北京師範大學

硕士学位论文

论文题目：基于关系路径的知识库补全算法研究

作　　者：　**黄勇**

导　　师：　**王志春 副教授**

系别年级：　**计算机软件与理论**

学　　号：　**201521210022**

学科专业：　**知识工程**

完成日期：　**2018年 3月**

北京师范大学研究生院

北京师范大学学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：　　日期：　　年　　月　　日

学位论文使用授权书

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

本人签名：_____　　日期：_____

导师签名：_____　　日期：_____

基于关系路径的知识库补全算法研究

摘　要

知识库补全基于现有知识库的三元组，预测知识库中实体和实体之间的关系。当前的知识库补全算法主要包括基于符号逻辑的知识库补全算法和基于表示学习的知识库补全算法，其中效果显著的一种符号逻辑的知识库补全算法是路径排序算法。路径排序算法抽取实体之间的关系路径特征，构建逻辑回归分类模型，对知识库中实体和实体之间关系进行预测。然而仅仅基于关系路径进行知识库补全，并没有结合实体和实体之间的属性特征关系，一些知识库进行知识库补全构建的正负实例对比例悬殊，进行模型训练很难抽取有效的路径特征，也难以进行模型训练。

本研究基于路径排序算法，在传统的关系路径特征基础上，提出了一种新的结合不同关系路径特征和不同类型实体属性特征，并将多种不同类型的关系路径特征、实体属性特征进行组合，极大的丰富了关系预测的特征，显著提升了知识库补全效果。其次考虑到知识库补全中正负例不平衡问题，为了获得更好的模型预测结果，在传统逻辑回归模型的基础上，本论文研究了基于学习排序算法的路径补全模型，通过学习基于排序的损失函数，学习正负实体对的排序，从而进行知识库关系补全预测。本毕业论文的研究创新点主要有：

- 抽取关系路径和实体属性作为知识库补全的特征，将两种不同类型特征进行组合，极大拓展了知识库补全系统的维度，增强了知识库补全效果。
- 基于学习排序算法预测知识库补全中实体和实体的关系，通过直接学习知识库补全中的实体对排序顺序，改进损失函数进行模型预测。
- 构建了基于逻辑回归排序、树方法排序以及基于深度神经网络排序算法的知识库补全模型，可以利用不同知识库数据特征选择合适的排序算法模型。

关键词：知识库补全，路径排序算法， 学习排序，符号逻辑、关系路径

Knowledge Base Completion by relational path methods

ABSTRACT

Knowledge base (KB) completion aims to predict new facts from the existing ones in KBs. There are many KB completion approaches, one of the state-of-art approaches is Path Ranking Algorithm (PRA), which predicts new facts based on path types connecting entities. PRA takes the relation prediction as a classification problem, and logistic regression or SVM is used as the classification model. In this paper, we consider the relation prediction as a ranking problem, learning to rank model is trained on relations to predict new facts. Besides, our model use both literal and relational facts as feature matrix， which is much more comprehensive. We propose to extract literal features from literal facts, and incorporate them with path-based features extracted from relational facts; predictive model is then trained to infer new facts with assembly features and bring higher precision scores in classification metrics and other ranking metrics. Experiments on YAGO show that our proposed approach outperforms approaches using relational features and classification models. This paper has three main contribution, including:

- KB completion feature type is much more comprehensive, we use both literal and relational facts to generate feature matrix, and in different KBs our features help to generate higher scores both in ranking tasks and classification tasks.
- KB completion was considered as a ranking method, we use learning to ranking model to rank entity pairs rather than simple taking it as a classification model.
- KB completion methods are easily explainable, with combined literal and relational facts we can easily predict relation in KBs, other deep neural network models and tree models are used with assembly features.

KEY WORDS: Knowledge Base Completion, Learning to Rank, PRA, symbolic model

目 录

第 1 章　绪论

1.1　引言

近些年来，随着互联网的快速发展，人工智能和大数据技术的兴起和应用，网络知识、信息和数据大量增长，人们面对的数据越来越多。传统的基于网页链接检索的信息存储、信息检索方式越来越难于进行查找学习知识，随着移动设备的增加，人们对于信息检索、知识获取的要求也越来越高，面对各种数据信息，人们查找检索知识的需求越来越迫切，越来越期望能更好更快的检索了解学习知识。随之知识库构建、推理和应用变得流行起来，越来越多的大学、科研机构和商业公司开始构建大规模知识库，如Google Knowledge Graph[1]，NELL[2]，YAGO[3]， Freebase[4]，DBpedia[5]等。这些知识库基于自动和半自动的信息抽取、众包、专家领域内知识等方式进行构建，把互联网、书籍、数据库等各处非结构化的文本中数据结构化、抽取构建知识库，获得了大量的实体、关系和属性信息，能为人们提供的方便快捷的信息查询、检索、知识学习途径。这类知识库的数据完整性、数据精确性和数据质量等衡量指标非常高，已经有很多基于知识库的系统被成功用于商业领域，如谷歌搜索引擎①，微软的必应搜索② 等。除了这些通用领域的知识库之外，也有很多领域知识库被构建和应用，知识库也被用于生物学[6]、金融学、教育学等各个领域，一些较为成熟的问答系统、个人手机助手如苹果公司的Siri、谷歌公司的Google Assistant、微软公司的手机助手小冰、出门问问等也将知识库集成其中。

常见的知识库可以用有向图进行可视化表示，图1是一个简单的关于北京师范大学的知识库，描述了和北京师范大学相关的实体、属性和关系。图中展示了实体、实体和实体之间关系，其中实体是描述现实中实际存在的事物、地点、人物、事件如：北京师范大学、北京等，关系是描述实体和实体之间存在的某种联系和特征，如北京师范大学和北京存在着位于的关系。除此之外，一个知识库中的三元组还可以分为关系型三元组和实体属性三元组。关系三元组典型的例子如：（北京师范大学，位于，北京）。其中“北京师范大学”和“北京”是实体，属性事实三元组如：（北京，人口，2150万）。其中“北京”是实体，“2150万”是描述实体属性的属性值，“人口”是北京这类实体具有的一种属性特征类型。通常实体关系三元组用来描述实体和实体之间的联系，实体属性三元组描述

① https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html
② https://blogs.bing.com/search/2013/03/21/understand-your-world-with-bing/

实体的属性事实特征，这些不同类型的三元组共同构成一个完整的知识库。

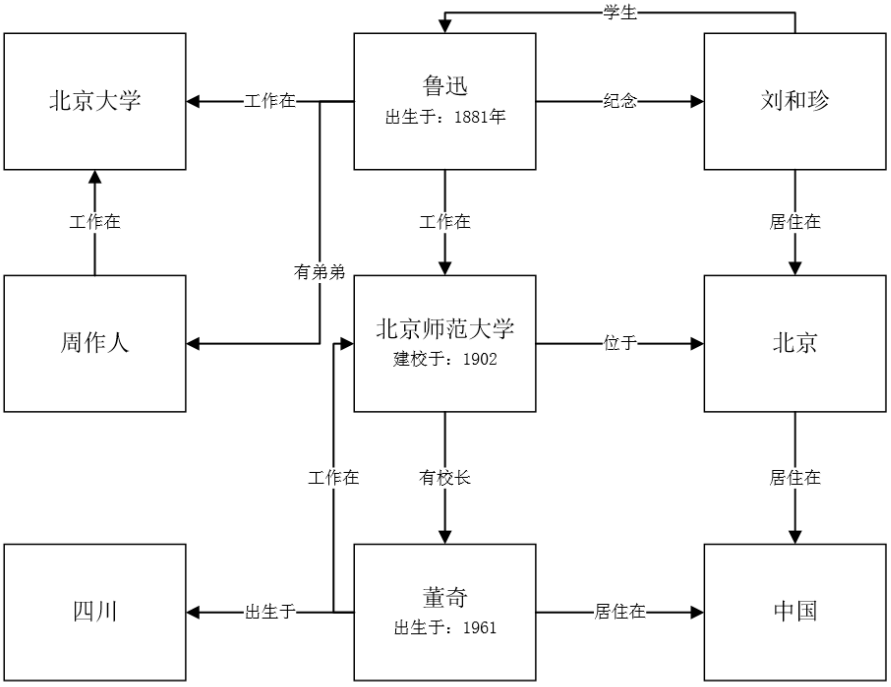

图 1　一个知识库例子

1.2　知识库补全问题

本节介绍知识库补全算法的主要研究问题，了解知识库补全算法相关的难点。尽管一些知识库的三元组数量巨大，但这些知识库仍然是不完备的，如很多人的出生地点并未包含在知识库中，一些演员是否出演过某些电影也是未知的。为了能发掘知识库中隐藏的实体和实体之间的关系，有很多知识库补全算法被提出。

1.2.1　结合关系路径和实体属性的知识库补全

虽然许多知识库的规模很大，但他们仍然是不完备的，如很多人的出生地点并未包含在知识库中，一些演员是否出演过某些电影也是未知的。为了解决这个问题，很多知识库补全的方法被提出来，这些方法基于知识库中已有的三元组预测新的三元组，如果将现有的知识库数据看做是多种关系构成的图，图的顶点是实体，图的边是实体对之间的关系，知识库补全可以看成是图中关系的预测。表1展示了YAGO和Freebase知识库中部分关系特征类型，和一些链接预测模型[7]不同，知识库补全需要处理多种不同关系类型的

关系预测，而多数链接预测只需要预测一种单一关系。为了解决预测实体和实体之间关系的问题，很多经典的知识库补全算法被提出，这些算法可以分为两类：基于逻辑符号推理的补全算法和基于表示学习的补全算法，有时候也被称为基于图特征和基于隐藏特征的补全算法。

表 1　YAGO和Freebase知识库部分关系类型

YAGO关系类型	Freebase关系类型
isCitizenOf	/location/country/form_of_governme
isAffiliatedTo	/tv/tv_program/regular_cast./tv/regular_tv_appearance/acto
wasBornIn	/media_common/netflix_genre/titles
playsFor	/award/award_winner/awards_won./award/award_honor/award_wi
isLocatedIn	/soccer/football_team/current_roster./sports/sports_team_roster/position
influences	/sports/sports_position/players./soccer/football_roster_position/t
hasWonPrize	/film/film/starring./film/performance/acto
dealsWith	/soccer/football_team/current_roster./soccer/football_roster_position/posi
hasChild	/film/actor/film./film/performance
graduatedFrom	/award/award_nominated_work/award_nominations./award/award_nomination/awar
isMarriedTo	/award/award_category/nominees./award/award_nomination/nominated_f
worksAt	/award/award_nominee/award_nominations./award/award_nomination/award_nomin
diedIn	/olympics/olympic_sport/olympic_games_cont
hasNeighbor	/music/performance_role/regular_performances./music/group_membership/role
happenedIn	/award/award_category/winners./award/award_honor/ceremony
livesIn	/film/film/release_date_s./film/film_regional_release_date/film_release_distribution_mediu
isPoliticianOf	/people/marriage_union_type/unions_of_this_type./people/marriage/s
participatedIn	/award/award_winning_work/awards_won./award/award_honor/award_winn
hasOfficialLanguage	/film/film/release_date_s./film/film_regional_release_date/film_release_re
owns	/film/film/languag
……	……
actedIn	/music/artist/genr

我们分析了常见的YAGO知识库，发现总共有四百多万实体关系数据被和三百多万属性事实的三元组。其中的实体关系数据在以往经典的知识库补全算法中被广泛使用，而属性事实数据尽管大量存在，却在知识库补全系统中并没有得到广泛应用，同时实体属性数据类型单位差别很大，难以进行统一有效的处理，将这些实体属性特征作为知识库补全的特征也十分困难。但可以预见在知识库关系预测中，实体属性特征会起着重要的作用，如何将实体属性三元组有效用于知识库补全是本论文的关键点之一。

分析可以发现实体属性特征的复杂性较大。在表2中，我们展示了YAGO和Freebase知识库中部分实体的属性特征。以YAGO知识库为例，我们可以发现，常见的属性特征信息可以不仅仅有数量类型的如：hasNumberOfPeople、hasArea等，也有日期类型的特征信息如：wasCreatedOnDate、wasBornOnDate，也有比值型的实体属性特征如：hasInflation、hasUnemployment等。如何整合这些不同类型的特征信息，如何将属性特征和实体关系特征进行组合，都是本研究的重点和难点工作之一。

