附件七：

大连理工大学

专业学位硕士研究生学位论文开题报告

论文题目：基于主动学习的知识图谱融合系统实现

姓　　名：严勇

学　　号：31817029

专业/领域：软件工程

培养类型：□ 全日制 □ 在职学习

指导教师：单世民

实践导师：祈贺

入学日期：2018 年 9 月

报告日期：2019 年 9 月

报告地点：综合楼二楼会议室

研究生院制表

说　明

学位论文开题考核是硕士研究生课程学习结束后开展学位论文工作的基本要求，是保证学位论文质量、工作进度和研究生培养质量的首要环节。专业学位硕士研究生学位论文可以是产品研发、工程设计、专题应用研究、工程/项目管理、调研报告等形式，具体要求应参照全国相关专业/领域专业学位教育指导委员会制定的专业学位标准。

一、考核内容：首先，考查硕士生对本专业/领域课程学习、校内实践（实验）情况，对专业知识、技能、标准规范等掌握程度；其次，考查学位论文工作准备情况，包括论文选题是否适合专业学位研究生培养、论文的目的意义及国内外发展状况、论文内容设置的合理性、方法的科学性、工作量与工作难度、预期成果的实用性和新颖性、文献阅读等；第三，还要考查学生校外企业实践情况及时间安排、研究生的学习和工作态度等；此外，还要考查该生实践环节的实践条件是否有保障、校内外实践安排是否合理等。

二、考核时间：硕士生的开题报告应在第 2 学期末或第 3 学期初进行。

三、报告撰写：开题报告正文字数不少于 5000 字；参考文献数量不少于 20 篇（其中外文文献不少于 40%）；正文及参考文献等撰写要求参见《大连理工大学硕士学位论文格式规范》。

四、考核办法：开题考核由学部（学院）集中组织 5 名以上本学科领域专家（至少一名专家来自企业，导师和企业导师除外）以答辩的方式进行。学生进行口头陈述时间不得少于 10 分钟。专家组给出考核成绩和是否通过的意见。

五、报告保存：开题报告一式两份，签字后分别由学部（学院）和学生保存。

六、信息登录：研究生开题后登录研究生信息管理系统上传开题报告（PDF 文档）及考核结果。

开题报告正文

撰写大纲：

1） 课程学习、技术标准学习及校内实践（实验）情况（附成绩单），参加科研和学术活动等情况；

2） 论文选题背景、目的和意义；

3） 国内外研究现状及发展动态分析；

4） 主要研究内容、研究目标、拟解决的关键问题；

5） 学位论文的研究方法、技术路线、试验手段、关键技术等论述；

6） 年度研究计划、预期成果及现有研究工作基础；

7） 校内外实习实践条件及时间安排；

8） 论文工作过程中可能遇到的困难、问题及对策。

9） 参考文献（不占字数）。

校内学习状况

在校期间，完成 20 学分的必修课程与 7 学分的选修课程，成绩如下：

必修课程	课程学分	选修学期	成绩	选修课程	课程学分	选修学期	成绩
工程伦理	1	1	86	研究生形势与政策	1	1	P
知识产权	1	1	85	可视计算	2	2	86
信息检索	1	1	P	数据分析理论与方法	2	1	69.4
中国特色社会主义理论与实践研究	2	1	85	实用机器学习	2	2	76
信息服务技术与标准	1	2	P				
现代人工智能	3	1	68				
软件工程发展前沿	1	2	P				
算法分析与设计	3	1	98				
阅读与写作 I（基础读写技能）	2	2	90				
矩阵与数值分析	3	1	81				
优化方法	2	2	71				

在校区间积极参与科研项目，目前主要参与知识图谱融合的研究与实现工作，也正是本论文的研究内容。

知识图谱融合背景、目的与意义

知识图谱以结构化的形式描述客观世界中概念、实体间的复杂关系，将互联网的信息表达成更接近人类认知世界的形式，提供了一种更好地组织、管理和理解互联网海量信息的能力。知识图谱给互联网语义搜索带来了活力，同时也在智能问答、推荐系统、信息检索中显示出强大威力，已经成为了互联网智能服务的基础设施。知识图谱与大数据和深度学习一起，已经成为推动人工智能发展的核心驱动力之一。

