Operations Research and Fuzziology 运筹与模糊学, 2024, 14(5), 230-237
Published Online October 2024 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2024.145465

Hans汉斯

一种基于相似度计算的实体数据关系归属方法

冒鸿宇，孙刘杰*，朱衍熹

上海理工大学出版学院，上海

收稿日期：2024年8月6日；录用日期：2024年9月20日；发布日期：2024年10月10日

摘　要

特定领域的数据蕴含了大量有价值的知识及关系划分，从中能正确将其进行关系划分一直是一个值得关注的话题。当前关系划分都依赖于大量样本模型进行训练得出，由于特定领域的实体数据关系样本数量较少，显然应用到特定领域中存在局限。因此本文针对该问题，提出一种基于相似度计算的实体数据关系归属方法，其中建立一个特定领域的少样本实体关系术语树，与待划分的实体数据进行相似度计算得到在树中具体位置，从而解决错误归属问题，显著减少人工管理成本，能够有效提升系统的可用性。

关键词

相似度计算，实体关系，实体归属，实体数据

A Relationship Attribution Method for Entity Data Based on Similarity Calculation

Hongyu Mao, Liujie Sun*, Yanxi Zhu

College of Publishing, University of Shanghai for Science and Technology, Shanghai

Received: Aug. 6th, 2024; accepted: Sep. 20th, 2024; published: Oct. 10th, 2024

Abstract

Domain-specific data contains a large amount of valuable knowledge and relationship delineation, from which it is always a topic of interest to be able to correctly perform relationship delineation. Currently, the relationship classification relies on a large number of sample models for training, due to the small number of domain-specific entity data relationship samples, it is obvious that there are limitations in applying to specific domains. Therefore, in this paper, we propose a similarity-based relationship attribution method for entity data, in which a domain-specific entity relationship term

*通讯作者。

文章引用: 冒鸿宇, 孙刘杰, 朱衍熹. 一种基于相似度计算的实体数据关系归属方法[J]. 运筹与模糊学, 2024, 14(5): 230-237.DOI: 10.12677/orf.2024.145465

tree with few samples is established, and the entity data to be partitioned is similarity-calculated to get the specific position in the tree, thus solving the problem of misattribution, significantly reducing the cost of manual management, and effectively improving the usability of the system.

Keywords

Similarity Calculation, Entity Relationship, Entity Attribution, Entity Data

1. 引言

随着大数据的高速发展，文本数据呈指数暴增，对于某种特定领域两篇文章内容的关系及关系归属确定，一直是一个值得关注的话题。关系归属需要进行实体抽取[1]、关系抽取[2]、关系划分等步骤去完成。在关系归属问题上通常采用的方式是通过机器学习、深度学习等进行训练样本得到抽取好的实体，从而进行实体关系抽取并划分，如成[3]等人根据互联网医疗健康平台用户生成的大量复杂信息内容实现语义发现与关系揭示，研究构建了基于改进 Casrel 实体关系抽取模型的在线健康信息语义发现模型，并在文本编码层引入更适用于医疗健康领域的 ERNIE-Health 预训练模型，有效提高了模型对在线健康信息语义发现的实体识别和实体关系抽取任务的 F1 值，但还需要数据集的扩充及疾病类型的健康信息的实证；王[4]等人针对金融文本语义特征难以准确提取的问题，提出使用预训练模型 FinBERT 对输入金融文本进行字、词粒度特征提取得到句子表示的方法；刘[5]等人发现土家民族器乐文本中实体位置、类型与实体关系具有强相关性特征，提出融合实体位置与类型特征的实体关系抽取模型，在完成命名实体识别任务后，通过特征拼接进行特征学习，最后进行关系分类学习。通过上述几个领域的文本实体抽取和关系划分得出，都是需要实体抽取和关系划分融合在一起直接放入模型进行训练得出关系，考虑到特定领域中的数据集样本数较少及不公开使用等特点，本文提出一种基于相似度计算的实体数据关系归属方法，将关系划分问题进行分而治之，即将实体抽取和关系划分进行模块化处理，并融入特殊领域的数据集的特点，根据少数样本建立一个特定领域的术语树，将待匹配的数据与术语树数据进行相似度计算[6]得出术语树中最匹配的节点关系划分，从而解决因为少样本导致错误归属问题，显著减少人工管理成本，能够有效提升系统的可用性。