1.2.2 基于学习排序算法的知识库补全

知识补全算法需要考虑不同类型的关系路径类型和实体属性类型外，如何构建合理有效的学习模型，预测知识库中实体对之间的关系，也是知识库补全算法的重要研究内容。通过研究可以发现知识库中的图结构是稀疏的，每个存在的正例三元组实体对在训练模型中，可能生成上百组负例三元组实体对，如何解决正负实体对不匹配的问题很关键。在正负实体对比例悬殊时，关系预测中仅靠传统算法中的打分，比如一些逻辑符号推理中的逻辑回归算法是不够的，这种评价中并未考虑候选实体对的顺序对预测结果的影响，也不关注候选实体的秩序关系，同时基于熵的损失函数过于简单，从而处理关系路径相关的特征时，不能有效的结合不同关系路径类型之间的组合关系，很难从多条关系路径类型中组合学习一些隐藏的关系路径特征，如何有效的解决这些问题，也是知识库补全算法中需要解决的重点。

通过研究知识库补全问题可以发现，对于知识库补全是一种对正负三元组进行排序的过程，只有当正例三元组排序结果好于负例三元组的排序结果时，知识库补全结果才是有效的，从而本研究提出了基于学习排序的知识库补全算法，并通过基于树模型、深度神经网络的排序算法，研究如何设定知识库补全中目标函数和优化方法。

在基于表示学习的知识库补全算法中，不同的模型有不同的目标函数，通过机器学习优化算法，学习不同实体、关系的低维度表示向量。对于基于符号逻辑的知识库补全算法来说，常见的知识库补全算法如路径排序和子图特征抽取算法都是采用二分类模型进行知识库补全算法模型构建。如路径排序算法中，通常采用逻辑回归算法或者支持向量机回归，学习实体和实体之间具有某种关系的值。这种算法通常可以看做一种基于pointwise的知识库补全算法，然而在一个知识库补全系统中，一组正负例实体对也可以当做一个整体进行优化，构建pairwise的知识库补全算法，进行模型优化，从而解决正负例实体对极其不平衡的问题。

1.3 论文研究工作

针对现有知识库补全技术不足，本研究将知识库中的关系路径特征和实体属性特征相结合，构建了一个更准确的知识库补全模型。首先，基于经典的路径排序模型抽取了关系路径特征；其次通过结合实体属性特征和关系路径特征，构建逻辑回归模型进行关系预测，从而进行知识库补全。除此之外，本研究提供一种基于学习排序算法的知识库补全技术，我们通过计算候选头实体和尾实体在关系预测中的位置排序，通过优化排序损失函数MAP来保证训练误差最小，从而获得最优的关系预测结果，并选择合适的模型评价指标来评估改进我们的预测结果。

1.3.1 结合关系路径和实体属性的知识库补全

本部分主要介绍如何结合关系路径特征和实体属性特征进行知识库补全算法的构建，具体来说：（1）抽取关系路径类型特征，包括如何抽取关系路径类型构建特征集合，如何抽取计算每个关系下的的实体对的关系路径特征向量；（2）抽取实体属性特征，包括如何获取实体属性类型集合，如何抽取每个关系下的实体对的实体属性集合，如何将实体对的实体属性集合进行标准化、归一化，如何结合不同类型的关系类型特征和实体属性特征，如何获取实体对和实体对之间实体属性特征关系。

1.3.2 基于学习排序算法的知识库补全

本部分主要介绍如何构建预测模型学习知识库中的实体对关系。具体包括如何选择目标损失函数，如何构建优化算法，如何学习不同实体对的排列秩序。通过探索不同类型的目标函数、预测模型。本部分将知识库补全算法进行优化组合，期望能结合学习排序模型提高特征的组合能力和模型的泛化能力，优化正负实体对的排列顺序，使得知识库补全中，正例三元组排序结果高于负例三元组，而非简单的对知识库中的三元组进行打分。

本部分的研究的工作重点有两点：（1）提出了基于排序的知识库补全算法，将知识库补全算法从优化二分类模型改进为优化一组相关实体对排序的顺序问题，通过构建pairwise模型，进一步提升模型的预测能力。（2）探索基于深度神经网络的知识库补全算法，结合wide & deep模型，将基于深度神经网络的排序算法用于知识库补全技术中，从而进一步提高模型的特征组合能力和模型泛化能力。

表 2　YAGO和Freebase知识库中的部分实体属性特征

YAGO实体属性特征	FB15K实体属性特征
hasNumberOfPeople	/tv/tv_program/air_date_of_first_episode
hasArea	/user/jg/default_domain/olympic_games/closing_date
wasCreatedOnDate	/user/jg/default_domain/olympic_games/opening_date
hasLongitude	/time/event/start_date
hasPopulationDensity	/time/event/end_date
hasLatitude	/user/maxim75/default_domain/dbpedia_import/geocode_checked
wasBornOnDate	/tv/tv_program/air_date_of_final_episode
hasHeight	/award/award_category/date_discontinued 16
wasDestroyedOnDate	/user/ktrueman/default_domain/international_organization/founded
diedOnDate	/base/usnris/nris_listing/significant_year
hasLength	/sports/pro_athlete/career_start
happenedOnDate	/tennis/tennis_player/year_turned_pro
hasUnemployment	/royalty/order_of_chivalry/date_founded 21
hasEconomicGrowth	/business/defunct_company/ceased_operations
hasRevenue	/government/legislative_session/date_began
hasGini	/government/legislative_session/date_ended
hasExpenses	/music/artist/active_start
hasInflation	/people/deceased_person/date_of_cremation
hasGDP	/base/cdnpolitics/legislative_assembly/founded
hasImport	/royalty/royal_line/ruled_to
hasExport	/royalty/royal_line/ruled_from
hasPoverty	/music/artist/active_end
hasBudget	/sports/pro_athlete/career_end
hasWeight	/base/lewisandclark/places_eastward/from
.....	.....
hasDuration	/base/yalebase/secret_society/founded

1.4 论文组织结构

本论文的组织结构如图2 所示。本研究首先提出知识库补全领域相关研究问题，并进行概述，其次介绍知识库技术国内外研究现状。接着介绍本人的研究工作，包括结合实体属性的知识库补全算法和基于学习排序的知识库补全模型技术。最后对本研究的相关工作和后续研究方向进行了总结和展望。

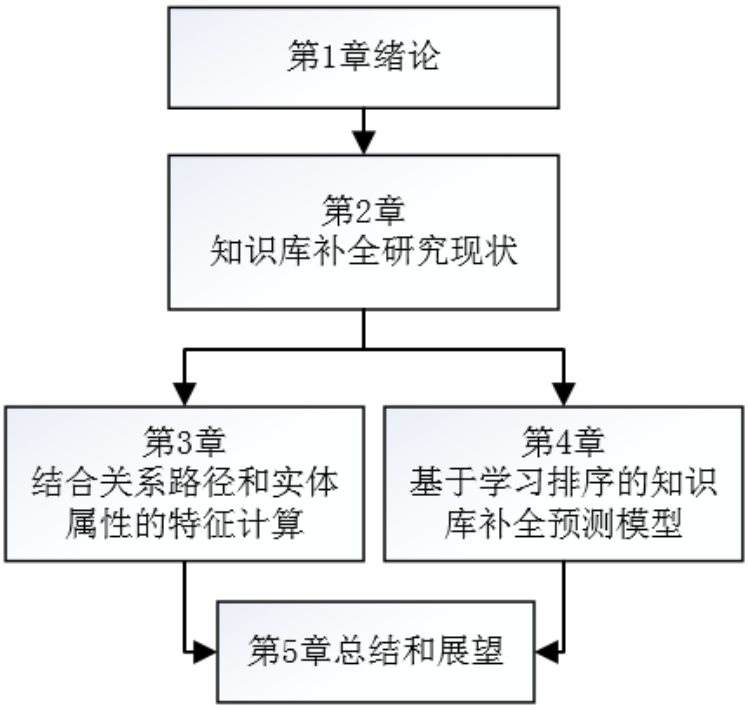

图 2　论文组织结构图

第2章 介绍当前知识库技术的研究现状。主要包括知识库构建、知识库应用和知识库补全和推理。

第3章介绍了基于关系路径的知识库补全算法研究框架，介绍了如何结合关系路径类型和实体属性类型进行知识库补全算法模型预测。

第4章介绍了如何构建学习排序的算法进行知识库补全模型预测，包括如何构建目标函数，如何进行特征优化。

第5章总结了本研究工作的重点和不足，提出了未来研究的一些思路和方向。

第 2 章　知识库补全研究现状

当前知识库相关的研究主要分为三个方面：（1）知识库的构建，各种基于互联网的、通用的、领域内的知识库构建。（2）关系机器学习，基于已经构建好的知识库进行关系预测，知识库补全推理等。（3）知识库应用系统。如基于知识库的问答系统，基于知识库的信息检索系统，生物、金融、教育等领域的本体知识库的应用等。其中研究的热点之一就是知识库补全推理，或者说如何发掘知识库中实体和实体之间隐藏的关系。

当前存在的知识库系统如NELL、Knowledge Vault等，尽管已经包含了大量的实体、关系和属性，但这类知识库中很多实体和实体之间的关系任然是缺失的。例如，尽管知识库中存在很多三元组如：（北师大，位于,北京）、（北京，位于,中国）。但是当预测北师大是否位于中国时，通常计算机很难直接进行这类常识性的推理。如何构建一个辅助计算机进行常识推理的系统，发现知识库中隐藏的实体和实体之间的关系，是知识库补全系统的重要任务之一。当前有很多基于表示学习和基于符号逻辑的知识库补全算法被提出，这类算法都是基于知识库中现有的关系类型去预测知识库中未知的关系。但是在知识库中存在大量的属性事实型三元组未被使用，而这些属性事实三元组在进行关系预测时有很重要的作用，如：（A，isMarriedTo，B）、（A，HasChild，C）。当预测B是否是C的父母时，这些相关实体的年龄、性别等实体属性也很重要，如何能处理不同类型的实体属性特征，将其用于知识库补全系统中，是知识库补全工作的一个重要探索点。

在构建知识库补全系统的过程中，基于符号逻辑的算法通常通过训练一个二分类的分类器来判断实体和实体之间是否具有某种关系。通常这个分类器的正例是通过抽取现有知识库中存在的三元组作为正例，或称之为正例三元组，而负例则是采用随机采样等方法生成负例三元组。通常在知识库补全系统中，每个正例三元组可以构建上百组不同种类的负例三元组，如何解决正负例三元组不均衡的问题，正确预测实体和实体之间的关系，也是知识库补全算法的一个重要优化方向。

2.1　知识库构建和应用

在知识库构建过程中，数据完整性、数据精确性和数据质量是构建知识库的重要指标，通常知识库可以由三元组组成。知识库的构建可以有多种方法：（1）通过机器学习和自然语言处理技术自动抽取三元组[8]进行构建。（2）半自动抽取的方法，通过从维基

百科等网站的infobox基于规则模式抽取三元组。（3）基于协同创作[?]的方法，通过众包的形式系统创建知识库。（4）基于专家知识构建的领域内知识库，这些领域知识库有较高的专业性。

2.1.1 YAGO知识库

YAGO是一个从维基百科上抽取的、包含地理名词、WordNet[9]等数据的知识库。YAGO将WordNet的词汇定义与维基百科的分类体系进行了融合集成[10]，并集成了多种地理类型的数据，使得YAGO具有更加丰富的实体分类体系。YAGO还考虑了时间和空间知识[11]，为很多知识条目增加了时间和空间维度的属性描述。目前，YAGO包含1.2亿条结合关系三元组和实体属性三元组的知识。YAGO也是IBM Watson[12]的后端知识库之一。YAGO2[13]是YAGO的一个实例，当前YAGO2包括超过260万的实体和超过1.2亿的关系三元组知识，我们使用了其中实体的关系型三元组和属性型三元组共有4,484,914条、 37种关系型三元组的事实描述，同时有3,353,659条、35种属性性三元组的事实描述。

