重庆师范大学硕士学位论文

基于深度学习的本体匹配方法及其优化算法研究

硕士研究生：　段旭祥
指 导 教 师：　吴至友　教授
学 科 专 业：　计算数学
所 在 学 院：　数学科学学院

重庆师范大学

2020 年5 月

A Thesis Submitted to Chongqing Normal University in Partial Fulfillment of the Requirements for the Degree of Master

Research on ontology alignment method and optimization algorithm based on deep learning

Candidate：DUAN Xuxiang
Supervisor：Professor Wu Zhiyou
Major：Computational Mathematics
College：School of Mathematical Sciences

Chongqing Normal University

May, 2020

基于深度学习的本体匹配方法及其优化算法研究

摘 要

关键词：

Research on ontology alignment method and optimization algorithm based on deep learning

ABSTRACT

Keywords:

目 录

1 绪 论

本章首先通过介绍知识图谱与本体匹配来说明本文的研究背景及研究意义，然后阐述本文的研究内容，最后告诉读者本文的论述结构与编排。

1.1 研究背景

主要介绍知识图谱，引入本体匹配

1.1.1 知识图谱

自1989年Tim $BernersLee$发明万维网后，人类便真正进入了信息爆炸式增长的时代，可以根据自己的需要在网页查阅任意内容，但原始的万维网都由网页互相链接而成，网页中的信息没有有效，规范化的组织，计算机无法找到明确的语义信息，因而计算机很难理解，并对其进行处理，最终导致人们很难从海量的信息中迅速检索，筛选得到自己需要的信息。所以$TimBerners-Lee$针对该问题，再次提出语义网，语义网通过RDF标准[1]对网页中的文档添加计算机语言的释义，计算机可以根据语义做出理解，判断，实现人与计算机之间的更好的无障碍沟通。并且语义网根据不同的语义关系建立对应的链接，于是开放链接数据[2]项目顺势而生，研究者对大量数据进行链接，发布，从而也形成了很多知名好用的知识库，如：$Freebase$[3],$DBpedia$[4]等，这也为知识图谱的应运而生埋下了伏笔。

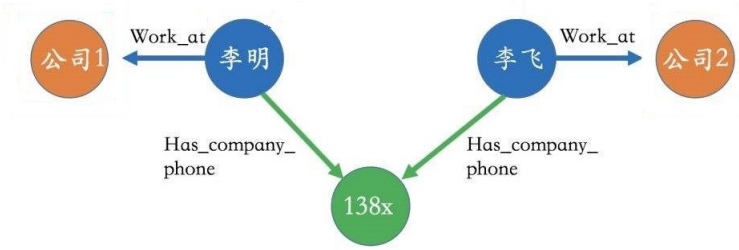

图 1.1 知识图谱三元组示例

2012年，谷歌提出知识图谱[5]这一概念。知识图谱通过多样的语义检索技术[6]，整合来自多方的信息，经过语义检索，自然语言处理技术，来增强其搜索引擎的搜索准确性。知识图谱源于语义网络，由语义网络发展而来，可以解释概念（实体）间的语义关系，进而实现了现实世界中各种知识的形式化描述。知识图谱可以从多种渠道收集信息来提升搜

[1]https://www.w3.org/RDF/
[2]https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
[3]https://developers.google.com/freebase/
[4]http://wiki.dbpedia.org/
[5]https://www.google.com/intl/bn/insidesearch/features/search/knowledge.html
[6]https://www.techopedia.com/definition/23731/semantic-search

索引擎效果，被广泛应用于语义搜索，知识问答，推荐系统等领域。知识图谱将知识库中的知识进行可视化，是一种基于图的数据结构，这种图结构中包含节点和有向线段，节点对应着知识库中的实体（概念），有向线段对应知识库中的关系。如图：图中的李明，李飞，分明表示两个不同的实体，其属性是“姓名”；公司1，公司2分别表示两个不同的工作地点的实体，其属性是“工作地点”，蓝色的有向边“work_at”表示关系，该关系表示“工作地点在哪里”，所以，如“李明work_at 公司1”就是一个完整的三元组。