现有的通用知识图谱的数据来源都非常相似，通常是从三大中文百科全书中提取结构信息，或者再结合一些领域数据，例如，音乐、电影、商业等。因此，知识内容必然有大量的重叠。再加上各机构使用的本体集，即数据模式不尽相同，必然会导致数据不均匀、歧义等问题。为了使各个独立的知识库能够统一成一个规范表达，知识库融合迫在眉睫，而解决这一问题的首要工作就是实体融合。由于不同知识图谱之间存在着数据的语义不均一性,实体和属性的表示有许多变种和歧义,这给实体融合技术带来了巨大的挑战。实体融合算法主要基于语言学特征、层次结构、属性值域、辅助数据源、机器学习、知识表示学习等。一般地,基于语言学相似度的算法比较难于应用在大规模的数据集当中,机器学习相对灵活,但是依赖于训练数据和优化算法,知识表示学习能够脱离实体的文本信息,根据 RDF 三元组之间的结构特征对实体进行编码。传统的实体融合工具提供的匹配算法通常非常有限,不能满足用户的多样性需求,且缺少友好的用户界面,对于普通用户来说,使用门槛较高。本文的贡献主要有：

1. 提出完整的图谱融合工程解决方案，确保融合后图谱的准确性并能够对外提供服务。图谱融合方案流程包括业务本体对齐，模型选择，属性配置，数据标注、自动匹配与匹配审核，环环相扣；
2. 优化与改进基于记录的匹配算法，引入和实现基于知识表示的匹配算法和基于同义词林的匹配算法；
3. 基于主动学习的标注方法以及优先审核策略，尽可能减少人工的操作。

国内外研究现状

目前，有针对特定领域以及通用的融合工具。在特定领域方面， Aoas 是依赖于生物学知识的融合系统，AgreementMaker 专门用来处理地理信息，GNAT 主要处理音乐数据，PKB-CRS 用来融合大学和学术会议。然后，知识图谱数据多种多样，要为每一个领域单独设计融合系统不切实际，因此通用融合工具的需求更加迫切。SAMBO 是经典的模型，不仅可以对齐实体，还能融合知识库。

SILK 也是通用系统，优势在于良好的交互界面，用户拖动式的算法编排更加方便与灵活。Limes 是典型的关系发现框架，使用三角不等式原理对融合的时间复杂度进行了优化。RDF-AI 实现数据预处理、对齐、融合、内部链接以及后期处理五个功能，是流程完整的工具。

综合来看，实体融合工具可分为以下 6 种：

1. 基于语言学信息：这类方法通常使用实体和关系的文本信息，例如，名称、同义词或者定义。实体之间的相似性计算主要根据文字描述，最常使用的方法包括简单的字符串匹配、信息检索算法。几乎所有的系统都会使用这类方法。

2．基于结构信息：这类方法利用数据集的语义结构 卩738]给出融合建议。一般地，可根据 RDF 三元组中的 is-a 和 part-of 关系来构建实体之间的有向无环图。因此，基于实体的邻近节点关系，就可以使用不同的方法计算它们之间的相似度。通常可采用节点的距离来判断相似性，例如，P.Mork 等人通过连接每个实体的直接子节点和父节点组成了一个简单的图结构，匹配了两个大规模医学数据集。

3．基于背景知识：使用专用、通用的辞典或者主题列表作为辅助信息，可提高实体融合的有效性。例如，将实体映射到 WordNet，就可以根据 WordNet 的同义词集合对齐实体。利用第三方数据集作为中间数据，以及重用之前的融合数据，也属于背景知识的范畴。许多工具采用了这类方法。

4. 基于限制条件：如果两个关系的主语集合以及宾语集合有大范围的重合，那么可以考虑这两个关系存在较大的相似性。或者两个实体的属性值与第三个都不相交，则它们之间也可能存在等价链接。单独使用限制条件能抽取的融合实体数很少，通常作为补充算法使用。

