李天宝

turingmac@hotmail.com

基于知识库的海量异构数据集成系统的设计与实现

实习周报

上海骇咕赛信息科技有限公司

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **本科生实习&毕业设计（论文）工作周报** | | | | | | | | |
| 学生姓名 | 李天宝 | 联系电话 | 15704600640 | | Email | turingmac@hotmail.com | | |
| 实习基地  (或实验室) | 上海骇咕赛信息科技有限公司 | | 基地导师 | 丁盛豪 | | | 校内导师 | 王宏志 |
| 报告日期 | 2016-11-27 | | 实习周次 | 第19周 | | | | |
| 报告区间 | 2016-11-21 到 2016-11-27 | | | | | | | |
| 工作重点 | 了解现有知识库 | | | | | | | |
| 工作内容及进度 | 1. 阅读freebase文档 2. 阅读yago文档 | | | | | | | |
| 遇到问题 | 1. 现有知识库不能良好的包含同义词相关概念 | | | | | | | |
| 自我小结 | 本周主要进行了知识库相关的了解和学习，由于知识库众多，之间的特性差异很大，不容易确定哪个最适合系统 | | | | | | | |
| 下周计划 | 查询同义词知识库 | | | | | | | |
| 导师意见 | 签字： | | | | | | | |

**说明：**

1. 上述表格大小可以根据文字多少调整；
2. 每周要写工作周报并发送给校内导师，实习基地导师可根据需要发送；
3. 校内实习学生每2周要书面报呈周报给导师，当面指导并签字。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **本科生实习&毕业设计（论文）工作周报** | | | | | | | | |
| 学生姓名 | 李天宝 | 联系电话 | 15704600640 | | Email | turingmac@hotmail.com | | |
| 实习基地  (或实验室) | 上海骇咕赛信息科技有限公司 | | 基地导师 | 丁盛豪 | | | 校内导师 | 王宏志 |
| 报告日期 | 2016-12-04 | | 实习周次 | 第20周 | | | | |
| 报告区间 | 2016-11-28 到 2016-12-04 | | | | | | | |
| 工作重点 | 查询同义词知识库 | | | | | | | |
| 工作内容及进度 | 1. 了解现有同义词知识库 2. 阅读Automatic meaning discovery using Google | | | | | | | |
| 遇到问题 | 1. 同义词知识库较少 2. 同义词知识库包含信息太少 | | | | | | | |
| 自我小结 | 本周主要研究了同义词知识库，但是不太适合我们的系统，最终决定使用良构的“is a”关系的知识库 | | | | | | | |
| 下周计划 | 了解字符串匹配算法 | | | | | | | |
| 导师意见 | 签字： | | | | | | | |

**说明：**

1. 上述表格大小可以根据文字多少调整；
2. 每周要写工作周报并发送给校内导师，实习基地导师可根据需要发送；
3. 校内实习学生每2周要书面报呈周报给导师，当面指导并签字。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **本科生实习&毕业设计（论文）工作周报** | | | | | | | | |
| 学生姓名 | 李天宝 | 联系电话 | 15704600640 | | Email | turingmac@hotmail.com | | |
| 实习基地  (或实验室) | 上海骇咕赛信息科技有限公司 | | 基地导师 | 丁盛豪 | | | 校内导师 | 王宏志 |
| 报告日期 | 2016-12-11 | | 实习周次 | 第21周 | | | | |
| 报告区间 | 2016-12-05 到 2016-12-11 | | | | | | | |
| 工作重点 | 了解字符串匹配算法 | | | | | | | |
| 工作内容及进度 | 1. 实现基本的集中字符串匹配算法 2. 学习KMP | | | | | | | |
| 遇到问题 | 1. 现有算法无法进行近似匹配 | | | | | | | |
| 自我小结 | 本周主要对学习了字符串近似算法，一开始想要修改KMP算法达成近似匹配，但是这种逐字匹配的算法对于我们的系统来说时间开销太大 | | | | | | | |
| 下周计划 | 学习编辑距离算法 | | | | | | | |
| 导师意见 | 签字： | | | | | | | |

