出了多特征融合表征方法。这些方法基于特征向量的形式完成数据表征工作，虽然基于特征特征向量的表征方法操作简单，但却存在特征离散的问题。与此同时，随着技术不断发展，数据特征维度越来越高，此时，特征向量难以表征数据的全部特征以及数据之间的关系，并且使用特征向量将所有的特征输入模型，不仅不能加快研究的进展，甚至还可能成为研究中的一大负担。此时，面对数据的高维度特征，如何对特征进行筛选、如何获得并表征特征之间关系成为研究的重点。在这基础上，由于图对关系的天然可表示性，使得多源数据融合与图表征、图嵌入等图表示方法联系在了一起。

表1.1 多源数据表征方法研究

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 场景  文献 | 基于向量表征 | | 基于图表征 | | 面向知识图谱 | |
| 效果 | 难度 | 效果 | 难度 | 效果 | 难度 |
| DCU[38] | ★ | ★ |  |  |  |  |
| NRC-Canada-2014[39] | ★ | ★ |  |  |  |  |
| 多特征融合  方法[40] | ★★ | ★★ |  |  |  |  |
| 图嵌入框架[41] |  |  | ★★★ | ★★ |  |  |
| 线性整数优化  方法[42] |  |  | ★★★ | ★★★ |  |  |
| 本体表示图  模型[43] |  |  | ★★★ | ★★★★ |  |  |
| PNFE[44] |  |  | ★★★ | ★★★ | ★★★ | ★★★ |
| CLMed[45] |  |  | ★★★ | ★★★ | ★★★ | ★★★ |
| 双层随机游走方法[45] |  |  | ★★★ | ★★★ | ★★★ | ★★★ |

Zhao[41]等人通过图嵌入的方式进行多源时空数据融合表征分析，该方法从高维的多源时空数据中提取低维特征表示，根据多源数据特征之间的关系构建图网络结构从而进行语义探索。Nikolay[42]等人提出了一种新颖的线性整数优化方法，用于识别不同基因之间的复杂生物学依赖性，并根据该依赖性生成图结构用于描述不同基因之间的关系。Yi[43]等人则基于语义关系设计了带有实体支持的多源数据融合框架，该框架本质上是一种基于本体表示的图结构，能够很好的支持实体匹配和多源数据融合表征工作。和特征向量表征方法相比，基于图的特征表示方法在表示数据本身特征的基础上能够更好地表示数据之间的关系，特征和特征之间不再离散存在。然而，相比于特征向量的线性结构，图结构更为复杂，因此在操作、管理和储存上难度较大，亟需一种方法对图表征下的图结构进行有效管理。

为了更好存储和利用多源数据表征后的图结构，多源数据图表征逐渐和知识图谱结合在一起，例如，Pan[44]等人基于多源数据融合表征构建了非功能需求的知识图谱来完成链接预测任务，Sheng[45]等人基于跨语言医学数据融合构建了大规模跨语言医学知识图谱用于提高医学知识检索质量。刘峤[45]等人利用知识图谱表征实体语义上的双向关系，进而完整知识推理。这些方法利用知识图谱在关系表征上的优点，对数据及其关系进行抽取、融合、加工、表示和存储，使得图的操作和管理更加高效、有效。然而，归根结底，知识图谱中的图是普通图结构。当知识图谱规模较大时，图谱中检索等操作依旧存在效率低下的问题。

**1.2.3 图检索优化方法研究现状**

图表征是目前数据分析中常用的方法之一，但随之而来的问题是在面对庞大的图结构时如何更好的检索或匹配到有用信息。针对这一问题，国内外进行了许多研究，这些研究可以主要分为两个方面，一个是对图结构进行优化，另一个是对图检索方式进行优化，表1.2分别列举了部分在图结构优化和图检索方式优化的相关研究。

对于图结构优化，Khalil[47]等人提出了强化学习和图嵌入的独特组合来优化图结构，Cui[48]等人提出了图贝叶斯优化框架，通过将该框架和图内核结合起来的方式充分利用隐式图结构特征补充图显式特征，从而实现图结构优化的目的。Yi[49]等人提出了一种新颖的框架，称为联合图优化和投影学习（JGOPL），用于基于图降维的图结构优化工作，这些方法在针对图结构进行优化后，实验结果均有一定的改善或提升，但却也存在着不同的问题，例如，Khalil[47]等人和Cui[48]等人的方法虽然可以优化图结构，但却使得图丧失了可解释性。

表1.2 图计算优化方法

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 方法  文献 | 图结构优化 | | 图检索方式优化 | |
| 效果 | 难度 | 效果 | 难度 |
| 图组合优化方法[47] | ★★★ | ★★ |  |  |
| 图贝叶斯优化[48] | ★★★ | ★★★ |  |  |
| JGOPL[49] | ★★★ | ★★★ |  |  |
| 图匹配模型优化[50] |  |  | ★★★ | ★★★ |
| Gromov-Wasserstein  学习框架[51] |  |  | ★★★ | ★★★ |
| 双层随机游走方法[45] |  |  | ★★★ | ★★★ |
| 图匹配网络[52] | ★★★ | ★★★★ | ★★★★ | ★★★★ |

在图检索优化方面，Yang[50]等人提出了基于邻接矩阵模型和基于高斯平滑的图匹配优化方法，Xu[51]等人提出了一种新颖的Gromov-Wasserstein学习框架，用于联合匹配（对齐）图并学习关联图节点的嵌入向量，Li[52]等人提出了一种新颖的图匹配网络模型，该模型在给定一对图作为输入的情况下，通过一种新的基于跨图关注度的匹配机制来进行图检索工作。刘峤[45]等人面对图结构上的检索工作，引入随机采样机制，在避免图检索中穷举搜索的同时，提高计算效率。这些方法虽然可以提高图检索的效率，但却忽视了对图结构本身的优化。当然，目前也有一些对图结构和图检索方式同时进行优化的工作，如Li[52]等人在提出的新颖图匹配网络模型在提升检索效率的同时，通过利用GNN训练了图结构，从而达到图结构优化的目的，但在多源数据融合的基础上，面向知识图谱对图结构和图检索方式进行优化的相关工作仍有待深入研究。

**1.3 本文主要研究内容**

本文针对面向知识图谱的多源数据图表征问题以及图表征下多源数据的快速图检索问题进行了研究，总体技术路线如图1.1所示：

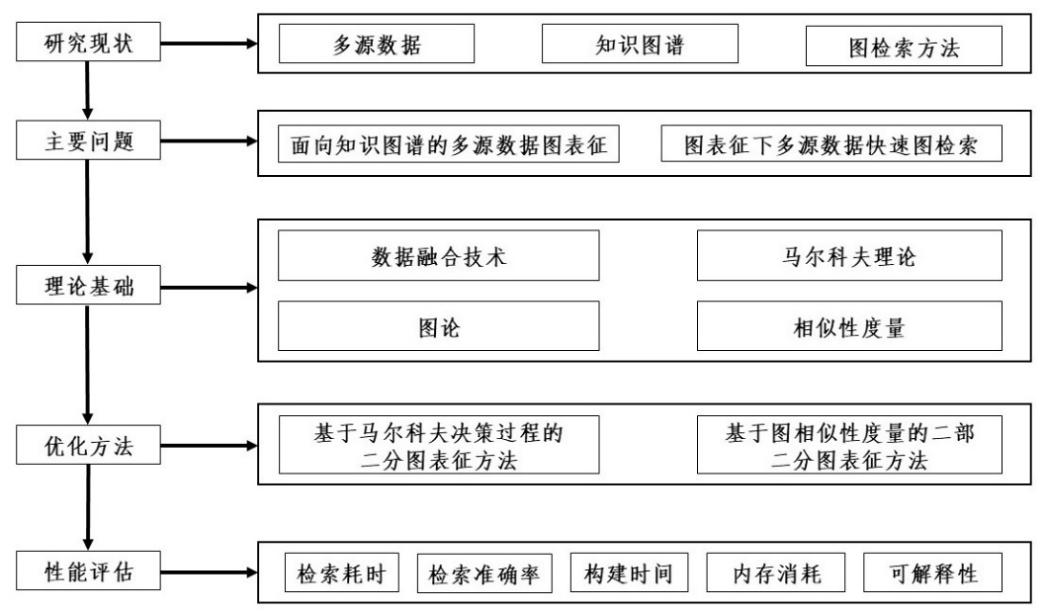


图1.1 本文总技术路线示意图

本文主要学习强化学习、自然语言处理和知识图谱中的部分先进技术，针对多源数据图表征和检索问题提出了两种方法，并将所提方法应用到问卷型医疗数据和电子病历型医疗数据的处理过程中，最后通过实验验证这两种方法的可行性和有效性。具体工作如下：

1. 基于马尔科夫决策过程的二分图表征方法

受强化学习中马尔科夫链的启发，综合考量多源数据条件下结构化与非结构化数据、实体间显性关系与隐性关系等多方面因素，本文提出基于马尔科夫决策过程的多源数据图表征方法，实现对数据时序关系和逻辑关系的有效提取，在表征数据显性关系的同时凸显数据的内部关联。与此同时，本方法利用二分图结构实现数据的表征优化，这一方法可以在保证表征方法可解释的前提下，优化多源数据图表征结构，从而实现提高数据检索效率的目的。

2. 基于图相似性度量的二部二分图表征方法

对于庞大且复杂的二分图结构，本文基于图相似性度量提出二部二分图表征方法。该方法依托有限马尔科夫二分图相似性度量算法，从多源数据内容相似度出发，综合考量图表征下多源数据的语义相似度和结构相似度，并在此基础上基于马尔科夫最佳决策获得原二分图结构中的top-k子图，进而在实现图结构优化的同时保证决策结果的相对最优，提高了检索的效率。与此同时，优化后的top-k子图结构与原二分图结构构成具有闭合环路的二部二分图，打破了原有图结构中静态性的限制，赋予了图结构动态演进的优势。

**1.4 本文组织结构**

本文主要用五个章节阐述面向知识图谱的多源数据图表征及检索优化方法研究工作，每个章节内容安排如下：

第一章：引言。主要介绍面向知识图谱的多源数据图表征及检索优化方法的研究背景和意义，根据现有研究内容，将研究问题划分为面向知识图谱的多源数据图表征问题以及图表征下多源数据的快速图检索问题两个子问题。针对子问题进一步详细阐述了多源数据融合以及图检索优化方法的研究现状，同时介绍了本文的研究内容和组织结构。

第二章：基础理论与相关技术。详细介绍了本课题所采用的基础理论和相关技术，主要包括知识图谱与图理论、马尔科夫理论以及相似性度量方法。

第三章：提出了基于马尔科夫决策过程的二分图表征方法。首先介绍了基于马尔科夫决策过程的图表征方法，然后在此基础上详细阐述了如何对该方法进行二分图表征优化，以及如何对优化后的图结构进行管理。

第四章：在第二章提出的图表征方法基础上，提出一种基于图相似性度量的二部二分图表征方法。首先介绍并定义了基于有限马尔科夫决策过程二分图度量算法，其次在此基础上构建top-k二分图表征方法，并进一步优化为二部二分图表征方法，然后讲解如何在该二部二分图上完成检索与动态演。

第五章：实验设计与分析。本章使用问卷型医疗数据和电子病历型医疗数据，分别对第三章提出的二分图表征方法和第四章中提出的二部二分图表征方法设计实验，并进行分析和评估。数据对本章方法进行实验分析与应用，并对分析应用结果进行评估。

第六章：总结与展望。对全文工作进行总结，并进一步讨论未来研究方向。

**第2章 基础理论与相关技术背景**

**2.1图的定义及相关关系**

知识图谱是表示客观世界实体以及实体间关系的语义网络，从其结构上看，知识图谱可以看作是一个大型多关系图，该图由顶点和边组成，顶点用于表示实体，实体包含内容与属性，边用于表示关系，关系展示实体之间的联系。基于此，图结构可以全面展示数据内部和数据之间的信息。与此同时，根据知识图谱与图的关系可以发现，知识图谱本质上是图的一个子集，图结构是知识图谱的具体表现形式。因此，要用图表示知识图谱，需要对图与图论有基本了解。基于此，本节将从图与二分图、图的表示以及图数据库三个方面介绍图的定义与内容、图的使用以及图的存储。

**2.1.1 图与二分图**

图论中，图（graph）并不等同于图像（image），它是指一种由顶点和边组成的数据结构，用于表示对象之间的关系，同时也是离散数据中的一个分支。一个图通常由点和边组成，因此，图可以定义为。其中，为点的集合，为边的集合。图根据边的性质可以划分为有向图和无向图。

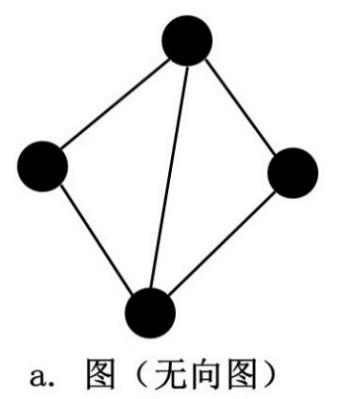
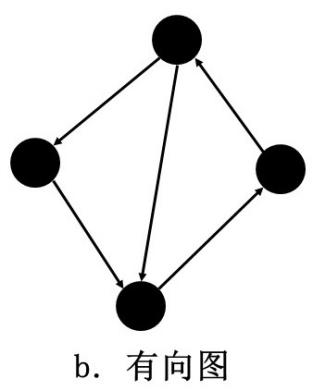
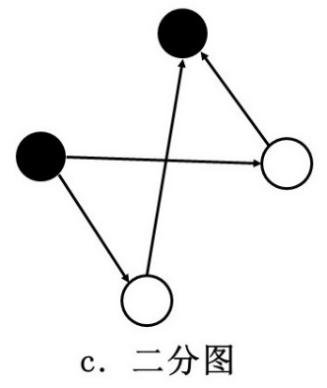
  

图2.1 图示例

图2.1中的子图a是一个简单的图结构，同时也是一个无向图结构。在这个图结构中，圆点为顶点，一般用于表示某个实体或某个事件，顶点和顶点之间的连线为边，同样表示实体和实体之间的联系。图2.1中的子图b为一个有向图结构。和无向图相比，有向图中的限制条件更多，同时也可以更好地表示实体或事件之间的前后关系或因果关系。

二分图是图的一个特殊分支。该结构将图中的顶点划分为两个互不相交的子集，两个子集之间的点相互连接，子集内部的点相互独立。图2.1中的子图c为一个二分图结构。图中黑色顶点属于一个子节点集，白色顶点属于另一个子节点集，黑色顶点至白色顶点之间以及白色顶点与黑色顶点之间通过边相连，黑色顶点之间无边相连，白色顶点情况亦然。基于上述描述可以得知，判断一个图结构是否为二分图，主要判断图中的节点是否可以分割为两个相互独立的点集。和普通图结构相比，二分图上的检索与匹配效率相对更高，因为与某一节点相连的下一节点必定在该节点所在子节点集之外的另一个子节点集中，从另一方面来看，二分图结构实际上可以**缩小匹配范围**。

**2.1.2 图的表达**

图常用的表示方法一般有邻接矩阵和邻接链表两种。对于一个有个顶点的图而言，邻接矩阵一般使用的矩阵表示图结构。若图为无向图，则矩阵对角线相等，且矩阵位置为1则表示两个点之间有边，若矩阵位置为0，则表示两个个之间无边相连。无向图邻接矩阵对称且唯一。对于有向图而言，邻接矩阵坐标中的值通常代表两点之间边的权重。和邻接链表相比，邻接矩阵可以更加快速的判断图的节点之间是否有边相连，但缺点在于无论图结构稀疏还是稠密，邻接矩阵需要一次性分配足够大的空间用于存储图结构，而当图为稀疏图时，会存在较大的空间浪费。

邻接链表在图表示方面更加灵活。该方法主要是对图中每一个顶点生成一组链表，链表的内容为该顶点以及与该顶点相连的其他顶点。和邻接矩阵相比，邻接链表的表示方法更加节省空间，它无需在存储结构时一次性分配大量空间，而只需要在表示过程中，一边遍历图结构一边分配。因此，当图较为稀疏，邻接链表可能是比邻接矩阵更好的表示方法。

**2.1.3 图数据库**

图数据库（Graph Database）是图结构管理中最常使用的工具之一。和关系型数据库不同的是，图数据库是一种非关系型数据库，主要使用图结构来表示、存储和查询数据。它可以通过点、边以及属性等明确且直观地展示出图结构中节点之间的关系，并在此基础上进行快速检索。因此，和关系型数据库相比，图数据库可以更加高效地表示或管理图结构。目前较为常见且流行的图数据库主要有以下几种：

（1）Neo4j图数据库

Neo4j是一个高度可扩展的本地图数据库，可以对数据网络进行存储、分析、映射等操作，从而发现数据内部的显性关系和隐性关系。除此之外，Neo4j图数据库支持智能实时应用程序、本地标签索引快速写入、符合索引快速读取以及多种语言驱动程序[53]。因此，Neo4j图数据库被广泛应用于实时推荐、欺诈识别、人工智能等多个领域[54]。虽然Neo4j数据库中的表示语言和关系型数据库相比较不直观，但由于其具有开源、轻量、访问速度快、用户生态完整以及高性能等特点，所以目前仍旧是最受欢迎的图数据库之一。

（2）HyperGraphDB

HyperGraphDB是一个用于存储超图的开源数据库，支持非阻塞并发写入与读取、图节点之间的N元高阶关系的建立与查询以及可扩展动态数据库模式。除此之外，HyperGraphDB还可以进行自定的索引编制与存储管理。目前该数据库常用于语义web项目和知识管理项目[55]。然而，HyperGraphDB的便携性相比于Neo4j图数据库而言较差，因为该数据库目前只支持32位的Linux系统。

（3）Titan图数据库

Titan是一个分布式图数据库，专注于紧凑的图形序列化，能够支持多个用户实时并发执行图操作，并且具有线性和弹性可扩展性，因此能够较好的满足不断增加的用户需求。除此之外，该数据库内部对大型图结构进行了存储与优化，并且可与多个大数据平台进行集成，从而能够支持全局图数据分析[56]。然而，Titan数据库的用户生态完整性和Neo4j数据库相比较差，且该数据库更新慢，其开源社区已处于无人维护状态。

**2.2 马尔科夫理论**

虽然图与知识图谱能够较好的表示数据和数据之间的联系，但却无法表示出数据之间的因果关系或逻辑关系。而要探究数据之间的前因后果，马尔科夫理论是很好的选择。马尔科夫理论的核心是马尔科夫链（Markov chain）[57]，该概念由数学家Andrey Andreyevich Markov提出，指的是在某个状态空间中，从某一个状态转换至另一个状态的随机过程[57]。该随机过程中，当前状态只能决定下一状态，下一状态只能由当前状态决定且与其他状态无关。在马尔科夫链中，称该种性质为“无记忆性”，也称为马尔科夫性。在马尔科夫性下，三个概念被逐步提出，分别是马尔科夫过程（Markov process）、马尔科夫奖励过程（Markov reward process）以及马尔科夫决策过程（Markov decision process），如图2.2所示，这三个概念用于描述状态空间中事件之间的关系。