2.1.2 Freebase知识库

Freebase[4]是一个协同创作的知识库系统，内容通过用户添加，所有的条目采用结构化的形式，将结构分为三次：领域-类型-主题，其中，每一个条目叫做一个主题，每个主题包含不同的属性类型，一些同类型的主题组成一个类型，所有相关的类型构成一个领域。这种通过协同创作方式创建了结构化的人类知识，截止2007年Freebase包含1.25亿条三元组，超过4000种类型和超过7000种属性知识。一些基于Freebase研究结构化知识的问答系统[14][15],关注Freebase中的信息抽取和语义解析[16],也有一些研究关注Freebase中的实体消除歧义问题[17]。

除了YAGO和Freebase知识库，也有很多重要的知识库如：Google Knowledge Graph，Wikidata，DBpedia等，这些通用的数据库都有着各自的特点和特色，表3展示了不同类型的知识库的实体、关系类型和三元组等数量，其中的M表示百万数量。本研究中的实验考虑到同时使用属性三元组和实体关系三元组，从而选择在YAGO和Freebase数据集中筛选属性特征较多的实体，构建两个子数据集进行模型训练。

在通过自动、半自动的信息抽取、专家知识构建知识库后，有着各种各样广泛的应用，并能辅助人们检索、学习通用知识；一些专家对领域内的知识进行有序化，规范化，辅助人们去了解学习一个领域的专业知识。

表 3　不同知识库数据统计

知识库	实体数量	关系类型数量	三元组数量
Knowledge Graph	570M	35000	18000M
YAGO2	9.8M	114	447M
DBpedia	4.6M	1367	538M
Wikidata	18M	1632	66M
Freebase	40M	35000	637M

知识库最广泛的应用是谷歌和微软等商业公司的信息检索系统，2014年谷歌提出的知识图谱[18]就是结合FreeBase、Wikidata以及各种互联网数据进行信息抽取、知识整合后的著名系统，谷歌也在其博客上讲述了如何将知识库应用在实际的检索中，微软的Satori也是一个类似的知识库系统,被必应搜索广泛应用。其他搜索公司如百度、搜狗等公司都在知识库构建和信息检索领域有各自的应用，对于每一个用户搜索的关键词，我们可以通过知识库系统来返回更丰富，更全面的信息。比如搜索一个企业法人的姓名，我们的智能搜索引擎可以返回与这个人相关的所有历史借款记录、联系人信息、行为特征等。另外我们通过可视化把复杂的信息以非常直观的方式呈现出来，使得用户对于对隐藏信息的来龙去脉一目了然。

基于知识库的问答系统也有广泛的发展，2000年就有人提出基于知识的问答系统[19]，通过构架知识库完成问答。也有很多方法基于知识库进行表示学习[20]，通过学习实体的低维度表示，学习知识库的实体和实体之间的关系，然后构建相关的知识库问答系统。还有一些基于标签语义解析的方法，构建知识库问答系统[21]。一些论文[22]研究结合深度学习、关系抽取和问答系统，从而获得更好的知识库问答效果。一些论文[23]研究基于开放的知识库构建开放领域的问答系统，期望能基于百科知识进行问答系统构建。也有基于知识库的系统通过结合智能计算[24]，在医疗诊断领域获得一些突破。也有一些研究结合了知识库和医疗图像信息[25]，对一些医疗疾病进行诊断治疗，期望能获得更好的治疗效果。

知识库在金融、教育等领域也可以有各种有效的应用。在进行金融风险检测时，不一致性验证可以用来判断一个借款人的欺诈风险，这个跟交叉验证类似。比如借款人张三和借款人李四填写的是同一个公司电话，但张三填写的公司和李四填写的公司完全不一样，这就成了一个风险点，需要审核人员格外的注意。除了贷前的风险控制，知识库也可以在贷后发挥其强大的作用。比如在贷后失联客户管理的问题上，知识库可以帮助我们挖掘出更多潜在的新的联系人，从而提高催收的成功率。

知识库也可以帮助我们分析用户和理解用户。通过结合多种数据源去分析实体之间的关系，从而对用户的行为有更好的理解。比如一个公司的市场经理用知识库来分析用户之间的关系，去发现一个组织的共同喜好，从而可以有针对性的对某一类人群制定营销策略。只有我们能更好的、更深入的理解用户的需求，我们才能更好地去做营销。

2.2 知识库补全和推理

当前知识库补全方法主要有两种方法：基于基于表示学习的知识库补全和基于符号逻辑的知识库补全。基于表示学习的方法是通过学习实体和关系的低维度向量表示，用向量的相似度计算预测实体之间的关系。常见的表示学习方法有TransE[26]、TransH[27]、TransR[28]等，也有基于矩阵张量分解的表示学习方法如RESCAL[29]等。基于符号逻辑的方法主要包括AMIE[30]、PRA[31]和SFE[32]等；其中，AMIE方法通过从知识库中挖掘关联规则进行知识库补全， PRA方法基于连接实体的关系路径来预测它们之间的关系，SFE在PRA框架的基础上，抽取更多隐藏的关系路径特征进行模型预测。

2.2.1 基于表示学习的知识库补全

基于学习表示的知识库补全技术是近年来的研究热点之一。近年来获得了极大的关注热度。这类知识库补全技术通过不同的目标函数，希望能学习到知识库中实体和关系的低维度向量表示。这些实体的向量维度通常是50-300维度之间，将高维度稀疏的图数据张量，学习获得低维度的连续向量表示。获得实体和关系的向量表示后，可以通过向量之间余弦距离计算实体-关系-实体相似度。

2.2.1.1 RESCAL

RESCAL是2011年在ICML上发表的一篇基于协同学习解决多关系知识库补全问题的方法。将实体-关系-实体构建三维的张量矩阵。RESCAL基于潜藏模型，提供了一种有效的表示学习方法。和其他表示学习模型不同，RESCAL这类模型更多借鉴推荐系统中张量分解算法，训练学习实体和关系的向量表示。详细来说，RESCAL基于R阶的分解模型，χ_k被分解为：

$$\chi_k \approx AR_kA^T, k = 1...m$$

其中A表示一个$n \times r$的矩阵，表示n个实体和r种关系组成矩阵，R_k是一个$r \times r$的矩阵。则只需要最小化函数：

$$\min_{A,R_k} f(A, R_k) + g(A, R_k)$$

其中$f(A, R_k)$是度量χ_k和AR_kA^T距离的函数。 $g(A, R_k)$是A, R_k的复杂度惩罚项。通过梯度下降等学习算法可以计算获得模型的参数。

其他类似进行知识库补全的方法还有张量分解机算法[33], 基于隐藏变量的张量分解机算法[34]，结构化的低维表示[35]、无结构化的低维表示等。基于张量分解模型和推荐系统等领域的协同过滤算法相似，都是从矩阵补全角度学习向量模型。不同的是在知识库中，矩阵其实是由实体-关系-实体组成的三维张量，而推荐系统中的用户-商品矩阵是二维向量。一些实验[36]表明在稀疏的知识库图中，使用张量分解模型有较好的预测效果。

2.2.1.2 TransE

TransE是2013年谷歌发表在NIPS上的一篇论文，研究考虑如何将知识库中的多种不同的关系、实体，学习获得它们的低维向量表示，期望能获得$h + r \approx t$效果，其中h和t分别是头实体和尾实体学习到的低维向量，r是头实体和尾实体之间的关系向量表示。TransE模型定义了间隔损失函数：

$$L = \sum_{(h,r,t)\in S} \sum_{(h',r,t')\in S'} [\gamma + d(h + r, t) - d(h' + r, t')]_+$$

其中γ是间隔超参数，s是真实存在的知识库,(h, r, t)是这个知识库中的三元组，s'是负例三元组组成的知识库实体，(h', r, t')表示负例三元组。$d(h + r, t)$表示头实体加上关系向量和尾实体之间的欧氏距离。通过定义间隔损失函数，TransE期望学习的模型能保证正实例对比负实例对距离小，这样就能保证正实例对比负实例对余弦相似度更高。

其他的算法如TransH和TransR都是基于TransE模型的基础上，通过学习更准确有效的损失函数来获得低维度向量表示。 TransH考虑一些关系中的映射属性知识，如一对一、一对多、多对多关系等，训练学习这些关系的低维度向量表示。 TransR[37]考虑实体和关系的多方面属性，在分割独立的向量空间中学习实体、关系的低维度向量表示。ProjE[38]基于神经网络的实体关系投影表示学习。

2.2.2 基于符号逻辑的知识库补全

基于逻辑符号的知识库补全算法也是多年来知识库补全和知识库推理中的关注热点之一。从90年代昆兰等人提出的规则归纳，将观察集数据中的知识以规则的形式提炼出来，到近年来热门的基于路径排序算法（PRA）和子图特征抽取（SFE）的知识库推理补全。这类进行知识库补全的算法多是利用符号逻辑，统计发现知识库图数据结构中存在的规则或规律，从而进行知识库补全预测。

2.2.2.1 规则挖掘

规则挖掘基于训练集知识库中的数据，通过挖掘知识库中隐藏的关联规则，进行知识发现和知识库补全。早期昆兰等人提出了一阶逻辑推理的算法FOIL[39]，通过学习知识库中的例子来构建霍顿子句。如：

$$MotherOf(A,C) \wedge MarriedTo(A,B) \Rightarrow FatherOf(B,C)$$

就是一个典型的一阶逻辑推理。FOCL[40]等算法也是基于一阶逻辑推理构建的知识库搜索算法。近年来随着互联网的快速发展，Schoenmackers[41]等人也将一阶逻辑逻辑推理用于互联网文本中。其他规则挖掘的典型算法包括AMIE[42],AMIE受关联规则启发，基于开放世界假设，在一阶逻辑推理的基础上，能更快更高效的处理大规模的知识库数据，作者此后对AMIE算法进行改进获得AMIE+[43]，这些算法在 YAGO知识库中获得了很好的效果。RDF2Rules[44]等算法也是基于逻辑推理，构建频繁的谓词圈，获得了更加高效和准确的逻辑推理算法。

2.2.2.2 路径排序算法

路径排序算法[31]是2010年，劳逆等人提出的知识库补全算法。在传统的一阶逻辑推理的基础上，路径排序算法基于随机游走，在由知识库构成的图数据中查找更多、更长的有效的路径进行知识库推理，和规则学习方法不同，路径排序算法在抽取图中的关系路径后，构建了分类学习器，通过分类器学习不同关系路径的权重。利用每种关系下三元组的不同关系路径权重，来进行新关系事实的预测，从而构建了新的知识库补全算法。此外，作者还构建了基于反向的随机游走图搜索算法[45]，进一步提升路径排序算法在知识库补全中的效果。考虑给定一个知识库关系r_i，我们抽取所有具有这种关系的实体对构成集合：

$$R_i = \{(h_{ij}, t_{ij}), (h_{ij}, r_i, t_{ij}) \in KB\}$$

随机游走查找路径后，我们抽取所有连接实体对(h_{ij}, t_{ij})的路径类型p_i构成分类特征：

$$P_i = \{p_i|(h_{ij}, p_i, t_{ij}) \in KB\}$$

从而我们可以选择合适的分类器模型进行关系路径的训练学习，利用合适的路径类型来进行关系预测，知识库补全。

子图特征抽取[32]提供了一种更加简便有效的关系路径计算方法。相比于路径排序算法，子图特征抽取能获得更多的潜藏路径，在进行关系预测的时候能获得更加显著有效的提升。此外，也有很多算法结合了子图特征抽取和表示学习算法[46]进行知识库补全。一些算法对相似的关系进行聚类，构建多任务的路径排序模型[47]。

2.3 知识库补全相关假设

知识库补全系统中，通常需要假定知识库中三元组正负实例对的正确性，或者在何种限制模式生成三元组负例。常用的有三种假设：（1）开放世界假说，（2）封闭世界假说，（3）局部封闭世界假说。我们给定一个三元组集合D^+表示正例：

$$D^+ = \{(h, r, t)|(h, t) \in KB\}$$

对于开放世界假说，给定一个知识库中的三元组集合，我们认为这些实际存在三元组是正例三元组，对于在知识库中不存在的三元组，开放世界假说认为这个三元组是不确定的，可以通过概率大小预测这个三元组的正确性。

