**哈尔滨工业大学**

**国家示范性软件学院**

**本科毕业设计(论文)中期报告**

**题 目：基于知识库的海量异构数据**

**集成系统的设计与实现**

**专 业 物联网工程**

**学 生 姓 名 李天宝**

**学 号 1133730206**

**联 系 方 式 15704600640**

**年 级 2013级**

**实 习 基 地 上海骇咕赛信息科技有限公司**

**基地指导教师 丁盛豪**

**联 系 方 式 13761793694**

**校内指导教师 王宏志**

**联 系 方 式 13069887146**

**中 检 日 期 2017.03.18**

**哈尔滨工业大学软件学院**

目 录

[1. 毕业设计（论文）内容概述 1](#_Toc477463996)

[1.1. 项目来源及开发目的和意义 1](#_Toc477463997)

[1.2. 主要开发任务（或总体设计内容及方案） 2](#_Toc477463998)

[1.3. 本人所承担任务（模块）说明 2](#_Toc477463999)

[1.4. 开发环境和开发工具 2](#_Toc477464000)

[1.5. 项目原定进度安排 3](#_Toc477464001)

[2. 中期完成情况说明 4](#_Toc477464002)

[2.1. 预定计划的执行情况 4](#_Toc477464003)

[2.2. 中期工作说明及成果汇报 4](#_Toc477464004)

[2.2.1. 系统功能结构 4](#_Toc477464005)

[2.2.2. 系统流程 5](#_Toc477464006)

[2.2.3. 系统用例详细描述 6](#_Toc477464007)

[2.2.4. 背景概念及定义 7](#_Toc477464008)

[2.2.5. 数据结构设计 9](#_Toc477464009)

[2.2.6. 核心算法 10](#_Toc477464010)

[2.2.7. 结果展示 13](#_Toc477464011)

[2.3. 存在的困难与问题 14](#_Toc477464012)

[2.4. 如期完成预定任务的可能性分析 16](#_Toc477464013)

[2.5. 后期工作安排（或进度和计划调整） 16](#_Toc477464014)

[参考文献 17](#_Toc477464015)

[附件：第1次中期检查记录表 20](#_Toc477464016)

[附件：导师意见邮件截图 21](#_Toc477464017)

# 毕业设计（论文）内容概述

## 项目来源及开发目的和意义

本项目来源于上海骇咕赛信息科技有限公司大数据平台数据集成子系统的最新需求。公司长期专注于企业级高并发高可用性技术解决方案和数据分析、数据挖掘领域，主要产品有高并发高可用性分布式缓存集群系统，该项目设计为大型计算集群、为分布式计算服务提供热缓存数据交换。在此系统之上，针对于海量数据整合、处理、分析、学习的需求，公司力主架构于该分布式系统，实现简便、高效的海量数据加工的功能，为该系统设计海量数据快速处理的API。作为数据处理平台，数据集成便是其中不可缺少的一项，本人的项目即基于此内容进行一些创新性的研发。

数据集成涉及将不同来源中的数据整合，并统一地向用户展示这些。数据集成在很多领域上非常重要，包括商业（当两个相似的公司需要合并它们的数据库）、科学（组合来自不同生物信息学数据库的研究结果）。 随着数据量和数据分享需求爆炸式的增长，数据集成出现的频率越来越高。它已成为大范围理论工作的焦点，然而许多问题仍然没有解决。

上文中，维基百科的介绍已经暗示了数据集成的内容与重要性。简言之，数据集成是一个将具有不同概念、上下文、逻辑关系的数据文本进行合并，形成一个具有统一模式的数据集。作为数据分析和运用的基础，计算机领域内数据集成具有重要的意义，包括数据清洗、模式识别、生物信息等等。然而，随着大数据时代的到来，数据每天都在被生成、分析并且大量的应用着，数据驱动的决策也成为了社会中不可或缺的一部分。现如今，互联网上分散存在着海量的数据，为了充分利用这些数据中的信息以及蕴含的价值，将这些数据有效的集成在一起成为了一个显著的需求。

大数据时代下的数据集成，与传统数据集成的考核点（数量、速度、多样性、真实性）有着一定程度上的不同。首先，单一数据源无法为单一目标储存如此大量的数据。其次，很多数据源是动态的，总会有很多数据源在生成、消失，或者某些数据源改变了储存方式。第三，在诸多数据源之间，我们很难保证他们是完全统一的，这就造成了异构数据源的产生，并且相同条目下的数据也可能重复或者存在精度、修改时间等方面的差异。

## 主要开发任务（或总体设计内容及方案）

为了将海量的异构、异源数据进行集成，首先需要做的就是将不同数据库的模式进行集成，生成一个全局的数据库模式，进而方便数据库记录的填充和数据库的融合。对于数据集成问题，传统的方法往往是预先指定一个全局的数据库模式，然而针对于海量数据的背景，人们难以在大量的数据中捕捉全局的信息来设计预定的全局模式，并且建立全局模式和每一个数据库模式之间的匹配关系也是耗时耗力的。