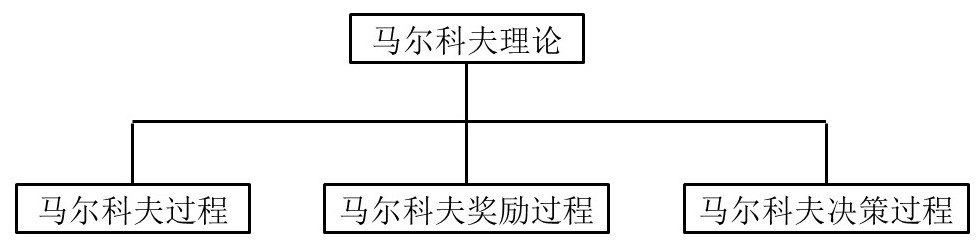


图2.2 马尔科夫理论示意图

**2.2.1 马尔科夫过程**

马尔科夫过程（MP）主要包括两个参数，分别是和。它是在马尔科夫性的基础上引入了状态转移概率的概念。其中，为智能体的有限状态集合，；为状态转移概率，指的是从当前状态跳转至下一状态的可能性，记为

因此，马尔科夫过程也可以记为。在马尔科夫过程中，仅考虑当前状态和下一状态之间的跳转以及跳转的概率，不考虑除这两个因素之外其他因素的影响。所以，该方法仅可以表示事件之间的简单关系，对复杂环境下事件的描述较不充分。

**2.2.2 马尔科夫奖励过程**

马尔科夫奖励过程（MRP）在马尔科夫过程的基础上提出奖励的概念。它认为从当前状态跳转至下一状态时，当前状态能够获得来自外部环境的奖励。该奖励能够随着跳转不断累积，且跳转次数越多，获得奖励的折扣越大。该折扣由衰减因子控制，也可以称为阻尼系数。该系数由用户定义，取值范围在0-1之间，取值越小，说明用户更在意眼前事件的奖励，取值越大，则说明未来奖励对用户更重要，且当取值越大时，马尔科夫奖励过程的收敛速度越快。基于上述概念，引入回报()来表示某个时刻的状态将具备的回报，它是马尔科夫奖励过程中从某一个状态开始采样直到终止状态时所有奖励的有衰减的之和：

在马尔科夫奖励过程中，使用值函数衡量每个状态的价值。值函数为奖励过程中回报的期望，公式如2.3所示。当状态值函数的值越高时，该状态的价值就越高，在整个奖励过程中越重要。

基于上述概念，马尔科夫奖励过程可以记为。相比于马尔科夫过程，马尔科夫奖励过程能够在一定程度上反映环境对事件的影响，且能够对状态的价值进行评定，所以它对事件的描述更为充分。

**2.2.3 马尔科夫决策过程**

在复杂环境下，事件的发生受多种因素的影响，仅用奖励对事件进行描述是不全面的。因此，马尔科夫决策过程（MDP）的概念应运而生。马尔科夫决策过程（MDP）的核心于1960年由Ronald Arthur Howard提出[58]，其中心思想是在具有马尔科夫性质的系统环境下模拟智能体可实现的策略与回报。其中，策略具有随机性。

马尔科夫决策过程涉及5个重要参数，分别是。其中，和马尔科夫过程以及马尔科夫奖励过程一样，表示智能体的有限状态集合，记。而则是智能体在环境中所能采取的一组动作，记；记为，它表示智能体在某一状态下采取动作s之后，转移到状态的概率。对于，可以描述为

还可以描述为

表示在状态s下采取动作a后转移到状态时所获得的奖励，可以描述为

记为，用户可以根据数据的事实际情况以及状态和动作的重要程度设置相应的值；与马尔科夫奖励过程中的含义相同，作用相似，用于控制马尔科夫决策过程的收敛速度。和马尔科夫奖励过程相同的是，马尔科夫决策过程中同样使用回报()来表示某个时刻的状态将具备的回报（公式2.2）。不同之处在于，马尔科夫决策过程中引入的是状态价值函数(Value function)的概念,记为：

它表示为已知当前的状态s，按照某种决策行动产生的长期回报期望，即状态价值函数是回报的期望。除此之外，由于马尔科夫决策过程比马尔科夫奖励过程多了“动作”的概念，因此，马尔科夫决策过程中还引入了动作价值函数，用于衡量动作对状态价值产生的影响。动作价值函数定义如下：

基于上述描述和定义，相较于马尔科夫过程以及马尔科夫奖励过程，马尔科夫决策过程能够在考虑多重因素的前提下，对状态空间内的事件进行更加充分的描述。这对数据表征工作的开展有着重要作用。

在马尔科夫决策过程中，最佳策略求解方法有策略迭代(Policy Iteration)和值迭代(Value Iteration)两种。其中，策略迭代是在决策未知的情况下，根据每次选择或转移所得到的奖励r学到的最优决策，而值迭代则是在已知决策和马尔科夫决策过程模型的情况下，根据决策获得最优值函数和最优策略。而在多数情况下，数据大多处于初始状态，此时决策未知，此时，值迭代可能是更合适的选择。值迭代下需要引入或定义几个参数。首先，回报()来表示某个时刻的状态将具备的回报，它是一个马尔科夫决策过程中从某一个状态开始采样直到终止状态时所有奖励的有衰减的之和:

从上面的式子可以看出，除非整个马尔科夫决策过程结束，否则无法获取所有的来计算出每个状态的回报。因此，需要引入一个状态价值函数(Value function)的概念,记为：

它表示为已知当前的状态，按照某种决策行动产生的长期回报期望，即状态价值函数是回报的期望。但价值函数只能用来衡量每一种状态的好坏，却没有考虑动作带来的价值影响，所以除了价值函数之外还需要引入动作价值函数：

根据、和的定义，可以将分别代入和进行推导，推导过程如下：

(2.12)

(2.13)

(2.14)

(2.15)

(2.16)

以上是状态价值函数的Bellman等式，此时时刻的状态和时刻的状态满足递推关系，同理可以动作价值函数的Bellman等式：

此时，将状态转移概率和奖励代入上述过程中得到的价值函数和Bellman等式，可以得到新的值函数的计算公式：

(2.18)

该公式在策略固定的情况下更新值函数直到值收敛，从另一个角度来讲就是为了更好地估计基于当前策略的价值以下最优策略。

在对一些参数进行基本的定义和描述之后，可以进行马尔科夫决策过程策略迭代。策略迭代方法分为两步，具体步骤如下：

步骤一：策略评估。使用Bellman等式评估当前策略下的马尔科夫决策过程的值函数，直到值函数稳定收敛。

步骤二：策略迭代，获得最佳决策。根据更新后的值函数更新每个状态下的策略直到策略稳定。采用的公式如下所示：

(2.19)

至此可以得到马尔科夫决策过程下每一个状态对应的最佳决策，即每个状态采取的最佳动作。在采取最佳动作之后，可以根据最大似然型的方法来选择下一个状态。此时所使用的决策是最佳决策，记最佳决策为。

**2.3 相似性度量方法**

通过马尔科夫理论，可以获得图与知识图谱中数据之间的关系以及图结构中的关键性结构，但在图结构和其他线性结构不同的是，当图结构较大时，图上的检索难度高，检索时间长。因此，缩小图检索范围，是提高图检索效率的方法之一。然而，缩小检索范围并不是盲目缩减图结构，而是应该在保留关键性结构的同时，尽可能给用户提供更多的可能性。其中，找到图结构中的关键性结构，马尔科夫理论可以提供较好的帮助，而如何基于关键性结构提炼出更多的可能性，相似性度量方法提供了解决思路。对于混合型文本数据而言，文本相似性度量中，一般从文本语义与文本结构两个方面进行考察。其中，语义相似度旨在衡量文本内容本身的相似性，而结构相似性则目的与考察文本内在逻辑，两者对文本相似性度量都有极为重要的意义。基于此，本节将从语义相似性度量方法和结构相似性度量方法两个方面进行阐述。

**2.3.1 语义相似性度量方法**

在语义相似度度量中，最常用的度量方法主要从词频和词向量两个方面入手。其中，基于词频的相似性度量方法中，较为常见的有词集模型、词袋模型、n-gram方法以及TF-IDF方法。这几种方法均是以词语之间相互独立为前提提出的，一般是通过收集文本中的词语构建词典，然后利用不同方式对文本进行切分，最后通过计数的方式统计不同词语在文本中的出现次数，根据词语出现的次数判断词语的重要程度，进而计算词语的相似性。这几种方法虽然计算方法简单，但却难以衡量相似词语之间的相似程度。

在基于词向量的相似性度量方法中，一般是通过使用如word2vec等方法将词语转换为词向量，然后通过距离衡量词向量之间的相似度[59]。词向量之间的距离越大，则表示两个词语之间的相似程度越低。在词向量距离的计算中，最常使用的距离有以下两种：

1）欧氏距离

欧氏距离也称为欧几里得距离，一般指n维空间中，两个点或两个向量之间的自然长度。计算公式定义如公式2.7所示，其中，u、u’代表两个向量，m为向量维度。

2）余弦距离

余弦距离实际上衡量的是两个向量之间夹角余弦值的大小，一般通过（1-余弦相似度）的方法计算余弦距离。其中，余弦相似度的计算公式如公式2.8所示，余弦距离计算公式如公式2.9所示。

两种距离计算方式的区别在于，余弦距离更关注向量方向上的相对差异，因此余弦距离反映的是向量之间的动态变化或动态趋势，而欧氏距离则更关注数值上的绝对差异，因此，当两个向量的表述标准一致时，欧氏距离可能会获得更加准确的度量结果。

**2.3.2 结构相似性度量方法**

结构相似性度量主要是指利用图拓扑结构信息衡量数据之间相似程度的普适性度量方法。在文本结构相似性度量中，通常是使用图论将文本的逻辑结构表示出来，然后使用结构相似性度量算法进行计算。其中，SimRank系列算法是较为常见的结构相似性度量算法之一。

SimRank是由Glen Jeh 和 Jennifer Widom在2002年提出的算法。该系列算法的主要做法是将数据抽象为一个二分图，该二分图由点集V与边集E组成，即SimRank二分图可以表示为），然后计算该二分图的结构相似性。在SimRank算法中，最为核心的定义如定义2.1所示。

**定义2.1** 在图G中，如果两个节点相似，则与这两个节点相连的节点相似；如果两个节点相连的节点相似，则这两个节点相似。

基于上述定义，若是图G中的两个节点，则这两点的相似度计算公式如下：

其中，分别代表和相连节点的集合，表示之间的相似度，c为常数。而的相似度可以用公式2.11计算。

其中，为常数，，分别代表和相连的节点集合，为，之间的相似度。根据上述公式可以发现，和通过来回迭代求解，通过来回迭代，可以得到SimRank二分图中节点之间的结构相似度。

根据语义相似度度量方法和结构相似性度量方法，我们可以从繁杂的图结构中提炼出关键性结构，从而实现检索图结构，缩小检索范围，提高图检索效率的目的。

**2.4 本章小结**

本章从三个部分介绍了本文使用的基础理论和相关技术背景，三个部分分别为图的定义及相关关系、马尔科夫理论以及图相似性度量方法。其中，图的定义及相关关系为本文所提方法的理论基础，马尔科夫理论和相似性度量方法用于对理论基础的进一步优化，三者相辅相成、环环相扣，为后续方法的提出打下基础。

**第3章 基于马尔科夫决策过程的二分图表征方法**

在特征级多源数据融合表征中，现有的工作大多是通过特征向量的方式来完成，向量的表示方法则一般依托于数组或矩阵来实现。在数组和矩阵中，通常使用数值进行填充。这种表示方法的优点在于简单、易操作且方便计算，但问题同样非常明显。问题主要表现在以下三个方面：数据特征离散、数据内部关联被忽视、可解释性差。针对上述问题，基于马尔科夫决策过程提出了二分图表征方法。和传统方法不同的是，本方法使用图结构而非向量的方式表示多源数据。主要思路是依据马尔科夫决策过程提取数据中的时序关系和逻辑关系，然后根据时序关系和逻辑关系中所涉及到的实体的特点对实体进行分类。最后依据提取到的时序关系、逻辑关系以及实体分类结果构建MDP（马尔科夫决策过程）二分图。实验结果表明，该方法不仅可以有效表征多源数据的显性关系和隐性关系，与其他模型（如图卷积神经网络模型）结合使用，可以实现较好的预测效果。

**3.1 基于马尔科夫决策过程的图表征**

根据马尔科夫决策过程的描述和定义，本文在马尔科夫决策过程的启发下，将马尔科夫决策过程（MDP）图记为，描述。该图具有马尔可夫性，即该图中的所有状态s均包含了与该状态相关的所有信息，只要当前状态s可知，那么该状态s之前的所有历史信息在后续的决策中不再需要，当前状态s所具备的信息可以决定未来。与此同时，根据马尔科夫决策过程的性质可知，基于马尔科夫决策过程的图表征具有时序性。

此时，该图表征方法所构建的图M为普通图结构。如图3.1所示，图3.1是基于马尔科夫决策过程的图表征的一个简单案例，该案例使用第5章5.2小节的牙科病例数据进行描述，数据集在5.2中详细描述。图3.1中，实线点“脓肿”为某一状态s，该状态经过动作a（面部检查）后，有0.5的概率转移至下一个状态s’，即绿色点“口腔溃疡”。在这个过程中能够获得0.2的R(Reward)。在此图结构上，要获得上述信息，用户最多需要经过5次检索才可以实现检索目的，即检索所有的绿色点。

但随着数据量的增加，图M的规模与复杂度也随之增加，实现检索目的所需要的检索次数也会随之增加。但图M的规模达到某一阶段时，图检索的效率极低，这样不利于后续的工作的开展。因此，本文在此基础上提出了基于马尔科夫决策过程的二分图表征方法。

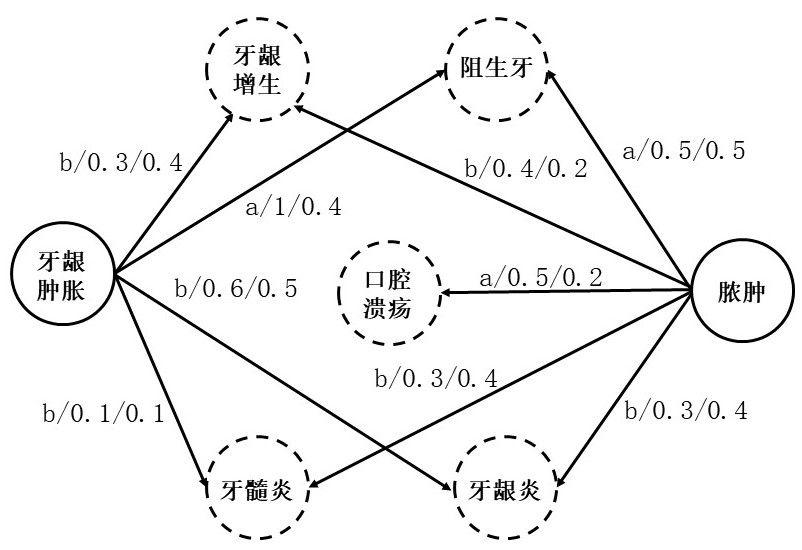


图3.1 基于马尔科夫决策过程的图表征的简单案例

实线点为状态点，虚线点为下一个状态点；a、b均表示动作；

（a/X/Y）表示（动作/状态转移概率P/奖励R）

**3.2 基于马尔科夫决策过程的二分图表征**

根据马尔科夫决策过程（MDP）图的描述，可以对基于马尔科夫决策过程的二分图做如下定义。定义内容如定义3.1所示。

**定义3.1** 基于马尔科夫决策过程的二分图是一个满足的图模型。它是一个具有状态节点集（）和动作节点集（）的有向二分图。其中，

1）状态节点集（）中包含若干子状态节点集。子状态节点集之间有序而内部无序，子节点集之间的顺序符合数据时序顺序和逻辑顺序。状态节点集（）中的状态节点记为，即。状态节点集（）可以作以下描述：

(3.1)

2) 动作节点集（）中包含若干子动作节点集。子动作节点集之间有序而内部无序，子动作集之间的顺序符合数据时序顺序和逻辑顺序。动作节点集（）中的动作节点记为，即。动作节点集（）可以作以下描述：

(3.2)

3) 是从状态节点到动作节点的边的集合，称状态节点到动作节点的有向边为决策边。为决策边集合。决策定义为在某一给定状态s下某一动作a的概率分布，即

4) 是动作节点到状态节点的边的集合，称动作节点到状态节点的有向边为转移边。为转移边集合。

5） 为某一时刻下，从状态节点经过动作节点之后到达下一个状态节点的概率。在基于马尔科夫决策过程的二分图中，为转移边的特征之一。。

6） 是某一时刻下，状态节点经过动作节点之后到达下一个状态节点时获得的奖励。在基于马尔科夫决策过程的二分图中，为转移边的特征值。。的具体取值由用户根据数据特性决定。

经过上述定义，可以将MDP图转化为MDP二分图。在MDP二分图中，保留了MDP图中状态节点，并按照状态节点之间的时序关系以及内在的逻辑关系对状态节点进行了分类。分类之后的状态节点集之间有序，而状态节点集内部无序。此外，和MDP图不同的是，MDP二分图增加了动作节点的概念。此时二分图中的“二分”，“一分”为状态，“一分”为动作。转移边和决策边的设置使得图结构更加清晰明了的同时，增加了图结构所蕴含的信息量，减少了MDP二分图中的检索次数，实现了类似剪枝的效果。