对于封闭世界假说，给定一个知识库三元组集合，我们认为只有知识库存在的三元组才是正例三元组。对于知识库中不存在的三元组，封闭世界假说认为这些三元组是负例三元组。但通常这样会产生正负实例极其不平衡情况，每个知识库中存在三元组都能生成数百例负例三元组。这些负例三元组组成负例集合D^-：

$$D^- = \{(h_j, r, t_i)|(h_i \neq h_j \wedge (h_i, t_i) \in KB)\}$$

$$\cup\{(h_i, r, t_j)|(t_j \neq t_i \wedge (h_i, t_i) \in KB)\}$$

局部封闭世界假说对封闭世界假说进行了改进。给定一个关系，知识库中这个关系的三元组记为正例三元组，随机在这个关系中替换头尾实体，生成这个关系下实体集合中错误的实体对，这样就可以大大减少所有负例三元组个数。基于局部封闭世界假设构

建的负例三元组集合D^-可以用如下公式表示：

$$D^- = \{(h_{ik}, r_i, t_{ij}) | (h_{ik} \neq h_{ij} \wedge (h_{ik}, t_{ik}) \in KB \wedge (h_{ij}, t_{ij}) \in KB)\}$$

$$\cup \{(h_{ij}, r_i, t_{ik}) | (t_{ij} \neq t_{ik} \wedge (h_{ik}, t_{ik}) \in KB \wedge (h_{ij}, t_{ij}) \in KB)\}$$

2.4 评价方式

知识库补全效果的评价方法和信息检索类似，主要采用MAP和MRR[46]指标进行模型评价。同时考虑到在一个给定关系下，给定一个头实体和一系列的候选尾实体，我们希望排在前面的实体对是正确的实体对，如果正例实体对排在负例实体对后面，则这个模型的预测效果有待改进。所以我们借鉴了信息检索领域相关的评价方法，除了采用关键的MAP评价指标，我们也采用了一些常见的AUC和Hit@1评价指标，希望能详尽的分析模型的有效性和可预测性。本部分介绍了四类评价知识库补全的模型评估指标，包括：（1）hit@1，描述排序在第一位的正例三元组比例，（2）平均准确率MAP，衡量所有正例三元组在整个排序列表中的排序情况，值越大说明整体排序情况越好，（3）平均秩倒排序MRR,每组三元组中，正例三元组秩序排名，值越大越好；（4）AUC，使用二分类模型评价函数来预测知识库补全效果。

在信息检索中，hit@1表示每组三元组对应的正负实例中，正实例在该组正负例三元组中排名第一的三元组比例，尽管这种评价指标较为简单，但可以有效衡量模型中正例三元组排在第一位比例。对于每个关系，我们定义的的Hit@1计算公式如下：

$$hit@1 = \frac{\sum_{q=1}^{n} hit_q}{n}$$

其中hit_q表示该关系下第q组实体对预测结果中，正例是否排名第一，如果正例排名第一则hit_q为1，否则为0。

平均准确率的英文全称是mean average precision。MAP的衡量标准比较单一，给定一个头实体和一系列候选尾实体，头尾实体对之间的关系非0即1，我们采用局部封闭世界假设，在知识库中存在的实体对标记为1，即为正例，否则为负例。 MAP核心是利用头实体对应的相关的尾实体出现的位置来进行排序算法准确性的评估。它反映系统在全部相关文档上性能的单值指标。系统模型计算出来的实体对越靠前(rank 越高)，MAP就应该越高。否则准确率默认为0。MAP的计算公式如下：

$$MAP = \frac{\sum_{q=1}^{n} AP(q)}{n}$$

$$AP(q) = \frac{\sum_{i=1}^{k} P(i) \times rel(i)}{k}$$

其中MAP表示n个头实体和其候选尾实体的平均准确度的均值。AP(q)是一个头实体和其K个候选尾实体排序结果的评价，如果正确的候选尾实体排在错误的候选尾实体之前，则AP值越高；如果所有的头实体和其候选尾实体AP值越高，则模型的MAP值也就越高。$rel(i)$在第i个实体对是预测结果时为1，否则为0。

MRR的全称是mean reciprocal rank。MRR通过定义第一个正确的候选尾实体来判断模型好坏，一个模型的MRR值越高，说明正确的候选尾实体在模型中排序越靠前。 MRR通常和MAP一起综合评价一个模型的好坏。 MRR的定义公式如下：

$$MRR = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{r_i}$$

其中n表示一个关系下所有不同头实体的数目，r_i是每个头实体对应的正确尾实体的秩序。 MRR在很多基于问答系统的知识库补全系统[48]中也是常见的评价方法之一。通常情况下，基于符号逻辑的知识库补全系统中MRR值较高，多数实验结果都在0.9左右，大部分关系中的MRR实验结果差别不显著，所以本研究的重点是关注如何显著提高知识库补全系统中关系补全的MAP指标。

除了常见的排序模型评价方式，我们在实验过程中也引入二分类分类器模型进行模型评价，同时也考虑模型对正负例排序能力。AUC（Area Under Curve）被定义为ROC曲线（tpr和fpr曲线函数）下的面积，显然这个面积的数值不会大于1，同时这个值一般大于0.5。其中tpr和fpr的计算公式如下所示：

$$tpr = \frac{tp}{tp + fn}$$

$$fpr = \frac{fp}{fp + tn}$$

当你随机挑选一个正样本以及一个负样本，当前的分类算法根据计算得到的Score值将这个正样本排在负样本前面的概率就是AUC值。对所有实体对的预测得分进行排序，选择不同的阈值对正负例进行切分计算tpr和fpr的得分，从而绘制出完整的ROC曲线，并计算得到AUC的面积大小。AUC取值越大，当前的分类算法越有可能将正样本排在负样本前面，即能够更好的分类，整个模型的排序效果也更好。AUC计算公式如下所示：

$$auc = \int_0^1 tpr * dfpr$$

在知识库补全预测中，无论是逻辑回归模型、支持向量机模型还是学习排序模型，都需要关注模型的二分类的分类器性能问题。对于知识库中的实体对排序、实体对打分预测，在评估一组实体对集合的排序秩序性能之外，也关注模型的分类性能好坏，这些对于知识库补全算法是十分必要的。

2.5 本章工作总结

本部分介绍了当前知识库技术相关领域的一些重要研究内容，从知识库构建技术到知识库应用和发展，我们学习研究了各种重要的知识库构建、推理、应用相关内容，介绍了如何评价知识库补全算法的相关指标。本部分重点关注了知识库补全算法相关的研究现状，基于符号逻辑和基于表示学习的知识库补全算法一直是相关领域的研究热点，本部分抓住这些研究中的重要公式、推理和相关假设，为构建我的知识库补全算法框架提供了坚实的基础。

第 3 章　结合关系路径和实体属性的知识库补全

大规模的知识库包含大量的实体、关系和属性值，使用三元组的形式对现实世界中实体的各种知识进行表示，三元组包括关系型和属性型两大类。关系型三元组形如（北京师范大学，位于，北京），其中“北京师范大学”和“北京”分别表示关系型三元组的头实体和尾实体，“位于”表示是关系路径特征；属性型三元组形如（北京师范大学，建校于，1902年），其中“北京师范大学”是头实体，“建校于”是实体属性特征，“1902年”是具体的实体属性特征值。虽然许多知识库的规模很大，但他们仍然是不完备的，如很多人的出生地点并未包含在知识库中，一些演员是否出演过某些电影也是未知的。为了解决这个问题，很多知识库补全的方法被提出来，这类方法的作用是基于知识库中已有的三元组预测新的三元组。

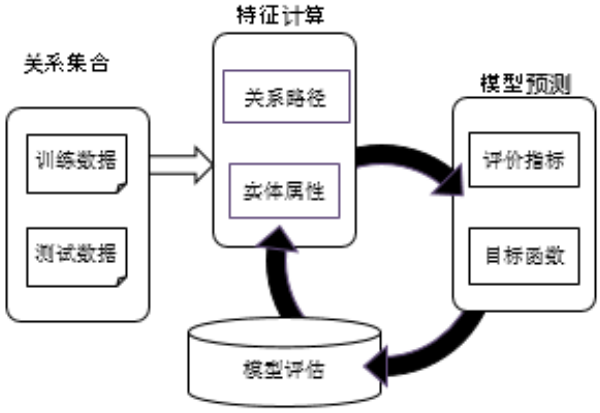

图 3　知识库补全计算框架

本章节提供一种结合关系路径特征和实体属性特征的知识库补全方法。通过提取知识库中实体的关系路径特征和实体属性特征，构建模型，进行知识库的实体关系预测。除此之外，本论文研究了基于学习排序算法进行知识库补全的技术，在传统基于分类器模型的基础上，提出了非线性的排序算法，对知识库候选实体对进行排序计算，获得更优的候选实体对排序集合。

图3展示了本部分基于逻辑符号的知识库补全算法基本框架。其中对于每个关系集合，将集合切分为训练集和测试集。模型的训练分为两个部分：（1）特征计算，包括关系路径特征计算和实体属性特征计算两个部分；（2）模型预测，包括基于分类模型的逻

辑回归算法，基于树模型的学习排序算法，基于深度神经网络的排序算法。根据对模型评价结果，不断调整预测模型参数，使得模型能够达到最优的结果。

3.1 问题引入

特征计算是知识库补全的特征抽取阶段，特征的数量和特征有效性是决定知识库补全效果的关键点之一。给定一个完整的知识库，我们将知识库中的实体和关系分别转化为图中的顶点和边，这样就能构建一个基于图模型的知识库补全系统。对于知识库中每种待预测的关系，我们抽取对应关系下的实体对，将实体对的头实体和尾实体在图中进行随机游走[49]，获得连接头实体和尾实体的关系路径特征，从而获得了知识库补全的关系路径特征。给定一个目标关系r和这个关系对应的实体对集合

$$I_s = \{(s_j, t_j) | < s_j, t_j > \in KB\}$$

我们期望找出连接头尾实体对的关系路径特征。由于连接头尾实体对之间的路径数量很大，通常需要限定关系路径的长度，并采用随机游走算法计算选择从头实体到尾实体的关系路径，将这些路径集合作为关系路径特征。对于能连接从头实体到尾实体的关系路径，我们记录这个关系路径的类型，作为模型预测的特征。我们计算0-1二值化后的的路径类型，作为关系路径特征的特征值。对于每个关系抽取的关系特征，我们最终记为$V_r\ (h_i, t_i)$作为关系路径特征，表示从头实体到尾实体有哪些关系路径进行连接。例如对于实体对（北京师范大学，位于，北京）三元组，我们可以基于上述的关系路径抽取算法获得（北京师范大学，位于，北京市海淀区，位于，北京）、（北京师范大学，有校长，董奇，居住在，北京）等多条路径来获得关系“位于”对应的关系路径类型特征。

本论文通过枚举不同维度的实体属性特征，计算这些实体对在这些实体属性特征下的特征值。属性特征抽取过程较为复杂，不仅需要考虑不同类型的属性信息不一致问题，也要研究如何处理缺失值问题。首先对于每个实体，有着不同类型的描述特征，如一个人的信息，不仅有出生年月这种时间类型的信息，也有年龄、性别等不同种类的属性信息。我们采用了对实体信息进行标准化的方法，将实体对的头实体和尾实体这些不同属性特征进行归一化处理范围限定在[0，1]之间。进行归一化后计算获得新的属性特征记为$V_l(h_i)$和$V_l(t_i)$，分别表示头实体和尾实体在所有熟悉特征下的实体属性特征向量，其中h_i 和t_i表示给定关系l的第i个头实体和尾实体，同时我们对于头实体和尾实体进行相减计算获得$V_l(h_i - t_i)$属性值。除此之外，对于很多实体在不同特征上的缺失值，我们将缺失值进行了补0处理，期望获得更优结果。如对于“北京师范大学”这个实体，我们抽取了他