因此，在上述背景下，通过设计一些合理匹配关系和高效的算法，省时、准确的生成一个全局数据库模式成为了迫切的需求。

## 本人所承担任务（模块）说明

本人在整个项目中的具体任务是属于海量数据快速处理平台的数据集成系统，具体的工作任务是设计并实现模式集成子系统。这项任务包括设计完成模式集成的高效、精准的算法，使用现有公司数据进行测试，系统实现并整合到公司系统中等。

模式集成的主要工作是针对异源、异构数据表的模式，在形式和语义的两个维度将它们进行集成，从而得到一个统一的模式，既能将多个数据源中的所有属性全部包含，又能保证产生的数据模式中属性彼此不重复，满足数据库的范式要求。

## 开发环境和开发工具

项目涉及到的开发环境如下：

操作系统：Windows/Linux/MacOS

处理器：Intel Core i7，2.40GHz主频或以上

内存：8GB 1333 MHz DDR3

程序运行环境：以Linux/MacOS为主，开发所涉及的语言和库应在其他各个平台（包括但不局限于Windows）通用，不受系统环境限制

注：以上开发环境均为前期开发要求，当集成到公司系统时，应使用更高性能的服务器，具体服务器配置在此保密

如表 1‑1 开发工具表所示，本系统实际开发过程使用到了这些开发工具。

表 1‑1 开发工具表

|  |  |  |
| --- | --- | --- |
| 工具类别 | 工具名称 | 作用 |
| 集成开发环境 | Visual Studio 2013  Xcode | 程序最主要的开发、调试平台 |
| 脚本开发工具 | Atom | 数据预处理、后处理脚本开发平台 |
| 界面设计工具 | Photoshop CC2015 | 图片素材设计 |
| 前端开发工具 | PyCharm | 前端界面的实现 |
| 版本控制软件 | Github | 版本控制 |

## 项目原定进度安排

项目原定进度安排见表 1‑2 项目原定进度工作计划表。

表 1‑2 项目原定进度工作计划表

|  |  |  |  |
| --- | --- | --- | --- |
| 起始时间 | 完成时间 | 计划工作内容 | 备注 |
| 2016.07.18 | 2016.07.31 | 配置公司要求的环境，熟悉集群 | 已完成 |
| 2016.08.01 | 2016.08.28 | 学习所需基础知识 | 已完成 |
| 2016.08.29 | 2016.09.11 | 分析现有系统，寻找创新点 | 已完成 |
| 2016.09.12 | 2016.09.18 | 选择研究方向 | 已完成 |
| 2016.09.19 | 2016.10.09 | 验证选题的可行性 | 已完成 |
| 2016.10.10 | 2016.11.06 | 研读相关工作现有论文，进行比较 | 已完成 |
| 2016.11.07 | 2016.11.20 | 撰写开题报告，准备开题答辩 | 已完成 |
| 2016.11.21 | 2016.12.04 | 寻找知识库，根据需要进行处理 | 已完成 |
| 2016.12.05 | 2017.01.01 | 完成形近整合部分算法的设计 | 已完成 |
| 2017.01.02 | 2017.02.05 | 完成语义整合部分算法的设计 | 已完成 |
| 2017.02.06 | 2017.03.05 | 使用少量数据进行试验，验证准确度 | 已完成 |
| 2017.03.06 | 2017.03.18 | 准备中期答辩 | 在进行 |
| 2017.03.19 | 2017.04.01 | 用大量数据进行试验，验证效率 | 未完成 |
| 2017.04.02 | 2017.04.22 | 根据实验结果调整系统 | 未完成 |
| 2017.04.23 | 2017.05.13 | 设计、实现前端界面 | 未完成 |
| 2017.05.14 | 2017.05.27 | 进行最后的验证，完善系统 | 未完成 |
| 2017.05.28 | 2017.06.20 | 撰写论文，参加毕业答辩 | 未完成 |

# 中期完成情况说明

## 预定计划的执行情况

开题至中期答辩前时间段的工作如表 2‑1开题至中期工作计划表。

表 2‑1开题至中期工作计划表

|  |  |  |  |
| --- | --- | --- | --- |
| 起始时间 | 完成时间 | 计划工作内容 | 备注 |
| 2016.11.21 | 2016.12.04 | 寻找知识库，根据需要进行处理 | 已完成 |
| 2016.12.05 | 2017.01.01 | 完成形近整合部分算法的设计 | 已完成 |
| 2017.01.02 | 2017.02.05 | 完成语义整合部分算法的设计 | 已完成 |
| 2017.02.06 | 2017.03.05 | 使用少量数据进行试验，验证准确度 | 已完成 |

在2016.11.21-2017.03.05的工作时间段期间，已经将开题答辩时计划的工作逐项完成，包括预处理知识库、设计形近整合算法、设计语义整合算法、筛选部分有代表性的数据进行测试。由于开题时对领域了解不足，有些情况未考虑在内，在实际的工作中添加了一些额外的工作，诸如数据结构设计、实验数据筛选等，这些具体的工作将在2.2中期工作说明及成果汇报中详细介绍。