**说明：**

1. 上述表格大小可以根据文字多少调整；
2. 每周要写工作周报并发送给校内导师，实习基地导师可根据需要发送；
3. 校内实习学生每2周要书面报呈周报给导师，当面指导并签字。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **本科生实习&毕业设计（论文）工作周报** | | | | | | | | |
| 学生姓名 | 李天宝 | 联系电话 | 15704600640 | | Email | turingmac@hotmail.com | | |
| 实习基地  (或实验室) | 上海骇咕赛信息科技有限公司 | | 基地导师 | 丁盛豪 | | | 校内导师 | 王宏志 |
| 报告日期 | 2016-12-18 | | 实习周次 | 第22周 | | | | |
| 报告区间 | 2016-12-12 到 2016-12-18 | | | | | | | |
| 工作重点 | 学习编辑距离算法 | | | | | | | |
| 工作内容及进度 | 1. 阅读Binary codes capable of correcting deletions, insertions, and reversals 2. 实现编辑距离算法 | | | | | | | |
| 遇到问题 | 1. 逐个单词计算编辑距离开销太大 | | | | | | | |
| 自我小结 | 本周主要对学习了编辑距离，这个衡量方式对于后续算法设计会有很大的帮助 | | | | | | | |
| 下周计划 | 设计形近整合算法 | | | | | | | |
| 导师意见 | 签字： | | | | | | | |

**说明：**

1. 上述表格大小可以根据文字多少调整；
2. 每周要写工作周报并发送给校内导师，实习基地导师可根据需要发送；
3. 校内实习学生每2周要书面报呈周报给导师，当面指导并签字。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **本科生实习&毕业设计（论文）工作周报** | | | | | | | | |
| 学生姓名 | 李天宝 | 联系电话 | 15704600640 | | Email | turingmac@hotmail.com | | |
| 实习基地  (或实验室) | 上海骇咕赛信息科技有限公司 | | 基地导师 | 丁盛豪 | | | 校内导师 | 王宏志 |
| 报告日期 | 2016-12-25 | | 实习周次 | 第23周 | | | | |
| 报告区间 | 2016-12-19 到 2016-12-25 | | | | | | | |
| 工作重点 | 设计形近整合算法 | | | | | | | |
| 工作内容及进度 | 1. 阅读Ed-Join: an efficient algorithm for similarity joins with edit distance constraints 2. 阅读Ed-Sjoin; An Optimal Algorithm for Similarity Joins with Edit Distance Constraints | | | | | | | |
| 遇到问题 | 1. 相关领域了解不够 | | | | | | | |
| 自我小结 | 本周主要对研究了Ed-join现有的实现 | | | | | | | |
| 下周计划 | 设计形近整合算法 | | | | | | | |
| 导师意见 | 签字： | | | | | | | |

**说明：**

1. 上述表格大小可以根据文字多少调整；
2. 每周要写工作周报并发送给校内导师，实习基地导师可根据需要发送；
3. 校内实习学生每2周要书面报呈周报给导师，当面指导并签字。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **本科生实习&毕业设计（论文）工作周报** | | | | | | | | |
| 学生姓名 | 李天宝 | 联系电话 | 15704600640 | | Email | turingmac@hotmail.com | | |
| 实习基地  (或实验室) | 上海骇咕赛信息科技有限公司 | | 基地导师 | 丁盛豪 | | | 校内导师 | 王宏志 |
| 报告日期 | 2017-01-21 | | 实习周次 | 第24周 | | | | |
| 报告区间 | 2016-12-26 到 2017-01-01 | | | | | | | |
| 工作重点 | 设计形近整合算法 | | | | | | | |
| 工作内容及进度 | 1. 阅读Answering approximate string queries on large data sets using external memory 2. 设计算法 | | | | | | | |
| 遇到问题 | 1. 基于外存的算法效率受磁盘影响较大 | | | | | | | |
| 自我小结 | 本周主要设计了形近整合的join算法 | | | | | | | |
| 下周计划 | 了解知识库 | | | | | | | |
| 导师意见 | 签字： | | | | | | | |

