汪洋, 杨伟宏, 孙剑,等. 情感障碍症知识图谱构建关键技术研究[J]. 智能计算机与应用,2025,15(1):38-45. DOI:10.20169/j.issn.2095-2163.24061805

情感障碍症知识图谱构建关键技术研究

汪 洋[1], 杨伟宏[2], 孙 剑[1], 周旭东[1,3], 田正卫[1]

(1 泸州职业技术学院, 四川 泸州 646000; 2 哈尔滨工业大学(深圳), 广东 深圳 518055;
3 国家技术转移西南中心 泸州分中心, 四川 泸州 646000)

摘 要: 情感障碍症知识图谱构建关键技术主要包括多源数据语料库构建、信息抽取、知识融合、知识图谱存储与可视化等。首先,采用多源头数据来整合构建情感障碍症知识语料库,经过数据处理得到归一化的语料库 Normal;其次,在信息抽取阶段提出一种混合最大熵马尔可夫模型-卷积神经网络(MEMM-CNN)的信息抽取方法,通过 MEMM 首次训练语料库数据集、CNN 进一步训练模型,重复迭代完成信息抽取中的 3 个任务单元,得到最终的三元组集;在知识融合阶段设计了一种基于 ρ 值的自适应加权估计知识融合算法(SAW),通过可信度函数计算三元组可信度并与可信度阈值进行比较,完成三元组极性识别,输出可信库三元组集,该算法相比较于 VOTE、ACCU 等传统数据融合算法能够显著提高可信库中三元组的识别数量和三元组极性识别准确率,同时具有抗噪性能好、响应时间短等优点;最后,利用 Neo4j 图形数据库动态生成情感障碍症知识图谱。

关键词: 知识图谱; 情感障碍症; 信息抽取; 知识融合; Neo4j 图形数据库

中图分类号: TP391.5 **文献标志码**: A **文章编号**: 2095-2163(2025)01-0038-08

Research on key technologies in constructing affective disorder knowledge graph

WANG Yang[1], YANG Weihong[2], SUN Jian[1], ZHOU Xudong[1,3], TIAN Zhengwei[1]

(1 Luzhou Vocational and Technical College, Luzhou 646000, Sichuan, China;
2 Harbin Institute of Technology (Shenzhen), Shenzhen 518055, Guangdong, China;
3 National Technology Transfer Southwest Center Luzhou Branch, Luzhou 646000, Sichuan, China)

Abstract: The key technologies for constructing a knowledge graph of affective disorder mainly include multi-source data corpus construction, information extraction, knowledge fusion, knowledge graph storage and visualization, etc. Firstly, multiple sources of data are used to integrate and construct a corpus of knowledge on affective disorder. After data processing, a normalized corpus, Normal, is obtained. Secondly, a mixed Maximum Entropy Markov Model - Convolutional Neural Network (MEMM-CNN) information extraction method was proposed in the information extraction stage. The corpus dataset was first trained by MEMM, and the model was further trained by CNN. Repeat the iteration to complete the three task units in information extraction and obtain the final triplet set. Then, in the knowledge fusion stage, an adaptive weighted estimation knowledge fusion algorithm based on ρ-value (SAW) was designed, which calculates the triplet credibility through the credibility function and compares it with the credibility threshold to complete triplet polarity recognition and output the trusted database triplet set. Compared with traditional data fusion algorithms such as VOTE and ACCU, this algorithm can significantly improve the recognition quantity and polarity recognition accuracy of triples in the trusted database. At the same time, it has advantages such as good noise resistance and short response time. Finally, use the Neo4j graphical database to dynamically generate a knowledge graph of affective disorder.

Key words: knowledge graph; affective disorder; information extraction; knowledge fusion; neo4j graphics database

0 引 言

知识图谱是现实世界中根据实体间关系相互连接起来所形的一种网络结构,用于呈现各类实体以及实体间的关联关系[1]。利用自然语言处理、机器学习、数据挖掘等人工智能技术将结构化、半结构化、非结构化的数据整合成知识图谱,可以完成某一知识领域大数据的快速分析和简化表示,最终实现

基金项目: 四川省科技计划(21CXJDPT0001); 泸州市科技计划(2021-JYJ-96); 泸州职业技术学院校级科研项目(LZZX-B-02)。

作者简介: 汪 洋(1991—),男,硕士,讲师,主要研究方向:人工智能,大数据医疗; 杨伟宏(1991—),男,博士,主要研究方向:自然语言处理,机器学习。

收稿日期: 2024-06-18

智慧搜索与智能交互[2-3]。2012 年，谷歌公司首次提出了知识图谱（Knowledge Graph，KG）的概念，目前知识图谱在中国医疗卫生领域得到了快速的发展和应用[4]。情感障碍症（抑郁症）目前已成为仅次于癌症的世界第二大健康“杀手”，但是精细化的情感障碍症知识图谱资源依旧匮乏稀缺[5]。