## 中期工作说明及成果汇报

本节中将对现有完成工作进行详细的说明，包括但不局限于系统设计、算法设计、对应模型、成果截图。

### 系统功能结构

本系统根据功能划分主要包含5个模块：用户界面模块、预处理模块、形式整合模块、语义整合模块和全局模式模块。具体的功能结构图如图 2‑1 系统功能结构图所示。



图 2‑1 系统功能结构图

### 系统流程

整个系统以多个数据库的模式为输入，以现有知识库为依托，经过形式整合、语义整合阶段，生成全局属性。系统通过前端界面对以上过程的结果及必要的中间结果进行体现。该系统的流程如图 2‑2 系统流程图所示。

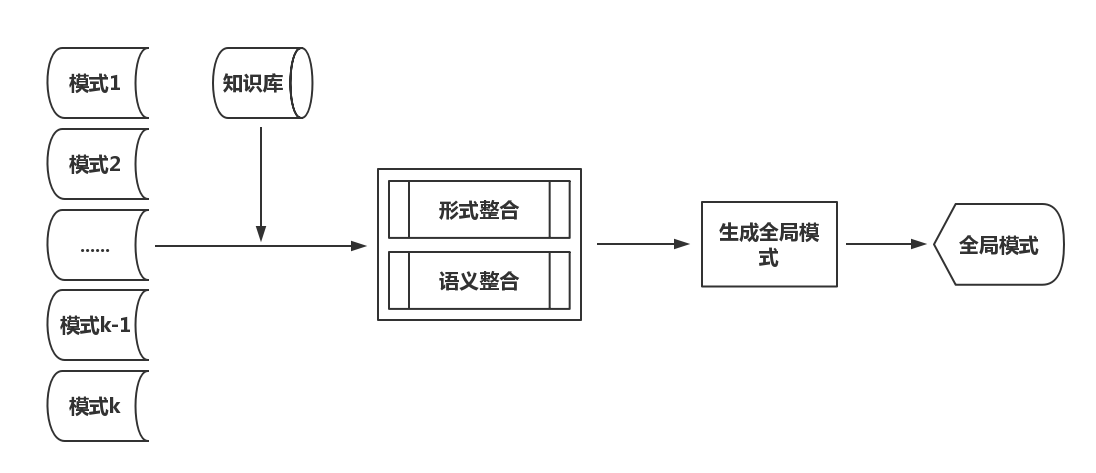


图 2‑2 系统流程图

系统的用户提供待集成的k个模式作为系统的输入，处理过程中经历了如下的流程：

1. 形式整合：将形式上相近的属性进行匹配，将在给定阈值范围下的属性归为一组（这种情况一般发生在拼写错误的情形下）
2. 语义整合：将语义上相近的属性进行匹配，将在给定阈值范围下的属性归为一组。这一过程利用了外部的知识库（经过预处理）
3. 生成全局模式：利用前两个步骤生成的匹配属性组，在每组相似的属性中选出一个代表属性，并将其他属性在给定的k个模式中去除并对模式取并，从而得到全局模式

### 系统用例详细描述

系统用例图如图 2‑3 系统用例图所示，描述了该系统实施模式集成功能时与用户的交互需求。

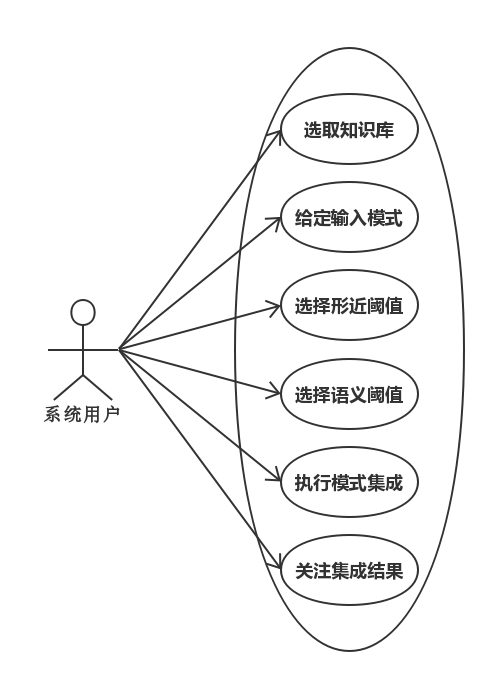


图 2‑3 系统用例图

用例的详细描述如下：

* 选取知识库：该系统的操作人员需要指定判断语义相似的知识库，知识库涵盖的语义范围以及精准度将直接决定系统的性能和集成效果。通常情况下待集成的模式之间应该有着必要的联系，往往是表达同一领域内的信息，因此知识库选取时更可能需要考虑对所研究范围的包含情况。目前完成的系统只支持*Freebase*，后期可根据需要添加知识库，甚至针对某领域编写需要的相似规则
* 给定输入模式：用户给出待集成的模式集合，为了保证集成的效果和实际意义，这些模式最好是来自接近的领域的，应有部分属性是相似的或者用来描述同一类实体的。输入的模式可能会包含很多属性，为了操作简便，我们只按照纯文本读取，暂不考虑其在数据库中的属性值以及XML等其他形式中属性之间的关系
* 选择形近阈值：用户在执行集成前需指定,用以限定拼写差异的上限。现实情况中，这一环节往往是用来应对未经过较系统的数据清洗的模式，加入此模块是用来排除输入属性具有未被发现的拼写错误而产生的干扰。由于这种情况出现几率不高，为了防止对系统总体的性能和准确率产生干扰，形近阈值往往取得比较低
* 选择语义阈值：用户在执行集成前需指定*γ*,用以限定语义差异的上限。这个阈值的确定很大程度上取决于输入模式和知识库的密度。知识库中相关领域的知识越密集，可能所需的阈值相对的就会大一些。阈值取的越大，进而匹配的结果包含的属性就会越多，花销时间越长，相对的全局模式可能会遗漏一些信息；而当阈值比较小时，时间开销较低，但得到的匹配结果较小，最后集成得到的全局模式越接近于输入k个模式的并集。因此这个阈值需要根据具体问题而定，目前阶段暂未有自动计算阈值的方法
* 执行模式集成：用户给定各个配置后，启动模式集成
* 关注集成结果：系统运行得到集成结果，用户需关注每一次得到的结果，并在小数据范围内评价该运行结果，并可能由于结果不良修改给定的阈值并重新执行