5．机器学习：实体融合实际上就是处理数据的各种特征，所以很自然地会联想到用机器学习来解决问题。常见的思路有学习融合表达式、训练分类模型、训练权重参数等。F.Duchateau 等人使用决策树模型作为聚合函数将多个融合算法组合使用。

6．知识表示学习：通过把知识图谱中的实体和关系等语义信息映射成实数空间的稠密低维向量，直接用数学表达式来计算各个实体之间相似度。这类方法不依赖任何的文本信息，获取到的都是数据的深度特征。因此，表示学习模型的好坏直接影响到数据特征的提取，好的模型能发现更优质的融合实体。

研究内容

研究目标是将多源大规模异构数据源导入知识库，经过存储格式转变、实体自动匹配、关系推理后对外提供增强后的能够提供服务的知识库。为了达到此目标，现阶段以及未来的研究内容包括：

1. 调研各类融合框架，作细致的横向比较，分析优缺点，重点阅读 dedupe 源码与论文；
2. 基于图数据存储模式，设计知识图谱融合方案，包括原型设计、数据处理流程，框架语言选择；
3. 解决多源数据导入问题，数据源包括来自关系型，文档型，时序型数据库；
4. 同义词库的维护；
5. 记录匹配算法改进与实现，包括基于记录的匹配、基于网络的匹配以及基于同义词林的匹配；
6. 在算法自动匹配后，分析具体的实体合并业务逻辑，以及合并后的撤销操作。

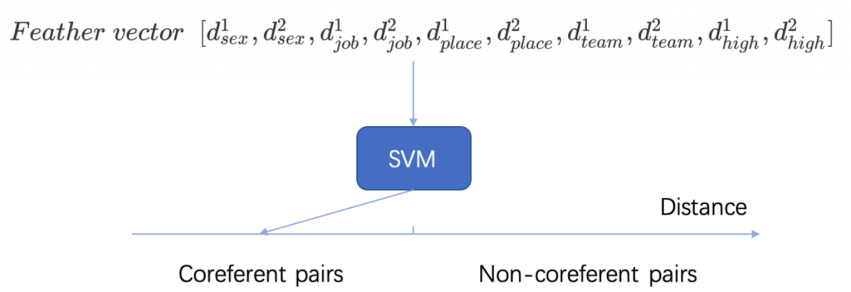
研究方法

本论文研究的重点是实体融合方法，采用的方法包括基于记录的匹配方法、基于结构的匹配方法和基于同义词的匹配方法。

基于记录的匹配方法

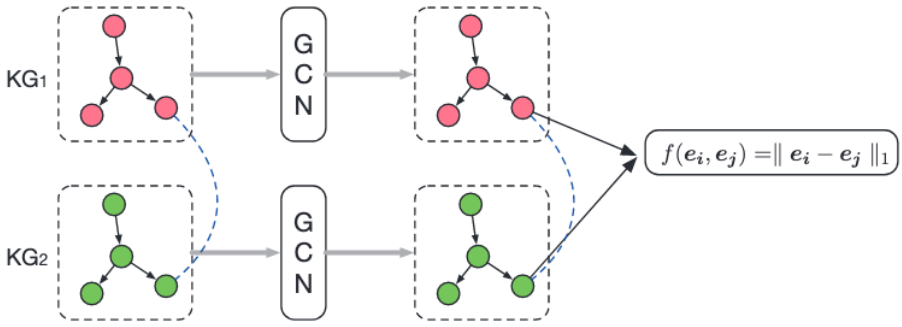
计算两个记录各个属性的相似度，再综合衡量记录相似度，最后根据主动学习的标注数据训练分类器模型。