的“建校时间”、“占地面积”等所有的实体属性特征，并计算每个实体对在这些特征下标准化、归一化等特征工程转化后的特征值，而且将缺失属性特征补0。

3.2 关系路径特征计算

特征计算是知识库补全的特征抽取阶段，特征的数量和特征有效性是决定知识库补全效果的关键点之一。给定一个完整的知识库，我们将知识库中的实体和关系分别看做为图中的顶点和边，这样就能构建一个基于图模型的知识库补全系统。对于知识库中每种特定待预测的关系，我们抽取对应关系下的实体对，将实体对的头实体和尾实体在图中进行随机游走[49]，获得连接头实体和尾实体的关系路径特征，从而获得了知识库补全的关系路径特征。给定一个目标关系r和这个关系对应的实体对集合

$$I_s = \{(s_j, t_j) | < s_j, t_j > \in KB\}$$

3.2.1 关系路径类型集合

我们期望找出给定关系下，所有连接头尾实体对的关系路径特征，最终获得一个关系路径类型的集合。在知识库中，由于连接头尾实体对之间的路径数量很大，通常需要限定关系路径的长度，一般限定图中的关系路径长度在2-6之间。

对于能连接从头实体到尾实体的关系路径，我们记录这个关系路径的类型，作为模型预测的特征。我们采用随机游走算法[50]计算选择从头实体到尾实体的关系路径，将这些路径集合作为关系路径类型特征集合。例如对于实体对（北京师范大学，位于(locatedIn)，北京）三元组，我们可以基于上述的关系路径抽取算法获得（北京师范大学，位于，海淀区，位于，北京）、（北京师范大学，有校长，董奇，居住在，北京）等多条路径来获得关系“位于”对应的关系路径特征。对于上述两条关系路径，我们抽取关系路径中的关系类型，生成一个关系路径类型集合，包括$\{M \rightarrow M\}$和$\{\Psi!^{\cdot\cdot} toEO()$等不同关系组成的关系路径类型，从将多种不同的关系路径类型结合而构成一个关系路径类型集合。对于反方向的关系路径我们采用$EO(^{-1}$表示。

3.2.2 关系路径特征向量

获取关系路径类型集合（记为S(r)）后，我们将给定关系r下的实体对(s_j, t_j)计算所有关系路径类型集合下的特征值。如果实体对(s_j, t_j)在集合S(r)中的某个关系路径s_i存在，则我们可以将该实体对的对应关系路径类型的特征值记为1，否则该实体对在这个关系路径

类型下的特征值记为0，这种算法简化了不同关系类型在知识库中的权重，但是一些实验表明[46] 进行0-1二值化可以简化关系路径类型特征和特征值计算过程，同时对于实验结果的影响并不显著。为了能在不影响模型效果的前提下，加速我们的计算过程，我们对自己的关系路径类型特征值计算进行了0-1二值化。

3.3 实体属性特征计算

除了3.2中计算得到的关系路径类型，考虑到在知识库中任然存在大量的实体属性特征并未被使用，本研究也对每个关系下的实体属性特征进行计算，获得这些实体属性特征下的特征值。实体属性特征获取较为简单，只需要枚举知识库中存在的不同实体属性类型，即可这些实体属性类型作为实体特征集合。实体属性特征的处理和计算是获取更有效特征、提升知识库补全算法的关键。

属性特征抽取过程较为复杂，不仅需要考虑不同维度的属性信息不一致问题，也要研究如何处理缺失值问题。本论文通过枚举不同维度的实体属性特征，计算这些实体对在这些实体属性特征下的特征值。首先对于每个实体，有着不同类型的描述特征，如一个人的信息，不仅有出生年月这种时间类型的信息，也有年龄、性别等不同种类的属性信息。我们采用了对实体信息进行标准化的方法，将实体对的头实体和尾实体这些不同属性特征进行归一化处理范围限定在[0，1]之间。进行标准化后计算获得新的属性特征记为$V_l(h_i)$和$V_l(t_i)$，分别表示头实体和尾实体在所有熟悉特征下的实体属性特征向量，其中h_i 和t_i表示给定关系l的第i个头实体和尾实体，同时我们对于头实体和尾实体进行相减计算获得$V_l(h_i - t_i)$属性值。除此之外，对于很多实体在不同特征上的缺失值，我们将缺失值进行了补0处理，期望获得更优结果。如对于“北京师范大学”这个实体，我们抽取了他的“建校时间”、“占地面积”等所有的实体属性特征，并计算每个实体对在这些特征下标准化后的特征值，而且将缺失属性特征补0。

在对每个关系下的实体对集合进行实体属性特征抽取计算后，我们获得了这个关系下头实体的实体属性特征、尾实体的实体属性特征、对头尾实体进行归一化后的归一化实体属性特征，以及头实体和尾实体差值计算得到的实体属性特征。

通过将3.2关系路径特征和3.3抽取的实体属性特征进行结合，我们获得了基于关系路径类型和实体属性类型的特征矩阵，并构建机器学习模型对生成的特征矩阵进行学习预测，获得知识库补全中的实体对的关系。

3.4 关系路径和实体属性特征实验

本研究通过抽取关系路径特征和实体属性特征，并对抽取特征进行计算后，构建了基于逻辑回归的预测模型，期望能进行研究分析，发掘结合关系路径和实体属性特征的新方法如图4所示，我们获取知识库中的关系型特征类型（关系路径特征类型）和非关系型特征类型（实体属性特征类型），将两种特征相互融合构建实体对的预测向量，从而获得知识库补全模型，能更好进行知识库补全预测。

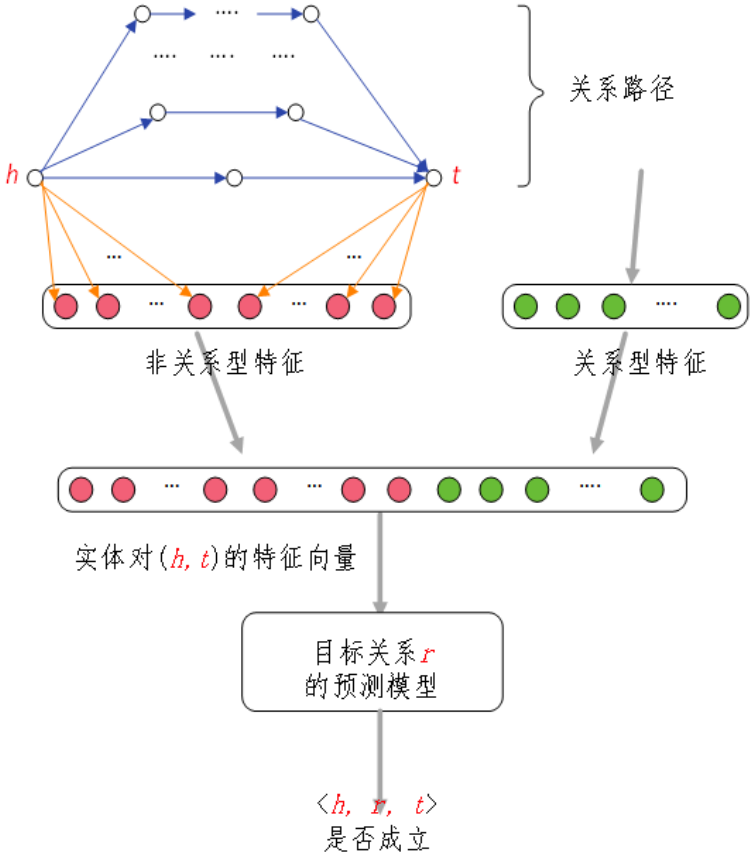

图 4　结合实体属性和关系路径知识库补全计算过程

本研究的实验分为两组，包括基于YAGO知识库和基于Freebase知识库的模型预测评估。尽管有很多不同类型的知识库，考虑到这些知识库需要包含关系三元组和实体属性三元组，我们选择了这些知识较为丰富的YAGO和Freebase知识库。对于这些知识库，我们抽取其中实体属性特征较多的三元组，作为一个训练预测数据，从而避免很多缺少实体属性三元组的实体。

知识库并采用MAP、MRR、AUC和hit@1多种不同的评估指标进行分析，其中MAP、MRR和hit@1是在排序模型中常见的模型评价指标，AUC是常见的分类和排序模型评价指标，这些指标也是很多常见知识库补全、信息检索、图链接预测任务的评估方法。本研究全面深入的研究了结合实体属性和关系路径的知识库补全算法，在不同知识库中如何结合，从而有效地提高模型效果。

3.4.1 YAGO知识库实验

本部分的效果展示了在YAGO知识库下进行知识库补全算法的结果，我们的方法和对比方法被分为两组。其中，SFE-literal（图中简称SFE-lit）和PRA-literal（图中简称PRA-lit）是结合关系路径特征和实体属性特征的方法，其MAP、MRR和AUC预测结果最好，总体的预测结果最稳定，效果最好，而未加实体属性特征的PRA和SFE特征较好，使用表示学习的TransE和TransR模型效果在MAP中预测较差，但是通过AUC和hit@1进行模型的二分类效果结果和基于符号逻辑的效果相似。结果显示，加入实体属性特征后，结合实体属性特征和关系路径特征的补全技术，相比只采用关系路径特征的补全技术有更高的准确性，在进行排序相关的预测效果评测时，有较好的效果。在采用二分类模型进行评价时候，无论是符号逻辑的路径排序相关算法还是表示学习的低维嵌入算法，都能有效的进行模型预测。表4中展示了YAGO和Freebase知识库的基本数据情况。

表 4　知识库基本数据统计

	FB15K	YAGO
实体	14951	58130
关系类型	1345	32
实体属性类型	336	25
属性三元组	25776	141606
关系三元组	592213	499350
训练集	483142	399480
测试集	59071	49132

相比只采用关系路径的知识库补全算法，结合实体属性特征和关系路径特征的知识库补全算法，在YAGO数据集合上，结果有较大的提升，相比原来的模型，SFE-literal和PRA-literal在计算MAP时，都获得了近1%的效果提升，模型提升效果统计显著。基于上述实验可以获得如下结论：（1）预测知识库中新三元组通过结合关系路径特征和实体属性特征能更加的精确有效，无论是以分类为优化目标的分类模型，还是以排序

为优化目标的排序问题。（2）由于在YAGO知识库集合中，有更多的属性事实进行关系预测，采用特征组合，能获得更多的不同类型组合特征，达到更好的优化结果，因此，结合属性事实和关系事实进行预测是非常重要的。（3）对于某些特殊的关系，进行标准化处理是非常有效的，但是并非对于所有的属性事实进行标准化有效，选择合适的属性特征和实体关系特征进行组合是十分必要的。（4）本研究的实验结果表明，对于多数YAGO2中的关系来说，我们的属性事实不仅可以用来预测关系事实，而且还能调整原来的关系特征的路径权重，使得模型预测更加合理。因此，结合属性事实和更丰富的关系特征能获得更好的知识库补全结果。

表 5　YAGO知识库实体属性和关系路径特征模型预测结果

	MAP	MRR	AUC	hit@1
SFE-lit	78.610%	100.000%	92.994%	45.328%
PRA-lit	78.170%	97.830%	93.496%	45.104%
PRA	77.610%	97.830%	93.258%	45.509%
SFE	77.360%	97.830%	93.225%	45.502%
transE	56.400%	91.200%	92.740%	45.642%
transR	46.470%	65.410%	92.986%	46.128%

我们进一步分析了每个关系下关系路径和实体属性特征的权重，表6显示了在YAGO知识库中的三种关系：Export、GraduateFrom、HasAcademicAdvisor，通过结合实体属性特征和关系路径特征，学习获得的重要权重。相比于路径排序算法这种只使用关系特征进行预测的方式，结合属性事实特征不仅仅能增加模型预测的全面性，将模型预测结果精度提高，同时也能调整逻辑回归算法中不同关系路径特征和不同属性事实特征的权重。通过将关系路径特征和属性事实特征进行结合，使得模型的预测结果更加可靠。如表所示，我们分析关系“graduateFrom”可以发现，除了常见的关系路径特征： isCitizenOf → isCitizenOf^{-1} → livesIn → isLocatedIn^{-1}、 isAffiliatedTo^{-1} → isCitizenOf → livesIn → isLocatedIn^{-1} 等。一些重要的实体属性特征如：wasCreatedOnDate、happenedOnDate等都在关系预测结果中有着重要的作用。通过分析这些关系路径的权重我们可以发现，实体属性特征不仅能提高关系预测中的精度，有较高的准确性，同时也能调整关系路径的权重，产生更多的有表达力的关系路径特征。