### 背景概念及定义

#### 语义相似判定方法

在进行模式集成的过程中，很重要的一个问题就是如何将相似的属性归为同一条（类）属性，避免在最后生成的全局属性中产生冗余，同时又不能将不同的属性划为一类使得全局属性在信息上有遗漏。这其中一个很关键的问题就是如何判定两个属性是相似的，这个概念对于人类来说很好理解，但是交给计算机来做却是非常困难的，其中一个重要原因是人类对于判断什么样的属性是相似的具有先验知识，这个知识是在人们的成长过程中逐步积累起来、加以自己的分析和理解而形成的。而计算机没有从如此复杂背景下提取的知识和做出这样判断感性能力，同时属性是否相似还与领域、上下文息息相关，这对于计算机来说很难达成。为了解决这个问题，在我们的模式集成系统中引入了外部的知识库，尽可能通过这种方式使计算机的集成过程模拟人脑的思维。

目前阶段我们采用了*Freebase*[[1]](#footnote-1)知识库，这是一个表示“is-a”关系的知识库，能够表示出相邻的两个知识概念就有类别从属或者包含的关系，相似的知识库还有*WordNet[[2]](#footnote-2)*, *Probase[[3]](#footnote-3)*和*YAGO[[4]](#footnote-4)*。这样的知识库整体的结构是一个图*G*，图中的每一个节点是一个概念如*a*和*b*（代表着客观存在的物体或者抽象的类别），边集*S*中的每条边*e*连接的两个节点（概念）具有从属或者包含的关系，该结构可以表示为：

虽然同义词数据库可以直接的表达我们希望得到的关系，但是现有的同义词数据库并不完善，并且在某些专业领域内，并不能简单地通过同义词来判定两个属性是相似的。而这种“is a”关系型的知识库一定程度上可以解决我们的需求（表达数据库表之间相关联的外键结构），尤其是*Freebase*等知识库在专业术语范围内有着很好的描述能力。当然采用这样形式的知识库不是唯一的，例如*NGD* (Normalized Google Distance)等均可以提供类似的信息。因此我们希望设计出一种统一的衡量方式和集成算法，在后期需要时可以很好地将这些知识库应用到我们的系统中。

基于上文中对语义知识库的介绍，为了将这种定性的知识间关系引入到我们的算法中，需要以一种定量的方式对此理解性的概念进行表示，这里对语义距离给出定义：

可以理解为*a*与*b*之间的距离可以表示为两点之间路径的长度（包含边的个数）。尽管这种定量表示方式可能不是非常合理，但是路径长度与语义差异程度正相关是毋庸置疑的，目前阶段取得的实验效果可以验证这种定义的定性准确性，后期的优化过程中可以试图为不同的边赋予不等的长度。基于这种定义，实际匹配的过程中需要用户指定语义近似阈值*γ*，当两个概念之间的距离小于*γ*时，可以看做这两个概念在含以上是近似的。

#### 形近相似判定方法

对于形式近似的情况，我们采用最基础的编辑距离算法，并由用户给定作为其阈值。执行编辑距离算法时，我们为给定模式中的属性和知识库中的概念按照*q*-gram生成倒排表（长度*q*由用户给定）。具体来说，我们将词以*q*为单位长度进行划分，当两个词有足够多的分片相同时，他们之间的差异便会足够小，可以认为他们是相似的。我们的方法采用*Count Filtering*的计算方式，具体来说*a*与*b*之间如果具有*LBab*个相同的*q*-gram分片，我们可以认为*a*与*b*在拼写上是相近的，其中

### 数据结构设计

#### Cluster Set

集成系统运行过程中需要考虑语义相似和拼写相似，而其中利用知识库进行语义相似判定的部分其实是更困难的。由于知识库巨大，我们不可能将知识库完全导入到内存中，在寻找相似概念的时候也不可能遍历整个知识库。因此我们的想法是在知识库上利用类似于数据库中的join算法，只在待集成模式中的属性附近去寻找语义相似的概念。同时为了减少遍历次数，尽量使基于拼写相似的集成一同完成，我们设计了如下的数据结构*Cluster Set*，这种结构适用于算法执行过程中的中间结果和系统最后的输出。

**[*Cluster Set*]**

*S* as the concepts set of the knowledge base, a cluster set is a set of pairs {*U*, *SU*}, where *U* is a set of attributes and