而实际应用中的知识图谱是由大量的节点通过有向边互相连接而成的大型的图网络。

如今，随着知识图谱的广泛应用，知识图谱作为搜索引擎的新帮手，可为用户提供更加全面和准确的信息，提升用户信息检索的体验。当我们在搜狗百科中搜素“特朗普”时，搜狗百科可以告诉用户准确的答案，“特朗普”是“唐纳德特朗普”，他是美国第45任总统，并且通过知识图将其相关的链接数据返回给用户，详细告诉用户其出生年月日，年龄，籍贯，教育背景，以及资产情况，政治背景等情况。

图 1.2 “特朗普”搜狗百科

谷歌，微软，百度，阿里，搜狗等纷纷推出了自己的知识图谱，知识图谱已经被广泛应用于语义搜索，智能问答和个性化推荐等方面，而知识图谱的应用更涉及电商，法律，医疗，金融等众多领域。下面以语义搜索，智能问答和个性化推荐为例，阐述知识图谱在现实中的应用情况。

语义搜索：知识图谱的应用使得搜索引擎不再只依赖于用户输入的关键词的字面本身，能够真正理解关键词的含义，并针对当前关键词找到知识图谱中与之相关联内容，将密切相关的内容根据相似度的高低呈现给用户。目前，谷歌，微软，百度等公司也都利于知识图谱来实现搜索效果的提升。

智能问答：智能问答从方法上其实是与智能搜索是类似的，首先对用户的问题作语义的解析，然后在已有的知识库的问答对中寻找与用户问题相类的问句，最后将知识图中已有的答句作为回答返回给用户，在实际生产中，智能问答也已经得到广泛的应用，诸如：苹果Siri，微软小冰和小娜，IBM 的Watson 系统[7]，亚马逊Alexa，百度度秘等智能问答产品。

个性化推荐：个性化推荐是知识推荐的一个延申，在知识推荐的基础上引入领域知识图谱，以商品为例，将商品的特征关键词作语义解析，对于用户购买或搜索的关键词相似度较高的商品向用户做推荐。可以较有效的克服“冷启动”的问题[8]。个性化推荐在阿里，京东的电商平台，百度的推荐系统等都有应用。

由此可见，知识图谱以其广泛，且必要的应用价值，已经成为当下时代发展的通用性工具，所以，进行知识图谱方面的研究具有重要且深远的意义。

1.1.2 本体匹配

知识图谱可以看做是服从于本体控制的知识单元的载体，就好比本体是制作蛋糕的模具，而知识图谱便是蛋糕，模具的好坏必然影响蛋糕的质量。知识图谱的构建过程大致可分为：知识建模，知识获取，知识融合，知识存储，知识计算，知识应用等步骤，其中知识融合作为必要且必须的步骤，直接影响知识图谱构建的质量。知识融合一定程度上可看作本体的匹配，所以，本体匹配的效果直接影响着知识图谱构建的质量。所以，知识图谱的研究是一个较大的课题，本体匹配是知识图谱研究中的一个子课题。接下来将重点介绍什么是本体，以及什么是本体匹配。

本体的概念最早是在哲学中出现，随着人工智能的发展，这一概念逐渐被引入到计算机科学中，自然也被赋予了不一样的定义，本体的定义最受认可的定义是1993年Grube提出的”一种形式化的，对于共享概念体系采用规范，明确而又详细的说明“。本体的构建是为了将有相关性的某一领域的知识，整合不同的群体对于该知识的共同的理解，从而将该含义的知识进行统一化的表示，并且对该领域中不同层次的统一化的知识表示给出不同词汇间的相互关系。

大型本体的构建是一项耗时，耗费人力的工作，但一个完整的领域本体便是一个完整的知识库，对该领域研究工作的推进具有非常重要的意义。但当多个人某一领域知识进行本体构建时，由于个人的差异，可能存在对同一个属性命名不相同，取值范围不同等问题，即是本体的异质现象，同样，不同的本体之间也会存在对同一含义的属性命名或取值不同的情况，因此，这就需要相关的技术手段，来解决这种问题，也就是本体匹配方法。本体匹配方法就是用于寻找存在异质本体之间的语义映射关系，克服本体异质问题的方法，而具体是通过计算不同本体中两个实体之间的相似度，通过相似度的高低来判断两个实体之间的语义关系。