3.4.2　Freebase知识库实验

本部分的效果展示了在Freebase知识库下进行知识库补全算法的结果，我们的方法

表 6　YAGO知识库关系路径和实体属性特征比较

Export		
PRA	imports→ hasMusicalRole^{-1}→ hasMusicalRole	
	livesIn^{-1}→ diedIn→ imports	
	livesIn^{-1}→ wasBornIn→ dealsWith^{-1} → imports	
	isCitizenOf^{-1}→ influences^{-1} → isCitizenOf^{-1} → imports	
	isCitizenOf^{-1}→ wasBornIn → dealsWith^{-1} → imports	
PRA-lit	hasCapital → hasCapital^{-1} → exports	hasGini
	imports → hasMusicalRole^{-1} → hasMusicalRole^{-1}	hasInfaltion
	isCitizenOf^{-1} → wasBornIn → hasCapital^{-1} → exports	hasEconomicGrowth
	exports → hasMusicalRole^{-1} → hasMusicalRole	hasPoverty
	isInterestedIn^{-1}→ wasBornIn → hasCapital^{-1} → exports	wasDestroyedOnDate
GraduateFrom		
PRA	isCitizenOf → isCitizenOf^{-1} → livesIn → isLocatedIn^{-1}	
	diedIn → happenedIn^{-1} → participatedIn^{-1} → isLocatedIn^{-1}	
	hasWebsite → hasWebsite^{-1} → livesIn → isLocatedin^{-1}	
	isAffiliatedTo → isAffiliatedTo^{-1} → isCitizenOf → isLocatedIn^{-1}	
	isAffiliatedTo^{-1} → isCitizenOf → livesIn → isLocatedIn^{-1}	
PRA-lit	isAffiliatedTo → isAffiliatedTo^{-1} → isCitizenOf → isLocatedIn^{-1}	happenedOnDate
	hasAcademicAdvisor → hasAcademicAdvisor^{-1} → graduated-From	wasDestroyedOnDate
	hasWebsite → hasWebsite^{-1} → livesIn → isLocatedin^{-1}	wasDestroyedOnDate
	isAffiliatedTo → isAffiliatedTo^{-1} → isLeaderOf → isLocatedIn^{-1}	wasCreatedOnDate
	isCitizenOf → isCitizenOf^{-1} → livesIn → isLocatedIn^{-1}	wasBornOnDate
HasAcademicAdvisor		
PRA	wasBornIn → happenedIn^{-1} → participatedIn^{-1} → livesIn^{-1}	
	diedIn → hasCapital^{-1} → isLocatedIn^{-1} → diedIn^{-1}	
	worksAt → graduatedFrom^{-1} → livesIn → wasBornIn^{-1}	
	hasAcademicAdvisor → hasAcademicAdvisor^{-1}→ influences^{-1} → hasAcademicAdvisor	
	livesIn→ diedIn → hasAcademicAdvisor	
PRA-lit	hasGender → hasGender^{-1}	wasDestroyedOnDate
	worksAt → graduatedFrom^{-1} → livesIn → wasBornIn^{-1}	hasHeight
	hasAcademicAdvisor → hasAcademicAdvisor^{-1} → diedIn → diedIn^{-1}	hasHeight
	diedIn → diedIn^{-1} → livesIn → livesIn^{-1}	wasBornOnDate
	graduatedFrom → worksAt → worksAt^{-1} → worksAt	diedOnDate

表 7　Freebase知识库实体属性和关系路径特征模型预测结果

	MAP	MRR	AUC	hit@1
transR	66.440%	95.500%	86.182%	44.145%
transE	72.910%	98.650%	89.118%	47.348%
SFE	86.490%	98.650%	97.087%	55.559%
SFE-lit	86.560%	100.000%	97.140%	55.636%
PRA	86.930%	98.200%	97.021%	55.527%
PRA-lit	87.140%	100.000%	97.108%	55.606%

和对比方法被分为两组，表7展示了在不同评价指标下Freebase知识库中部分关系的预测结果。其中，SFE-literal和PRA-literal是结合关系路径特征和实体属性特征的方法，其MAP、MRR和AUC预测结果最好，尤其在采用AUC进行模型评价时候，相比于表示学习方法，基于符号逻辑的知识库补全算法特征有非常高的提升。从基于MAP和MRR的评价指标来看，加入实体属性特征的总体预测结果最稳定，效果最好，而未加实体属性特征的PRA和SFE特征较好，使用表示学习的TransE和TransR模型效果在MAP中预测较差，结果显示，加入实体属性特征后，结合实体属性特征和关系路径特征的补全技术，相比只采用关系路径特征的补全技术有更高的准确性，在进行排序相关的预测效果评测时，有较好的效果。在采用二分类模型进行评价时候，无论是符号逻辑的路径排序相关算法还是表示学习的低维嵌入算法，都能有效的进行模型预测。

总体上看，使用MAP、MRR、AUC以及hit@1评价指标时，基于符号逻辑的知识库补全算法效果较基于表示学习的算法性能有较大的提升，而结合实体属性特征和结合关系路径特征的知识库补全算法相比仅仅使用关系路径特征的模型效果更好。这说明无论采用分类作为模型的优化方向，还是采用实体对排序的序列作为知识库补全优化目标，使用符号逻辑相比于近似的表示学习方法效果都有明显的提升。

分关系来看，从表8可以分析发现，在Freebase的15种关键关系中，我们通过采用结合关系路径特征和实体属性特征的知识库补全算法，相比于只采用关系路径的知识库补全算法，我们的模型效果在MAP排序指标上有很大的提升。如/tv/tv/genreprograms、/film/actor/film/film/performance/film和 /film/director/film、/media/commonnetflix/genretitles等关系使用MAP评测都有较大的模型效果提升。

我们以关系/people/person/nationality和关系/film/actor/film/film/performance/film进行分析，研究关系路径特征和实体属性特征在路径排序算法中的特征表现。通过分析MAP指标来看，TransE和TransR这两个模型在各个关系上的效果较为稳定，符号逻辑相关算法在这些不同的关系上效果差别较大，通过分析相关的关系路径和实体属性权重，我们可以

表 8　FB15K部分关系MAP得分

MAP-score	PRA-lit	PRA	SFE-lit	SFE	transE	transR
/people/person/profession	12.150%	11.902%	12.359%	12.096%	77.134%	71.784%
/film/actor/film/film/performance/film	28.309%	26.179%	22.121%	20.560%	87.007%	78.431%
/media/common/netflix/genre/titles	50.135%	48.281%	49.919%	49.546%	63.788%	65.603%
/people/ethnicity/people	57.143%	57.143%	57.143%	57.143%	60.402%	50.396%
/film/film/genre/films/in/this/genre	57.285%	55.650%	41.435%	41.828%	68.762%	64.254%
/music/instrument/instrumentalists	58.268%	58.268%	58.268%	58.268%	53.702%	52.452%
/music/genre/artists	69.514%	69.514%	69.514%	69.514%	77.745%	65.328%
/location/location/time/zones	77.479%	77.130%	77.479%	77.479%	82.790%	72.245%
/tv/tv/genre/programs	80.952%	80.952%	80.952%	80.952%	78.140%	60.484%
/people/person/nationality	81.119%	79.616%	81.774%	80.770%	75.448%	73.757%
/people/cause/of/death/people	82.653%	82.653%	82.653%	82.653%	52.979%	44.508%
/music/record/label/artist	83.704%	83.704%	83.704%	83.704%	55.037%	41.867%
/film/production/company/films	85.561%	85.561%	85.561%	85.561%	68.758%	59.189%
/film/director/film	100.000%	100.000%	100.000%	100.000%	94.077%	88.385%
/film/film/directed/by	100.000%	100.000%	100.000%	100.000%	95.254%	85.679%

发现，这些关系路径特征相对于YAGO知识库来说，关系路径权重区别并不明显，甚至在结合实体属性特征后，对于/film/actor/film/film/performance/film这个关系来说，所有的关系路径和实体属性路径特征综合从44120减少到42787种特征。说明基于随机游走抽取的关系路径特征并不能很好的区分Freebase中实体和实体之间的关系，但是当加入关系路径特征后，能很好的惩罚这些加入区分度不好的关系路径特征。

其次尽管部分关系相对于表示学习来说，预测结果的效果较差，但是，基于图模型的知识库补全系统中，大部分关系的预测相对于基于表示学习效果提升很多。这一方面说明基于表示学习是一种近似的模型预测，而基于符号逻辑的关系路径学习排序则是一种精确而有效的学习方式。另一方面也说明，在知识库构建的图模型中，部分关系较为稀疏的实体之间使用表示学习，预测结果较为理想，而采用结合实体属性和关系路径的符号逻辑补全算法，则能基于有效的关系路径，获得十分有用的预测模型结果。

其次，由于Freebase知识库总共包含超过1345种不同类型的关系，这些关系覆盖了电影、人种、电视剧、音乐等不同类型的关系，这些不同的关系中部分关系具有对称性，一些关系和关系之间差别很大，如何能将这些不同类型互不影响或者相互对称的关系区分出来，构建一个合理的图，基于这种合理的图模型进行关系路径的抽取和实体属性特征的计算，是未来研究的重要研究步骤。

3.5 本章工作总结

本部分通过抽取了实体属性类型、关系路径类型作为知识库补全的特征向量类型，并通过计算每个关系下实体对的实体属性特征值和关系路径特征值。将这些特征值进行组合，并构建了简单的逻辑回归算法模型，通过再不同知识库中的实验，评估了不同类型特征组合的情况下，我们结合关系路径特征和实体属性特征相比传统的关系路径特征计算问题，在分类预测、排序预测评估指标下的优势。

在接下来的章节中，本研究将通过构建不同的机器学习模型，比较逻辑回归分类算法、基于学习排序的树模型、深度学习模型等来评估知识库补全算法应该如何选择预测模型。

第 4 章　知识库补全预测模型

4.1　问题引入

当前的知识库基于打分模型进行知识库补全有很大不足。一是知识库中正负实体对比例差别很大，对于每个在知识库中实际存在的三元组正实例，可能有成千上万条不存在的三元组负实例相对应，如三元组（北京师范大学，位于，中国）这个三元组在知识库中实际存在，是一条正实例，而（北京师范大学，位于，美国）和（北京师范大学，位于，日本）等上百条负实例与之对应，如何解决正负实体对不匹配的问题很关键，正负实体对比例悬殊，关系预测中仅靠打分是不够的。二是相关的方法都是通过评价三元组得分高低来预测结果的，而并未考虑候选实体对的顺序对预测结果的影响，也不关注候选实体的秩序关系，而基于学习排序的算法可以解决候选实体的秩序关系。

本部分研究基于抽取知识库中的关系路径特征进行模型预测。模型预测主要分为两种方法：分类模型和排序模型。传统的基于逻辑回归的分类算法对正负实体对进行分类模型学习，利用二分类算法学习正负实体对的得分，依据得分大小将正负实体对进行划分，从而进行关系预测。排序模型通过构建排序算法，对一组正负实体对进行学习排序，通过学习实体对的秩序排名，期望能将正实体对排在负实体对之前，这样既可获得比传统分类回归算法更好的结果。本研究构建了基于树排序方法的知识库补全算法和基于神经网络排序的知识库补全算法。

本研究构建了一种新的知识库补全的模型。其知识库补全的技术关键点在于：（1）给定一个输入关系，将输入关系切分为训练集（包括验证集）和测试集；（2）对于训练集和测试集中的正实体对基于局部封闭世界的假设，生成对应比例的负实体对；（3）将正负实体对构成的集合在由知识库构成的图中抽取特征，基于随机游走的算法抽取从头实体对到尾实体对的路径类型作为关系预测的特征；（4）构建学习排序模型，将生成的正负实体对进行模型训练，并将训练获得的模型为新的关系特征提供预测；（5）基于MAP和MRR进行模型评价，改进模型的参数，更新模型获得更好的预测效果。