*Cluster Set*实际上是一个用来储存经过匹配后的集合，其中*U*集合储存着被下文中算法认定为相似的属性，而为了加速这个算法在拼写相近匹配阶段的运行，*SU*中储存的一些属性，每个至少和*U*中的一个元素距离小于。这些属性是可能在接下来的集成过程中归并到集合*U*中的，相比于每次在知识库上进行查找和匹配，提前将可能用到的属性以较小的代价储存下来，以空间换时间，能够提升算法的运行效率。

#### Neighbor Table

由于数据量巨大（模式中的属性和知识库中的概念目前共计在100,000,000条以上），很难将数据长期储存在内存中，因此该系统和内部的算法都是基于外存的，所以如何将数据读取的时间控制在一定范围内将是一个核心问题。通常情况下，外存的数据系统能够高效的运行取决于数据在磁盘上如何组织，hash和B树是比较常见的选择。在该数据集成系统中，为了方便高效，采用了hash的方法将知识库中的边*e*(*a*, *b*)储存，具体来说是分别对应起点*a*和*b*，在其hash值对应的空间内储存对应终点*b*和*a*以及其他的如边的长度等信息。

由于我们的算法实际上是在知识库中进行边的连接，形成多条路径 (*start*,*end*,*length*)，对于不同的路径，只要它们拥有共同的终点*end*，便可以与边*e*(*end*, *b*)进行连接。因此可以将拥有共同终点的路径归为一类，多条路径只需做一次相同的扩展以提升效率，系统中使用数据结构*Neighbor Table*，定义如下：

**[*Neighbor Table*]**

*t* is an attribute and *P* is the set of all paths in the knowledge base. *Hk*(*t*) is a table on the disk indexed by hash value of string *t*, s.t.

*Neighbor Table*接收知识库中的一个概念*t*，并返回所有以*t*为终点、长度为*k*的路径，这样的结构可以以*t*为关键字形成一个hash表，并将生成和访问的时间保持在一个较低的数量级，边之间的连接延长可以引申为*Neighbor Table*的延展。类似的，边集*E*在这里可以表示为*H1*(*t*)。

### 核心算法

本系统处理的核心问题是在知识库中寻找语义和形式上相近的属性，对于模式中被判断为相似的属性归并成一条，去除冗余并得到最终的全局模式。然而知识库中所含有的概念众多，在每次搜索时将其遍历一遍是不现实的，知识库中绝大多数的信息在单次匹配中是用不上的。因此我们借鉴了数据库中的join运算符，将其思路应用到我们的系统中。

为了对语义相似的属性进行匹配，本系统首先定义了一个Semantic Join操作符，主要任务是将至少满足如下三条之一的属性认为是语义相似的属性，将这些属性进行归并，并根据新生成的*U*集的结果维护*SU*，其定义如下：

**[*Semantic Join Operator*]**

Given two families of cluster sets *R*, *T*, and a threshold *d*, two elements (*U1*, *S1*) and (*U2*, *S2*) are from *R* and *T*, respectively are semantically joined if they satisfy one of the following constraints.

The result of Semantic join on (*U1*, *S1*) and (*U2*, *S2*) is a pair (*U*, *SU*), where

这里的*γ*是用户根据数据具体情况给定的一个语义阈值，具体来说是用来限定知识库上被认定为语义相似概念的距离上限。目前阶段这个值只能通过对知识库进行分析而给出，未来希望可以通过机器学习、关联挖掘等手段进行计算。从另一个角度来说，在给定目标属性集*R*和知识库边集*E*的前提下，Semantic Join算法的功能需求是将知识库上*γ*距离内的属性进行归并，其过程可抽象为下面的公式：

将上述功能映射到知识库的图结构时，可以理解为是从目标属性集*R*起始，将边逐步进行连接成长度在之间的路径，这些路径的起始于*R*中的属性，终点属性可以看作是与*R*中属性相似的属性。