北京大学计算语言研究所和郑州大学自然语言处理实验室基于大规模医疗文本数据，利用自然语言处理和文本挖掘技术研发的中文医学知识图谱 CMe KG2.0 涵盖了超过 1 万种疾病、近 2 万种药物、1 万余个症状、3 千种诊疗手段，在中国健康管理、疾病风险预测、辅助诊疗、病历结构化等智慧医疗领域发挥了巨大作用，但是 CMe KG2.0 针对情感障碍症生成的疾病图谱比较简略和粗糙[6-7]。中国中医科学院中医药信息研究所牵头完成的中医药知识图谱（TCMKB）目前实现了中医药知识自动问答以及辅助决策等应用，但是根据医学界针对情感障碍症治疗的普遍共识表明目前尚无有效中药能够治疗此类疾病[8]。中国医科大学赵雪娇[9]利用自然语言处理技术对妇产科教材中的医学知识进行抽取和表示，构建了妇产科知识图谱，存在的主要问题是构建的知识医疗知识图谱的数据源单一，说服力较弱。

国内目前已构建的中文医学知识图谱有关情感障碍症的内容较为粗糙、资源匮乏。本文通过情感障碍症知识图谱构建关键技术研究，尝试构建出完整的精细化情感障碍症知识图谱，不仅可以用于对情感障碍症疾病的普适性教育，还可以用于情感障碍症患者自我诊断和辅助医生临床决策等，具有重要的现实意义。

1　情感障碍症知识图谱技术架构

1.1　构建方案

情感障碍症知识图谱自顶向下整体构建方案如图 1 所示，主要包含多源数据语料库构建、信息抽取、知识融合、知识图谱存储与可视化 4 个阶段。

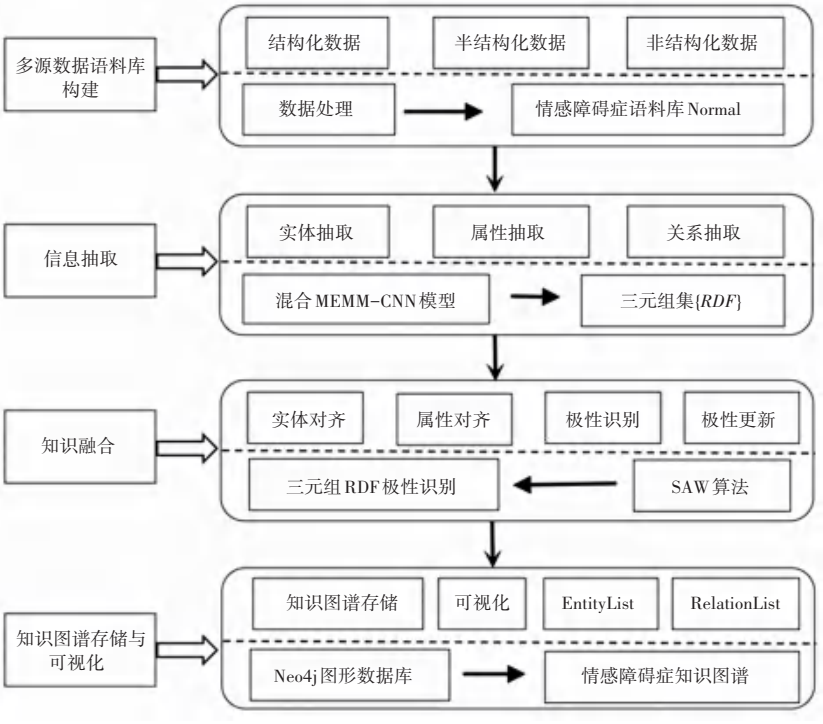


图 1　情感障碍症知识图谱整体构建方案

Fig. 1　Overall construction plan for the knowledge graph of affective disorder

1.2　多源数据语料库构建阶段

通过各个渠道获取结构化、半结构化以及非结构化的情感障碍症医疗文本大数据，保证了数据来源的丰富性和全面性，是构建情感障碍症知识图谱的基础[10]。本文主要选取 4 个源头数据。

（1）医学书籍。主要选取医学院使用的教材、百强出版社出版的权威书籍等，这类资源具有很高的权威性，是构建情感障碍症知识图谱的重要数据源之一。

（2）医学论文。主要从知网、维普、万方三大期刊检索“情感障碍症”、“抑郁症”等关键词获取中文核心期刊科研论文，此部分数据同样具有较高的权

威性和真实性。

(3)医院电子病历。与精神卫生中心等医院以科研项目合作的方式,获取的诊疗数据必须经过脱敏处理,此部分数据获取难度较大,得到的数据信息较为有限。

(4)互联网数据资源。通过网络爬虫技术选取较权威的网站进行数据爬取,此部分数据量巨大,不可避免地存在部分不可靠数据,是动态补充情感障碍症语料库的重要数据来源。

将以上多源头数据经过数据清洗、分割切片、数据整合、数据转换等数据处理操作[11],得到归一化的情感障碍症语料库 Normal,达到进行信息抽取等后续步骤的初始条件。