[7]https://www.ibm.com/watson/
[8]https://www.jianshu.com/p/193fea0a7004

本体匹配作为知识图谱研究范畴的一部分，作为构建知识图谱中的必要步骤，所以与知识图谱一样，也被广泛应用于电商，法律，机器翻译以及语义信息集成等众多领域，本体匹配作为本文的主要研究对象，这里就不对其应用作具体介绍了。

图 1.3 “猪八戒”相关的百科词条

1.2 研究意义

当下，深度学习技术已经被验证在知识图谱的多个研究方向中都是非常有效的技术手段，且是必备技术手段，而本体匹配的相关研究已经开展多年，但多数都是利用传统的方法，而利用深度学习技术进行本体匹配研究的工作还不多，所以本文从深度学习的角度，开展对本体匹配的研究，来丰富采用深度学习技术解决本体匹配任务的研究。

但深度学习的运用，自然需要解决对本体中实体与关系的表示问题，所以本文结合知识图谱表示学习方法来作为实现实体或关系的嵌入方法，从深度学习角度，利用表示学习思想和方法，为本体匹配的进一步研究供新的思路与方法。

此外，本体自身的结构具有其特殊性，而前人的研究尚未考虑到本体自身结构的特点，也一直是本体匹配研究尚未解决的问题。所以本文充分考虑本体自身及本体与本体之间的结构特点，来尝试解决这一亟待克服的困难。

虽然深度学习技术十分有效，但对于模型的参数的调节却是一项耗时耗力的工作，参数对模型结果的影响不言而喻，本文中，表示学习模型效果的好坏直接影响着后续匹配的效果，所以本文采用了带有全局思想的黑箱优化算法–HORD算法来实现对模型超参数的半自动优化，来提升最终的匹配效果，也为几类表示学习模型的超参数调节提供统一可行的框架。

1.3 研究内容

本文的主要研究内容包含以下四个方面：

1.将带有全局思想的黑箱优化算法–HORD算法，用以对几类表示学习模型的超参数半自动优化。表示学习模型针对不同的数据集进行训练时，需要手动调节模型的超参数，这是一项比较耗时的工作，并且得到的最终效果也是一个“经验值”，所以采用带有全局思想的HORD算法对表示学习模型的超参数进一步调节，从而进一步提升表示学习的效

果。

2.利用表示学习中基于距离的建模思想及其方法，并且充分考虑本体自身的结构信息，以及上下文信息，将结构信息作为新的模型约束，上下文信息来获得更好的嵌入表示，来搭建新的本体匹配模型。

3.在内容2的基础上，进一步针对内容2中模型的上下文信息不足的问题，进一步改进，采用效果更好的BERT模型，作为本体匹配模型的预训练嵌入表示，并且将已经在内容1 中验证有效的HORD算法，来进一步优化模型的超参数，提升本体匹配模型的效果。

4.在基准数据集上，利用常用的评估指标对模型效果进行评估，并将本文中的模型的效果与其他较新的模型结果进行对比，分析，验证本文工作的有效性。

1.4 文章结构安排

本文共分为六个章节，本文的结构编排如下：

第一章：一方面，简述本文的研究背景，介绍知识图谱，引出本体匹配，并介绍知识图谱相关概念；另一方面，从理论与实际应用角度阐述本文的研究意义，并对本文的具体研究内容进行简述。

第二章：对本体匹配的相关工作，方法从传统的方法，以及基于深度学习的方法两个方面对已有工作进行综述，总结。

第三章：详细介绍本文的具体研究内容，首先对本章的预备知识，HORD算法，几类表示学习模型，BERT模型进行说明；然后介绍基于HORD算法的几类表示学习模型超参数优化方法，其次，提出基于多视角的生物医学本体匹配模型，最后，阐述基于BERT模型和HORD算法对MuitiOM的改进工作。