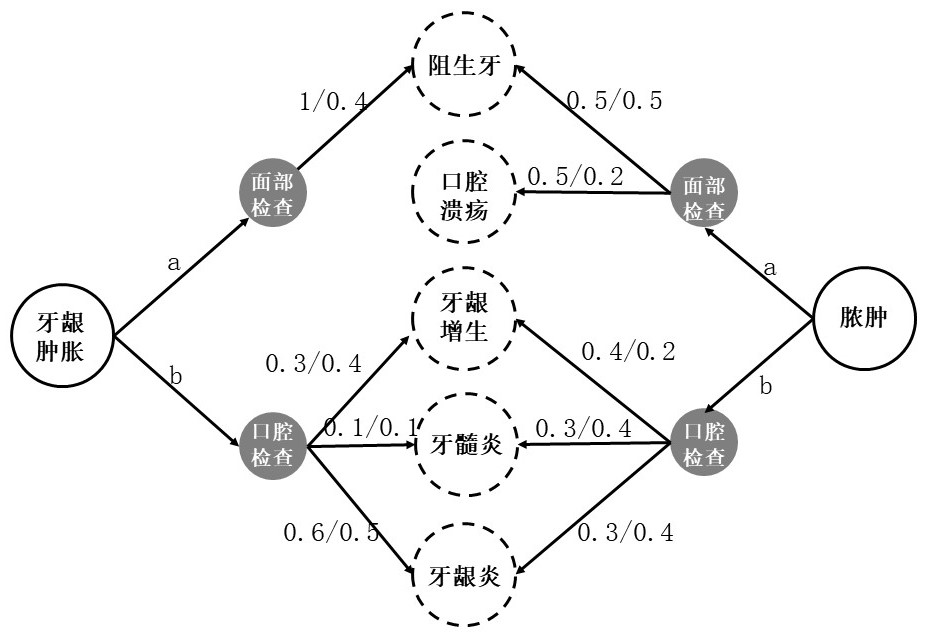


图3.2 基于马尔科夫决策过程的二分图表征简单案例

（X/Y）表示（状态转移概率/奖励r）

图3.2是图3.1经过基于马尔科夫决策过程的二分图表征之后的结果。此时MDP图转换为了MDP二分图。二分图中的节点为2类，一类是状态节点（实线点和虚线点），另一类是动作节点（灰色点）。在此图上完成3.2节中的检索工作最多需要检索4次。其中，在第一次检索时，当发现检索结果为动作b而非动作a后，可以直接省略对分支b的相关检索。此时相当于完成了对MDP二分图的剪枝。当图结构越庞大且复杂时，相比于MDP图而言，此工作检索效率越高。图3.1转换为图3.2具体步骤如下：

步骤一：提取状态节点之间的时序关系和逻辑关系。按照上述关系将状态节点划分到对应的子状态节点集中。图3.1中绿色的状态节点在实线状态节点之后发生，所以若实线节点的集合记为，那么绿色节点的集合记为。

步骤二：提取动作节点之间的时序关系和逻辑关系。操作如同步骤一。

步骤三：提取状态节点和动作节点之间的逻辑关系。完成状态节点和动作节点之间的有向边（决策边）的连接。

步骤四：提取动作节点和状态节点之间的逻辑关系。完成动作节点和状态节点之间的有向边（转移边）的连接。之后根据公式计算对应的状态转移概率，并赋予转移边设定的奖励。

在时间复杂度方面，本章方法和定义可以得知，本章方法在给定基于马尔科夫决策过程二分图定义的情况下，该方法主要耗时在于节点的找寻以及节点之间关系的构建，时间复杂度为，其中，是状态节点集中状态节点的数量，是动作节点集中动作节点的数量。

**3.3 面向知识图谱的基于马尔科夫决策过程的二分图表征**

通过上述操作，可以根据数据的特点完成基于马尔科夫决策过程的二分图的构建工作。下一步，图结构的存储和管理是所面临的问题之一，这是因为图的特性使得图结构的存储和管理相对其他结构（如线性结果等）而言更为复杂。此时，做好图结构的存储和管理是提高基于马尔科夫决策过程的二分图表征方法可操作性的重点内容。而又因为图结构与知识图谱有着天然的契合，所以基于马尔科夫决策过程的二分图能够很轻易的转化为知识图谱并加以使用。转化为知识图谱之后，该二分图能够轻松的表征各种原始数据类型（结构化数据、半结构化数据、非结构化数据），并能够基于知识图谱更加顺利地完成后续的操作。

在知识图谱的管理中，图数据库是主要的存储和管理工具。图数据库（如Neo4j等）能够将数据存储在图网络上而非表上，并且能够轻松的完成对图本身的各项操作。在图数据的管理之下，基于马尔科夫决策过程的二分图表征可以更加丰富。这是因为在此二分图中，使用者通过图数据库可以在节点和边中表达多重特征，并且特征的形式不再仅仅局限于向量。这样能够使得数据中所传达出来的信息得到更加充分的使用，进而提高下游工作的性能（如预测准确率等）。在本章中，我们首先通过上述基于马尔科夫决策的二分图表征方法得到有关二分图表征的相关信息和数据，然后在图数据库中构建二分图结构，最后把得到的信息和数据按照对应关系赋给而数据库中的二分图结构，进而完成整个二分图表征工作。此时获得的二分图实际上也是一个知识图谱。

除此之外，和传统向量表征方法相比，基于马尔科夫决策过程的二分图表征方法在图数据库的管理以及知识图谱的作用下，能够表征或获得更多数据之间的关系，这也从侧面增加了基于马尔科夫决策过程的二分图表征方法的可解释性。

**3.4 本章小结**

本章提出了基于马尔科夫决策过程的二分图表征法。在马尔科夫决策过程的基础上提出了基于马尔科夫决策过程的图表征方法，并进一步将图表征方法优化为二分图表征方法。该方法通过提取数据之间的时序关系和逻辑关系，将数据分为状态和动作，并将状态和动作之间的关系分为决策和转移，从而实现对数据的有效表征。在二分图结构的存储和管理上，将基于马尔科夫决策过程的二分图结构与图数据库联系起来，用图数据库管理和操作二分图结构及特征，实现了多源数据及其关系的合并表征与使用，图数据库的使用也为图结构的可扩展性提供了机会。我们使用问卷型医疗数据对该方法进行实验分析与应用，具体实验结果将在第5章的5.1节中展示。

然而，当数据量较大时，所构建的基于马尔科夫决策过程的二分图较为庞大，此时不论是在此二分图上进行检索，还是对二分图进行存储，都对时间和空间具有较大要求。除此之外，此时的二分图为静态图谱，如果后期知识进行更新，那么图谱的适用性就会降低。因此，本文在后续的工作中提出基于相似性度量的二部二分图表征方法，该方法在对图表征和图检索方法进行优化的同时，在一定程度上解决了知识图谱的静态性问题。

**第4章** **基于图相似性度量的二部二分图表征方法**

对于多源异构数据特征离散、数据内部关联性差以及传统表征方法可解释性低的问题，本文提出的基于马尔科夫决策过程的图表征方法已经能够在一定程度上环节上述问题。但基于马尔科夫决策过程的图表征方法所构建的图结构为普通图结构，在此结构上进行检索等工作难度较大。所以本文紧接着提出了基于马尔科夫决策过程的二分图表征方法。该方法在解决上述问题中，相比于基于马尔科夫决策过程的图表征方法能够取得更好的解决效果。但不可否认的是，基于马尔科夫决策过程的二分图表征方法在图构建以及图检索过程中仍旧需要耗费大量的时间和空间。尤其是在利用图数据库管理该二分图结构时，这个问题尤为突出，具体表现为内存溢出以及Neo4j图数据库（网页版）页面崩溃等问题。因此，基于马尔科夫决策过程的二分图表征方法更适合于数据量较小时的数据表征工作。除此之外，基于马尔科夫决策过程的二分图表征方法所构建的图谱为静态图谱，这一点使得图谱在后续的长期使用中受到限制，并且根据反馈，该方法的可解释性仍旧存在一定的可提升空间。因此，本章在前一章工作的基础上，利用马尔科夫决策过程（MDP）最佳决策算法和图相似性度量算法对方法进行优化改进，提出了基于图相似性度量的二部二分图表征方法在结构较大时检索困难的问题，本文提出基于图相似性度量的二部二分图算法，从而解决上一章所提方法中的不足。

**4.1 有限马尔科夫决策过程二分图相似性度量**

从相似性度量方法上看，相似性度量可以分为语义相似性度量和结构相似性度量。而在基于马尔科夫决策过程的二分图中，结构相似性度量可以分为状态相似性度量和动作相似性度量。因此，本节将从基于马尔科夫决策过程二分图相似性定义入手，对有限马尔科夫决策二分图的语义相似性和结构相似性进行介绍，最后得出有限马尔科夫决策二分图相似性度量方法。

**4.1.1 基于马尔科夫决策过程二分图的相似性**

在文本的相关研究中，主要通过语义相似性和结构相似性来度量文本的相似性[60][61]。本文所提出的表征方法主要是将多源文本的相关数据通过图的方式进行融合表征，此时，文本相似性通过图相似性来度量。在本文的表征方法中，图总共包含四个元素，分别是状态点、动作点、转移边和决策边。其中，文本内容或语义通过点来表示，文本中或文本之间的时序关系或逻辑关系通过点边之间的联系来表示。基于此，本章将文本相似性中的语义相似性和结构相似性延伸到基于马尔科夫决策过程二分图的相似性之中。其中，基于马尔科夫决策过程二分图的语义相似性做出以下定义：

**定义4.1** 基于马尔科夫决策过程二分图的语义相似性为二分图中状态节点之间的语义相似度，记为。的定义如下：

其中，为状态节点和之间的语义距离，在本章中采用余弦距离公式（公式2.21）进行计算。

根据第3章对基于马尔科夫决策过程的二分图的定义可知，此二分图中状态节点集和动作节点集二分。因此，基于马尔科夫决策过程的二分图中，状态节点的邻居始终是动作节点，动作节点的邻居始终是状态节点。基于此，本章引用SimRank中的思想对基于马尔科夫决策过程二分图的结构相似性做出以下定义[61]：

**定义4.2** 在基于马尔科夫决策过程的二分图结构中，如果两个节点的邻居结构相似，那么这两个节点的结构相似。

与此同时，可以对基于马尔科夫决策过程二分图的相似性做出如下定义：

**定义4.3** 在基于马尔科夫决策过程的二分图结构中，如果两个节点的邻居相似，那么这两个节点的相似。

定义4.2和定义4.3均通过递归实现，因此该方法仅可用在基于马尔科夫决策过程的有限二分图中。除此之外，根据该定义可以发现，在基于马尔科夫决策过程的二分图中，状态节点的相似性与动作节点的相似性紧密相关，反之亦然。因此，接下来本文将对有限马尔科夫决策过程二分图的状态相似性和动作相似性进行描述和定义。本文中分别将有限马尔科夫决策过程二分图的状态相似度和动作相似度记为和。

**4.1.2 二分图状态相似性度量**

根据上一小结的描述可知，在有限马尔科夫决策过程二分图中，状态相似性的度量本质上是在计算与状态相连的邻居——动作的相似度。对于状态节点集中的任意两个状态节点、，记它们的邻居节点集分别为和，这两个邻居节点集同时也是动作节点集。当邻居点集之间的距离越小时，两个动作节点之间的相似性越高。本节中引用Hausdorff距离计算邻居点集和之间的距离，记Hausdorff距离为，Hausdorff距离的计算公式如下：

(4.2)

其中，为中的子节点，为的子节点。为和中所有节点之间距离的最小值。

得到Hausdorff距离之后，就能够进一步得到邻居点集和之间的相似度，该相似度实际上就是与其相连的状态节点和之间的相似度。本章中相似度的计算公式如下：

其中，是状态相似度影响因子，，该参数类似于马尔科夫决策过程中的衰减因子。和一样，参数的值由使用者定义，可被用于调整有限马尔科夫决策过程二分图中的状态相似性大小。

综上，有限马尔科夫决策过程二分图的状态相似度度量流程图如图4.1所示。

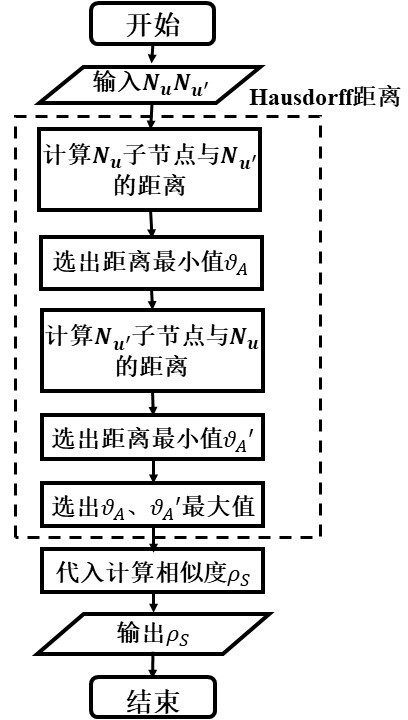


图4.1 状态相似度度量流程图

状态相似度计算步骤具体如下：

步骤一：计算中的子节点和中所有节点之间的距离，并选出所有距离中的最小值；

步骤二：交换和以及对应参数，重复步骤一；

步骤三：对比两次距离计算结果，选出两个最小值中的最大值，即Hausdorff距离，记为；

步骤四：将上一步骤中得到的代入公式4.3，计算有限马尔科夫决策过程二分图的状态相似度。

**4.1.3 二分图动作相似性度量**

根据上一小节中的步骤和公式可以得知，有限马尔科夫决策过程二分图的动作相似度是计算有限马尔科夫决策过程二分图的状态相似度的关键参数。在有限马尔科夫决策过程二分图的动作相似度度量中，本节引入EMD距离用于计算[62]。EMD距离的公式定义如下：

其中，对于动作节点集中的任意两个动作节点、，记它们的邻居节点集分别为和，这两个邻居节点集同时也是状态节点集，状态节点。为状态节点语义相似度。为状态节点之间的语义距离。在EMD中视作代价或者权重。在本章中，该参数的作用等同于状态转移概率的作用。因此，本章中将此参数看作是状态转移概率，并将对应状态转移概率的值赋给该公式中的参数。参数主要在函数发挥作用。本章对函数作以下定义：

**定义4.3** 对于任意两个点,当均属于状态节点集时，

对于任意一个点属于状态节点集时，

其中，表示动作节点所连状态节点采取动作后转移至状态节点时的概率；对于任意一个点属于状态节点集时，

其中，表示动作节点所连状态节点采取动作后转移至状态节点时的概率。

在有限马尔科夫决策过程二分图中还涉及到了奖励这一参数，本章将在有限马尔科夫决策过程二分图的动作相似度度量中加入对这一参数的使用，具体使用方法是计算奖励r之间的距离，计算公式定义如下：

其中，表示采取动作的奖励的回报。

基于上述公式和定义，本章对动作节点的相似度计算公式定义如下：

其中，是动作相似度影响因子，，该参数类的作用类似于，参数的值由使用者定义，可被用于调整有限马尔科夫决策过程二分图中的动作相似性大小。

**4.1.4 有限马尔科夫决策过程二分图相似性度量方法**

综合上述章节的定义和公式，可以给出有限马尔科夫决策过程二分图相似性度量方法，度量方法流程如图4.2所示。

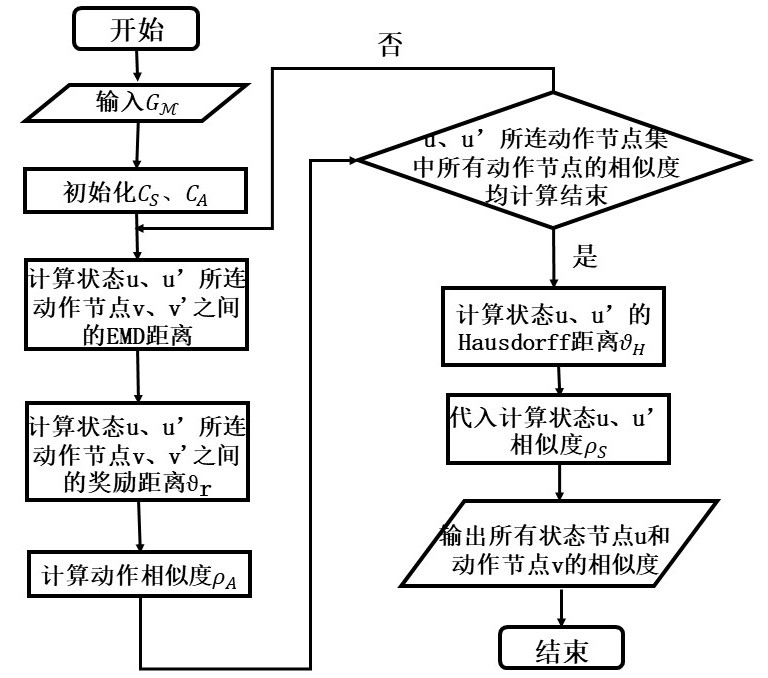


图4.2 有限马尔科夫决策过程二分图相似性度量方法流程图

具体度量步骤如下所示：

步骤一：输入基于马尔科夫决策二分图；

步骤二：初始化状态相似度影响因子和动作相似度影响因子，其中，；

步骤三：根据公式4.4计算状态、所连动作节点之间的EMD距离；

步骤四：根据公式4.8计算状态、所连动作节点之间的奖励距离；

步骤五：将步骤三中的计算结果代入公式4.9中，计算状态,所连动作节点之间的动作相似度；

步骤六：重复步骤三至步骤五，直到状态、所连动作节点集中所有动作节点的相似度均计算结束；

步骤七：将步骤六的最终结果代入公式4.2，计算状态、的Hausdorff距离；

步骤八：将步骤七中获得的Hausdorff距离代入公式4.3，计算状态节点、的相似度；

步骤九：重复步骤三至步骤八，直到所有状态节点集和动作节点集中所有点均被计算完成；

输出：所有状态节点和动作节点的相似度与。

**4.2 基于最佳决策及相似性度量的top-k二分图表征**

在本节中，我们基于马尔科夫决策过程最佳决策以及有限马尔科夫决策过程二分图相似性度量方法提出top-k二分图表征方法，用于提取基于马尔科夫决策过程二分图的top-k子结构，从而实现简化图结构，缩小检索范围的目的。该方法的核心思想是通过有限马尔科夫决策过程二分图相似性度量方法，计算与马尔科夫决策过程最佳决策最为相似的top-k决策，然后根据top-k决策，得到与其相连的k个最相似状态，最后以此类推完成整个二分图的构建与表征。该二分图结构构建流程如图4.3所示。