**说明：**

1. 上述表格大小可以根据文字多少调整；
2. 每周要写工作周报并发送给校内导师，实习基地导师可根据需要发送；
3. 校内实习学生每2周要书面报呈周报给导师，当面指导并签字。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **本科生实习&毕业设计（论文）工作周报** | | | | | | | | |
| 学生姓名 | 李天宝 | 联系电话 | 15704600640 | | Email | turingmac@hotmail.com | | |
| 实习基地  (或实验室) | 上海骇咕赛信息科技有限公司 | | 基地导师 | 丁盛豪 | | | 校内导师 | 王宏志 |
| 报告日期 | 2017-01-08 | | 实习周次 | 第25周 | | | | |
| 报告区间 | 2017-01-02 到 2017-01-08 | | | | | | | |
| 工作重点 | 了解知识库 | | | | | | | |
| 工作内容及进度 | 1. 修改知识库格式 | | | | | | | |
| 遇到问题 | 1. 知识库部分结构混乱 | | | | | | | |
| 自我小结 | 本周主要对知识库进行预处理 | | | | | | | |
| 下周计划 | 应用知识库 | | | | | | | |
| 导师意见 | 签字： | | | | | | | |

**说明：**

1. 上述表格大小可以根据文字多少调整；
2. 每周要写工作周报并发送给校内导师，实习基地导师可根据需要发送；
3. 校内实习学生每2周要书面报呈周报给导师，当面指导并签字。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **本科生实习&毕业设计（论文）工作周报** | | | | | | | | |
| 学生姓名 | 李天宝 | 联系电话 | 15704600640 | | Email | turingmac@hotmail.com | | |
| 实习基地  (或实验室) | 上海骇咕赛信息科技有限公司 | | 基地导师 | 丁盛豪 | | | 校内导师 | 王宏志 |
| 报告日期 | 2017-01-15 | | 实习周次 | 第26周 | | | | |
| 报告区间 | 2017-01-09 到 2017-01-15 | | | | | | | |
| 工作重点 | 应用知识库 | | | | | | | |
| 工作内容及进度 | 1. 知识库与模式的匹配 2. 挖掘知识库中属性相关信息 | | | | | | | |
| 遇到问题 | 1. 知识库中数据有限 2. 知识库与属性之间有偏差 | | | | | | | |
| 自我小结 | 本周主要进行了知识库与模式之间的，两者之间还是有着一定差异的，现阶段为保证进度先进行算法设计，后期应设计更好的对应方法 | | | | | | | |
| 下周计划 | 语义整合算法设计 | | | | | | | |
| 导师意见 | 签字： | | | | | | | |

**说明：**

1. 上述表格大小可以根据文字多少调整；
2. 每周要写工作周报并发送给校内导师，实习基地导师可根据需要发送；
3. 校内实习学生每2周要书面报呈周报给导师，当面指导并签字。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **本科生实习&毕业设计（论文）工作周报** | | | | | | | | |
| 学生姓名 | 李天宝 | 联系电话 | 15704600640 | | Email | turingmac@hotmail.com | | |
| 实习基地  (或实验室) | 上海骇咕赛信息科技有限公司 | | 基地导师 | 丁盛豪 | | | 校内导师 | 王宏志 |
| 报告日期 | 2017-01-22 | | 实习周次 | 第27周 | | | | |
| 报告区间 | 2017-01-16 到 2017-01-22 | | | | | | | |
| 工作重点 | 语义整合算法设计 | | | | | | | |
| 工作内容及进度 | 1. 语义整合算法设计 2. 阅读 String similarity measures and joins with synonyms | | | | | | | |
| 遇到问题 | 1. 利用知识库进行集成的先例较少 | | | | | | | |
| 自我小结 | 本周主要进行了语义整合算法设计 | | | | | | | |
| 下周计划 |  | | | | | | | |
| 导师意见 | 签字： | | | | | | | |