第四章：利用常用的模型评价指标，标准数据集，在上一章基础上，对三个研究内容进行实验评估，并将实验结果与其他方法作比较，进行分析。

第五章：对本文的工作进行总结，并对未来工作进行展望。

第六章：本文的相关参考文献。

2 相关工作

2.1 传统的本体匹配方法

待完善

2.1.1 基于元素层面的方法

待完善

2.1.2 基于结构层面的方法

待完善

2.2 基于深度学习的本体匹配模型

待完善

2.2.1 引入外部同（反）义词典的方法

待完善

2.2.2 引入预训练向量的方法

待完善

3 本文工作

3.1 预备知识

待完善

3.1.1 常用表示学习模型

待完善

3.1.2 超参数优化算法–HORD算法

待完善

3.1.3 BERT模型

待完善

3.2 基于HORD算法的几类表示学习模型超参数优化

待完善

3.2.1 HORD-OPENKE算法及结构

待完善

3.2.2 算法学习

待完善

3.3 基于多视角的生物医学本体匹配模型–MuitiOM

待完善

3.3.1 上下文视角的嵌入

待完善

3.3.2 本体结构视角的嵌入

待完善

3.3.3 外部资源视角的嵌入

待完善

3.3.4 负采样方法及匹配算法

待完善

3.3.5 模型学习

待完善

3.4 基于BERT模型和HORD算法对MuitiOM的改进

待完善

3.4.1 微调BERT模型参数获得本体的实体表示

待完善

3.4.2 HORD算法调节模型超参

待完善

4 实验与评估

4.1 数据集

待完善

4.1.1 表示学习数据集

待完善

4.1.2 医学本体匹配数据集

待完善

4.2 评估指标

4.2.1 表示学习模型评估指标

待完善

4.2.2 本体匹配评估指标

待完善

4.3 基于HORD算法的几类表示学习模型超参数优化的评估结果

待完善

4.3.1 实验设置

待完善

4.3.2 预测结果与分析

待完善

4.4 基于多视角的生物医学本体匹配模型–MuitiOM的评估结果

待完善

4.4.1 实验设置

待完善

4.4.2 本体匹配结果与分析

待完善

4.5 基于BERT模型和HORD算法对MuitiOM的改进的评估结果

待完善

4.5.1 实验设置

待完善

4.5.2 本体匹配结果与分析

待完善

5 结论及展望

5.1 本文工作总结

5.2 未来工作展望

6 参考文献

致 谢

值此论文完成之际，我由衷地感谢我的导师吴至友教授。从我读硕士开始已经将近三年时间，这期间吴老师在学习上给予了我极大的帮助和鼓励，在我的学业和论文的研究工作中无不倾注着导师辛勤的汗水和心血。吴老师渊博的知识、严谨的治学态度、一丝不苟的工作作风和无私的奉献精神都影响并激励着我。能够成为吴老师的学生是我人生中的一件幸事！同时也非常感谢杜学武教授对我的指导和帮助，杜老师锐意进取的精神、豁达的胸襟、乐观的心态都是我今后需要不断学习的。感谢白富生教授，白老师对人真诚、热情，我的进步也离不开白老师的关心和帮助。感谢李国权教授在我对硕士期间对我的帮助，不论是科研方面的合作还是生活上的相处都是非常愉快的事情。感谢院系领导老师对我的支持和帮助。

衷心感谢东南大学漆桂林教授，对我的关心和帮助。漆老师为人十分谦逊，对人热情。在我研究生期间漆老师给予了我很多的帮助，在此对他表示衷心的感谢！特别感谢高桓博士，高桓博士待人真诚，做事严谨。在我研究生期间给予我思想上，科研态度，以及方法上的指导，在此对他表示真心的感谢！也感谢李炜卓博士，李炜卓博士为人实在，踏实，在我研究生期间，引导我踏入了知识图谱相关研究，在此对他表示衷心的感谢。

感谢在学习和生活上给予我关心、鼓励的同门学们。

最后我要感谢我的父母、爱人和家人对我的理解和支持!

段旭祥

2020 年5 月6 日

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。除了文中特别加以标注和致谢的地方外，论文中不包含他人已经发表或撰写过的研究成果，也不包含为获得重庆师范大学 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。

学位论文者签名：　　签字日期：　年　月　日

学位论文版权使用授权书

本学位论文作者完全了解重庆师范大学有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权重庆师范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、汇编学位论文。

学位论文作者签名：　　签字日期：　年　月　日