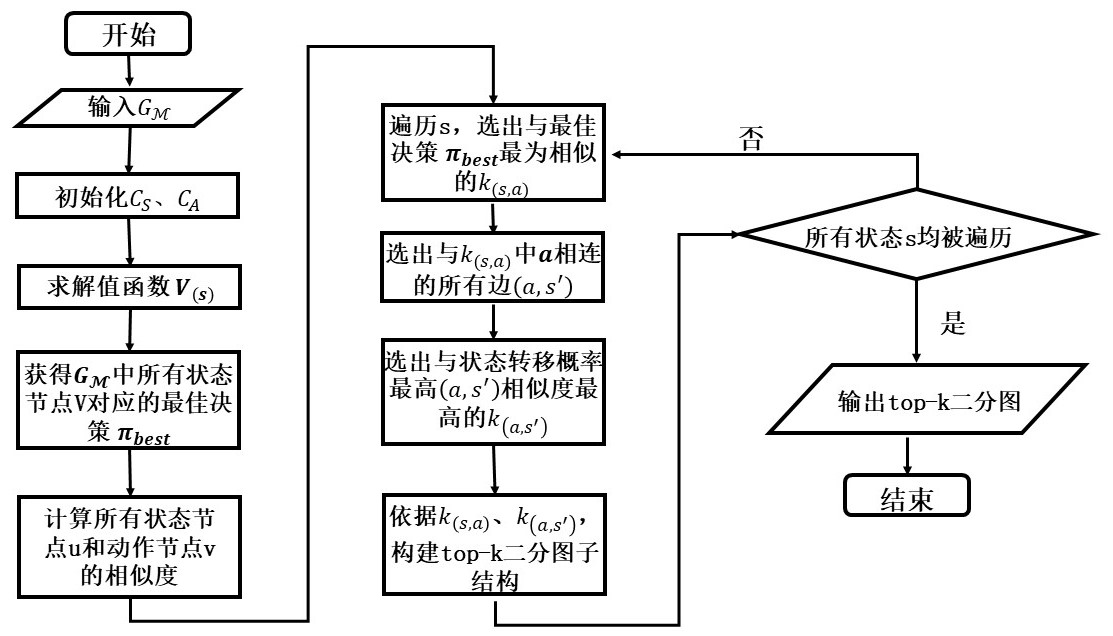


图4.3 基于最佳决策及相似性度量的top-k二分图结构构建流程图

基于最佳决策及相似性度量的top-k二分图具体实现步骤如下：

步骤一：输入有限马尔科夫决策过程二分图

步骤二：初始化状态相似度影响因子和动作相似度影响因子，其中，；

步骤三：依据马尔科夫决策过程的策略迭代方法求解值函数，获得有限马尔科夫决策过程二分图 中所有状态节点对应的最佳决策；

步骤四：依据有限马尔科夫决策过程二分图相似性度量方法，计算所有状态节点和动作节点的相似度；

步骤五：依据最佳决策，对比状态及与其相连的动作的相似度，选出与最佳决策最为相似的k个对，记最为相似的最佳决策个对为；

步骤六：遍历，对比每组（）中动作和与相连的状态相似度，选出与状态转移概率最高的相似度最高的个对，记为；

步骤七：依据、，选择对应的状态节点和动作节点,构建top-k二分图子结构；

步骤八：重复步骤五至步骤七，直到所有状态均被遍历，top-k二分图结构构建完成。

基于上述步骤，可以完成top-k二分图的构建。此时的top-k二分图可以看作是基于马尔科夫决策过程二分图的一个子结构，和原二分图图结构相比，top-k二分图更加的简洁且高效。因为，此top-k二分图中既保留了最优决策及与其最相似的k个决策，还保留了决策后与最可能实现的状态最相似的个状态。既在一定程度上为使用者提供了更优、更可能的决策方法和可能性，又大大简化了图结构，降低了图检索的时间。其中，步骤五和步骤六是整个构建过程中的重点，我们称这两步为top-k的构建。这两步主要完成了基于马尔科夫决策过程的二分图结构优化。

**4.3 基于图相似性度量的二部二分图表征及其动态演进**

在上述的章节中，我们提出了基于最佳决策及相似性度量的top-k二分图表征方法，该方法虽然可以优化图结构、提高图检索效率，但利用该方法构建的图结构是一个静态图结构。随着数据量的不断增加和变化，该图所提供的检索结果可能会和预期产生一定的差异，甚至是误差。这一点是上述方法中需要注意且解决的问题之一。因此，在本节中，我们调整图结构的表征方法，使得所构建的图结构具有动态演进的特性，从而提高方法的可用性。

**4.3.1 基于最佳决策和图相似性度量的二部二分图表征**

在基于马尔科夫决策过程的top-k二分图表征中可以发现，基于最佳决策及相似性度量的top-k二分图结构实际上是基于马尔科夫决策过程二分图的子图。它是从基于马尔科夫决策过程的二分图中，挑选出与最佳决策边最相似和与最可能的转移边最相似的边，以及与这些边所连的点，构建的图结构。基于此，本小结将此top-k二分图和原二分图（基于马尔科夫决策过程二分图）结合起来，构建一个二部二分图结构。该二部二分图如图4.4所示。在此结构中，原二分图为底层图结构，该图结构依据马尔科夫决策过程和图相似性度量生成顶层的top-k二分图。之后，数据可以在顶层的top-k二分图中进行匹配或者检索，匹配或检索结果直接向外输出。

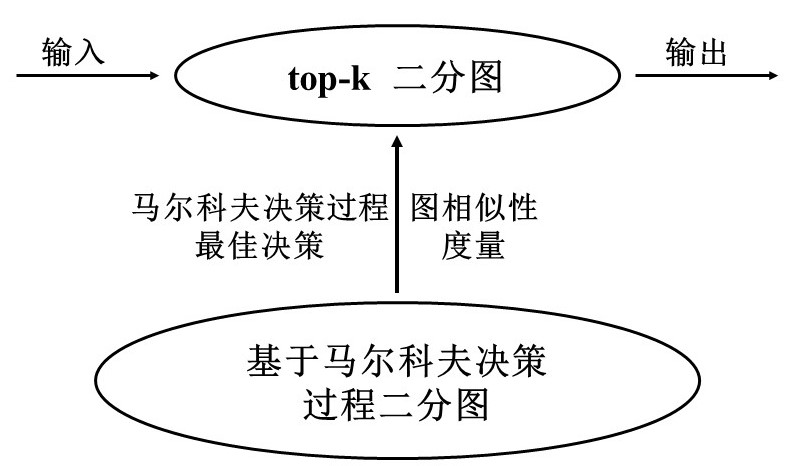


图4.4 二部二分图结构示意图

此时，二部二分图结构构建完成，但此二部二分图仍旧为静态图结构，不具备动态演进的功能。因此，还需对该二部二分图结构进行进一步的优化与调整。

在时间复杂度方面，本章方法构建过程中耗费时间最多之处在于图相似度的计算，而该份工作中Hausdorff距离和EMD距离的计算是其中的关键。其中Hausdorff距离的时间复杂度为，其中，，二分图状态节点的最大出度。EMD距离的时间复杂度是，其中，，二分图动作节点的最大入度。因此，本章算法构建时的时间复杂度为。虽然该方法时间复杂度比二分图表征方法的时间复杂度更高，但该方法可以有效缩小检索范围以及在图数据库上的开销，基于该方法在图数据库上构建图结构的时间复杂度为。其中为二分图中的取值，为二分图中状态节点的数量，为二分图中动作节点的数量，，。

**4.3.2 二部二分图的动态演进**

基于最佳决策和图相似性度量二部二分图的动态演进，主要是通过向底层二分图反馈匹配信息的方式实现。实现过程如图3.5所示。二部二分图动态演进过程具体实现步骤如下：

步骤一：基于马尔科夫决策过程二分图生成top-k二分图；

步骤二：输入待匹配数据，在top-k二分图中完成数据的检索和匹配；

步骤三： 输出和反馈。top-k二分图中的匹配与检索结果一方面向外数据给用户，另一方面反馈至底层二分图；

步骤四：底层二分图的更新。底层二分图依据顶层二分图反馈的结果，对马尔科夫决策过程中的状态转移概率进行更新。

步骤五：新top-k二分图的生成与更新。状态转移概率更新后，会进一步的影响马尔科夫决策过程最佳决策、状态节点相似度以及动作节点相似度。更新相应参数数据，依据更新后的数据生成新的top-k二分图，并将原上层top-k二分图更新为新的top-k二分图；

步骤六：重复步骤二至步骤五，直至没有待匹配数据的输入。

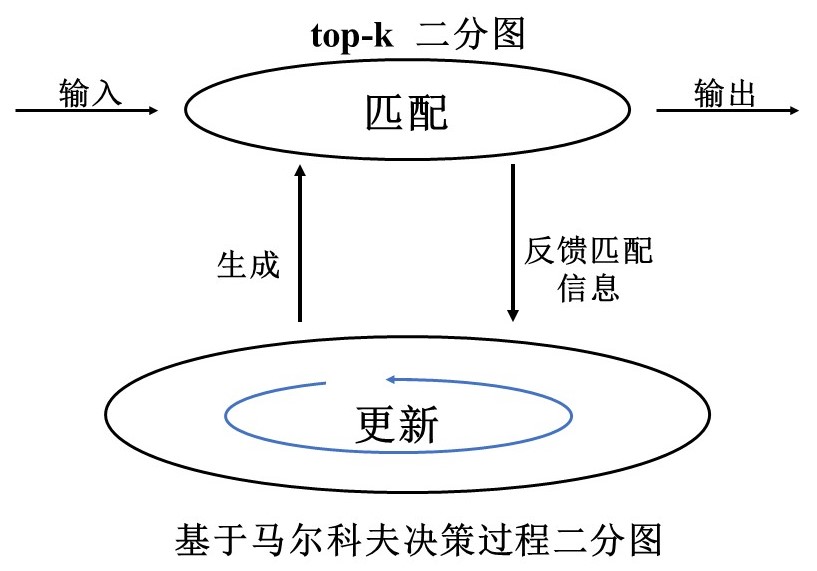


图4.5 二部二分图动态演进示意图

考虑到图操作的难度与复杂性以及本章方法的时间复杂度和空间复杂度，在经过一系列实验和操作之后，建议使用者根据需求更新所构建二部二分图。使用者可以采用批数据的处理方式，匹配一批或多批数据，并对匹配结果进行汇总，然后再向下反馈匹配信息。因为根据状态转移概率的计算公式可以发现，当原数据量总量较大时，新数据中小规模的状态转移次数的变化，不会对状态转移概率值产生较大的改变。除此之外，减少二部二分图动态演进的次数可以减少实验的开销。

**4.4 本章小结**

本章介绍了基于图相似性度量的二部二分图表征方法。受图相似性度量算法的启发，将有限马尔科夫图相似性度量算法与马尔科夫决策过程最佳决策算法结合，应用于基于马尔科夫决策过程二分图表征方法中。在基于马尔科夫决策过程二分图表征方法的基础上，首先利用马尔科夫决策过程最佳决策算法求解值函数，获得马尔科夫决策过程最佳决策，然后应用有限马尔科夫图相似性度量算法计算与最佳决策top-k相似的决策边，以及与最大转移概率边top-k相似的转移边。提取上述所获内容，重构二分图，并与原二分图形成二部二分图结构。该二部二分图中顶层图结构的检索结果向下反馈至底层二分图，底层二分图利用反馈信息进行迭代更新，然后将更新结果向上反馈至顶层二分图，实现顶层二分图以及整个二部二分图的动态更新。

**第5章 实验设计与分析**

**5.1 二分图表征方法实验分析与应用**

基于马尔科夫决策过程的二分图表征方法实际上是数据和下游算法（如图检索、图分类等）之间的桥梁。通过该方法能够使得下游算法的可操作性和可解释性得到更好的提升。而产前抑郁是一项公共卫生问题，这一点在高、中、低收入国家中都有充分的不良结果证明[63]。怀孕期间，抑郁症的患病率在高收入国家和地区为7-15%，在中低收入国家中为19-25%[64]。妊娠期抑郁症不仅影响孕妇的工作、运动、饮食和药物使用，还影响低出生体重（LBW）、早产 （PTB）等分娩结局和子代的宫内生长限制（IUGR）[65][66]。基于此，本章在基于马尔科夫决策过程的二分图表征方法下使用产前孕妇心理状态数据完成对孕妇分娩结局（婴儿出生结局）的预测，并通过实验结果证明基于马尔科夫决策过程的二分图表征方法的有效性与可解释性。除此之外，本节将通过与传统向量表征方法下的经典算法（决策树和随机森林）的预测结果进行对比来体现基于马尔科夫决策过程二分图表征方法的优势。

**5.1.1 实验数据描述**

（1）研究人群

本实验使用的数据来自中国孕妇队列研究（CPWCS）[67]。该项目综合考虑了许多因素，例如地理，经济和人口。它选择分布在中国15个省的24家医院作为研究地点并收集数据。CPWCS已在临床试验注册网站ClinicalTrails.gov（NCT03403534）上注册。该项目的研究人群是中国的孕妇。孕妇的数据来自两个方面，即医生一方和孕妇一方。其中，医生方面包括孕妇的基线数据、新生儿问卷（单胎和双胎）和分娩结果问卷。孕妇包括四个部分：孕早期问卷、孕中期问卷、孕晚期问卷和产后六周问卷。所有信息均以问卷或电子表格的形式存在。每个阶段的数据包含十多个方面，例如孕妇的基本信息，环境安全，运动，饮食习惯，心理和睡眠。数据涵盖范围广泛，来源复杂且形式多样，属多源数据。与此同时，由于数据收集时间跨度非常大，跨越了孕妇的整个怀孕期间以及产后六周，因此不可避免地会出现数据丢失和填充错误等问题。其中，数据缺失和遗漏问题尤为明显。这些问题进一步导致了数据质量低和数据关系模糊的问题。这给数据处理带来了很大的困难。为了更好地进行后续数据处理工作，本实验对数据质量进行了基本评估。在筛选了丢失和填写错误（填写结果超过问卷阈值）等数据之后，本实验计算了不同问卷数据的完整性。结果如图5.1所示。图5.1中灰色部分数据来自于产妇端，斜线部分数据来自于医生端。根据图片可以发现，随着孕期的增加，孕妇数据的完整性逐渐降低。至于来自医生的数据，与“结果”（分娩结局和新生儿结局）有关的数据相对不完整。最终，本实验选择了妊娠期数据与分娩结局（或新生儿出生结局）数据完整的患者，并排除了数据超过问卷阈值的患者。本章研究中总共包括345名患者的数据。

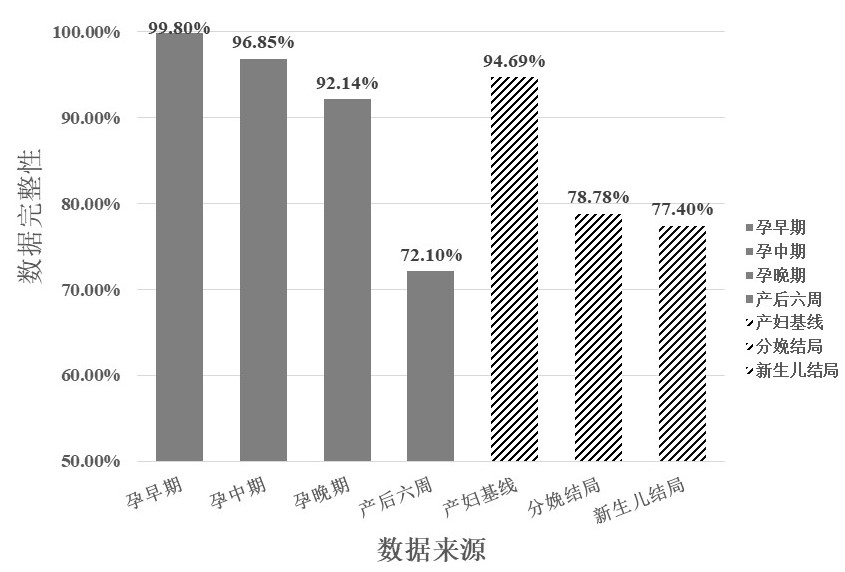


图5.1 数据完整性评估图

（2）预测变量

为了聚焦研究的重点内容，本实验从CPWCS中提取**爱丁堡妊娠抑郁量表（EPDS）数据**和结局数据（分娩结局或新生儿出生结局）。EPDS数据反映了孕妇在早期，中期和晚期的精神状态。

其中，EPDS数据包括10个问题以及孕妇对相应问题的回答。每个问题都有4级得分（0-3分）。EPDS的总分为0-30分。在传统的研究中，一般认为EPDS得分越高者的精神状态越差[68]。在中国地区的医学研究中，一般将EPDS作为产前抑郁的标准[69]。除此之外，因为孕妇每个阶段的EPDS是独立、顺序完成的，所以阶段内孕妇的心理应该是稳定的。即一个调查问卷反映的是该时期孕妇的某一个心理状态。其次，因为孕妇的妊娠周期长，所以不同妊娠阶段孕妇的心理状况可能会发生改变，即阶段间孕妇的心理状态可能是变化的。也就是说，每个孕妇的EPDS数据可能是阶段内稳定、阶段间变化。将每一位孕妇每一个阶段的每一个EPDS问答看作是离散事件的话，那么整个EPDS数据集可以看作是离散事件的动态系统状态空间。对于离散事件的动态系统状态空间的研究，经典的方法之一就是马尔科夫决策过程。

出生状况在所使用的数据集中包括8种状态：正常、早产、过期产、低出生体重婴儿、巨大儿、出生缺陷、死胎死产和其他。在这种情况下，本实验将生产状态分为两大类，即正常生产和异常生产。其中，异常产生包括除了正常状态之外的7个状态。