本文主要贡献如下：

1) 在特定领域实体数据中缺乏大量样本的情况下构建少样本关系的术语树索引结构。旨在不依赖于大量数据进行样本训练的问题，对特定领域保密性及特殊性具有重要意义。

2) 在关系待划分实体数据中进行术语树节点相似度计算得到带划分实体数据在术语树中的位置，从而进行准确关系归属划分。为如何不依赖于大量实体数据关系进行有效划分的课题中提供了新视角。

3) 在通用数据上进行验证研究所提技术可以对关系待划分实体数据进行及时有效的分析处理，从而将提高准确率，同时减少人工成本。

2. 方法流程

实体关系归属方法明确核心模块，即实现对领域数据中实体的精准识别与归属判定。为实现这一目标，设计多个相互独立又紧密协作的模块，每个模块负责处理特定的任务，如数据处理、数据关系确立、数据关系显示等。

在数据处理模块中，利用了预训练模型等进行文章数据的实体抽取并用 MySQL [7]数据库进行存储。此外在实体中进行术语树的建立，通过额外的术语树进行关系归属。

在数据关系确立模块中，由于提前建立了额外的术语树，因此只需要考虑已经抽取好的实体与术语树中的术语的匹配程度。在该方法中使用了多种计算相似度方法进行融合充分表达两者之间的相似程度进行关系确立。