1.3 信息抽取和知识融合阶段

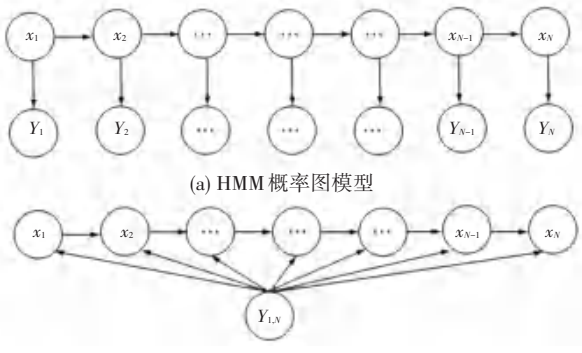
信息抽取阶段包含实体抽取、属性抽取、关系抽取等任务单元[12]。实体抽取是从 Normal 语料库中提取出命名实体;属性抽取是从语料库中抽取实体的属性信息并构造实体的属性列表;关系抽取是从语料库中进一步提取各个实体的关联关系。信息抽取最终目标是输出三元组集{RDF},RDF 为单个<实体 A,属性/关系,实体 B>。本文综合最大熵马尔可夫模型(Maximum Entropy Markov Model,MEMM)在动态文本分类领域精准有效的特点和卷积神经网络(Convolutional Neural Network,CNN)在特征提取和数据分类领域的优异表现,提出了一种混合最大熵马尔可夫模型-卷积神经网络(MEMM-CNN)的信息抽取方法。

知识融合阶段核心任务是三元组的极性识别(可信为 1,不可信为-1),主要包括实体对齐、属性对齐、极性识别、极性更新等任务单元[13]。实体对齐主要用于消除实体冲突、实体命名多元指代等问题[14]。属性对齐主要是消除实体的属性列表中元素存在命名冲突、指代冲突等问题。由于 Normal 语料库中数据来源众多,可靠数据和不可靠数据交织在一起,导致数据抽取阶段输出的三元组集{RDF}中可能存在极性结果相反的三元组,所以针对三元组的极性识别计算很重要。本文设计了一种基于 ρ 值的自适应加权估计知识融合算法(SAW),用来完成三元组集中三元组的极性计算,可以最大限度的输出可信三元组集。

2 混合 MEMM-CNN 模型

2.1 HMM 和 MEMM 模型

隐马尔可夫模型(Hidden Markov Model,HMM)和最大熵马尔可夫模型(Maximum Entropy Markov Model,MEMM)都属于经典的马尔可夫模型,可以广泛应用在语音识别、词性自动标注、知识抽取、概率文法等各个自然语言处理等领域[15]。HMM 优点是计算简单,缺点是只依赖于每一个状态和其对应的观察对象,目标函数和预测目标函数不匹配。而最大熵马尔可夫模型(MEMM)是在隐马尔可夫模型的基础上应用最大熵模型思想,将一个概率生成模型转化为概率判别模型,结合上下文依赖,通过直接判别减少建模负担。HMM 和 MEMM 概率图模型如图 2 所示。



(a) HMM 概率图模型

(b) MEMM 概率图模型

图 2 HMM 和 MEMM 概率图模型

Fig. 2 HMM and MEMM probability graph models

MEMM 条件概率 $P(X \mid Y)$ 公式如下式:

$$P(X \mid Y) = \prod_{i=1}^{N} P(y_i \mid y_{i-1}, x_{1:N}) \tag{1}$$

其中,X 为状态变量,Y 为观测变量,$P(y_i \mid y_{i-1}, x_{1:N})$ 通过最大熵分类器建模,具体计算如下:

$$P(y_i \mid y_{i-1}, x_{1:N}) = \frac{\exp\left(\sum_{k=1}^{M} \lambda_k f_k(y_i, y_{i-1}, x_{1:N})\right)}{Z(y_{i-1}, x_{1:N})} \tag{2}$$

其中,$f_k(y_i, y_{i-1}, x_{1:N})$ 为特征函数;$Z(y_{i-1}, x_{1:N})$ 为归一化函数;λ_k 是特征函数的权重;M 表示特征函数的个数。

合并公式(1)和公式(2),得到下式:

$$P(X \mid Y) = \prod_{i=1}^{N} \frac{\exp\left(\sum_{k=1}^{M} \lambda_k f_k(y_i, y_{i-1}, x_{1:N})\right)}{Z(y_{i-1}, x_{1:N})} \tag{3}$$