4.2　基于逻辑回归的补全模型

本研究的特征计算模块采用随机游走算法抽取给定关系下的路径类型特征。对一个

给定的关系下的实体对，我们需要抽取在知识库中，能在有限路径长度下，从头实体到达尾实体的路径信息。对于如图1知识库中实体对(北京师范大学，中国)，我们在知识库中可以抽取路径类型信息如：(校长-生活在)、(位于-位于)、$\Psi' f^{-1}$-位于-位于)、(校长-出生-相邻-位于)等关系路径信息，并将这些关系路径信息作为学习排序的特征。通常，路径长度被限制在3-6跳之间，过高则关系路径太多，计算复杂度太高，而小于3跳的关系路径则使得获得关系路径类型信息太少，不能有效提供特征。

对于抽取到的特征，传统方法构建了一个分类器模型，学习每个关系和这个关系包含的实体对集合，将预测关系问题转化成一个分类预测问题，学习每个实体对具有某种关系的概率。

$$E_r = \{(h_i, t_j), y_i\}_{i=1}^{N}$$

表示关系r所有的实体对集合，其中$y_i \in \{0, 1\}$，其中0表示负实体对，即知识库中并不是实际存在的三元组， 1表示正实体对，表示在知识库中实际存在的实体对，通过对知识库中的实体对进行分类器模型学习，我们可以获得测试集合中实体对的打分情况。通常这个分类器采用逻辑回归算法进行模型的训练。具体来说，对于每个关系的实体对，传统模型采用逻辑回归学习得到的关系路径特征向量V_r和实体属性特征向量V_l。并定义了如下的逻辑回归函数，对每个关系下的实体对集合进行评价打分。

$$f(v, w) = \frac{1}{1 + e^{w(V_r \oplus V_l)}}$$

其中w表示关系路径特征和实体属性特征的学习权重参数。经典论文采用对数似然函数进行最大似然估计，并通过随机梯度下降算法学习这些模型的参数，除此之外，还考虑到模型的过拟合和参数正则化表示，本论文定义了如下的学习目标函数：

$$L_r = \frac{1}{N} \sum_{i=1}^{N} \{(y_i log(f(v, w_r)) + (1 - y_i) log(1 - f(v, w_r))\} + \alpha \|w_r\| + \beta \|w_r^2\|$$

其中，L_r表示给定关系r的目标函数，α和β 分别是l_1 和l_2正则化惩罚项的权重，对于每个关系我们采用随机梯度下降算法使得整个训练集对数损失最小，同时结合l_1 和l_2防止过拟合。最终我们可以学习得到每个关系下的关系路径特征和实体属性特征的权重。

4.3 基于学习排序的树模型

对于中抽取的关系路径特征和抽取的实体属性特征，我们需要构建基于学习排序算法的知识库补全模型进行关系训练和关系预测。传统的路径排序算法中，当通过随机游

走计算获得路径类型信息，并获得这些关系路径的值后，再通过基于分类或者回归的算法，计算的到每个实体对的打分值，打分高的实体对排在打分低的实体对之前，表示更可能是实际存在的实体对。而本技术不仅仅考虑实体对的打分高低，更关系实体对之间的排序关系，正实体对总需要排序在负实体对前面，这样就能保证在预测的候选实体对中，总是排在前面的实体对是好的结果。更具体的，我们采用一种学习排序的算法进行知识库补全。通过学习最小化实体对的pairwise损失函数，直接优化MAP训练损失函数来进行模型参数更新，从而获得更好的模型预测结果。我们使用基于LambdaMART的树的学习排序算法。对于一个给定的关系r，定义目标函数：

$$F(x_i|w,c) = \sum_{i=1}^{K} \alpha_i \pi(f_i) + \sum_{i=1}^{N} l(f_i(x), f_i'(x)) + \frac{C}{2} WW^T$$

其中第一项中的$\sum_{i=1}^{K} \alpha_i \pi(f_i)$是描述树复杂度的函数，总共有K个树进行模型训练。而第二项中$l(f_i(x), f_i'(x))$是模型的训练误差函数，其中$f_i(x)$是每个实体对的实际分数，而$f_i'(x)$是通过模型学习得到的预测值，共训练了N轮，训练误差函数可以根据实际需要改变，常见的训练误差函数可以选择MAP、AUC、NDCG等不同排序指标，通常学习排序中MAP评价指标最为常见。考虑到我们目标函数是pairwise损失函数最小化，我们也选择MAP作为训练损失函数进行模型训练。第三部分$\frac{C}{2} WW^T$是模型的惩罚项，是防止模型在训练数据中过拟合的L2惩罚函数。

4.4 基于学习排序的知识库补全实验

本研究构建了一个面向YAGO2的知识库补全实例。YAGO是一个从维基百科上抽取的、包含地理名词、WordNet等数据的知识库，YAGO2是YAGO的一个实例。当前YAGO2包括超过千万的实体和超过1.2亿的实体知识我们使用了其中实体的关系型三元组共有4,484,914条、37种关系类型。我们按照37种不同的关系类型，将知识库切分为37份不同关系的实体对集合。对于每个关系，我们首先基于知识库中的三元组抽取正实体对，对于每个由（头实体，尾实体）构成实体对，基于局部封闭世界假设，使用当前关系中的其他实体随机替换生成10个负实体对，其中5个随机替换头实体， 5个随机替换尾实体。生成正负实体对后，我们按照4：1的比例将这些实体对切分为训练集合和测试集合，这样就完成了知识库数据的预处理。

我们展示了基于YAGO2的总体MAP和MRR进行模型评价的结果。我们可以看到基于学习排序的知识库补全技术相比传统的路径排序算法在MAP上有很大的提升，基于学习排序的算法相比传统的算法在YAGO2数据集上有近50%的效果提升；而四种方法

在MRR指标上效果相当，基于学习排序算法的MRR并未比传统路径排序算法有显著下降。

我们详细分析了37种YAGO2中关系的MAP指标，并在表 9展示了部分关系的MAP值。分析可以发现，大部分关系采用学习排序算法后，预测结果有较大的提高，而在不同的关系类型预测中， MAP差别较大，如：“playFor”和“isConnectedTo”等关系预测有较大的提高，而在“isInterestedIn”等关系中关系预测提升较差。总体来说，有超过30种关系的MAP预测获得了显著的提升，只有不到5种关系MAP预测结果并未有统计上的显著提升，这样的实验说明基于学习排序算法的知识库补全技术相比传统的分类回归打分模型有非常大的效果提升。

表 9　学习排序补全算法部分关系MAP值

relation	PRA	SFE	rankPRA	rankSFE
actedIn	0.3379	0.3496	0.6222	0.6252
created	0.2523	0.2532	0.3089	0.3128
dealsWith	0.1729	0.1411	0.1265	0.1572
graduatedFrom	0.2646	0.2726	0.5607	0.5726
hasCapital	0.5637	0.6014	0.7275	0.7304
hasChild	0.5004	0.5078	0.6674	0.6757
influences	0.2946	0.2932	0.5771	0.5836
isAffiliatedTo	0.6364	0.6538	0.7816	0.7840
playsFor	0.6538	0.6606	0.8020	0.8036
wasBornIn	0.3661	0.3742	0.6053	0.6070
worksAt	0.2343	0.2281	0.5022	0.5014
wroteMusicFor	0.3488	0.3621	0.6144	0.6161

4.5　本章工作总结

本章中，我们主要比较了基于逻辑回归分类模型和基于学习排序的树模型，通过优化知识库补全中的模型构建，我们使用基于pairwise的树排序模型，更好的提升了知识库补全中实体对之间的排序问题，将传统的基于逻辑回归的分类问题转化为基于学习排序的模型优化问题。本研究通过对YAGO知识库进行试验发现，不同知识库补全模型中，基于排序的模型优化算法可以大幅提高正实体对在所有实体对中的秩序，从而能改变原有的知识库补全优化中仅仅对实体对进行打分的模式，能够获得更好的预测效果。对于多数关系而言，基于学习排序的树模型算法效果明显优于传统的逻辑回归分类算法。

第 5 章　总结和展望

随着知识库在信息检索、手机助手、对话系统、人工智能等各个领域的广泛发展，知识库的完整性和真实性越来越重要，而如何对知识库进行补全，验证实体和实体之间关系的真实性，都是提高知识库可用性的重要问题。本研究从基于路径排序算法的知识库补全进行研究，选择实体和实体之间的关系路径、头尾实体中不同的属性值进行组合、并构建模型、选择合理有效的目标函数进行知识库补全。同时，本研究也结合了一些新的深度神经网络知识，将不同的特征进行深度融合，获得了更好的知识库补全效果。然而，基于符号逻辑的路径排序算法，补全效果和知识库图的拓扑结构有关，如何能在稀疏的图中进行知识库补全，如何更好的结合关系路径特征和实体属性特征以及基于表示学习获得的实体、关系向量表示特征，是未来进行知识库补全工作的重点。

5.1　论文工作总结

知识库补全算法是知识库构建和知识库应用的重要桥梁，基于信息抽取获得的三元组，需要通过知识库补全技术提高知识库的可用性，进而才能提供各种有效的知识库应用系统。本论文从知识库补全的特征构建开始，结合实体属性特征和关系路径特征，将不同类型的关系路径特征，不同类别的实体属性特征结合，能提高知识库补全的模型的可预测性，可拓展性。此外，本研究拓展了原有的知识库补全模型，将原来基于pointwise的二分类模型，拓展为基于学习排序算法的pairwise模型，提供了更多的模型优化方法，作为原有基于逻辑回归中的log损失函数，或者基于支持向量机的hinge损失函数的补充。这些不同的损失函数作为机器学习的目标，在进行知识库补全工作中，可以依据具体任务不同而选择合适的模型进行预测。

本研究的实验发现了很多重要的结论，能为后续的知识库补全工作提供研究思路。（1）知识库中的关系路径系统是稀疏的，将稀疏的关系路径处理好来预测实体和实体之间的关系，是知识库补全的重要步骤。（2）基于符号逻辑和关系路径进行知识库补全，预测效果和图的拓扑结构密切相关。对于一些关系的实体，这些实体和其他实体之间关系很少，是一类比较孤立的顶点。这导致了基于随机游走过程中，部分实体对很难抽取有表达力的关系路径特征，同样对于部分实体对，他们的实体属性特征也很少，这样会导致知识库补全算法很难进行模型预测。如何结合能学习稀疏图中的表示实体向量模型，能有效将这些实体关系模型进行预测。（3）知识库中的实体属性特征和关系路径

特征能相互影响，对于部分关系路径表达力差的特征，好的实体关系路径特征能进行模型惩罚，减少这样关系特征的权重，而对于关系路径表达力强的特征，在加入实体属性特征后能增加这些特征的权重。（4）除了传统的基于逻辑回归、支持向量机算法，通过优化hinge损失函数，log损失函数学习二分类模型，本研究展示了其他学习模型的优势。通过基于pairwise的模型优化方法，本研究能将一组正负实体对作为一个整体序列进行模型优化，学习这些正负实体对的秩序，从而将知识库补全作为一个排序结果进行训练。这种算法的另外一个优势就是能避免模型的正负例不平等问题，传统的算法随机选择若干比例的负例进行模型优化，而基于排序的算法可以将三元组中多种正例三元组排在负例三元组之前，提高模型的可用性。

5.2 未来研究展望

尽管本研究基于符号逻辑算法开始，拓展了关系路径和实体属性等不同类型的特征，增加了模型预测的优化方法，优化了目标函数，但是未来的研究还是有很多重要的工作，如何结合符号逻辑算法和表示学习算法，如何将表示学习中的优化目标函数和符号逻辑中的优化目标函数相互统一。