在数据关系显示模块中，将两个实体进行关系组成存储成 JSON 格式并发送至前端界面，并通过关系名词进行连线组成小型文章网络图谱。具体方法流程如图 1 所示。

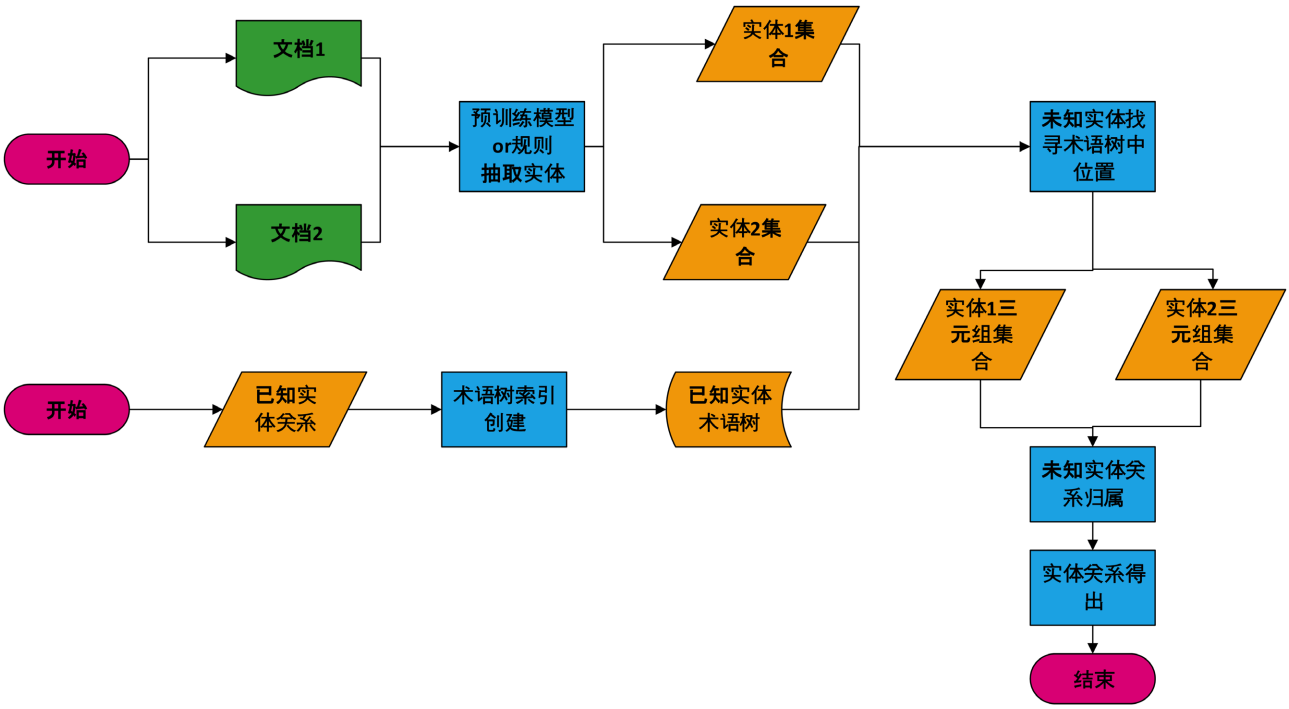

Figure 1. Framework flowchart
图 1. 框架流程图

3. 实体处理

3.1. 实体抽取

实体抽取结果的好坏直接决定之后进行实体关系归属确立操作效果的准确率。首先采用基于规则的方法进行一系列规则来识别文章中的实体成分，但考虑到基于规则的方法需要基于领域知识和数据特性进行系统确认，才能够准确地将文章中实体进行抽取出。随着数据的不断扩充，该方法的缺点逐渐显示，因此引入了机器学习算法来优化实体抽取的性能。通过训练模型，可以让系统自动学习并抽取实体。但考虑领域数据的不透明性与特殊性，文章将用最简单方法进行实体抽取。

3.2. 术语树建立

由于实体众多，但基于特定领域，如果是基于深度学习的方法进行实体归属，没有特定的数据集不可以集中训练，因此将已知的实体分类进行术语树索引结构表示，计算之后即可知道未知实体对已知实体关系，通过已知关系进一步判别归属。术语树索引结构创建示意图如图 2，步骤如下。

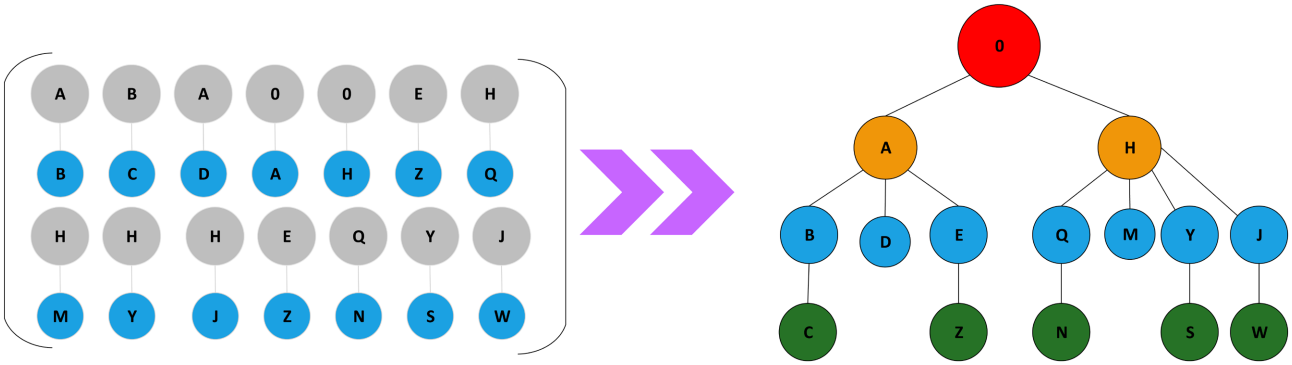

Figure 2. Schematic diagram for building a terminology tree
图 2. 术语树建立示意图

算法 1 术语树索引结构算法

输入：两个已知实体关系归属表
输出：已知实体关系归属索引结构

步骤：

1) 读取数据库存储的两张已知实体数据关系表三元组[8] (**ID , Entity, ParentID**)

2) **For each** 实体数据 **DO**

3) 根据四元组(**ID, Entity, ParentID, GrandID**)将两张表进行合并成一个实体关系归属索引结构

4) **repeat**

5) **util** 实体数据关系归属合并完

6) 将实体关系归属索引结构至内存备用

7) **end for**

8) **end**

3.3. 实体相似度计算

由于大部分实体相似度很高，难以分辨种类且没有一个具体数据集进一步提高准确性和效率。因此在关系归属确立中，我们根据该领域的实体数据特征，使用了多种计算相似度方法融合进行相似度计算从而进行关系归属。