2.2 MEMM-CNN 模型

2.2.1 模型概述

混合 MEMM-CNN 模型综合了 MEMM 和 CNN 在文本分类和信息抽取领域的各自优势。在训练阶段首先使用 MEMM 训练语料库数据集,然后针对 MEMM 存在实体关系抽取代价高、可扩展性差等弱点,使用 CNN 进一步训练模型。CNN 通常由输入

层，卷积层，池化层，全连接层，SoftMax 层等构成[16]。卷积层用来提取特征，池化层用来保留主要特征从而达到降维的目的，同时防止过拟合，提高模型的泛化能力。在 CNN 中使用两个卷积层和两个池化层有助于提高训练速度，改善精度[17]。在测试阶段，采用训练阶段同样“先 MEMM 后 CNN”分类方法，完成信息抽取中的实体抽取、属性抽取、关系抽取 3 个任务单元，输出最终的三元组集合。

2.2.2　训练阶段

从情感障碍症 Normal 语料库中抽取数据构建原始训练数据集，建立最优训练模型并输出结果。混合 MEMM-CNN 训练阶段如图 3 所示，具体步骤如下：

步骤 1　从情感障碍症 Normal 语料库中抽取数据，经过数据处理，包括数据清洗、数据标准化和数据分割等，使其达到训练要求，得到训练数据集 Normal $-x$；

步骤 2　使用 MEMM 训练 Normal $-x$，尝试完成信息抽取单元任务；

步骤 3　MEMM 通过梯度下降法训练权值，使 Normal $-x$ 的对数似然值 L 达到最大，输出中间结果三元组集 $\{RDF(1)-x\}$；

步骤 4　使用 CNN 进一步训练模型；

步骤 5　重复步骤 2～步骤 4 迭代过程直至完成信息抽取中的 3 个任务单元：实体抽取、属性抽取、关系抽取；

步骤 6　输出三元组集 $\{RDF(2)-x\}$，即为 MEMM-CNN 训练最终输出结果。

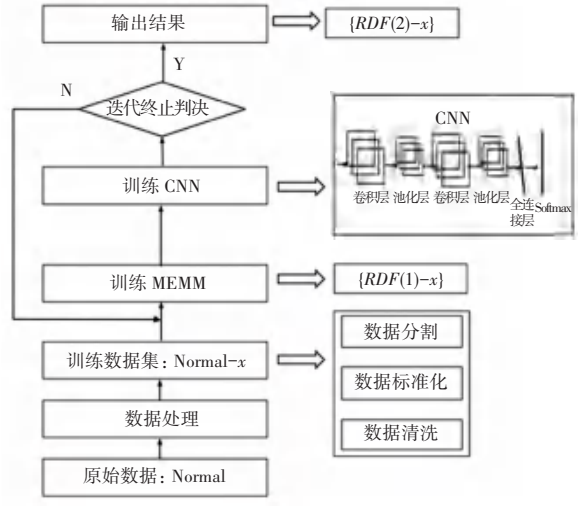


图 3　混合 MEMM-CNN 训练阶段

Fig. 3　Training stage of hybrid MEMM-CNN

2.2.3　测试阶段

在测试阶段，使用情感障碍症 Normal 语料库中未使用过的测试数据集应用于训练阶段建立的模型。采用跟训练阶段相同流程，首先对原始数据进行数据处理得到测试数据集 Normal $-c$，在 MEMM 阶段进行第一次分类，输出三元组集 $\{RDF(1)-c\}$；在 CNN 阶段完成第二次重新分类，输出三元组集 $\{RDF(2)-c\}$，重复迭代上述过程最终完成信息抽取的任务单元，三元组集 $\{RDF(2)-c\}$ 即为混合 MEMM-CNN 模型的最终输出结果。

2.3　信息抽取任务单元

实体抽取是从情感障碍症 Normal 语料库中提取出命名实体 Q，比如情感障碍症类型、治疗药物名称、临床症状名称等。

属性抽取是从 Normal 语料库中抽取实体的属性信息，构造实体 Q 的属性列表 $L=\{L_1,\cdots,L_i\}$，实体 Q 和属性列表 L 中任意元素 L_i 构成二元组$<Q,L_i>$，即<实体，属性>。比如治疗双向情感障碍药物的注意事项、副作用和禁忌等可以构成的二元组关系有<治疗双向情感障碍药物，注意事项>，<治疗双向情感障碍药物，副作用>，<治疗双向情感障碍药物，禁忌>。经过 MEMM 和 CNN 两个步骤，抽取得到二元组$<Q,L_i>$，再次去匹配 Normal 语料库，可以得到第一部分三元组<实体 A，属性，实体 B>，放入三元组集 $\{RDF\}$ 中。