Semantic Join算法的具体流程如图 2‑4 Semantic Join算法所示。

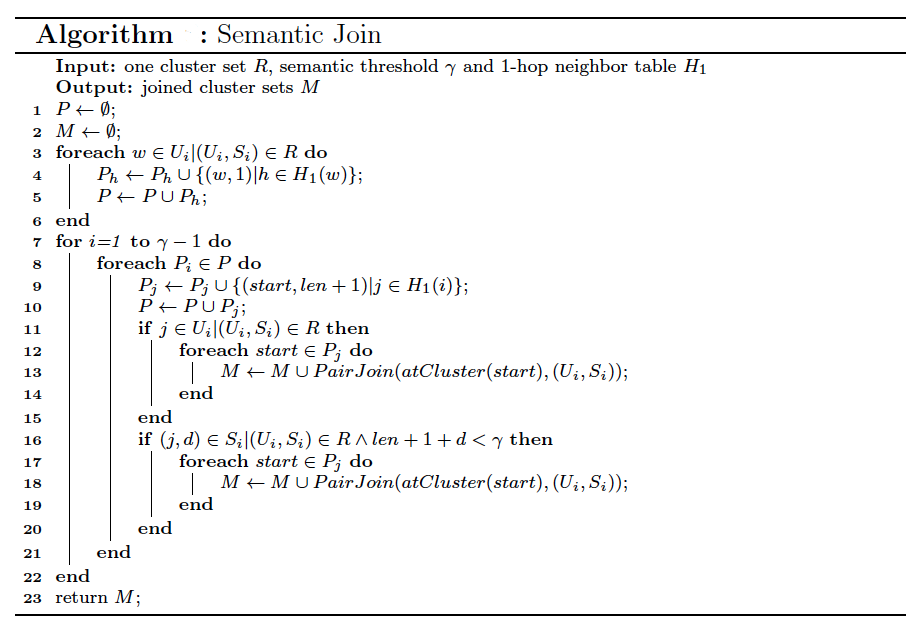


图 2‑4 Semantic Join算法

Semantic Join算法接收的输入变量有：

* *R*：待集成的属性集合（以cluster set形式储存）
* *γ*：用户指定的语义距离阈值
* *H1*：知识库的边集（以neighbor table形式储存）

该算法大体可以分为以下三个阶段：

1. 初始化（1-6行）：这一阶段扫描输入的属性集*R*，并将以*R*中的每一个属性*r*起始的路径加入到路径集合*P*中，其中*P*的定义如下：

**[*Path Set*]**

*Pa* is a path set, all paths in which share the same end node *a*, s.t.

当前过程得到的路径集合只包含了以目标属性开头的边，并按照其终点进行分组

1. 路径扩展（7-10行）：这一阶段在知识库中每执行一次join操作，即对于现有得到的属性集合再向外扩展一个单位距离。整个算法共计执行次join操作，使得能够包含与*R*中属性相距至多为*γ*的概念，即满足给定阈值范围的相似属性
2. 集合维护（11-20行）：这一阶段是将相似属性所处的集合合并，即已知*R1*中的*r1*和*R2*中的*r2*相似，需要将*R1*和*R2*合并，包括合并*U*集合和维护对应的*Su*集合

### 结果展示

现阶段已经完成了对核心算法的设计，并在小数据集上对算法的精度进行了实验。由于模式集成的效果如何，直接取决于对属性匹配的结果，这里对Semantic Join算法的准确度进行了实验。

#### 实验设置

* 实验环境：当前阶段采用Windows10 64-bit，Intel Core i7 2.4GHz，内存8GB
* 数据集：知识库采用*Freebase*，模式来自于NYC OpenData[[5]](#footnote-5)和SF OpenData[[6]](#footnote-6)数据库（主要涉及范围为城市建设、基础设施等），我们从网上获取到公开的数据集并加以我们的处理
* 阈值参数：参考2.2.3系统用例详细描述中用户对形近阈值和语义阈值的用例，这里的阈值应先期指定。根据现有的数据和知识库，此处暂且取

#### 实验指标

为了衡量算法的效果，采用了如下的通用指标：

* 准确率
* 召回率

具体到实验情景而言，以上两个公式中，*SA*是算法找到的匹配的属性，*ST*是实际上相似的属性（由于相似的概念比较模糊，这里的判定参考人们的直观概念），*ST*∩*SA*是算法找出且真实相似的属性。

#### 实验结果

准确率结果如表 2‑2 实验准确率所示。

表 2‑2 实验准确率

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Attribute |  |  |  | Recall | Precision |
| name | 76 | 61 | 57 | 0.934426 | 0.750000 |
| year | 93 | 64 | 58 | 0.906250 | 0.617021 |
| type | 73 | 58 | 53 | 0.913793 | 0.726027 |
| number | 79 | 68 | 65 | 0.955882 | 0.822785 |
| category | 12 | 13 | 15 | 0.923077 | 0.800000 |

由于实验准确度的测定与人对属性之间相似度的判断有关，这里引入了人工判断作为对比组，因此这部分的实验不能有较大的数据量。对于五组中的每一组实验，都只选取了一个属性构成了待集成的模式，并验证其在使用的数据集中匹配的情况。从结果中可以看到recall和precision是依据输入属性而变化的，recall的平均值为0.9266862，precision的平均值为0.7431666。

#### 实验分析

如上的实验是以人的判断作为绝对准则，人是通过多年积累的背景知识进行判断的。Semantic Join算法作为人工判断的一种模拟手段，由于不具有较广的背景知识、较灵活的知识应用，固然和人工匹配有一定差距，但是在某些程度上还是有着不错的表现，具体分析和改进如下：

* 关于recall的分析：recall这一指标大体的反映了算法结果与人工匹配的吻合程度，简言之，人们判定相似的属性有多少被算法识别了出来。实验结果中recall大体在0.9左右浮动，可以看出算法的识别率比较高，不能识别的很多是因为知识库和模式中的词语不能完全对应，这一点将在2.3存在的困难与问题进行讨论
* 关于precision的分析：precision这一指标大体的反映了算法结果的准确度，简言之，算法认为匹配的属性有多少是真正相似的。相比于recall，实验结果中precision的值是相对较低的。有些不准确的可能是知识库分布不均匀导致的，这一点将在2.3存在的困难与问题进行讨论