1) 相似性度量[9]-[11]：即综合评定两个事物之间相近程度的一种度量。两个事物越接近，它们的相似性度量也就越大，而两个事物越疏远，它们的相似性度量也就越小。相似性度量的给法种类繁多，一般根据实际问题进行选用。常用的相似性度是有：相关系数(衡量变量之间接近程度)，相似系数(衡量样品之间接近程度)，若样品给出的是定性数据，这时衡量样品之间接近程度，可用样本的匹配系数、一致度等。用数量化方法对事物进行分类，就必须用数量化方法描述事物间的相似程度。一个事物常常需要用多个变量来刻画，如对一群用 p 个变量描述的样本点进行分类，则每个样本点可看成是 p 维空间的一个点，很自然的想到用距离来度量样本点间的相似程度。

相似性的度量方法很多，有的用于专门领域，也有的适用于特定类型的数据，如何选择相似性的度量方法是一个相当复杂的问题，考虑到该领域的特殊性：没有情感倾向和区分度不高的特点，本文利用

编辑距离[6]和 Jaccard [12]系数融合进行实体关系归属。

2) 编辑距离(Edit Distance)，也称为 Levenshtein 距离，是衡量两个字符串之间相似程度的指标。它表示将一个字符串转换成另一个字符串所需的最少编辑操作次数。常见的编辑操作包括插入一个字符、删除一个字符、替换一个字符。用 $ed_{a,b}(i,j)$ 来表示 a,b 字符串的编辑距离，其中 i 代表 a 的长度，j 代表 b 的长度公式如下：

$$ed_{a,b}(i,j)=\begin{cases}\max(i,j) & if\ \min(i,j)=0\\ \min\begin{cases}ed_{a,b}(i-1,j-1)+1\\ ed_{a,b}(i-1,j)+1\\ ed_{a,b}(i,j-1)+1\end{cases} & \text{otherwise}\end{cases}$$

$ed_{a,b}(i,j)$ 表示 a 的前 i 个字符串与 b 的前 j 的字符串的编辑距离。其中 i 和 j 都是从 1 开始的下标。$ed_{a,b}(i,j)$ 值的描述如下：

- $\min(i,j)=0$ 代表有一个字符串为空，编辑距离就是另一个非空字符串的长度；
- 当 $\min(i,j)\neq 0$ 的时候，编辑距离如下三种情况的最小值：

  $ed_{a,b}(i-1,j-1)+1$ 删除 a_i；

  $ed_{a,b}(i-1,j-1)+1$ 插入 b_j；

  $ed_{a,b}(i-1,j-1)+1$ 替换 b_j。

3) Jaccard 系数：杰卡德相似系数(Jaccard similarity coefficient)：两个集合 A 和 B 的交集元素在 A，B 的并集中所占的比例，称为两个集合的杰卡德相似系数，用 $J(A,B)$ 表示：

$$J(A,B)=\frac{|A\cap B|}{|A\cup B|}=\frac{|A\cap B|}{|A|+|B|-|A\cap B|}$$

当集合 A，B 都为空时，$J(A,B)$ 定义为 1。

杰卡德距离(Jaccard Distance)：与杰卡德相似系数相反，用两个集合中不同元素占所有元素的比例来衡量两个集合的区分度，杰卡德距离越大，两个样本相似度越低。

$$d_j(A,B)=1-J(A,B)=\frac{|A\cup B|-|A\cap B|}{|A\cup B|}=\frac{|A\Delta B|}{|A\cup B|}$$

其中对参差(symmetric difference)：$A\Delta B=|A\cup B|-|A\cap B|$。

利用多种计算相似度融合算法步骤如下。

算法 2 多种计算相似度融合算法

输入：两个未知实体
输出：两个未知实体对于已知实体的关系归属索引结构

步骤：

1) **For each** 未知实体 **DO**

2) **For each** 已知实体 **DO**

3) 将已知实体和未知实体进行编辑距离计算相似度

4) 将已知实体和未知实体进行 **Jaccard** 系数计算相似度

5) 两者相似度进行权重相加得到综合相似度

续表

6) **end for**

7) 将未知实体与综合相似度最高的已知实体的关系信息进行赋值

8) **end for**

9) **end**

3.4. 实体关系归属确立

在得到未知实体与已知实体关系匹配之后，就需要将文档中所有的未知实体进行关系归属，其原理是根据现有已知的实体关系归属确定未知实体的关系归属具体算法流程如下图 3 所示，步骤如下表。