**说明：**

1. 上述表格大小可以根据文字多少调整；
2. 每周要写工作周报并发送给校内导师，实习基地导师可根据需要发送；
3. 校内实习学生每2周要书面报呈周报给导师，当面指导并签字。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **本科生实习&毕业设计（论文）工作周报** | | | | | | | | |
| 学生姓名 | 李天宝 | 联系电话 | 15704600640 | | Email | turingmac@hotmail.com | | |
| 实习基地  (或实验室) | 上海骇咕赛信息科技有限公司 | | 基地导师 | 丁盛豪 | | | 校内导师 | 王宏志 |
| 报告日期 | 2017-01-29 | | 实习周次 | 第28周 | | | | |
| 报告区间 | 2017-01-23 到 2017-01-29 | | | | | | | |
| 工作重点 | 春节假期 | | | | | | | |
| 工作内容及进度 | 1. 阅读 Entity matching: How similar is similar | | | | | | | |
| 遇到问题 |  | | | | | | | |
| 自我小结 |  | | | | | | | |
| 下周计划 |  | | | | | | | |
| 导师意见 | 签字： | | | | | | | |

**说明：**

1. 上述表格大小可以根据文字多少调整；
2. 每周要写工作周报并发送给校内导师，实习基地导师可根据需要发送；
3. 校内实习学生每2周要书面报呈周报给导师，当面指导并签字。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **本科生实习&毕业设计（论文）工作周报** | | | | | | | | |
| 学生姓名 | 李天宝 | 联系电话 | 15704600640 | | Email | turingmac@hotmail.com | | |
| 实习基地  (或实验室) | 上海骇咕赛信息科技有限公司 | | 基地导师 | 丁盛豪 | | | 校内导师 | 王宏志 |
| 报告日期 | 2017-02-05 | | 实习周次 | 第29周 | | | | |
| 报告区间 | 2017-01-30 到 2017-02-05 | | | | | | | |
| 工作重点 | 春节假期 | | | | | | | |
| 工作内容及进度 | 1. 阅读 Learning string transformations from examples | | | | | | | |
| 遇到问题 |  | | | | | | | |
| 自我小结 |  | | | | | | | |
| 下周计划 |  | | | | | | | |
| 导师意见 | 签字： | | | | | | | |

**说明：**

1. 上述表格大小可以根据文字多少调整；
2. 每周要写工作周报并发送给校内导师，实习基地导师可根据需要发送；
3. 校内实习学生每2周要书面报呈周报给导师，当面指导并签字。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **本科生实习&毕业设计（论文）工作周报** | | | | | | | | |
| 学生姓名 | 李天宝 | 联系电话 | 15704600640 | | Email | turingmac@hotmail.com | | |
| 实习基地  (或实验室) | 上海骇咕赛信息科技有限公司 | | 基地导师 | 丁盛豪 | | | 校内导师 | 王宏志 |
| 报告日期 | 2017-02-12 | | 实习周次 | 第30周 | | | | |
| 报告区间 | 2017-02-06 到 2017-02-12 | | | | | | | |
| 工作重点 | 语义整合算法优化 | | | | | | | |
| 工作内容及进度 | 1. 完成算法设计 2. 进行算法优化 3. 阅读 A survey of approaches to automatic schema matching | | | | | | | |
| 遇到问题 | 1. 如何对于基于磁盘的算法进行加速 | | | | | | | |
| 自我小结 | 本周主要完成算法设计，并进行优化，由于数据储存在磁盘上，这个对于效率的限制比较大 | | | | | | | |
| 下周计划 | 基于磁盘进行加速 | | | | | | | |
| 导师意见 | 签字： | | | | | | | |