## 存在的困难与问题

现阶段已完成了表 2‑1开题至中期工作计划表中的相关工作，实际的实现过程中存在着如下的问题，需要在后续的阶段中解决或是提出替代方案，具体如下：

1. 由于知识库和使用的开源数据集是来自于不同数据源，知识库中的概念和数据集中的属性之间的差异远远大于我们的预测，含义相同的词在表现形式上很多是完全不一样的，包括缩写、俗语、词-词组不对应、分隔符等，如何解决这个问题并进行更为合理的匹配直接影响到recall的取值
2. 对于知识库上的语义距离，现阶段我们只是考虑的边的个数，即在知识库中存在的每一条边长度均相等（为1），根据对知识库进行具体的分析，这一规定是不尽合理的。显然，同样是“is a”关系起权重可能并不相同，如知识库上“Peter”和“first name”都与“name“构成了“is a”关系，显然这两者在知识库中的分量是不一样的，也不能简单判定这两者之间的关系是怎样的。其次，知识库实际上是一个图结构，而简单的“is a”关系构成的是树结构，这意味着知识库中的边必然存在跨层的连接，这也在一定程度上影响着不同概念的权重。因此，是否能够通过某些途径，为知识图谱上的边赋予一定的权值，使得语义距离的取值更为合理影响着算法的准确度
3. 现阶段系统只考虑了同一时间段给出大量的模式进行集成，然而在实际的生产环境中，这样的情况往往只发生一次或者很少的几次，更多情况是基于这次的结果进行增量的集成。因此，基于现有算法设计增量集成的算法也是很有必要的
4. 由于编辑距离只是针对拼写的差异，尤其在单词长度较短时很大几率会产生假阳性的情况，例如“work”和“word”在时就会判定为相似，这无疑影响了系统的准确度。因此需要设计出一套多个角度的规则进行验证，来防止这种假阳性对集成的影响
5. 系统在磁盘上采用了hash的方式进行组织，如何合理的利用这种结构对系统进行加速成为了下一阶段的终点。我们的算法将图上的拓展、匹配通过join操作符近似归化成了*Neighbor Table*与*Path Set*的连接（这两种数据结构一定是储存在磁盘上的）。如果能够将在同一个table/set中的属性在磁盘上合理组织，以较少的磁盘读取次数尽可能多的读到每次同时使用的数据，决定着系统的性能

## 如期完成预定任务的可能性分析

从开题至中期的工作重点在于对数据集成、模式集成领域的了解以及语义集成核心算法的设计，目前来看这部分已经很好的完成，在准确度方面看到了不错的效果。后期任务主要是在大数据集上进行时间维度的实验并对现有算法进行改进。由于涉及到和数据集、知识库选取相关的困难，还需要设计一些方式处理手中的数据。目前来看，经过努力是能够如期完成既定任务的。

## 后期工作安排（或进度和计划调整）

后半期的工作进度、时间安排如表 2‑3 后期工作计划表所示。

表 2‑3 后期工作计划表

|  |  |  |  |
| --- | --- | --- | --- |
| 起始时间 | 完成时间 | 计划工作内容 | 备注 |
| 2017.03.19 | 2017.04.01 | 用大量数据进行试验，验证效率 | 未完成 |
| 2017.04.02 | 2017.04.22 | 根据实验结果调整系统和数据集 | 未完成 |
| 2017.04.23 | 2017.05.13 | 设计、实现前端界面 | 未完成 |
| 2017.05.14 | 2017.05.27 | 进行最后的验证，完善系统 | 未完成 |
| 2017.05.28 | 2017.06.20 | 撰写论文，参加毕业答辩 | 未完成 |