综上，EPDS数据为模型中的预测变量。本实验用基于马尔科夫决策过程的二分图表征方法表征预测变量，结合图卷积模型预测出生结局。

**5.1.2 实验过程描述**

实验按照图5.2所示流程进行。整个实验过程主要可以分为数据预处理过程、图结果的构建、图卷积与图分类三个部分。以下对这三个部分进行详细描述。

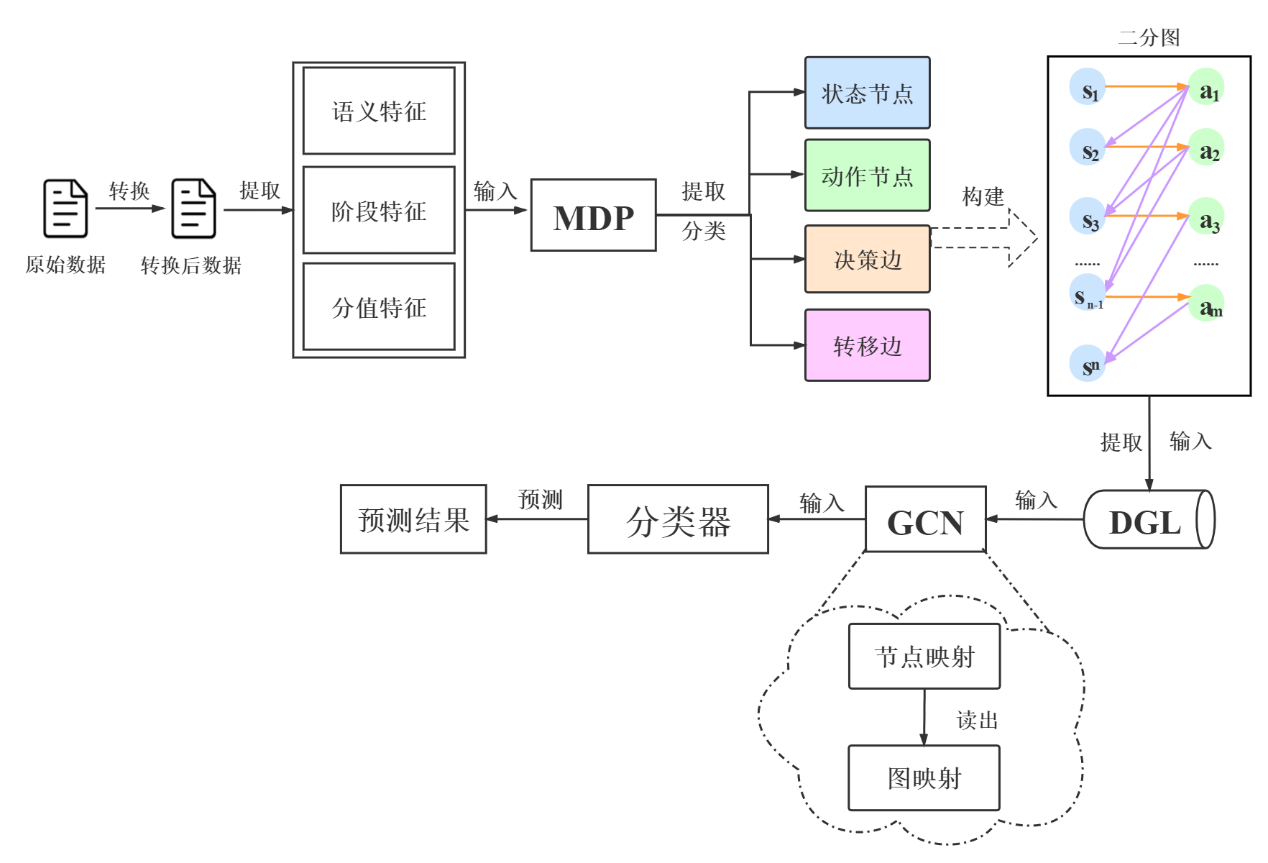


图5.2 实验流程图

（1）数据预处理过程

数据处理过程主要包括三方面内容，分别是问卷选项语义转换、句向量与特征提取以及状态转移概率的计算。

1）问卷选项语义转换

在传统EPDS问卷中，选项的最低分为0分，最高分为3分。选项与选项之间的分值设定是等距的，问卷问题与问题之间没有特殊的关联，即相互独立。但从自然语言语义方面出发，选项与选项之间的距离不可能严格相等。因为自然语言是具有情感性的，而且自然语言的表达具有可变性[70]。除此之外，问题内容的设置可能会在语义方面隐性的体现提问者的某些偏好。因此，为了更好的利用EPDS问卷所表露出来的信息，我们根据EPDS中的问题和选项所表达的含义，将问题与选项融合，每个选项还原成对应问题的完整回答。图5.3是还原过程及还原结果的示例。EPDS共有10个问题条目，每个问题有4个回答，经过问卷选项语义还原，能够将原EPDS转换成40个语义完整的问题短句。转换后的问题短句从自然语言角度上更容易被人类理解，写在后续的语义分析中，模型能够获得更多的信息。

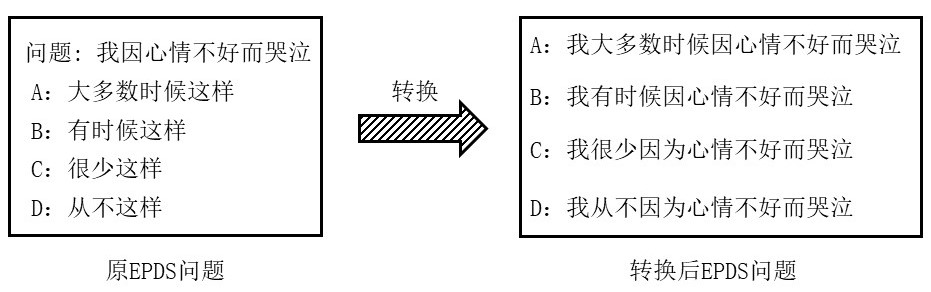


图5.3 问卷选项语义还原示例

2）句向量与特征提取

由于本次实验的数据是小样本数据，问卷的语义相对而言比较简单。传统句向量的获取方式采用词向量的方式求加权平均，这种方法更适合与大样本的复杂文本数据。本次实验的数据并不适合这种方法。因此，考虑到本次实验数据的特性，在实验中引入Bert预训练模型[71]。Bert预训练模型的全称是Bidirectional Encoder Representation from Transformers。它是Google于2018年提出的一个使用Transformer的双向编码器[72]。该编码器主要是利用大规模无标注语料进行训练，通过将不同长度的句子进行预训练模型编码，从而将其映射到一个固定长度的向量，最后获得包含丰富且较为准确语义信息的语义表示[72]。该语义表示在某些自然语言处理任务中微调之后能够直接应用于下游的自然语言处理任务[73]。

我们将上一步骤中得到的40个短句分别输入Bert模型，获得每个短句的句向量。之后，将句向量作为对应短语的特征。此时可以得到句向量特征矩阵。除此之外，为了更好的利用数据，在这个阶段中我们还将采集数据的阶段特征（孕期阶段）和分值特征（EPDS对应短句分值）。这些特征将和语义特征一起用于丰富图的节点特征和边特征，进而给图分类工作提供更好的支持。

3）状态转移概率计算

在本次实验中，将EPDS的短句语义部分和分值部分分别看作是状态和动作。孕妇每选择一个EPDS的选项可以获得该选项对应的语义，然后根据该语义能够采取查阅EPDS分值表并获得对应分值这一动作，最后能够在这一基础上进入下一个EPDS问题。这个过程实际上完成了一次状态转移。因此依据基于马尔科夫决策过程的二分图表征方法的定义，本次实验收集每一个孕妇在时刻选择某一选项的条件下在（）时刻选择另一个选项的概率，并将其记为状态转移概率。该概率可以反映这批数据中孕妇心理状态可能的变化趋势。在图结构构建阶段，本实验将把概率作为特征赋给对应的转移边。

（2）图结构的构建

在做完上述准备工作之后，可以根据需求构建基于马尔科夫决策过程的二分图结构。构建二分图的方式有很多种，既可以选择将所有的状态节点和动作节点融合在一张二分图上，然后在该图完成预测，也可以针对每一个孕妇构建一个二分图，然后采用图分类的方法完成预测。本实验选择采用图分类的方式完成工作。而为了使图结构的操作和管理更顺利地进行，本实验引入了Neo4j图数据库和DGL框架[53][74]。其中，Neo4j作为高性能NOSQL图形数据库，能够将数据存储在图网络上而非表上，并且能够轻松的完成对图本身的各项操作。而DGL作为图神经网络计算框架，能够帮助图结构更好的应用在下游的图神经网络算法中[75]。本实验中图结构构建的具体步骤如下：

步骤一：利用Neo4j面向孕妇EPDS数据构建基于马尔科夫决策过程二分图，并将上述步骤中得到的语义信息、分值信息、阶段信息、状态转移概率等分别作为状态节点特征、动作节点特征、边特征赋给对应的点和边。此时的二分图实际上也是一个小规模的知识图谱；

步骤二：利用DGL框架构建多图结构。本实验利用DGL框架构建了345个线性有向图。每个有向图有60个节点，59条边，节点之间顺序相连。本实验用二元组（）来记录图结构，其中，为图，为图标签。图用来表示，为的头结点集，为的尾结点集。顺序相连，连接路径为图的边。此时图和图标签的信息为空，即二元组（）为空；

步骤三：在Neo4j知识图谱上查询每个孕妇的数据，并将其赋给DGL中的线性有向图。赋值后的有向图变为基于马尔科夫决策过程的二分图，每个二分图实际上是知识图谱中的一个子图，也可以看作是知识图谱中的一条路径。本实验将每个孕妇的分娩结局作为在DGL框架中有向图的图标签。

（3）图卷积与图分类

本实验在对345名孕妇的EPDS数据进行二分图表征之后，基于图卷积神经网络（GCN）完成图分类。整个过程主要包括以下三部分内容，分别是图卷积、读出和分类。其中，图卷积和读出操作将在图卷积神经网络中完成。整个图卷积和图分类操作将在DGL框架中完成。

1）图卷积

本实验将上一步中得到的基于二分图输入到图卷积层中。图卷积过程采用的模型为公式5.1。其中，为l层的节点特征，为第l层的权重，为第l层的截距。本次实验将l层的节点特征乘以的权重，然后对计算的结果应用线性变换和ReLU激活函数。最后将节点特征更新为激活函数计算后的结果。本次实验使用了两层图卷积层，因为实验所使用的数据集较小，且在很多GCN相关的研究中表示2层GCN已经能够得到比较好的结果[76]。

2）读出

在读出操作中，输入的是图中所有节点的表示，输出的是整个图的表示。实验利用DGL框架所提供的读出函数接口，对图中所有节点表示取平均值，并将其作为图的表示。计算公式为公式5.2。

3）分类

将读出层中获得的图表示输入分类器进行分类和预测。在这个过程中使用Softmax计算每个类别的概率，并使用向后传播来更新参数。本实验在使用Softmax之前，对读出层获得的图表示做了一次线性变换，用于获得每一类在Softmax之前的logits。

**5.1.3 实验结果分析**

本节共做了两组实验对比。其中，第一组是在仅具备EPDS分值特征时，特征向量表征方法和基于马尔科夫决策过程的二分图表征方法的预测结果对比；第二组是基于马尔科夫决策过程的二分图表征方法下，不同特征对预测结果的影响对比。

（1）第一组实验结果对比

本组实验中共使用了三个模型，分别是决策树模型、随机森林模型以及图卷积神经网络模型。其中，决策树和随机森林在表征数据时使用的是特征向量表征方法，图卷积神经网络模型在表征数据时使用的是基于马尔科夫决策过程的二分图表征方法。三个实验均只使用EPDS分值特征。EPDS分值是判断孕妇心理状态的重要指标，一般情况下，EPDS分值越高代表该阶段孕妇心理状态越差。在以往的研究中，分值是判断孕妇心理状态的重要特征。部分研究中通常只利用EPDS的分值对数据进行分类和预测。

基于此，本部分实验以分值作为分类的唯一特征模拟EPDS评分过程，尝试不同的表征方法下孕妇生产情况预测结果准确性的影响。其中，决策树模型依据EPDS的分值构建一个树形结构，然后在该树形结构上完成决策过程[77]。其决策过程类似于EPDS问卷的填写过程。而随机森林则是在数据上构建一组决策树，然后对每个决策树的预测取平均[78]。与单个决策树模型相比，随机森林模型降低了过拟合的概率。本组实验的实验结果如表5.1所示。

表5.1 分值特征条件下不同表征方法预测结果对比

|  |  |
| --- | --- |
| **方法** | **准确率** |
| 特征向量表征方法（决策树） | 0.788 |
| 特征向量表征方法（随机森林） | 0.817 |
| 基于马尔科夫决策过程的二分图  表征方法（GCN） | 0.834 |

根据上表可以发现，在本组实验条件下，决策树模型预测的准确率仅有0.788。随机森林的预测准确率略高于决策树模型，但也只达到了0.817。在仅使用EPDS分值信息的情况下，我们的方法同样略微优于决策树模型和随机森林模型，准确率基本维持在0.834以上。从数据表征方式上来看，**决策树和随机森林在表征数据时使用的是特征向量表征方法，而我们的方法使用的是图表征方法**。这两种方法在使用单个特征（分值特征）时，性能并没有显著差异。从侧面看，该结果还反映出仅EPDS评分可能不足以预测产妇的生产情况。

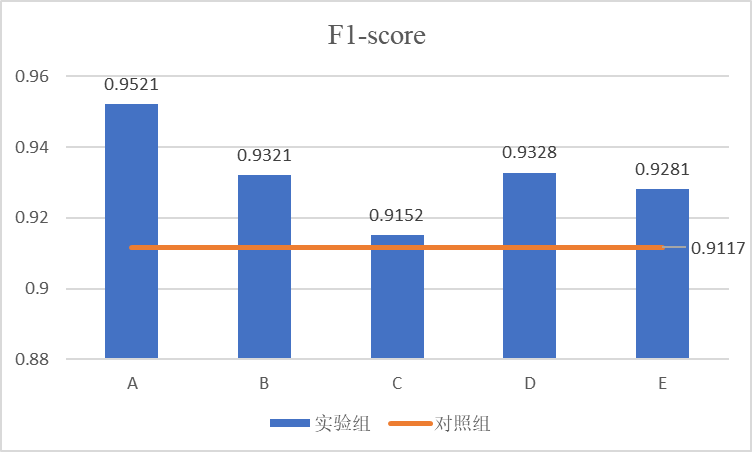
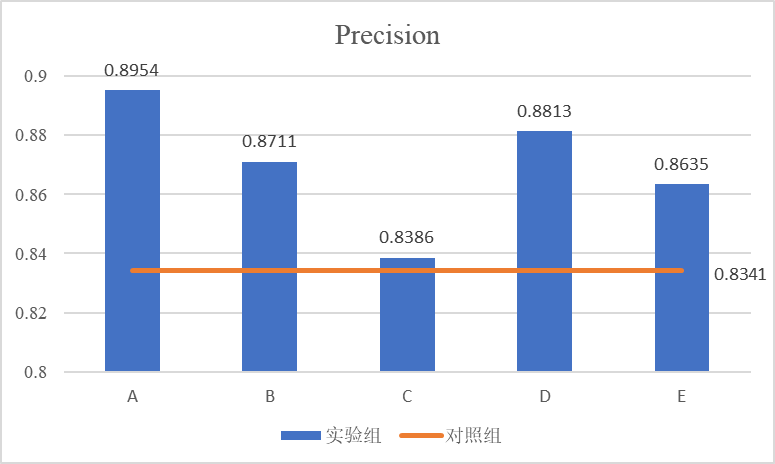
（2）第二组实验结果对比

在特征条件不仅仅只有分值时，我们在基于马尔科夫决策过程的二分图表征的前提下进行了预测。表5.2列出了不同特征条件下的实验结果。

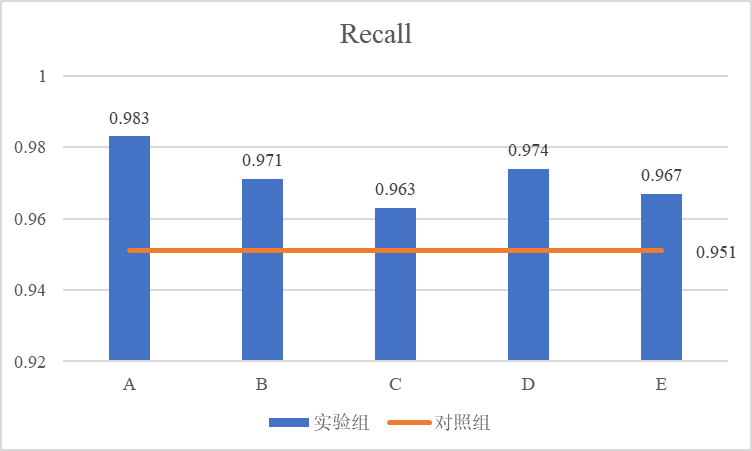
表5.2 不同特征条件下二分图表征方法实验结果对比

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **表现** | A | B | C | D | E | F |
| Recall | 0.983 | 0.971 | 0.963 | 0.974 | 0.967 | 0.951 |
| Precision | 0.8954 | 0.8711 | 0.8386 | 0.8813 | 0.8635 | 0.8341 |
| F1-score | 0.9521 | 0.9321 | 0.9152 | 0.9328 | 0.9281 | 0.9117 |

本组实验一共做了6个（A.B.C.D.E.F）实验。其中，实验A是在使用全部特征条件的情况下得出的实验结果。即实验A使用了语义特征、阶段特征（孕期阶段的标志）、分值特征（EPDS得分）以及边特征（状态转移概率）。除实验A以外的5组实验，相比于实验A而言，均缺少了部分特征条件。例如，实验B缺少了边特征，实验C缺少了语义特征，实验D缺少了阶段特征，实验E缺少了分值特征。而实验F则仅仅含有分值特征。此处的实验F等同于第一组对比实验（和特征向量表示法的对比实验）中的第3个实验。将实验F看作为对照组，可以得到图5.3所示对比结果。



a. Precision 对比图 b. F1-score对比图



c. Recall对比图

图5.4 实验结果对比图

根据图5.4可以得知，当具有所有特征条件时，实验结果相对最优。此时（实验A）可以获得最高的准确率和F1-score，分别为0.8954和0.9521。当缺少阶段特征时，实验结果仅次于实验A的结果。当实验缺少语义特征（实验C）时，实验结果在前5组实验中的表现最差。准确率仅有0.8386，F1-score仅有0.9152。这也可以体现出充分利用EPDS语义信息对孕妇分娩结局预测的重要性。和对照组相比，这五组实验结果均优于对照组，由此也可以看出，仅凭传统EPDS分值难以获得较好的预测结果，通过图表征的形式对EPDS问卷信息进行充分表征的方法在新生儿出生结局预测上非常重要。

综上所述，在分析如EPDS等问卷型数据时，不应仅仅只使用离散的评分数据，还应该考虑数据本身包含的语义信息。当语义信息混合且质量不高时，可以考虑使用图表征的形式表示数据，而非传统的特征向量表示方法。因为图表征的方式不仅可以同时考虑多方面特征，还可以充分表达数据之间的内部关系，从而获得更好的分析和预测结果。

**5.1.4 可解释性评估**

在方法的可解释性评估方面，本实验选取了20名医师对该方法的可解释性进行评估。20名医师中既包括初级医师和高级医师，也包括跨学科医师和专门做数据分析相关工作的医师。其中，一些初级医师和跨学科医师表示“能够在一定程度上理解这个方法”和“比其他人工智能方法更容易理解一些”。而高级医师，和专门做数据分析相关工作的医师则对方法中的一些细节存在疑问，认为“虽然比传统人工智能方法更好理解一些，但希望细节部分能够解释的更加清楚”。以下是部分医师对该方法的评论：

* 我是一名妇产科医生，我经常给我的患者使用EPDS。这种方法可以帮助我们更全面地了解孕妇的真实心理状态，并及时给予指导或治疗。
* 它的预测结果能够给我提供很好的参考。通过这个方法能够很好的了解孕妇妊娠期的心理波动。这能够帮助我更多的关注到孕妇的心理状态。
* 是一个不错的方法。以前的人工智能方法很难让人理解结果产生的原因。这个方法能够让我清晰的看到不同心理状态在决策中的权重。图形的表示方法也更容易理解。
* 虽然比之前的一些方法好，但是可理解性仍旧有提升的空间。预测结果可能需要用更多的数据、做更多的实验会更有说服力。