来源	性别	工作	出生地	球队	身高
数据源1	男	CBA主席	上海市	湖人队	1.9m
数据源2	男	职业篮球	沪	火箭队	2m



基于结构的匹配方法

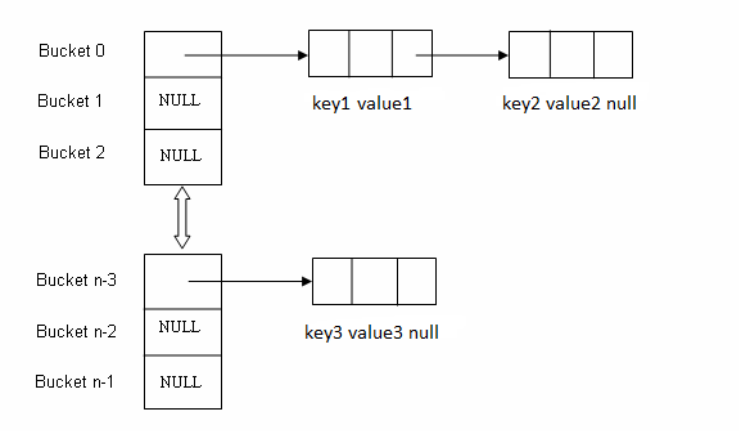
结构匹配是指，在记录匹配的基础上考虑图的拓扑结构信息，包括实体关系和属性关系。我们考虑使用图神经网络来完成这一任务。利用图卷积神经网络（GCN）进行实体匹配，充分利用节点的属性信息、图的结构信息以匹配描述同一事物的实体，[4]在跨语言的匹配工作中已经实现非常高的精确度。



论文使用两个GCN分别训练获得包含实体的结构信息的向量与包含实体的属性信息的向量，对与 连接后计算相似度。

基于同义词的匹配方法

维护一个全局 HashMap，计算主同义词的哈希码，根据哈希码计算其在数组中的位置，非主同义词则存放在主同义词所在的链表上，当链表长度超过一定阈值，链表转为红黑树以进行快速检索。



主动学习

主动学习是机器学习的一个子领域，在统计学领域也叫做查询学习、最优实验设计。该算法包括两个基本模块：学习模块和选择策略。主动学习通过“选择策略”主动从未标注样本集中挑选部分样本，交给相关领域专家进行标注，然后将标注样本增加到训练数据集给“学习模块”进行训练。当学习模块满足终止条件后停止，否则不断重复获得更多的标注样本进行训练。主动学习的优点是能够很好地处理较大的训练数据集，减少人工标注成本。

半监督学习和主动学习都是从未标记样例中挑选部分价值量高的样例标注后补充到已标记样例集中来提高分类器精度，降低领域专家的工作量，但二者的学习方式不同：半监督学习一般不需要人工参与，是通过具有一定分类精度的基准分类器实现对未标注样例的自动标注；而主动学习有别于半监督学习的特点之一就是需要将挑选出的高价值样例进行人工准确标注。半监督学习通过用计算机进行自动或半自动标注代替人工标注，虽然有效降低了标注代价，但其标注结果依赖于用部分已标注样例训练出的基准分类器的分类精度，因此并不能保证标注结果完全正确。相比而言，主动学习挑选样例后是人工标注，不会引入错误类标。

主动学习分两个阶段：

1）初始化阶段：随机从 U（未标注样本）中选取小部分，由监督者 S 标注，作为训练集建立初始模型。

2）循环查询阶段：S 从 U 中，按照某种查询标注（Q），选取一定为标注的样本进行标注，并计入到训练样本集 L 中，重训训练分类器，直至达到训练停止标准为止。

主动学习作为一种新的机器学习方法，其主要目标是有效地发现训练数据集中高信息量的样本，并高效地训练模型。与传统的监督方法相比，主动学习具有如下优点：能够很好地处理较大的训练数据集，从中选择有辨别能力的样本点，减少训练数据的数量，减少人工标注成本。

主动学习模型可以表示为 A = (C, L, S, Q, U)。其中 C 表示分类器、L 表示带标注的样本集、S 表示能够标注样本的专家、Q 表示当前所使用的查询策略、U 表示未标注的样本集。具体步骤如下：

（1）选取合适的分类器，记为 current_model ，选择主动选择策略。将数据划分为用于训练的带标注的样本 train_sample、用于验证的带标注样本 validation_sample、未标注的数据集 active_sample；