参考文献

1. Arasu, S. Chaudhuri, and R. Kaushik. Learning string transformations from examples. Proceedings of the VLDB Endowment, 2(1):514-525, 2009
2. D. Beneventano, S. Bergamaschi, S. Castano, A. Corni, R. Guidetti, G. Malvezzi, M. Melchiori, and M. Vincini. Information integration: the momis project demonstration. In Vldb, pages 611-614, 2000
3. S. Castano and V. De Antonellis. Global viewing of heterogeneous data sources. IEEE Transactions on Knowledge and Data Engineering, 13(2):277-297, 2001
4. R. Cilibrasi and P. Vitanyi. Automatic meaning discovery using google. In Dagstuhl Seminar Proceedings. Schloss Dagstuhl-Leibniz-Zentrum fur Informatik, 2006
5. A. Doan, P. Domingos, and A. Y. Halevy. Reconciling schemas of disparate data sources: A machine-learning approach. In ACM Sigmod Record, volume 30, pages 509-520. ACM, 2001
6. A. Doan, P. M. Domingos, and A. Y. Levy. Learning source description for data integration. In WebDB (Informal Proceedings), pages 81-86, 2000
7. U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. AI magazine, 17(3):37, 1996
8. L. Gravano, P. G. Ipeirotis, H. V. Jagadish, N. Koudas, S. Muthukrishnan, L. Pietarinen, and D. Srivastava. Using q-grams in a dbms for approximate string processing. IEEE Data Eng. Bull., 24(4):28-34, 2001
9. M. L. Lee, T. W. Ling, and W. L. Low. Intelliclean: a knowledgebased intelligent data cleaner. In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 290-294. ACM, 2000
10. V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In Soviet physics doklady, volume 10, pages 707-710, 1966
11. L. Li, H. Wang, J. Li, and H. Gao. Ed-sjoin; an optimal algorithm for similarity joins with edit distance constraints [j]. Journal of Computer Research and Development, 46:319-325, 2009
12. W.-S. Li and C. Clifton. Semantic integration in heterogeneous databases using neural networks. In vldb, volume 94, pages 12-15, 1994
13. W.-S. Li and C. Clifton. Semint: A tool for identifying attribute correspondences in heterogeneous databases using neural networks. Data & Knowledge Engineering, 33(1):49-84, 2000
14. X.-M. Lin and W. Wang. Set and string similarity queries: A survey. Jisuanji Xuebao(Chinese Journal of Computers), 34(10):1853-1862, 2011
15. J. Madhavan, P. A. Bernstein, and E. Rahm. Generic schema matching with cupid. In vldb, volume 1, pages 49-58, 2001
16. R. Mihalcea and P. Tarau. Textrank: Bringing order into texts. Association for Computational Linguistics, 2004
17. T. Milo and S. Zohar. Using schema matching to simplify heterogeneous data translation. In vldb, volume 98, pages 24-27. Citeseer, 1998
18. P. Mitra, G. Wiederhold, and J. Jannink. Semi-automatic integration of knowledge sources. Proceedings of Fusion'99, July 1999, 1999
19. P. Mitra, G. Wiederhold, and M. Kersten. A graph-oriented model for articulation of ontology interdependencies. In International Conference on Extending Database Technology, pages 86-100. Springer, 2000
20. L. Palopoli, D. Sacca, and D. Ursino. An automatic technique for detecting type conicts in database schemes. In Proceedings of the seventh international conference on Information and knowledge management, pages 306-313. ACM, 1998
21. L. Palopoli, D. Sacca, and D. Ursino. Semi-automatic, semantic discovery of properties from database schemes. In Database Engineering and Applications Symposium, 1998. Proceedings. IDEAS'98. International, pages 244-253. IEEE, 1998
22. E. Rahm and P. A. Bernstein. A survey of approaches to automatic schema matching. the VLDB Journal, 10(4):334-350, 2001
23. G. Salton, E. A. Fox, and H. Wu. Extended boolean information retrieval. Communications of the ACM, 26(11):1022-1036, 1983
24. K. Sparck Jones. A statistical interpretation of term specicity and its application in retrieval. Journal of documentation, 28(1):11-21, 1972
25. H. Wang, J. Li, and H. Gao. Efficient entity resolution based on subgraph cohesion. Knowledge and Information Systems, 46(2):285-314, 2016
26. C. Xiao, W. Wang, and X. Lin. Ed-join: an efficient algorithm for similarity joins with edit distance constraints. Proceedings of the VLDB Endowment, 1(1):933-944, 2008
27. 陈跃国, 王京春. 数据集成综述[J]. 计算机科学, 2004, 31(5): 48-51
28. 俞红奇, 丁宝康. 多数据库环境下的模式集成及查询分解[J]. 计算机工程, 2000, 26(10): 124-126
29. 甄玉钢, 刘璐莹, 康建初. 基于 XML 的异构数据库集成系统构架与开发[J]. 计算机工程, 2006, 32(2): 85-87
30. 张正明, 佟俐鹃. 异构数据库集成的研究与实现[J]. 航空计算技术, 2004, 34(2): 33-36
31. 瞿裕忠, 胡伟, 郑东栋, 等. 关系数据库模式和本体间映射的研究综述[J]. 计算机研究与发展, 2008, 45(2): 300-309
32. 黎建辉, 佘怀化, 阎保平. 基于元数据的关系数据库语义集成方法[J]. 计算机工程, 2008, 34(6): 54-56

# 附件：第1次中期检查记录表

|  |  |  |  |
| --- | --- | --- | --- |
| 学生姓名 | 李天宝 | 学 号 | 1133730206 |
| 专 业 | 物联网工程 | 中检日期 | 2017.03.18 |
| 指导教师 | 王宏志 | 职 称 | 教授 |
| 设计（论文）题目 | 基于知识库的海量异构数据集成系统的设计与实现 | | |
| 指导教师评语：  **该生前期毕业设计工作表现：**□非常突出 □积极努力 □较积极 □一般 □差；  **项目完成工作量约：**□70%以上 □60%以上 □50%以上 □低于50% □严重不足；  **中期报告撰写：**内容□充实 □较充实 □一般 □不够充分；  格式□规范 □较规范 □一般 □不规范。  □同意 □不同意 □报告修改后同意 参加中期检查。  签字： | | | |
| 答辩记录： | | | |
| 是否通过中检： □通过 □警告 □不通过 | | 成绩：□A □B □C | |
| 答辩组长签字： | | 答辩组成员： | |
| 答辩秘书签字： | |

# 附件：导师意见邮件截图





1. https://www.freebase.com/ [↑](#footnote-ref-1)
2. http://wordnet.princeton.edu/ [↑](#footnote-ref-2)
3. http://research.microsoft.com/en-us/projects/probase/ [↑](#footnote-ref-3)
4. http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/ [↑](#footnote-ref-4)
5. https://data.cityofnewyork.us/ [↑](#footnote-ref-5)
6. https://data.sfgov.org/ [↑](#footnote-ref-6)