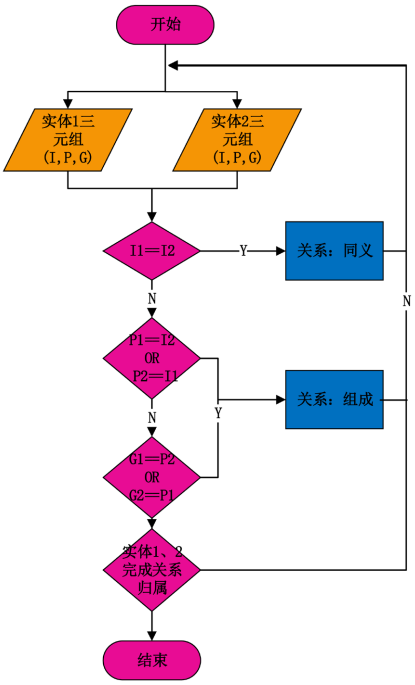

Figure 3. Relationship determination attribution flowchart
图 3. 关系确定归属流程图

算法 3 实体关系归属算法

输入：两个未知实体信息
输出：两个未知实体的关系归属

步骤：

1) 输入所有实体信息

2) **For each** 两个实体信息 **DO**

3) 两个实体的 **ParentID,GrandID** 进行对比：

4) **IF ParentID1= ParentID2**：

5) 实体关系：同义

续表

6) **ELSE IF GrandID1= ParentID2 OR ParentID1 =GrandID2**:

7) 实体关系：组成

8) 以 **JSON** 格式存储关系信息

9) **end for**

10) **end**

4. 关系归属显示

将实体关系归属完成之后存入 JSON 格式，并通过 Restful 接口进行实现前后端解耦，显示在界面上。前端使用的是 Vue，拿到 JSON 数据之后进行关系显示，实体作为节点关系作为边进行连线，显示其实体关系归属图。

5. 案例研究

根据所述方法，本文对实体关系归属方法进行系统编写代码，并利用两篇文档进行测试，测试结果准确率较高。多次测试两篇文档在该系统实体处理及关系显示的平均用时为 25 s，相较于人为进行标注实体、人为进行实体关系归属及图表展现的平均时间 30 min，降低至少 90%以上。两篇文章的输入如下图 4 所示，图 4 高亮部分为识别到的实体，关系输出结果如下图 5 所示。