关系抽取是从 Normal 语料库进一步提取各个实体的关联关系。经过 MEMM 和 CNN 两个步骤，将实体间的关系抽取出来，存放到关系库中；通过将多种零散的实体链接起来，从而得到第二部分三元组<实体 C，关系，实体 D>，放入三元组集 $\{RDF\}$ 中。

第一部分三元组<实体 A，属性，实体 B>和第二部分三元组<实体 C，关系，实体 D>经过归一化和去重处理后，得到最终的三元组集 $\{RDF\}$，三元组 RDF 示例见表 1。

表 1　三元组 *RDF* 示例

Table 1　Example of triplet *RDF*

实体 A	关系/属性	实体 B
双相情感障碍	典型症状	狂躁和抑郁交替
碳酸锂片	治疗	双相情感障碍
盐酸帕罗西汀片	副作用	恶心
心理治疗	具体方法	森田疗法

3　基于 ρ 值的自适应加权估计知识融合算法(SAW 算法)

3.1　实体对齐和属性对齐

由于情感障碍症语料库 Normal 数据来源包括

医学书籍、医学论文、医院电子病历、互联网数据资源等,以上数据源都不可避免的会出现知识重复、冗余、歧义、关联性冲突等问题,尤其是从互联网上爬取到的数据资源很可能充斥着不可靠和错误数据信息,可能导致信息抽取阶段输出的三元组集 $\{RDF\}$ 出现极性识别错误,所以针对三元组极性识别算法的设计尤为关键。

实体对齐主要用于消除实体冲突、实体命名多元指代等问题。例如常用来治理重度抑郁症的药物“盐酸帕罗西汀片”,其别名和通用名还包括“赛乐特”、“乐友”、“舒坦罗”、“Paroxetine”等,但其实都是指代同一种药物,这时就需要通过实体对齐来消除实体名称指代冲突。

属性对齐主要是消除实体 Q 的属性列表 $L=\{L_1,\cdots,L_i\}$ 中元素存在命名冲突、指代冲突的问题,完成实体对齐和属性对齐后更新三元组集 $\{RDF\}$,得到三元组集 $\{RDG\}$。

3.2 三元组极性识别

对 $\{RDG\}$ 中三元组计算可信度 R,可信度函数 R_q^Q 计算公式:

$$R_q^Q=\sum_{i=1}^{N}\alpha_i A_i^q+\sum_{t=1}^{M}\beta_t B_t^q \tag{4}$$

其中,Q 表示目标三元组集中关系三元组的总数量;R_q^Q 表示三元组的总数量为 Q 的目标三元组集中第 q 个三元组的可信度;N 表示第 q 个三元组所在的三元组子集中正向极性三元组的数量;M 表示第 q 个三元组所在的三元组子集中负向极性三元组的数量;A_i^q 表示第 q 个三元组所在的正向极性三元组子集 i 的极性值;B_t^q 表示第 q 个三元组所在的正向极性三元组子集 t 的极性值;α_i 表示 A_i^q 的可信度权值;β_t 表示 B_t^q 的可信度权值。

根据公式(4)计算出的 R_q^Q 值大于等于可信度阈值 ρ,则三元组 q 极性置为 1,三元组 q 放进可信库 D_1 中;R_q^Q 值小于可信度阈值 ρ,则三元组 q 极性置为-1,三元组 q 放入不可信库 D_2 中。

可信度阈值 ρ 的设定将直接影响到三元组进入到可信库 D_1 或不可信库 D_2 的最终概率,所以在可信度函数 R_q^Q 的基础上,设计了与三元组的数据来源有关的自适应加权可信度阈值估计方法,具体的可信度阈值的函数公式如下:

$$\rho_q=\begin{cases}\dfrac{N([\alpha]_{q-\max}+[\beta]_{q-\min})}{2(N+M)}, & M\geqslant N\\ \dfrac{M([\alpha]_{q-\max}+[\beta]_{q-\min})}{2(N+M)}, & M<N\end{cases} \tag{5}$$

其中,ρ_q 表示第 q 个三元组可信度阈值;$[\alpha]_{q-\max}$ 表示第 q 个三元组所在三元组子集中正向极性三元组的最大可信度权值;$[\beta]_{q-\min}$ 表示第 q 个三元组所在的三元组子集中负向极性三元组的最小可信度权值。

3.3 算法描述

基于 ρ 值的自适应加权估计知识融合算法(SAW)描述见表 2。

表 2 基于 ρ 值的自适应加权估计知识融合算法(SAW)描述

Table 2 Description of adaptive weighted estimation knowledge fusion algorithm based on ρ-value(SAW)