**5.2 二部二分图表征方法实验分析与应用**

基于相似性度量二部二分图表征法是基于马尔科夫决策过程二分图表征法的改进与优化。它在基于马尔科夫决策过程二分图表征法的基础上，利用马尔科夫决策过程最佳决策以及图相似性度量算法，**提取二分图结构中的top-k决策路径，从而在保证原二分图优势的同时实现对图结构的精简与优化**。与传统向量表征方法相比，基于相似性度量二部二分图表征方法在与知识图谱等工具结合使用时，能够更加清晰、透明、可解释地体现数据内部和数据之间的显性关系与隐性关系。这对分析混合型文本数据非常有帮助，电子病历型医疗数据是典型的混合型文本数据。医疗诊断的早期研究中广泛使用贝叶斯定理分类法[79]。该方法在处理患者病例数据时，将病例数据中的每个属性或症状视为独立离散变量，依据统计与概率等方法推测疾病的发生,忽略了属性与症状之间的关联性。这一点不仅与口腔诊断的实际情况相悖，而且会导致诊断推理出现不必要的偏差[80]。除此之外，该方法仅适用于解决小规模数据问题，且方法性能受数据集规模影响较大。针对上述问题，近年来出现了利用自然语言处理技术或神经网络技术挖掘并分析医疗数据的方法[81][82]。这些方法大多针对数据语义本身进行研究，且可解释性差。而基于知识图谱的分析方法虽然在一定程度上缓解了以往方法可解释性差的问题，却又存在图谱构建时间长、检索效率低等现象[83]。相比之下，我们的方法不仅可以对数据之间的相互关系形式化描述，生成贴近人类思维方式以及医生诊断逻辑的语义网络，提升诊断推理过程的可解释性与透明性，还可以解决传统知识图谱中的构建与检索问题。综上，本节选取电子病历型口腔诊断数据，利用本章提出的基于图相似性度量二部二分图表征方法，面向知识图谱对数据进行推理与预测，从而验证所提方法的有效性。

**5.2.1 实验数据描述**

本实验使用的数据为电子病历型口腔诊断数据（即牙科病例数据），数据来源于合作医院的牙科部门，由于收集时间、存储位置等因素影响，该数据集属于多源数据且为混合型文本数据。数据共包括5066个牙科病例。每个病例由七部分组成，分别是主诉、现病史、既往史和其他病史、查体、辅助检查、诊断、处理。表5.3为数据集的数据分布。根据数据分布可以发现，数据集中存在着不同程度数据缺失。其中，现病史与既往史和其他病史两个部分数据缺失最多，主诉部分数据几乎无缺失。辅助检查的缺失表示部分患者病情可能较轻，不需要采取过多的检查方法就可以完成治疗。处理结果的缺失则在一定程度上表示可能有患者未完成整个治疗过程。除此之外，数据分布表也体现出实验所用数据整体质量不高，需要对数据采取预处理操作。

表5.3 牙科病例数据分布

|  |  |
| --- | --- |
| **组成部分** | **数据量（/条）** |
| 主诉 | 5066 |
| 现病史 | 1451 |
| 既往史和其他病史 | 1401 |
| 查体 | 4718 |
| 辅助检查 | 3262 |
| 诊断 | 4665 |
| 处理 | 4435 |

图5.5为本次使用数据中的一个病例样本。根据图片可以发现，单个牙科病例样本实际上是一次牙科就诊过程，符合牙医的诊断逻辑。文本七个部分之间具有一定的时序关系的和逻辑关系。病例七个组成部分中，前六个部分可以看作是推理与分析过程，第七个部分“处理”可以看作是推理与分析结果，即前六个部分为最后一部分服务。本次实验目的是利用基于图相似性度量二部二分图表征方法推理分析上述牙科病例数据，模拟牙医诊断过程，根据分析结果为患者提供科学合理的处理方法。

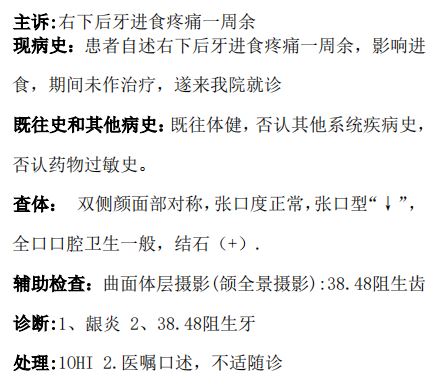


图5.5 牙科病例数据样本示意图

**5.2.2 实验过程描述**

实验流程如图5.6所示。整个实验过程主要包括数据预处理过程、二部二分图的构建以及动态知识图谱的检索与更新三个部分。以下对这个三个部分进行详细描述。

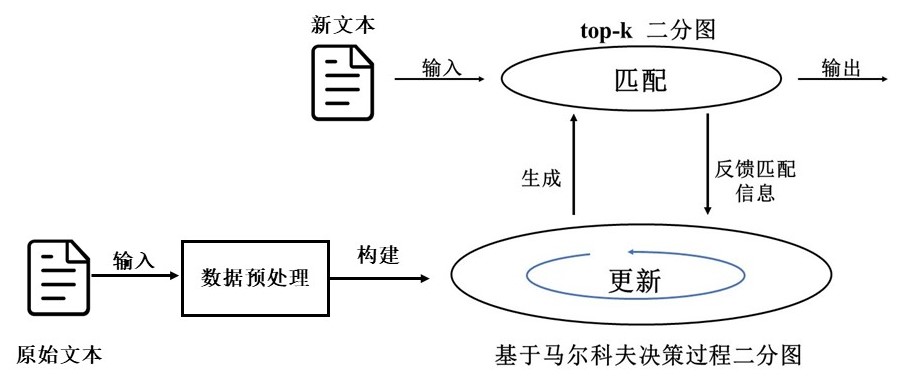


图5.6 实验流程图

（1）数据预处理过程

本项目使用的牙科病例数据属于混合型文本数据。针对该类型数据，需要进行语义分析和语义提取工作，从而将数据从混合型转换成结构型。其中，由于人工标注的时间成本较高且牙科病例数据之间具有较高的相似性，所以本次实验采用小样本标注+Bert预训练模型的方式减少人工标注的工作量，并以此为基础设计整个数据预处理流程。数据预处理流程图如图5.7所示。

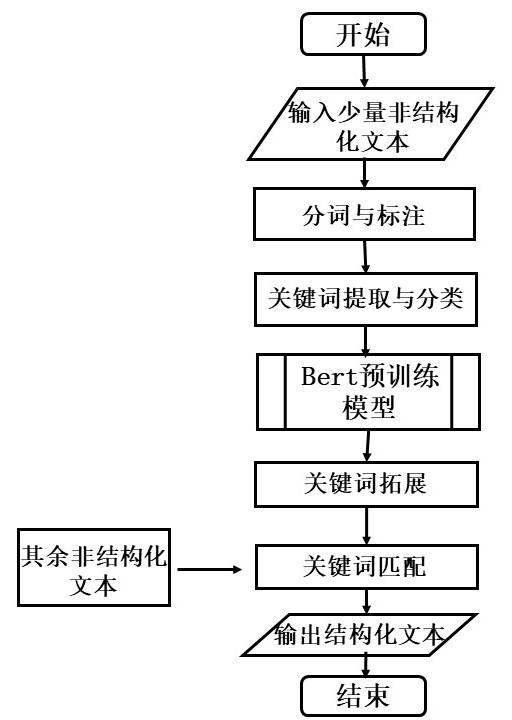


图5.7 数据预处理流程图

整个数据预处理流程主要可以分为以下几个部分：

1）分词与标注

本实验使用jieba分词器进行分词操作。jieba分词是一款中文开源分词包，具有高性能、高准确率、可扩展性等特点[84]。使用该分词器的主要原因是因为实验数据集中存在大量医学专用术语，传统中文分词工具在该数据集上的分词效果并不理想，而jieba分词支持自定义词表，所以可以很好的解决这一问题。本实验将医学专用词汇和常见药品名整合成词典作为分词器的补充词表，从而提高分词的准确率。本实验从数据集中随机选取10%（507个）的数据样本进行分词，并在分词之后对分词结果使用Brat文本标注工具进行标注[85]。

2）关键词提取与分类

基于图相似性度量的二部二分图表征方法需要将数据分为状态与动作两个部分，因此，针对这一点，实验将数据的7个部分进行重新划分，将其大致分为5大类别。其中，主诉、现病史、既往史和其他病史中的部位关键词为第一类；主诉、现病史、既往史和其他病史中的疾病类关键词为第二类；查体、辅助检查中的检查项目关键词为第三类；查体、辅助检查以及诊断中的结果关键词为第四类；处理中的方法关键词为第五类。这五类中，一、三、五为状态，二、四为动作。实验依据该划分对关键词进行提取和分类。

3）关键词拓展

实验使用Bert预训练模型，通过利用预训练中的词向量和词向量相似度概念对提取到的关键词进行拓展与延伸，从而获得更加丰富且完整的关键词数据集，以此减少人工标注关键词的工作量。其中，实验将词向量相似度阈值设为0.9。Bert预训练模型是由谷歌开发的用于预训练语言表示的自然语言处理模型。由于该模型规模较为庞大，所以实验操作中实际使用的是腾讯人工智能实验室开源的Bert-as-service服务，它允许用户以调用服务的方式使用 Bert模型而不需要关注模型实现细节[86]。实验将分类后的关键词按类别分别输入该服务中，实现关键词数据集的拓展。

表5.4 拓展后五类关键词数量情况

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 第一类 | 第二类 | 第三类 | 第四类 | 第五类 |
| 189 | 380 | 210 | 106 | 191 |

关键词拓展之后，可以获得五类关键词的内容与数量，该数量会对后续k值的选取产生一定的影响。表5.4中展示了拓展后五类关键词的数量情况。

4)关键词匹配

在使用Bert预训练模型获得拓展后关键词数据集后，将未标记的混合型文本与五类关键词进行匹配。实验中使用的匹配算法为Flashtext算法。

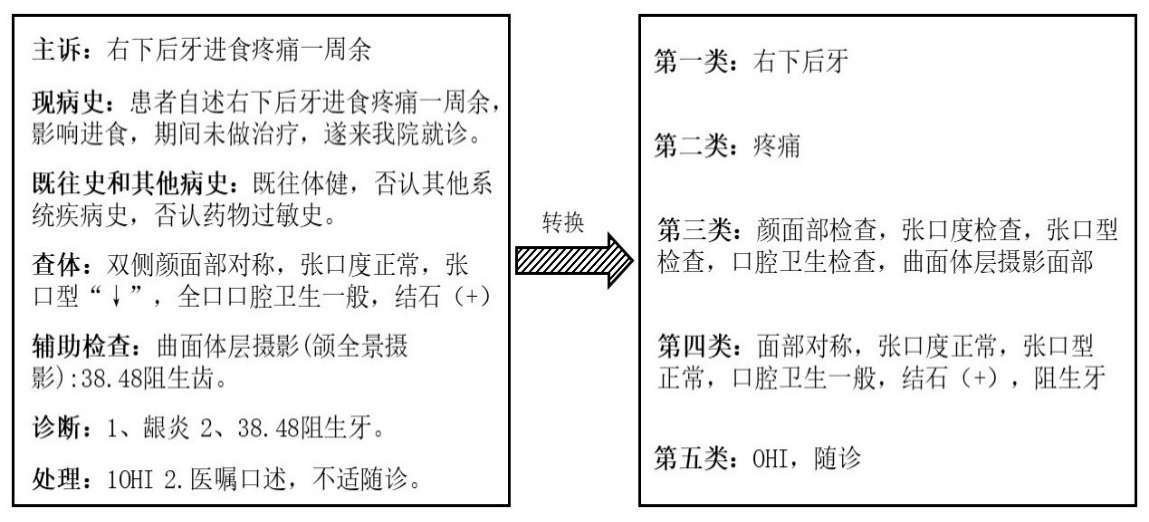


图5.8 混合型文本转换示意图

Flashtext算法是一个高效字符搜索和替换算法。实验使用Flashtext算法而非正则匹配法的原因是因为该算法的时间复杂度不依赖于搜索或替换字符的数量，即时间复杂度为O(N)，因此匹配效率比正则匹配法高得多。匹配完成后，按照对关键词的五类划分，对混合型文本进行转换或重构，使得原混合型文本转换为结构化文本，转换后的结构化文本以csv文件形式储存。图5.8为数据预处理前后混合型文本转换案例。

（2）二部二分图的构建

获得结构化文本之后，依据基于图相似性度量二部二分图表征方法构建图结构。其中，实验将第一、三、五类节点视作状态节点，将第二、四类节点视作动作节点，五类节点依据数据情况顺序相连。具体操作步骤如下：

步骤一：基于马尔科夫决策过程二分图表征方法构建底层二分图。其中，状态节点集中有三个子节点集，分别为第一、三、五类节点；动作节点集中有两个子节点集，分别为第二、四节点。图结构共有两类边，其中，第一类节点至第二类节点、第三类节点至第四类节点之间的边为决策边；第二类节点至第三类节点、第四类节点至第五类节点之间的边为转移边。

步骤二：遍历所有病例数据，计算各节点之间的状态转移概率，并将状态转移概率赋给对应转移边。

步骤三：依据马尔科夫决策过程最佳决策算法计算各状态的最佳决策边，并依据相似度度量算法计算与最佳决策边top-k相似的(k-1)条决策边。

步骤四：依据相似度度量算法计算与最大转移概率边top-k相似的（k-1）条转移边。

步骤五：提取上述边以及与边相连的节点，构建top-k二分图。

其中，在使用相似性度量算法中，我们对重要参数的设定进行了简单的实验。我们设置为0.95并调整从0.8到0.99，步长为0.01。针对每一组（，），将其代入似性度量算法计算平均相似度。之后，交换和的角色并重复上述实验。图5.9(a)和图5.9(b)为实验结果，该实验结果反映了参数对相似度的影响。根据图片所示可以发现，当不变时，随着的增加，平均相似度也在增加。相反，当不变时，值的增加显示出相反的效果。实验综合考量后选择多数实验中最常选择的值，即、为[58]。

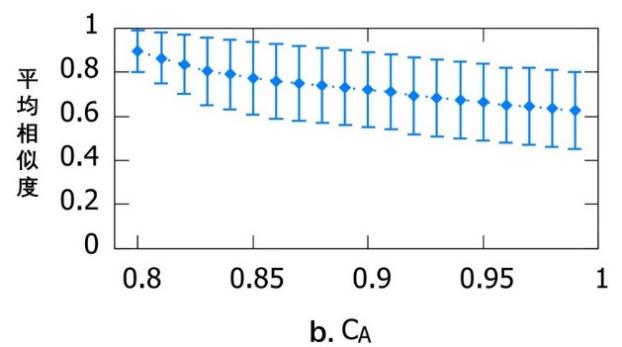
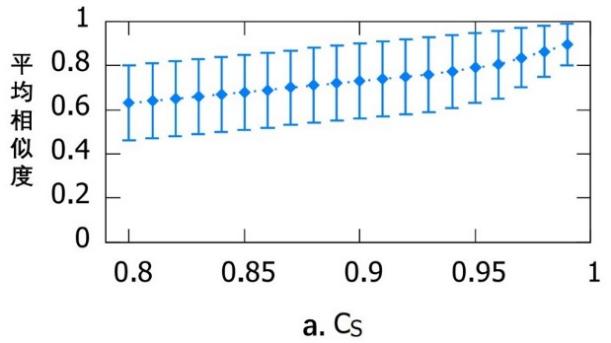


图5.9 取值对相似度的影响

完成上述步骤后，由于二部二分图整体结构较大，为了降低Neo4j图数据库的管理与操作压力，实验仅将顶层二分图（top-k二分图）构建在图数据库上，顶层二分图与底层二分图之间通过脚本程序构建联系。将顶层二分图构建在Neo4j图数据库上之后，可以得到如图5.10所示知识图谱。图中所示为顶层二分图知识图谱的部分结构。其中，橙色点为状态节点，蓝色点为动作节点。顶层二分图中k值可以根据数据的使用情况进行调节。

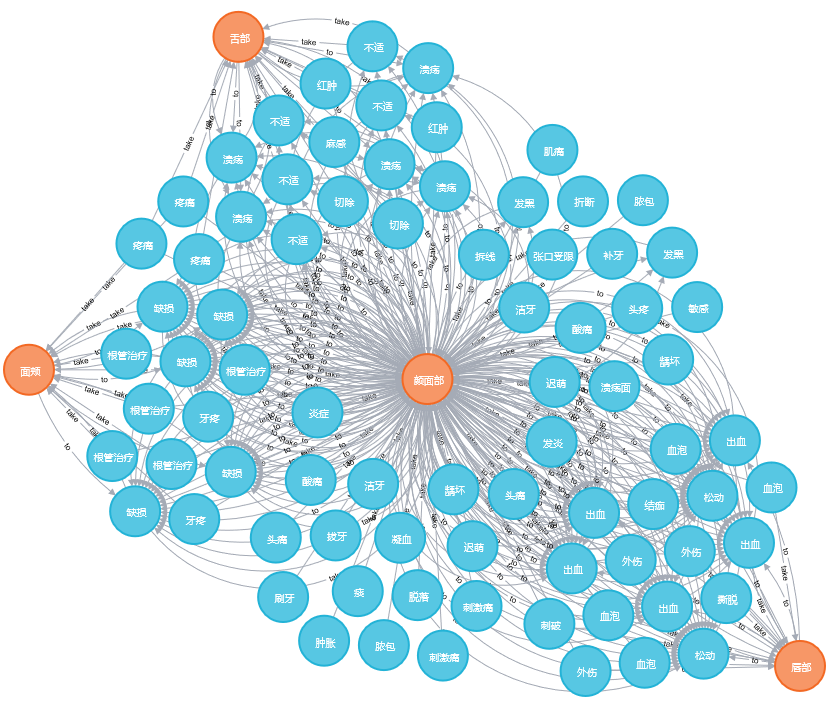


图5.10 顶层二分图示意图

（3）动态知识图谱的检索与更新

二部二分图结构构建完成之后，在顶层二分图中进行检索与匹配。检索结果一方面向外输出，另一方面记录在计数器中。当一批数据匹配完成后，将计数器的结果反馈给底层二分图，对底层二分图中的状态转移概率进行更新。在本实验中，若更新后底层二分图的状态转移概率与原状态转移概率之间的最大差距小于3%，则选择不更新顶层二分图结构，反之则对顶层二分图进行更新。

**5.2.3 实验结果分析**

实验使用5066个牙科病例数据进行实验，其中，训练集包括4066个病例，测试集分为两个部分，测试集一和测试集二分别包括500个病例。

实验所使用的计算机配置为Windows 10系统，CPU配置为Intel Core i5，1.60GHz主频，具有NVIDIA GeForce MX250显卡，16GB内存以及256GB外存。所使用的深度学习框架为Pytorch，开发环境为Python 3.5.6。

我们利用训练集构建二部二分图结构，并利用测试集一对不同k值下顶层二分图知识图谱构建时间、内存消耗、匹配准确率以及匹配时间等4个指标进行测量与对比。其中，由于单个病例可能有多个症状，不同症状可能有不同治疗路径，每个决策路径至少包括5个节点。因此，在本实验中，我们定义若一个病例中超过4节点能够在图中被匹配，那么就认为此次匹配是准确的，匹配的路径和最终结果将向外输出。