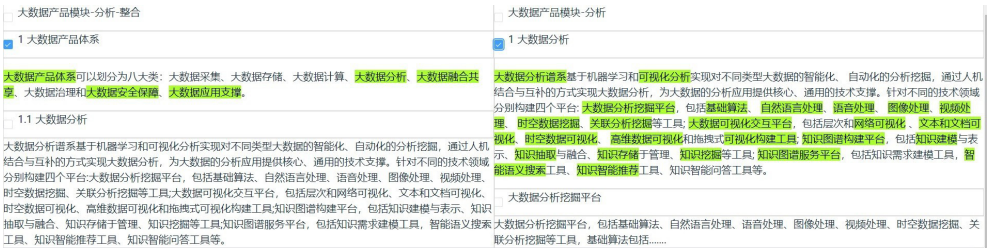

Figure 4. Two documents input
图 4. 两篇文档输入

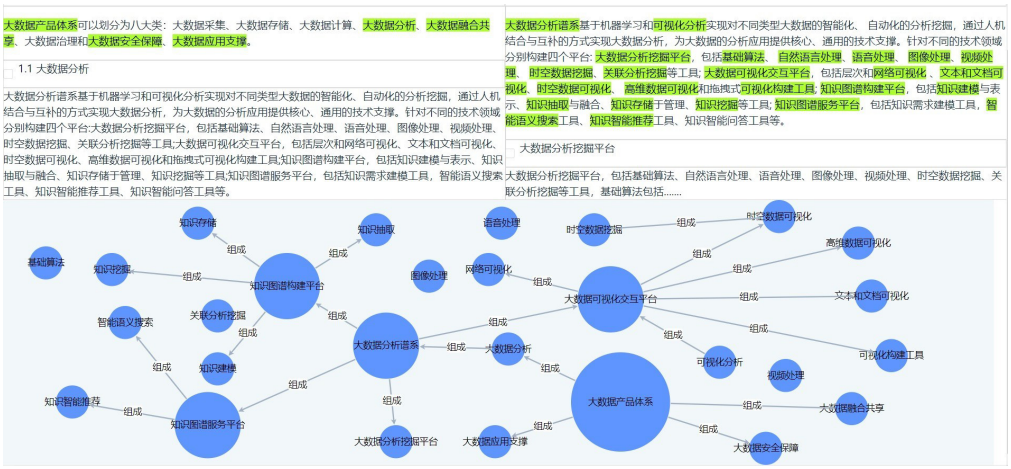

Figure 5. Document relationship output
图 5. 文档关系输出

6. 总结

本文在实体关系归属课题中将功能具体化，综合考虑各方面成本，结合现代技术与传统方法，提出预训练模型进行实体识别、建立术语树来进行关系匹配并将匹配得出的实体关系归属实时在前端页面显示。利用多篇文档进行系统测试，验证该方法具有可行性及有效性等特点。同时该系统的实现为后续工作提供可靠的理论基础。该方法真正应用至特定领域中将有效提高实体关系归属处理速度，实现数据的快速管理、降低人工管理成本，并且该方法在解决少样本数据训练准确率低的难题时打开了一个新视角——根据少样本(原有)数据样本进行术语树结构创建，用现有关系代替未知关系，降低了关系划分问题的复杂度。

参考文献

[1] 丁泓馨, 邹佩聂, 赵俊峰, 等. 一种基于主动学习的文本实体与关系联合抽取方法[J]. 计算机科学, 2023, 50(10): 126-134.

[2] 沈依宁, 王一然, 吴聪. 基于深度学习的关系抽取研究进展[J/OL]. 电子科技: 1-11. https://doi.org/10.16180/j.cnki.issn1007-7820.2025.07.006, 2024-06-23.

[3] 成全, 蒋世辉, 李卓卓. 基于改进 Casrel 实体关系抽取模型的在线健康信息语义发现研究[J/OL]. 数据分析与知识发现: 1-17. http://kns.cnki.net/kcms/detail/10.1478.g2.20231114.1648.004.html, 2024-06-23.

[4] 王欢, 王兴芬, 吕金娜. 面向金融文本的实体关系抽取方法[J]. 计算机工程与设计, 2023, 44(11): 3345-3351.

[5] 刘清堂, 蒋如意, 吴林静, 等. 融合实体位置与类型特征的土家民族器乐实体关系抽取研究[J/OL]. 数据分析与知识发现: 1-16. http://kns.cnki.net/kcms/detail/10.1478.G2.20240524.1021.002.html, 2024-06-23.

[6] 魏嵬, 丁香香, 郭梦星, 等. 文本相似度计算方法综述[J/OL]. 计算机工程: 1-19. https://doi.org/10.19678/j.issn.1000-3428.0068086, 2024-06-21.

[7] 赵停停. 基于 MySQL 数据库技术的 Web 动态网页设计研究[J]. 信息与电脑(理论版), 2023, 35(17): 174-176.

[8] 徐新黎, 卢齐林, 杨旭华, 等. 多任务特征交互的三元组抽取方法[J/OL]. 小型微型计算机系统: 1-10. http://kns.cnki.net/kcms/detail/21.1106.TP.20240529.1529.008.html, 2024-06-21.

[9] 温雨, 王琦, 严武军. 基于相似度融合的中文文本相似性度量方法研究[J]. 信息技术与信息化, 2023(10): 36-39.

[10] 杨政, 方正云, 李天骄, 等. 基于分层深度语义的科研项目文本相似度度量方法[J]. 计算机与数字工程, 2024, 52(3): 795-801, 851.

[11] Lamurias, A., Ruas, P. and Couto, F.M. (2019) PPR-SSM: Personalized Pagerank and Semantic Similarity Measures for Entity Linking. *BMC Bioinformatics*, **20**, Article No. 534. https://doi.org/10.1186/s12859-019-3157-y

[12] 王冠南, 郭丽娟, 彭曙蓉, 等. 基于正则表达式和 Jaccard 系数的智能变电站录波通道同源匹配[J]. 浙江电力, 2024, 43(1): 20-27.