SAW Algorithm

输入 三元组集 $\{RDF\}$, 可信度阈值 ρ, 语料库 Normal

输出 动态可信库 D_1, 不可信库 D_2

```
1. Begin;
2. Initialization;  //初始化
3. for E_i ∈ E ← aligned;  //实体对齐操作
4.    for L_i ∈ L ← aligned;  //属性对齐操作
5.       RDG = update (RDF);  //更新 RDF 三元组
6. end
7. for q in RDG
8.       R_q^Q = Σ_{i=1}^{N} α_i A_i^q + Σ_{t=1}^{M} β_t B_t^q ;  //计算关系三元组可信度 R 值
9.       if ( R ≥ ρ_q ) then  // R 值大于等于可信度阈值 ρ_q
10.          p_q = 1;  //三元组 q 极性置为 1
11.          D_1 ← q;  //三元组 q 放进可信库 D_1 中
12.      else if ( R < ρ_q ) then  // R 值小于可信度阈值 ρ_q
13.          p_q = -1; //三元组 q 极性置为-1
14.          D_2 ← q; //三元组 q 放进不可信库 D_2 中
15.      end if
16. end
17. if Normal←Normal∪{N_update} then//语料库动态更新
18.      update RDG ;  //更新 RDG 三元组
19.      start the iteration process;  //迭代过程;
20.      update D_1 ,D_2 ;  //更新动态可信库 D_1 , 不可信库 D_2
21. end
```

由于情感障碍症知识图谱的数据来源是动态更新的,情感障碍症语料库 Normal 中的内容也在不断更新。为提高三元组集 $\{RDG\}$ 极性识别结果的准确度和有效性,SAW 算法采取定时迭代执行策略,即每间隔 24 个小时重新计算三元组可信度 R 值,如果 R 值与可信度阈值 ρ 比较关系发生变化,则自适应更新可信库 D_1 和不可信库 D_2。

3.4 对比实验

选取可信库 D_1 中三元组集数量、三元组极性识别准确率、算法产生的噪声值、算法响应时间等评价指标来设计知识融合算法对比实验,对比实验结果如图 4~图 7 所示。可以看出 SAW 算法相比较于 VOTE (VOTE Algorithm)[18]、ACCU (Association -

based Conceptual Clustering and Updating Algorithm)[19]等传统数据融合算法表现出更好的性能，当三元组集 $\{RDG\}$ 中三元组数量为 8 000 时，SAW 算法输出的可信库 D_1 中三元组数量达到 7 792 个，算法识别三元组极性准确率达到 98.5%，显著高于 VOTE、ACCU 算法，同时 SAW 算法能够有效地抑制噪声，响应时间最短。