（2）初始化：随机初始化或者通过迁移学习初始化；如果有 target domain 的标注样本，就通过这些标注样本对模型进行训练；

（3）使用当前模型 current_model 对 active_sample 中的样本进行逐一预测，得到每个样本的预测结果。

（4）专家对选择的样本进行标注，并将标注后的样本放至 train_sapmle 目录下。

（5）使用当前所有标注样本 train_sample 对当前模型 current_model 进行 fine-tuning，更新 current_model；

（6）使用 current_model 对 validation_sample 进行验证，如果当前模型的性能得到目标或者已不能再继续标注新的样本，结束迭代过程。否则，循环执行步骤（3）。

研究计划

研究计划分为系统设计，算法研究、和系统实现三个部分。

10 月 1 日 - 10 月 31 日，确定初步的系统设计问题，如融合区存储问题，本体对齐方式问题，数据导入方案等，同时实现算法原型。

11 月 1 日 - 11 月 30 日，正式进入系统实现问题，并解决算法在实际应用中的问题。

困难、问题与对策

目前出现的问题如下：

1. 融合区如何存储，分布式采集，不同源的数据如何区分？
2. 分布式采集都是用这同一个系统吗？采集系统和提供服务的系统是否为一套东西？
3. 融合区数据迭代，新的数据如何更新旧的数据，包括易变属性问题？
4. 融合区的实体如何设置唯一标志？
5. 一轮融合结束的标志是什么？
6. 整棵树实体类不重复，多继承问题；
7. 融合和产品区域如何同步，即用融合区更新产品区的策略问题；
8. 如何取产品区和融合区的快照？
9. 自动匹配算法调用问题。

这些问题的解决依赖于和团队的密切讨论，以及在阅读文献时受到启发。

参考文献

[1] Bilenko, Mikhail Yuryevich. Learnable similarity functions and their application to record linkage and clustering. Diss. 2006.

[3] Wu, Zonghan, et al. "A comprehensive survey on graph neural networks." arXiv preprint arXiv:1901.00596 (2019).

[4] Wang, Zhichun, et al. "Cross-lingual knowledge graph alignment via graph convolutional networks." Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018.

[5] https://github.com/dedupeio/dedupe

[6] https://github.com/1049451037/GCN-Align

[7] https://github.com/liuhuanyong/CrimeKgAssitant.git

[8] García, Salvador, Julián Luengo, and Francisco Herrera. Data preprocessing in data mining. New York: Springer, 2015.

[9] 罗丹. 一种基于表示学习的知识图谱融合算法与系统实现[D].浙江大学,2018.

[10] DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia[J] . Jens Lehmann,Robert Isele,Max Jakob,Anja Jentzsch,Dimitris Kontokostas,Pablo N. Mendes,Sebastian Hellmann,Mohamed Morsey,Patrick van Kleef,Sö,ren Auer,Christian Bizer.Semantic Web . 2015 (2)

[11] Cross-lingual entity matching and infobox alignment in Wikipedia[J] . Daniel Rinser,Dustin Lange,Felix Naumann.Information Systems . 2012

[12] Matching large ontologies: A divide-and-conquer approach[J] . Wei Hu,Yuzhong Qu,Gong Cheng.Data & Knowledge Engineering . 2008 (1)

[13] A visual tool for ontology alignment to enable geospatial interoperability[J] . Isabel F. Cruz,William Sunna,Nalin Makar,Sujan Bathala.Journal of Visual Languages and Computing . 2007 (3)
[14] SAMBO—A system for aligning and merging biomedical ontologies[J] . Patrick Lambrix,He Tan.Web Semantics: Science, Services and Agents on the World Wide Web . 2006 (3)
[15] 知识表示学习研究进展[J]. 刘知远,孙茂松,林衍凯,谢若冰.计算机研究与发展. 2016(02)
[16] Krishnakumar A. Active Learning Literature Survey[J]. 2007.
[17] Liu K, Qian X. Survey on active learning algorithms[J]. Computer Engineering & Applications, 2012.
[18] Zhou Z, Shin J, Zhang L, et al.Fine-Tuning Convolutional Neural Networks for Biomedical Image Analysis: Actively and Incrementally[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2017:4761-4772.
[19] Zhu J J, Bento J. Generative Adversarial Active Learning[J]. 2017.
[20] Konyushkova K, Sznitman R, Fua P. Learning Active Learning from Data[J]. 2017.
[21] Maystre L, Grossglauser M. Just Sort It! A Simple and Effective Approach to Active Preference Learning[J]. Computer Science, 2017.
[22] 韩明皓. 基于知识图谱的关系推理算法研究[D].电子科技大学,2018.