从知识库补全的关系路径开始拓展，本研究在路径排序算法中加入子图特征路径，拓展知识库的关系路径特征。同时也考虑了实体属性特征在知识库补全中的重要应用，以及如何结合关系路径和实体属性特征，从而增加模型的可预测性。另外，本研究也拓展了不同的优化目标函数，从实验中发现结合pairwise的知识库补全算法可以有效的优化一组正负实体对集合，很好的解决知识库中正负实例对不平衡的问题。

本研究未来希望能关注如何有效结合关系路径特征、实体属性特征、实体向量特征，不仅仅通过不同的算法抽取关系路径特征，转换不同类型实体属性，而且构建了深度学习模型，将知识库补全算法的不同特征进行融合，更好提高了机器学习模型的效果。通过将这些不同的实体、关系预测统一构建一个合理有效的目标函数，优化目标函数获得更好的模型系统，从而将知识库补全算法进行统一优化、改进，从而模型预测变得更加智能。

本研究未来关注的另外一个重要如何结合关系路径特征、实体属性特征、实体向量特征预测实体的一些属性值。尽管很多知识库中，如何预测实体和实体之间的关系是一个重要的问题，但是我们在实验中也非常明显的发现，很多实体的属性值都是缺失的，如很多重要人物的出生日期、很多地点的地理位置等信息存在明显的缺失或者错误，如何从实体、实体属性类型预测实体属性值，对于解决实体属性值的缺失或错误问题，有

着重要的影响。这个问题对于未来的知识库补全、知识库拓展和应用也都是十分重要的问题，对于本人的未来研究之路，也是我需要进一步思考和探索的。

除了在理论和算法方面的研究工作之外，进行关系路径的知识库补全算法的一个不足时对于实验机器要求较高。通常需要耗费大量的内存来构建图，查找图中的关系路径，如何能降低关系路径搜索中对于机器硬件的要求也是一个可以优化的方向，如果能有更高效，内存使用更合理的随机游走路径搜索算法，也是很多领域关注的热点问题。

参考文献

[1] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, et al. From Data Fusion to Knowledge Fusion[J]. PVLDB, 2014, 7:881–892.

[2] T. Mitchell, W. Cohen, E. Hruschka, et al. Never-Ending Learning[C]. Proceedings of Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15), 2015.

[3] Fabian M. Suchanek, Gjergji Kasneci, Gerhard Weikum. Yago: A Core of Semantic Knowledge[C]. Proceedings of Proceedings of the 16th International Conference on World Wide Web, New York, NY, USA: ACM, 2007. 697–706.

[4] Kurt D. Bollacker, Colin Evans, Praveen Paritosh, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]. Proceedings of SIGMOD Conference, 2008.

[5] Christian Bizer, Jens Lehmann, Georgi Kobilarov, et al. DBpedia - A Crystallization Point for the Web of Data[J]. Web Semant., 2009, 7(3):154–165.

[6] Michel Dumontier, Alison Callahan, Jose Cruz-Toledo, et al. Bio2RDF Release 3: A Larger Connected Network of Linked Data for the Life Sciences[C]. Proceedings of Proceedings of the 2014 International Conference on Posters & Demonstrations Track - Volume 1272, Aachen, Germany, Germany: CEUR-WS.org, 2014. 401–404.

[7] Linyuan Lu, Tao Zhou. Link Prediction in Complex Networks: A Survey[J]. CoRR, 2010, abs/1010.0725.

[8] Gerhard Weikum, Martin Theobald. From information to knowledge: harvesting entities and relationships from web sources[C]. Proceedings of PODS, 2010.

[9] Gjergji Kasneci, Maya Ramanath, Fabian M. Suchanek, et al. The YAGO-NAGA approach to knowledge discovery[J]. SIGMOD Record, 2008, 37:41–47.

[10] Fabian M. Suchanek, Gjergji Kasneci, Gerhard Weikum. YAGO: A Large Ontology from Wikipedia and WordNet[J]. J. Web Sem., 2008, 6:203–217.

[11] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, et al. YAGO2: exploring and querying world knowledge in time, space, context, and many languages[C]. Proceedings of WWW, 2011.

[12] David A. Ferrucci, Eric W. Brown, Jennifer Chu-Carroll, et al. Building Watson: An Overview of the DeepQA Project[J]. AI Magazine, 2010, 31:59–79.

[13] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, et al. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia[J]. Artif. Intell., 2013, 194:28–61.

[14] Xuchen Yao, Benjamin Van Durme. Information Extraction over Structured Data: Question Answering with Freebase[C]. Proceedings of ACL, 2014.

[15] Xuchen Yao. Lean Question Answering over Freebase from Scratch[C]. Proceedings of HLT-NAACL, 2015.

[16] Xuchen Yao, Jonathan Berant, Benjamin Van Durme. Freebase QA: Information Extraction or Semantic Parsing?[C]. 2014.

[17] Zhicheng Zheng, Xiance Si, Fangtao Li, et al. Entity Disambiguation with Freebase[J]. 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, 2012, 1:82–89.

[18] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, et al. Knowledge vault: a web-scale approach to probabilistic knowledge fusion[C]. Proceedings of KDD, 2014.

[19] Ulf HERMJAKOB, Eduard H. HOVY, Chin-Yew LIN. Knowledge-Based Question Answering[C]. 2000.

[20] Min-Chul Yang, Nan Duan, Ming Zhou, et al. Joint Relational Embeddings for Knowledge-based Question Answering[C]. Proceedings of EMNLP, 2014.

[21] Wen tau Yih, Matthew Richardson, Christopher Meek, et al. The Value of Semantic Parse Labeling for Knowledge Base Question Answering[C]. Proceedings of ACL, 2016.

[22] Mo Yu, Wenpeng Yin, Kazi Saidul Hasan, et al. Improved Neural Relation Detection for Knowledge Base Question Answering[C]. Proceedings of ACL, 2017.

[23] Organisers Key-Sun, Choi, Jin-Dong Kim, et al. Open Knowledge Base and Question Answering Workshop (OKBQA)[C]. 2016.

[24] Babita Pandey, R. B. Mishra. Knowledge and intelligent computing system in medicine[J]. Computers in biology and medicine, 2009, 39 3:215–30.

[25] Ethan J Halpern, Eric L Gingold, Hugh White, et al. Evaluation of coronary artery image quality with knowledge-based iterative model reconstruction.[J]. Academic radiology, 2014, 21 6:805–11.

[26] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, et al. Translating Embeddings for Modeling Multi-relational Data[C]. In: C. J. C. Burges, L. Bottou, M. Welling, et al., (eds.). Proceedings of Advances in Neural Information Processing Systems 26. Curran Associates, Inc., 2013: 2787–2795.

[27] Zhen Wang, Jianwen Zhang, Jianlin Feng, et al. Knowledge Graph Embedding by Translating on Hyperplanes[C]. Proceedings of AAAI, 2014.

[28] Wenhao Huang, Ge Li, Zhi Jin. Improved Knowledge Base Completion by the Path-Augmented TransR Model[C]. Proceedings of KSEM, 2017.

[29] Maximilian Nickel, Volker Tresp, Hans-Peter Kriegel. A Three-Way Model for Collective Learning on Multi-Relational Data[J]. 28th International Conference on Machine Learning, 2011. 809—-816.

[30] Luis Galárraga, Christina Teflioudi, Katja Hose, et al. AMIE: association rule mining under incomplete evidence in ontological knowledge bases[C]. Proceedings of WWW, 2013.

[31] Ni Lao, William W. Cohen. Relational retrieval using a combination of path-constrained random walks[J]. Machine Learning, 2010, 81(1):53–67.

[32] Matt Gardner, Tom Mitchell. Efficient and Expressive Knowledge Base Completion Using Subgraph Feature Extraction[J]. Proceedings of EMNLP, 2015, (September):1488–1498.

[33] Steffen Rendle. Factorization Machines[J]. 2010 IEEE International Conference on Data Mining, 2010. 995–1000.

[34] Steffen Rendle. Factorization Machines with libFM[J]. ACM TIST, 2012, 3:57:1–57:22.

[35] Embedding Large Subgraphs into Dense Graphs[C]. 2009.

[36] Xin Luna Dong, Evgeniy Gabrilovich, Geremy Heitz, et al. Knowledge Vault : A Web-Scale Approach to Probabilistic Knowledge Fusion[J]. 2014. 601–610.

[37] Yankai Lin, Zhiyuan Liu, Maosong Sun, et al. Learning Entity and Relation Embeddings for Knowledge Graph Completion[C]. Proceedings of AAAI, 2015.

[38] Baoxu Shi, Tim Weninger. ProjE: Embedding Projection for Knowledge Graph Completion[C]. Proceedings of AAAI, 2017.

[39] J. Ross Quinlan, R. Mike Cameron-Jones. FOIL: A Midterm Report[C]. Proceedings of ECML, 1993.

[40] MICHAEL J. PAZZANI, CLIFFORD A. BRUNK. Detecting and correcting errors in rule-based expert systems: an integration of empirical and explanation- based learning[C]. 1991.

[41] Stefan Schoenmackers, Oren Etzioni, Daniel S. Weld, et al. Learning First-order Horn Clauses from Web Text[C]. Proceedings of Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. 1088–1098.

[42] Luis Antonio Galárraga, Christina Teflioudi, Katja Hose, et al. AMIE: Association Rule Mining Under Incomplete Evidence in Ontological Knowledge Bases[C]. Proceedings of Proceedings of the 22Nd International Conference on World Wide Web, New York, NY, USA: ACM, 2013. 413–422.

[43] Luis Galárraga, Christina Teflioudi, Katja Hose, et al. Fast rule mining in ontological knowledge bases with AMIE $$+$$ +[J]. The VLDB Journal, 2015, 24:707–730.

[44] Zhichun Wang, Juan-Zi Li. RDF2Rules: Learning Rules from RDF Knowledge Bases by Mining Frequent Predicate Cycles[J]. CoRR, 2015, abs/1512.07734.

[45] Ni Lao, Einat Minkov, William W. Cohen. Learning Relational Features with Backward Random Walks[C]. Proceedings of ACL, 2015.

[46] Matt Gardner, Partha P. Talukdar, Jayant Krishnamurthy, et al. Incorporating Vector Space Similarity in Random Walk Inference over Knowledge Bases[J]. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014. 397–406.

[47] Quan Wang, Jing Liu, Yuanfei Luo, et al. Knowledge base completion via coupled path ranking[C]. Proceedings of Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016. 1308–1318.

[48] Robert West, Evgeniy Gabrilovich, Kevin Murphy, et al. Knowledge Base Completion via Search-Based Question Answering[J]. 2014.

[49] Ni Lao, Teruko Mitamura, Tom Mitchell, et al. Efficient Random Walk Inference with Knowledge Bases[J]. PhD thesis, 2012.

[50] László Lovász. Random Walks on Graphs: A Survey[C]. 1993.

致　谢

衷心感谢导师王志春教授对本人的精心指导，他的言传身教将使我终生受益。

在北京师范大学学习的三年中，承蒙王志春老师不辞劳苦的指导和帮助，让我从一个无知的年青人，慢慢学到很多做人做事做学问的道理，这些知识不仅仅帮助我在学术的道路上探索向前，也帮我成长，学会发现更大的世界。

感谢我们实验室的所有师妹师弟们，感谢李楚同学不辞劳苦的帮助我解决Linux系统问题，感谢郑伟师妹帮我一起合作Java代码项目，感谢孙铭晨师妹经常为我们出谋划策，带给我们鲜美的水果零食，感谢吴妍蓉师妹帮我解决学术问题，带给我欢声笑语。最由衷的感谢我的父母，他们无私的为我付出，支持我在贫寒的学术道路上又坚持了三年，让我始终坚持自己的想法，做自己应该做的事情。

感谢清华的薛瑞尼及相关同学，他们制作维护的清华学位论文模板极大的方便了LaTeX用户的论文写作。

最后附上苏轼的一首诗：人生到处知何似，应似飞鸿踏雪泥。泥上偶然留指爪，鸿飞那复计东西。老僧已死成新塔，坏壁无由见旧题。往日崎岖还记否，路长人困蹇驴嘶。 人生匆匆，雪泥鸿爪。希望大家都能活的开心，做学问也开心。

黄勇

2018年 3月