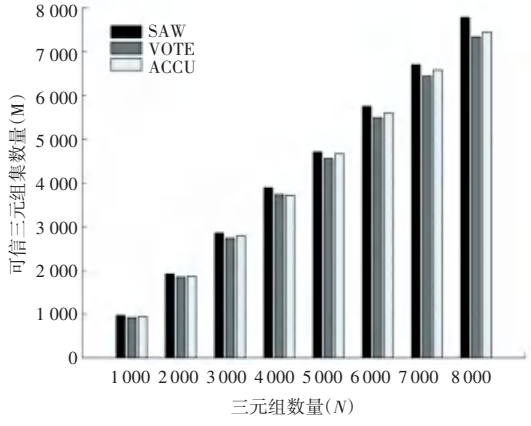


图 4　可信三元组集数量对比

Fig. 4　Comparison of the number of trusted triplet sets

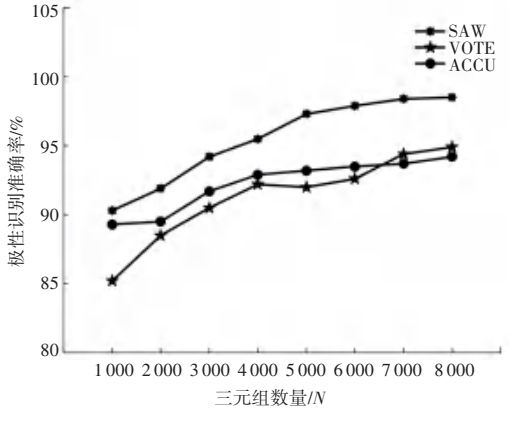


图 5　三元组极性识别准确率对比

Fig. 5　Comparison of accuracy in identifying triplet polarity

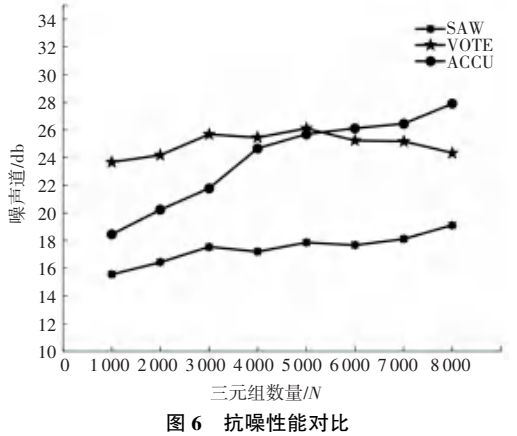


图 6　抗噪性能对比

Fig. 6　Comparison of noise resistance performance

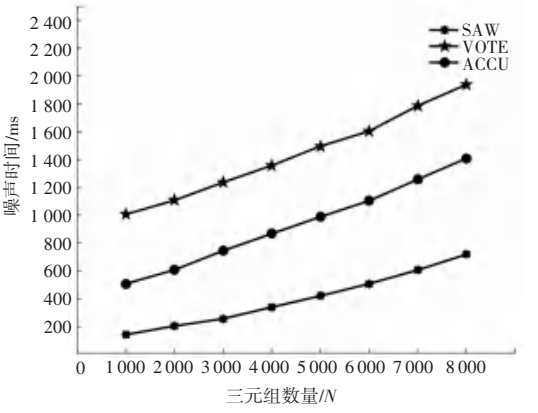


图 7　响应时间对比

Fig. 7　Comparison of response time

4　知识图谱存储与可视化

4.1　知识图谱存储

Neo4j 作为一种流行的开源的、易扩展的图形数据库，既能存储节点信息，又能直观地呈现实体之间的各种关系，是目前主流的知识图谱可视化构建工具[20]。本文借助 Neo4j 生成最终的情感障碍症知识图谱，情感障碍症知识图谱主要由实体节点信息表(EntityList)和关系信息表(RelationList)构成，其中实体节点信息表由 id(编码)和 name(名称)等字段构成，不同实体节点之间形成的链接关系构成关系信息表，由不同实体节点 id 等字段构成。

例如治疗药物实体节点(Medicine Entity)信息表和药物副作用实体节点(Reaction Entity)信息表示见表 3。治疗药物实体节点和药物副作用实体节点之间形成链接关系，形成的关系信息表中共构成 9 个三元组，见表 4。

表 3　实体节点信息表示例

Table 3　Example of entity node information representation

Entity 名称	来源数据表	id 字段	name 字段
治疗药物	Medicine. csv	301	碳酸锂片
		302	盐酸帕罗西汀片
药物副作用	Reaction. csv	401	恶心
		402	口干
		403	多尿
		404	头晕
		405	便秘
		406	乏力

表 4 关系信息表示例

Table 4 Example of relationship information representation

Entity1	Entity2	关系	Entity1 字段 id	Entity2 字段 id	构成三元组
治疗药物（Medicine）	药物副作用（Reaction）	治疗药物的副作用 Medicine_to_Reaction(SE)	301	401	<301,SE,401> <碳酸锂片,副作用,恶心>
			301	402	<301,SE,402> <碳酸锂片,副作用,口干>
			301	404	<301,SE,404> <碳酸锂片,副作用,头晕>
			301	405	<301,SE,405> <碳酸锂片,副作用,便秘>
			301	406	<301,SE,406> <碳酸锂片,副作用,乏力>
			302	401	<302,SE,401> <盐酸帕罗西汀片,副作用,恶心>
			302	402	<302,SE,402> <盐酸帕罗西汀片,副作用,口干>
			302	403	<302,SE,403> <盐酸帕罗西汀片,副作用,多尿>
			302	404	<302,SE,404> <盐酸帕罗西汀片,副作用,头晕>

4.2 知识图谱可视化

借助 Cypher 语言，将表 3 中 Medicine Entity、Reaction Entity 两个实体节点信息和表 4 中 SE 关系信息导入 Neo4j 生成知识图谱，可视化效果示例如图 8 所示。

本文经过多源数据语料库构建、信息抽取、知识融合等步骤共获取情感障碍症知识图谱相关有效实体节点共 1 409 个，有效链接关系 185 个，有效三元组数量 9 240 个。利用 Neo4j 生成的情感障碍症知识图谱（部分）如图 9 所示。

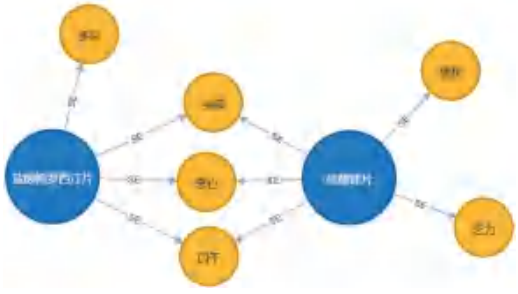


图 8 <Medicine,SE,Reaction>知识图谱示例

Fig. 8 Example of <Medicine,SE,Reaction> knowledge graph

图 9 Neo4j 生成的情感障碍症知识图谱（部分）

Fig. 9 Knowledge graph of affective disorder generated by Neo4j (partial)