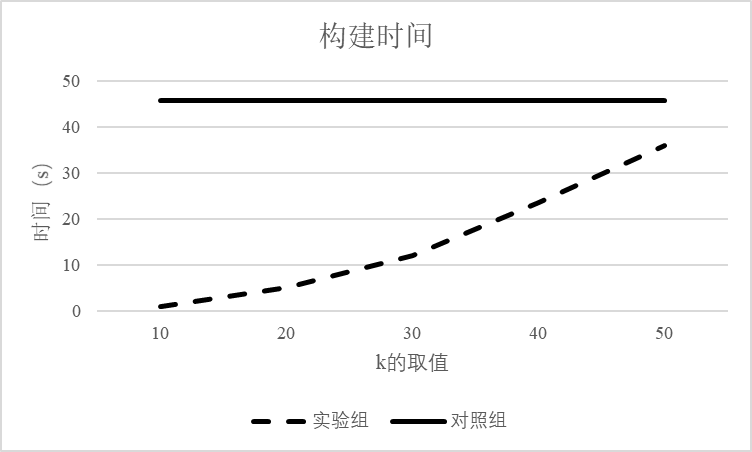
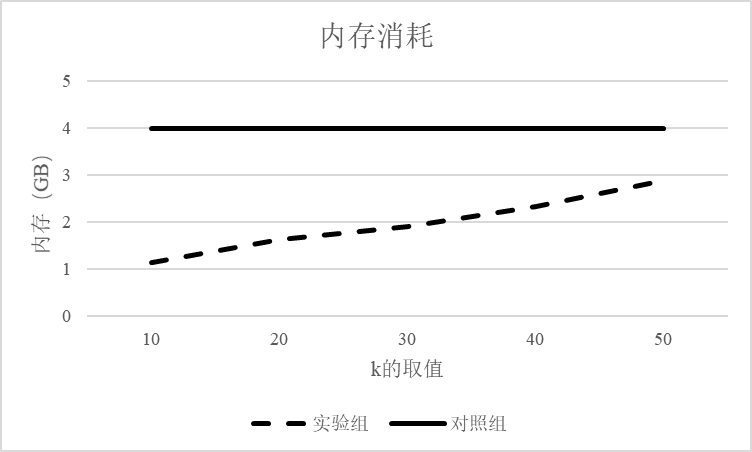
在本次实验中，我们将k值设置为10到50，每个k值之间的距离为10。当k=50时，top-k二分图的结构已经接近底层二分图的结构。因为大部分状态节点并没有一次性和50个动作节点相连，这中情况在动作节点中同样存在。基于上述设定，测试集一实验结果如表5.5所示。

表5.5 不同k值下的实验结果（测试集一）

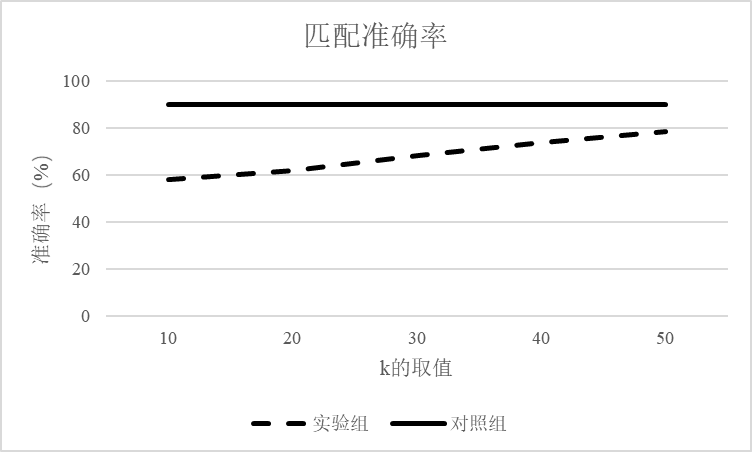
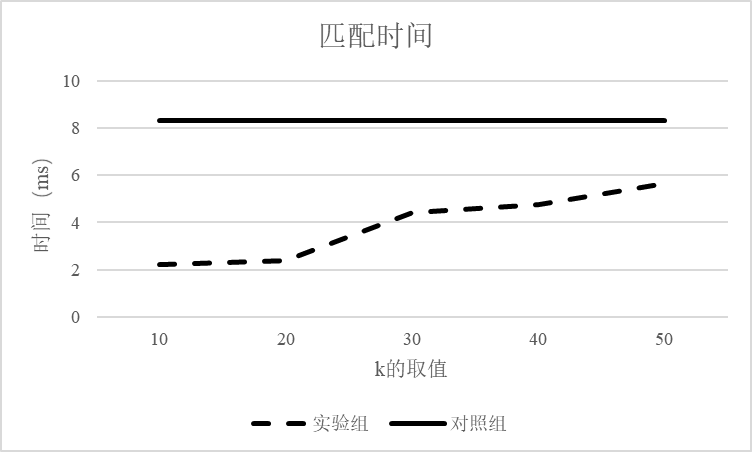
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| K | **构建时间**  **(s)** | **内存消耗**(GB) | **匹配准确率**  (%) | **匹配时间**  **(ms)** |
| 10 | 1.126 | 1.51 | 57.932 | 2.214 |
| 20 | 5.230 | 1.63 | 61.896 | 2.389 |
| 30 | 12.136 | 1.90 | 68.163 | 4.410 |
| 40 | 23.662 | 2.33 | 73.771 | 4.772 |
| 50 | 36.013 | 2.89 | 78.238 | 5.638 |

根据表5.5可以得知，随着k值不断增加，顶层二分图知识图谱的构建时间不断增加。在k的所有取值中，当k从20上升至30时，匹配时间的增长率最高，超过了84.5%，但此时，匹配准确率上升幅度同样最大，上升幅度超过了6%。当k取值为50时，所有指标均最高，当k值取60时，由于实验条件等的限制，已无法在图数据库Neo4j中建立top-k知识图谱，也无法完成相关匹配工作。与此同时，当时，top-k二分图已比底层二分图结构更大，此时构建top-k二分图已无意义。

为了更好的体现出二部二分图在检索效率等方面的优势，我们将底层二分图构建为一个独立的知识图谱，并在此图谱上进行检索。将图谱的构建时间、内存消耗、匹配准确率以及匹配时间作为对照组，将其与二部二分图上的对应结果进行对比分析，其中，底层二分图的构建时间为45.78s，内存消耗为3.98GB，匹配准确率为89.79%，匹配时间为8.34ms。对比结果图如图5.11所示。根据对比图可以发现，top-k二分图中k取值为50时，构建时间、内存消耗和匹配时间三个方面的数值均大大低于底层二分图，但在匹配准确率上却和底层二分图相近，两者之间的差距在10%左右。由此可以看出，二部二分图上的检索虽然会损失一部分的准确率，但却可以大大提升构建与检索的效率并降低检索过程中对内存的需求。

a. 构建时间对比图 b. 内存消耗对比图

c. 匹配准确率对比图 d. 匹配时间对比图

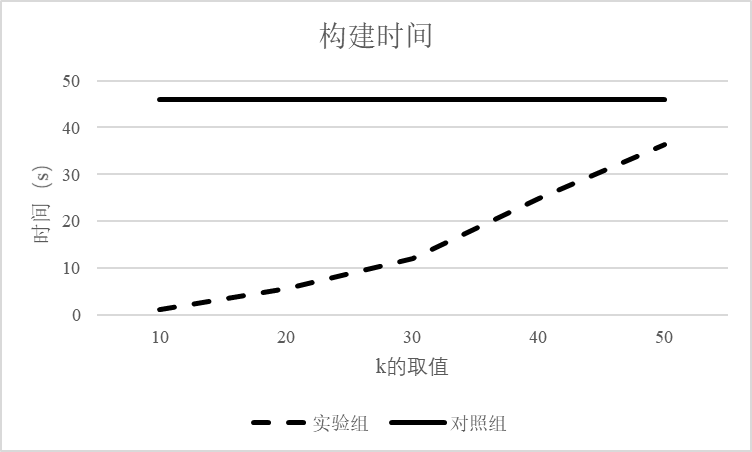
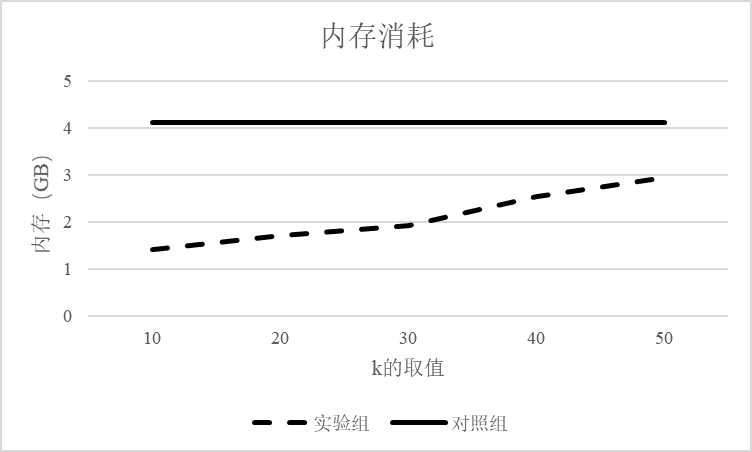
图5.11 二部二分图实验对比图

之后，我们将测试集一的匹配结果反馈至底层二分图，对底层二分图以及顶层二分图进行更新，并使用测试集二在更新后的二部二分图进行匹配测试。表5.6为测试集二在更新后top-k二分图上的实验结果。从表中可以发现，二部二分图更新之后，匹配准确率有一定的提升，而构建时间、内容消耗以及匹配时间并没有较大的改变。这是因为构建时间、内存消耗以及匹配时间的大小于图的结构，即k的取值有直接关系。而匹配准确率则与图谱和这批数据之间的贴合情况有更密切的关联。一般情况下，一批数据，尤其是一批医疗数据会反映某个阶段、某个地区、某类人群的某种共性，而这种共性会在图谱的不断更新之下被隐性表现出来，从而导致匹配准确率的小幅提升。

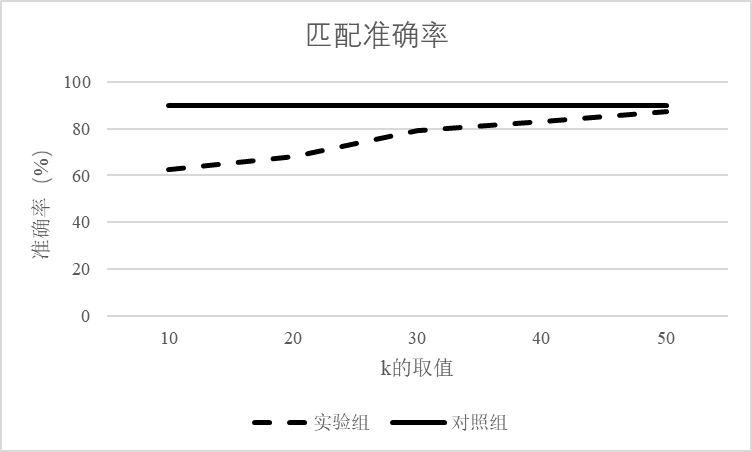
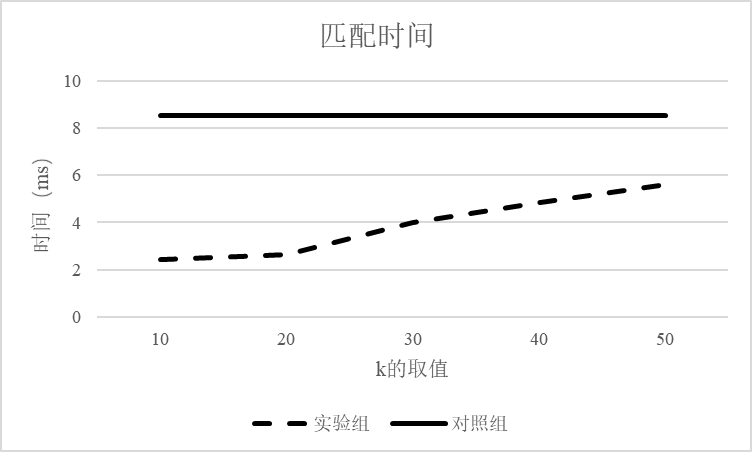
除此之外，我们更新底层二分图的知识图谱，并在更新后的底层二分图知识图谱上进行实验。其中，更新后底层二分图的构建时间为45.90s，内存消耗为4.11GB，匹配准确率为89.97%，匹配时间为8.53ms。将实验结果作为对照组和表4.4中的结果进行对比分析。对比图如图5.12所示。根据对比图可以发现，更新后，当k取值为50时，二部二分图匹配准确率极为接近底层二分图的匹配准确率，而在构建时间、内存消耗以及匹配时间方面，底层二分图的数据仍旧远远高于二部二分图。由此可见，二部二分图的表征方法在检索效率提升方法有较大优势。

表5.6 更新后不同k值下的实验结果（测试集二）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **K** | **构建时间**  (s) | **内存消耗**(GB) | **匹配准确率**  (%) | **匹配时间**  **(**ms) |
| 10 | 1.112 | 1.41 | 62.322 | 2.421 |
| 20 | 5.642 | 1.71 | 68.162 | 2.629 |
| 30 | 11.993 | 1.93 | 79.213 | 3.986 |
| 40 | 24.752 | 2.54 | 82.971 | 4.852 |
| 50 | 36.311 | 2.95 | 87.431 | 5.598 |

a. 构建时间对比图 b. 内存消耗对比图

c. 匹配准确率对比图 d. 匹配时间对比图

图5.12 动态演进后二部二分图实验对比图

除此之外，通过该方法构建的图结构相比传统方法而言更具有可解释性与透明性，整个匹配过程几乎可以模拟患者的治疗过程，匹配得到的路径与结果也更加符合医生的诊断逻辑。例如，通过匹配可以得到如下检查路径和检查结果。匹配过程如图5.13所示，图中蓝色节点为节点类别，其中，第一类节点对应“部位”，第二类节点对应“症状”，第三类节点对应“检查手段”，第四类节点对应“诊断”，第五类节点对应“处理”。图中灰色节点以及与灰色节点相连的有向实线为检索结果和检索路径，检索路径与检索结果为“右下后牙🡪疼痛🡪卫生检查🡪牙结石🡪随诊、OHI”。检索结果与检索路径模拟了医生看诊的过程，体现了表征方法的透明性与可解释性。通常情况下，复杂病例可能有多条检索路径与多个检索结果。

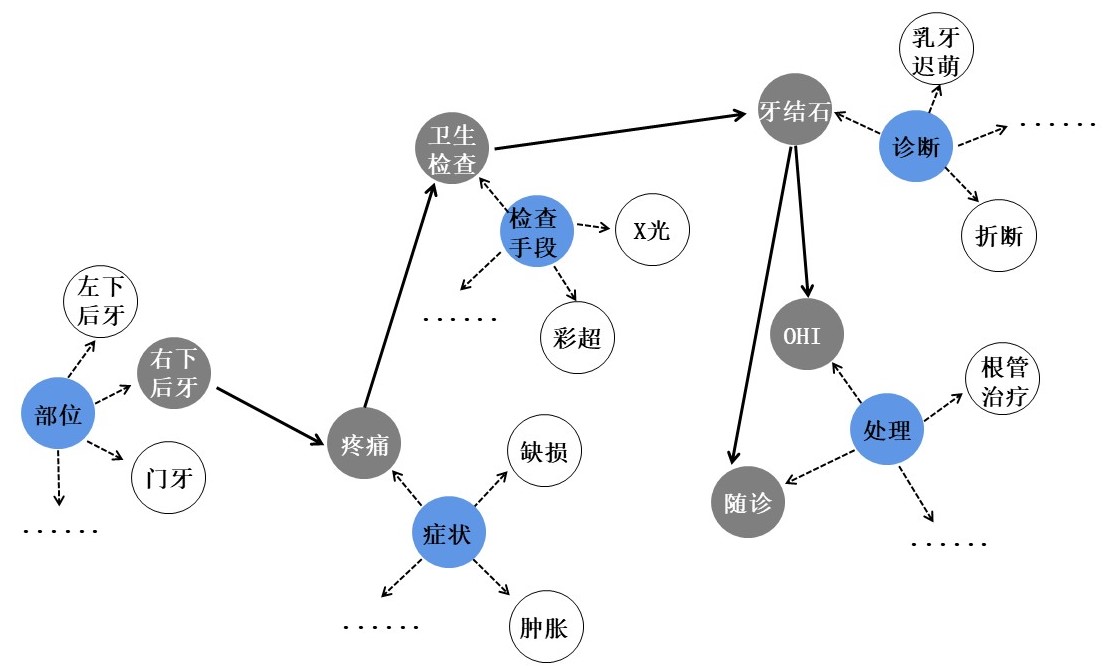


图5.13 检索路径与检索结果示意图

**5.3 本章小结**

本章使用问卷型医疗数据和电子病历型医疗数据，分别对基于马尔科夫决策过程的二分图表征方法以及基于图相似性度量的二部二分图表征方法进行实验验证与分析。实验结果显示，和传统特征向量表征方法相比，基于马尔科夫决策过程的二分图表征方法在数据特征以及数据关系的表征方面更占优势。数据在经过基于马尔科夫决策过程的二分图表征方法表征之后，使用图卷积网络对数据进行预测，预测结果的准确率和传统方法相比最大能够提升13.6%。专业医师的评估也体现了该方法的可解释性。

在基于图相似性度量的二部二分图表征方法实验方面，和基于马尔科夫决策过程的二分图表征方法相比，二部二分图表征方法构建时间平均减少65.8%，内存消耗平均减少50.2%，匹配时间平均减少53.4%。二部二分图动态演进后，虽然构建时间、内存消耗以及匹配时间无较大提升，但匹配准确率平均提升8%。

**第6章 总结与展望**

近年来，随着互联网以及相关技术的快速发展，多源数据的融合与处理引起了国内外研究者的广泛关注与讨论。其中，特征级融合是常用融合方法。这种方法一般采用特征向量的方式表征数据的特征，但这种方法表达的特征较为离散且单一，无法很好的反映数据之间的显性关系与隐性关系。目前，也有一些研究采用了图的方法表征数据，但却存在图结构复杂、图操作难度大以及静态性等问题。因此，本文针对多源数据中的混合型文本表征方法展开研究，从特征级融合的角度设计了两种新颖的数据图表征方法。具体工作如下：

设计并实现了基于马尔科夫决策过程的二分图表征法。该方法在马尔科夫决策过程的基础上提出了基于马尔科夫决策过程的图表征方法，并进一步将图表征方法优化为二分图表征方法。该方法首先通过提取数据之间的时序关系和逻辑关系，将数据分为状态和动作；然后将状态和动作之间的关系分为决策和转移，通过转移与决策表征数据之间的关系；最后将这些关系以二分图的方式表示出来，并通过特征赋予或加权的形式丰富数据之间的关系，从而实现对数据的有效表征。该方法通过图数据库管理二分图结构以及二分图所表示的特征。