大连理工大学专业学位硕士研究生学位论文开题报告评审意见表

<table>
<tr><td>学　号</td><td>31817029</td><td>学生姓名</td><td>严勇</td><td>专业/领域</td><td>软件工程</td></tr>
<tr><td colspan="3">第一次开题 □</td><td colspan="3">第二次开题 □</td></tr>
<tr><td colspan="6">实 践 导 师 信 息</td></tr>
<tr><td>姓　名</td><td>祁贺</td><td>性　别</td><td>男</td><td>职称/职务</td><td>项目经理/工程师</td></tr>
<tr><td>专　业</td><td>软件工程</td><td>所在单位</td><td>翔云创想（大连）科技有限公司</td><td>联系电话</td><td>18624359959</td></tr>
<tr><td>通讯地址</td><td colspan="3">辽宁省大连高新技术产业园区七贤岭产业化基地高新街 3 号五层 501</td><td>E_mail</td><td>qihe@beyondcloud.cn</td></tr>
<tr><td colspan="6">校内导师考核意见（对课程学习情况、校内实践实习情况、参加学术活动情况、学位论文工作及开题报告撰写情况、企业实践实习情况及计划、学生的学习和工作态度等进行考查，给出考核成绩和具体改进意见和建议）：
1） 考核成绩：□ 优秀，□ 良好，□ 中等，□ 及格，□ 不及格
2） 是否通过：□ 通过，□ 不通过
3） 关于开题报告撰写质量及学位论文工作的具体意见（可加页）：

选题来源于实际项目，有应用背景，题目有一定的难度。进行了查阅了较多的相关文献，进行了比较详尽的调研，所提出方法及所使用技术具有可行性，内容及技术方法方面均适合做专业学位硕士学位论文，同意开题。

导师签字：
年　　月　　日</td></tr>
</table>

实践导师考核意见（对学位论文工作及开题报告撰写情况、企业实践实习情况及计划、学习和工作态度等进行考查，给出考核成绩和具体改进意见和建议）：

1） 考核成绩：☐ 优秀，☐ 良好，☐ 中等，☐ 及格，☐ 不及格

2） 是否通过：☐ 通过，☐ 不通过

3） 关于开题报告撰写质量及学位论文工作的具体意见（可加页）：

该同学给出的基于主动学习图谱融合方法，解决棘手的图谱融合问题，查阅了相关文献，进行了相关调研，具备可行性，同意开题。

导师签字：

年　　月　　日

		姓名	职称	学科专业	是否博导	签字
评议专家组	组长	刘宇	教授	软件工程	是	
	成员	单世民	副教授	软件工程	否	
		徐秀娟	副教授	软件工程	否	
		赵哲焕	讲师	软件工程	否	
		祁贺	工程师	软件工程	否	
		吴昊奇	工程师	软件工程	否	

专家组评审意见（对课程学习情况、校内实践实习情况、参加学术活动情况、学位论文工作及开题报告撰写情况、企业实践实习情况及计划、学生的学习和工作态度等进行考查，给出考核成绩和具体改进意见和建议）：

1）选题是否属于本学科领域（含交叉学科）：☐ 是，☐ 不是（须重新开题）

2）选题是否符合专业学位论文要求：☐ 是，☐ 不是（须重新开题）

3）考核成绩：☐ 优秀，☐ 良好，☐ 中等，☐ 及格，☐ 不及格

4）是否通过：☐ 通过，☐ 不通过

5）关于开题报告撰写质量及学位论文工作的具体意见（可加页）：

选题来源于实际工程问题，有应用背景，调研情况比较详实，查阅了大量相关文献，从内容到技术均比较适合于做工程硕士轮文。开题准备比较充分，可以开展论文工作。

组长签字：

年　　月　　日

点长意见：   点长签字： 年　　月　　日