图 9 中包含的实体节点有疾病类型(Type Entity)、疾病症状(Symptoms Entity)、治疗药物(Medicine Entity)、药物副作用(Reaction Entity)、治疗方案(Plantreat Entity)、心理治疗具体方法(Xinli Entity)等。实体节点之间相互构成的链接关系包括 SE(治疗药物的副作用)、TM(药物治疗的疾病类型)、DT(药物治疗对应的药物名称)、PT(治疗方案对应的疾病类型)、XT(心理治疗对应的心理治疗方法)、SYM(疾病类型对应的疾病症状)等。各实体节点之间通过关系链接最终构成可视化的网状交织结构，并不断向外拓展。随着数据源的补充更新，相关实体节点个数和链接关系个数也将增加，生成的情感障碍症知识图谱也将动态变化。

5　结束语

本文重点研究了情感障碍症知识图谱构建过程中自顶向下的 4 个关键技术。采用多源头数据构建情感障碍症知识语料库，保障了数据来源的丰富性和全面性；信息抽取阶段提出的混合 MEMM-CNN 模型和知识融合阶段提出的 SAW 算法均具备较强的创新性和实用性；最后利用 Neo4j 图形数据库动态生成了精细化的情感障碍症知识图谱，该知识图谱目前可以应用于对情感障碍症的普适性教育、情感障碍症患者自我诊断、辅助医生临床决策等，同时本文取得的研究成果为未来情感障碍症实现智能科普、智能诊疗等提供了支撑。

参考文献

[1] 张吉祥，张祥森，武长旭，等. 知识图谱构建技术综述[J]. 计算机工程，2022，48(3)：15.

[2] 魏晓，王晓鑫，陈永琪，等. 基于自然语言处理的材料领域知识图谱构建方法[J]. 上海大学学报：自然科学版，2022，28(3)：13.

[3] 郑增亮，蔡晓琼，苏前敏，等. 知识图谱在医学领域的研究现状分析[J]. 智能计算机与应用，2023，13(5)：32-39.

[4] 王彩云，郑增亮，蔡晓琼，等. 知识图谱在医学领域的应用综述[J]. 生物医学工程学杂志，2023，40(5)：1040-1044.

[5] 尹一淑，刘军莲，王佳平，等. 抑郁症相关发病机制研究进展[J]. 医学综述，2022，28(12)：2368-2372.

[6] LAN G，HU M，LI Y，et al. Contrastive knowledge integrated graph neural networks for Chinese medical text classification[J]. Engineering Applications of Artificial Intelligence，2023，122：106057.

[7] 孙梦雅. 基于多源数据的房颤知识图谱构建与研究[D]. 郑州：郑州大学，2022.

[8] 王松，李正钧，杨涛，等. 中医药知识图谱研究现状及发展趋势[J]. 南京中医药大学学报，2022，38(3)：272-278.

[9] 赵雪娇. 妇产科知识图谱构建研究与实现[J]. 中国数字医学，2019，14(1)：3-5.

[10] LI X，ZHANG J，FAN L，et al. Construction and analysis of knowledge graphs for multi-source heterogeneous data of soil pollution[J]. Soil Use and Management，2023，39(3)：1036-1039.

[11] ZHANG X，HUANG X，XU W. Matrix-based multi-granulation fusion approach for dynamic updating of knowledge in multi-source information[J]. Knowledge-Based Systems，2023(28)：1-21.

[12] 曾江峰，庞雨静，高鹏钰，等. 基于 Lattice LSTM 的中医药古文献命名实体识别与应用研究[J]. 情报工程，2023，9(5)：112-122.

[13] 胡晨馨. 面向医学知识图谱构建的多源知识融合方法研究[D]. 郑州：郑州大学，2022.

[14] 迟棠，车超. 融合迭代式关系图匹配和属性语义嵌入的实体对齐方法[J]. 计算机科学，2023，50(S2)：81-86.

[15] AZERAF E，MONFRINI E，VIGNON E，et al. Hidden markov chains，entropic forward-backward，and part-of-speech tagging[J]. arXiv preprint arXiv，2005.10629，2020.

[16] 龙慧，马家庆，吴钦木，等. 基于小波变换 CNN 的电机运行状态识别研究[J]. 智能计算机与应用，2023，13(5)：122-125.

[17] 张振，张师榕，赵转哲，等. 混合 CNN-HMM 的人体动作识别方法[J]. 电子科技大学学报，2022，51(3)：444-451.

[18] UPADHYAY D，MANERO J，ZAMAN M，et al. Intrusion detection in SCADA based power grids：Recursive feature elimination model with majority vote ensemble algorithm[J]. IEEE Transactions on Network Science and Engineering，2021，8(3)：2559-2574.

[19] JIN C，ZHOU Y，YING S，et al. A knowledge-fusion ranking system with an attention network for making assignment recommendations[J]. Computational Intelligence and Neuroscience，2020(4)：6748430.

[20] 袁丹灵. 基于 Neo4j 的道路交通安全知识图谱构建及应用[D]. 长沙：中南大学，2023.