设计并实现了基于图相似性度量的二部二分图表征方法。该方法使用最佳决策算法与图相似性度量算法，在第一种方法的基础上，提炼原二分图中的 top-k结构形成top-k二分图，并与原二分图形成二部二分图结构。该二部二分图由基于马尔科夫决策过程二分图（底层二分图）以及基于图相似性度量的top-k二分图（顶层二分图）组成。底层二分图为顶层二分图提供支撑，顶层二分图用于检索并将检索结果反馈至底层二分图，从而在优化图结构、缩小图检索范围、提升检索效率的同时，实现图结构的动态循环与演进。

本文使用问卷型医疗数据和电子病历型医疗数据对方法进行实验设计与分析评估。评估结果表明，与传统向量表征方法相比，本文中的二分图表征方法具有更优的可解释性和表征效果；与二分图表征方法相比，二部二分图表征方法构建时间平均减少65.8%，内存消耗平均减少50.2%，匹配时间平均减少53.4%。二部二分图结构动态演进后，检索准确率平均提高8%。

本文提出的方法中仍旧存在一些不足之处，例如，如何降低基于图相似性度量二部二分图表征方法中图相似性度量算法时间复杂度，如何确定顶层二分图中的k值，动态演进多久进行一次等，这些问题需要开展更多的实验才可能得到更好的答案。

**致 谢**

时光飞逝，岁月如梭，三年研究生生涯也即将迎来尾声。回首三年时光，很幸运地是，我学习和收获了很多，也遇上了一群温暖又善良的人，这一路的相知相伴，让这三年满满都是美好的回忆。在这即将毕业之际，我要向每一位给予我帮助的人致以最诚挚的感谢。

首先，我要感谢的是我的导师徐子晨教授。徐教授治学严谨、学识渊博、眼光独到，像一盏明灯指引着我不断前行。正是徐教授的指引让我找到了合适且喜爱的研究方向，也正是徐教授的指导让我能够在这个研究方向上一直走下去。感谢徐教授在这期间对我的悉心教导，为我指点迷津，提供实验条件与科研环境。再次向徐教授致以最崇高的感谢与最诚挚的祝福。

其次，我要感谢一直相互帮助、相互关心的伙伴们。感谢实验室的同门小伙伴，与你们相伴，让我三年研究生时光不再乏味，转而变得充实、温馨又快乐。感谢宿舍的小伙伴，与你们同处，让我收获了深厚的友谊。祝愿所有小伙伴们都能学业顺利、心想事成。

最后，我要特别感谢父母二十多年对我的支持与关爱。是你们在背后的鼓励与理解，让我能够无忧无虑的享受学习生活，让我能够坚定地走自己喜欢的路。

再一次感谢身边所有人。毕业在即，新的旅途即将开启，我将带着你们的祝福继续前行。愿我们相识于微，重逢于炬！

程婕

2021年4月

**参考文献**

1. Suchanek F, Weikum G. Knowledge harvesting in the big-data era[C]. Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. 2013: 933~938.
2. Paulheim H. Knowledge graph refinement: A survey of approaches and evaluation methods[J]. Semantic web, 2017, 8(3): 489~508.
3. Qi G L, Gao H, Wu T X. The research advances of knowledge graph[J]. technology intelligence engineering, 2017, 3(1): 4~25.
4. Zhang Y, Dai H, Kozareva Z, et al. Variational reasoning for question answering with knowledge graph[C]. Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32(1).
5. Schlichtkrull M, Kipf T N, Bloem P, et al. Modeling relational data with graph convolutional networks[C]. European Semantic Web Conference. Springer, Cham, 2018: 593~607.
6. Paulheim H. Knowledge graph refinement: A survey of approaches and evaluation methods[J]. Semantic web, 2017, 8(3): 489~508.
7. Leighton B B, Palmer N P, Schmid P R, et al. Intelligent search tool for answering clinical queries: U.S. Patent 8,838,628[P]. 2014-9~16.
8. Sun Z, Yang J, Zhang J, et al. Recurrent knowledge graph embedding for effective recommendation[C]. Proceedings of the 12th ACM Conference on Recommender Systems. 2018: 297~305.
9. Yu T, Li J, Yu Q, et al. Knowledge graph for TCM health preservation: design, construction, and applications[J]. Artificial Intelligence in Medicine, 2017, 77: 48~52.
10. Zhao X, Jia Y, Li A, et al. Multi-source knowledge fusion: a survey[J]. World Wide Web, 2020, 23(4): 2567~2592.
11. Wang J, Lan Y, Zhang S, et al. Knowledge Graph for Multi-source Data Fusion Topics Research[C]. 2020 International Conference on High Performance Big Data and Intelligent Systems (HPBD&IS). IEEE, 2020: 1~5.
12. Psyllidis A. Ontology-based data integration from heterogeneous urban systems: A knowledge representation framework for smart cities[C]. CUPUM 2014: Proceedings of the 14th International Conference on Computers in Urban Planning and Urban Management, Cambrigde, USA, 7-10 July 2015. MIT, 2015.
13. Le-Phuoc D, Quoc H N M, Quoc H N, et al. The Graph of Things: A step towards the Live Knowledge Graph of connected things[J]. Journal of Web Semantics, 2016, 37: 25~35.
14. Paulheim H. Knowledge graph refinement: A survey of approaches and evaluation methods[J]. Semantic web, 2017, 8(3): 489~508.
15. Salahuddin M A, Al-Fuqaha A, Guizani M, et al. Softwarization of internet of things infrastructure for secure and smart healthcare[J]. arXiv preprint arXiv:1805.11011, 2018.
16. Gyrard A, Gaur M, Shekarpour S, et al. Personalized health knowledge graph[J]. 2018.
17. Tsai W H, Fu K S. Error-correcting isomorphisms of attributed relational graphs for pattern analysis[J]. IEEE Transactions on systems, man, and cybernetics, 1979, 9(12): 757~768.
18. Pujara J, Miao H, Getoor L, et al. Knowledge graph identification[C]. International Semantic Web Conference. Springer, Berlin, Heidelberg, 2013: 542~557.
19. Xu J, Chen K, Qiu X, et al. Knowledge graph representation with jointly structural and textual encoding[J]. arXiv preprint arXiv:1611.08661, 2016.
20. Tao Qi, Chuhan Wu, Fangzhao Wu, Suyu Ge, Junxin Liu, Yongfeng Huang, Xing Xie,Fast Neural Chinese Named Entity Recognition with Multi-head Self-attention[C]. CCKS 2019: 98~110.
21. Fenia Christopoulou, Makoto Miwa, Sophia Ananiadou,A Walk-based Model on Entity Graphs for Relation Extraction[C]. ACL (2) 2018: 81~88.
22. Ji S, Pan S, Cambria E, et al. A survey on knowledge graphs: Representation, acquisition and applications[J]. arXiv preprint arXiv:2002.00388, 2020.
23. Chen W, Huang J, Luo S, et al. Research on space-time evolution model of Xiangshan Culture knowledge graph based on named graph[C]. 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). IEEE, 2017: 673~678.
24. Xu C, Nayyeri M, Alkhoury F, et al. Temporal knowledge graph embedding model based on additive time series decomposition[J]. arXiv preprint arXiv:1911.07893, 2019.
25. De S, Gupta K, Stanley R J, et al. A comprehensive multi-modal NDE data fusion approach for failure assessment in aircraft lap-joint mimics[J]. IEEE Transactions on Instrumentation and Measurement, 2013, 62(4): 814~827.
26. Zhang J. Multi-source remote sensing data fusion: status and trends[J]. International Journal of Image and Data Fusion, 2010, 1(1): 5~24.
27. Garcia F, Mirbach B, Ottersten B, et al. Pixel weighted average strategy for depth sensor data fusion[C]. 2010 IEEE International Conference on Image Processing. IEEE, 2010: 2805~2808.
28. Gan Q, Harris C J. Comparison of two measurement fusion methods for Kalman-filter-based multisensor data fusion[J]. IEEE Transactions on Aerospace and Electronic systems, 2001, 37(1): 273~279.
29. 唐骥锋,刘国栋.基于多传感器融合的机器人蒙特-卡洛定位决策[J].传感器与微系统,2012,31(03):18~21.
30. Zhang W, Zhang Y, Zhai J, et al. Multi-source data fusion using deep learning for smart refrigerators[J]. Computers in Industry, 2018, 95: 15~21.
31. 杜炎,吕良福,焦一辰.基于模糊推理的模糊原型网络[J/OL]. 计算机应用:1~7[2021-04-06].http://kns.cnki.net/kcms/detail/51.1307.TP.20201229.1443.004.html.
32. Majumder S, Pratihar D K. Multi-sensors data fusion through fuzzy clustering and predictive tools[J]. Expert Systems with Applications, 2018, 107: 165~172.
33. Zhang Y Q, Fraser M D, Gagliano R A, et al. Granular neural networks for numerical-linguistic data fusion and knowledge discovery[J]. IEEE transactions on neural networks, 2000, 11(3): 658~667.
34. El Faouzi N E, Klein L A, De Mouzon O. Improving travel time estimates from inductive loop and toll collection data with Dempster–Shafer data fusion[J]. Transportation research record, 2009, 2129(1): 73~80.
35. Khan M N, Anwar S. Time-domain data fusion using weighted evidence and Dempster–Shafer combination rule: Application in object classification[J]. Sensors, 2019, 19(23): 5187.
36. Castanedo F. A review of data fusion techniques[J]. The Scientific World Journal, 2013, 2013.
37. Jaramillo V H, Ottewill J R, Dudek R, et al. Condition monitoring of distributed systems using two-stage Bayesian inference data fusion[J]. Mechanical Systems and Signal Processing, 2017, 87: 91~110.
38. Wagner J, Arora P, Cortes S, et al. Dcu: Aspect-based polarity classification for semeval task 4[J]. 2014.
39. Kiritchenko S, Zhu X, Cherry C, et al. NRC-Canada-2014: Detecting aspects and sentiment in customer reviews[C]. Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014). 2014: 437~442.
40. 饶元淇,赵旭俊,蔡江辉.轨迹数据的多特征融合及检测方法[J].小型微型计算机系统,2021,42(02):264~270.
41. Zhao L, Deng H, Qiu L, et al. Urban Multi-Source Spatio-Temporal Data Analysis Aware Knowledge Graph Embedding[J]. Symmetry, 2020, 12(2): 199.
42. Nikolay F, Pesavento M. Learning directed-acyclic-graphs from multiple genomic data sources[C]. 2017 25th European Signal Processing Conference (EUSIPCO). IEEE, 2017: 1877~1881.
43. Yi S. Learning ontologies for geographic entity matching and multi-sources data fusion[C]. 2013 21st International Conference on Geoinformatics. IEEE, 2013: 1~5
44. Pan Z, Zhuang X, Ren J, et al. Pattern-Based Knowledge Graph Embedding for Non-functional Requirements[C]. 2019 6th International Conference on Dependable Systems and Their Applications (DSA). IEEE, 2020: 407~412.
45. Sheng M, Zhang H, Zhang Y, et al. CLMed: A Cross-lingual Knowledge Graph Framework for Cardiovascular Diseases[C]. International Conference on Web Information Systems and Applications. Springer, Cham, 2019: 512~517.
46. 刘峤,韩明皓,江浏祎,刘瑶,耿技.基于双层随机游走的关系推理算法[J].计算机学报,2017,40(06):1275~1290.
47. Khalil E, Dai H, Zhang Y, et al. Learning combinatorial optimization algorithms over graphs[C]. Advances in Neural Information Processing Systems. 2017: 6348~6358.
48. Cui J, Yang B. Graph bayesian optimization: Algorithms, evaluations and applications[J]. arXiv preprint arXiv:1805.01157, 2018.
49. Yi Y, Wang J, Zhou W, et al. Joint graph optimization and projection learning for dimensionality reduction[J]. Pattern Recognition, 2019, 92: 258~273.
50. Yang X, Yang J, Li Y, et al. Effiecient Graph Matching by Adjacency Matrix Based Model and Gaussian Smoothing Based Optimization[J]. Aust. J. Intell. Inf. Process. Syst., 2019, 16(4): 82~89.
51. Xu H, Luo D, Zha H, et al. Gromov-wasserstein learning for graph matching and node embedding[J]. arXiv preprint arXiv:1901.06003, 2019.
52. Li Y, Gu C, Dullien T, et al. Graph matching networks for learning the similarity of graph structured objects[J]. arXiv preprint arXiv:1904.12787, 2019.
53. Webber J. A programmatic introduction to neo4j[C]. Proceedings of the 3rd annual conference on Systems, programming, and applications: software for humanity. 2012: 217~21.
54. Lu H, Hong Z, Shi M. Analysis of film data based on Neo4j[C]. 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS). IEEE, 2017: 675~677.
55. Iordanov B. Hypergraphdb: a generalized graph database[C]. International conference on web-age information management. Springer, Berlin, Heidelberg, 2010: 25~36.
56. Chang C, Moon B, Acharya A, et al. Titan: a high-performance remote-sensing database[C]. Proceedings 13th International Conference on Data Engineering. IEEE, 1997: 375~384.
57. Markov A A. The theory of algorithms[J]. Am. Math. Soc. Transl., 1960, 15: 1~14.
58. Wang H, Dong S, Shao L. Measuring Structural Similarities in Finite MDPs[C]. IJCAI. 2019: 3684~3690.
59. Goldberg Y, Levy O. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method[J]. arXiv preprint arXiv:1402.3722, 2014.
60. Islam A, Inkpen D. Semantic text similarity using corpus-based word similarity and string similarity[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2008, 2(2): 1~25.
61. Jeh G, Widom J. Simrank: a measure of structural-context similarity[C]. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. 2002: 538~543.
62. Ling H, Okada K. An efficient earth mover's distance algorithm for robust histogram comparison[J]. IEEE transactions on pattern analysis and machine intelligence, 2007, 29(5): 840~853.
63. Stein A, Pearson R M, Goodman S H, et al. Effects of perinatal mental disorders on the fetus and child[J]. The Lancet, 2014, 384(9956): 1800~1819.
64. Gelaye B, Rondon M B, Araya R, et al. Epidemiology of maternal depression, risk factors, and child outcomes in low-income and middle-income countries[J]. The Lancet Psychiatry, 2016, 3(10): 973~982.
65. Grote N K, Bridge J A, Gavin A R, et al. A meta-analysis of depression during pregnancy and the risk of preterm birth, low birth weight, and intrauterine growth restriction[J]. Archives of general psychiatry, 2010, 67(10): 1012~1024.
66. Ghimire U, Papabathini S S, Kawuki J, et al. Depression during pregnancy and the risk of low birth weight, preterm birth and intrauterine growth restriction-an updated meta-analysis[J]. Early Human Development, 2021, 152: 105243.
67. Gao M, Hu J, Yang L, et al. Association of sleep quality during pregnancy with stress and depression: a prospective birth cohort study in China[J]. BMC pregnancy and childbirth, 2019, 19(1): 1~8.
68. Bergink V, Kooistra L, Lambregtse-van den Berg M P, et al. Validation of the Edinburgh Depression Scale during pregnancy[J]. Journal of psychosomatic research, 2011, 70(4): 385~389.
69. Al-Azri M, Al-Lawati I, Al-Kamyani R, et al. Prevalence and risk factors of antenatal depression among Omani women in a primary care setting: cross-sectional study[J]. Sultan Qaboos University Medical Journal, 2016, 16(1): e35.
70. Achananuparp P, Hu X, Yang C C. Addressing the variability of natural language expression in sentence similarity with semantic structure of the sentences[C]. Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, Berlin, Heidelberg, 2009: 548~555.
71. Clark K, Khandelwal U, Levy O, et al. What does bert look at? an analysis of bert's attention[J]. arXiv preprint arXiv:1906.04341, 2019.
72. Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
73. Munikar M, Shakya S, Shrestha A. Fine-grained sentiment classification using BERT[C]. 2019 Artificial Intelligence for Transforming Business and Society (AITB). IEEE, 2019, 1: 1~5.
74. Wang M, Yu L, Zheng D, et al. Deep Graph Library: Towards Efficient and Scalable Deep Learning on Graphs[J]. 2019.
75. Zheng D, Wang M, Gan Q, et al. Learning Graph Neural Networks with Deep Graph Library[C]. Companion Proceedings of the Web Conference 2020. 2020: 305~306.
76. Li Q, Han Z, Wu X M. Deeper insights into graph convolutional networks for semi-supervised learning[C]. Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32(1).
77. Loh W Y. Classification and regression trees[J]. Wiley interdisciplinary reviews: data mining and knowledge discovery, 2011, 1(1): 14-23.
78. Breiman L. Random forests[J]. Machine learning, 2001, 45(1): 5-32.
79. Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective[J]. Artificial Intelligence in medicine, 2001, 23(1): 89~109.
80. Siegel M A , Firriolo F J , Finkelstein M W . Computer applications in oral diagnosis[J]. Dental Clinics of North America, 1993, 37(1):113~131.
81. Zhou L, Hripcsak G. Temporal reasoning with medical data—a review with emphasis on medical natural language processing[J]. Journal of biomedical informatics, 2007, 40(2): 183~202.
82. Ren Y, Shi Y, Zhang K, et al. Medical Treatment Migration Prediction Based on GCN via Medical Insurance Data[J]. IEEE Journal of Biomedical and Health Informatics, 2020, 24(9): 2516~2522.
83. Shi L, Li S, Yang X, et al. Semantic health knowledge graph: semantic integration of heterogeneous medical knowledge and services[J]. BioMed research international, 2017, 2017.
84. Li X, Shi T, Li P, et al. BiLSTM-CRF Model for Named Entity Recognition in Railway Accident and Fault Analysis Report[C]. Proceedings of the Asia-Pacific Conference on Intelligent Medical 2018 & International Conference on Transportation and Traffic Engineering 2018. 2018: 1~5.
85. Stenetorp P, Pyysalo S, Topić G, et al. BRAT: a web-based tool for NLP-assisted text annotation[C]. Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. 2012: 102~107.
86. Ma X, Wang Z, Ng P, et al. Universal text representation from BERT: an empirical study[J]. arXiv preprint arXiv:1910.07973, 2019.

**攻读学位期间的研究成果**

[1] Xu Z, **Cheng J**, Wang Y, et al. Developing (Almost) Free Distributed System Labs Using Container-based Technique[C]. Proceedings of the ACM Turing Celebration Conference-China. 2020: 101-106.

[2] 徐子晨，**程婕**，王玉皞，饶泓.基于胜任力模型的新工科在线系统实验教学研究[J]. 高等工程教育研究.（已录用，2021年7月见刊）