**说明：**

1. 上述表格大小可以根据文字多少调整；
2. 每周要写工作周报并发送给校内导师，实习基地导师可根据需要发送；
3. 校内实习学生每2周要书面报呈周报给导师，当面指导并签字。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **本科生实习&毕业设计（论文）工作周报** | | | | | | | | |
| 学生姓名 | 李天宝 | 联系电话 | 15704600640 | | Email | turingmac@hotmail.com | | |
| 实习基地  (或实验室) | 上海骇咕赛信息科技有限公司 | | 基地导师 | 丁盛豪 | | | 校内导师 | 王宏志 |
| 报告日期 | 2017-02-19 | | 实习周次 | 第31周 | | | | |
| 报告区间 | 2017-02-13 到 2017-02-19 | | | | | | | |
| 工作重点 | 基于磁盘进行加速 | | | | | | | |
| 工作内容及进度 | 1. 详读Database system implementation相关章节 2. 详读A first course in database systems相关章节 | | | | | | | |
| 遇到问题 | 1. 选取适当的数据在磁盘上的组织结构，如hash，B树等 | | | | | | | |
| 自我小结 | 本周主要对数据在磁盘上的组织进行了研究，现阶段决定使用hash进行实现 | | | | | | | |
| 下周计划 | 整理准确度实验数据 | | | | | | | |
| 导师意见 | 签字： | | | | | | | |

**说明：**

1. 上述表格大小可以根据文字多少调整；
2. 每周要写工作周报并发送给校内导师，实习基地导师可根据需要发送；
3. 校内实习学生每2周要书面报呈周报给导师，当面指导并签字。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **本科生实习&毕业设计（论文）工作周报** | | | | | | | | |
| 学生姓名 | 李天宝 | 联系电话 | 15704600640 | | Email | turingmac@hotmail.com | | |
| 实习基地  (或实验室) | 上海骇咕赛信息科技有限公司 | | 基地导师 | 丁盛豪 | | | 校内导师 | 王宏志 |
| 报告日期 | 2017-02-26 | | 实习周次 | 第32周 | | | | |
| 报告区间 | 2017-02-20 到 2017-02-26 | | | | | | | |
| 工作重点 | 整理准确度实验数据 | | | | | | | |
| 工作内容及进度 | 1. 搜索开源数据集 2. 验证作为实验数据的可行性 | | | | | | | |
| 遇到问题 |  | | | | | | | |
| 自我小结 | 本周主要确定了作为准确度实验的模式数据，最后选取了NYC OpenData 和SF OpenData，并根据需要进行简单地清洗和处理 | | | | | | | |
| 下周计划 | 进行准确度实验 | | | | | | | |
| 导师意见 | 签字： | | | | | | | |

**说明：**

1. 上述表格大小可以根据文字多少调整；
2. 每周要写工作周报并发送给校内导师，实习基地导师可根据需要发送；
3. 校内实习学生每2周要书面报呈周报给导师，当面指导并签字。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **本科生实习&毕业设计（论文）工作周报** | | | | | | | | |
| 学生姓名 | 李天宝 | 联系电话 | 15704600640 | | Email | turingmac@hotmail.com | | |
| 实习基地  (或实验室) | 上海骇咕赛信息科技有限公司 | | 基地导师 | 丁盛豪 | | | 校内导师 | 王宏志 |
| 报告日期 | 2017-03-05 | | 实习周次 | 第33周 | | | | |
| 报告区间 | 2017-02-27 到 2017-03-05 | | | | | | | |
| 工作重点 | 进行准确度实验 | | | | | | | |
| 工作内容及进度 | 1. 进行试验 | | | | | | | |
| 遇到问题 | 1. 现有的代码过于零散 2. 数据集中有非法数据 | | | | | | | |
| 自我小结 | 本周主要进行了准确度试验，使用recall和precision进行衡量 | | | | | | | |
| 下周计划 | 准备中期答辩 | | | | | | | |
| 导师意见 | 签字： | | | | | | | |

**说明：**

1. 上述表格大小可以根据文字多少调整；
2. 每周要写工作周报并发送给校内导师，实习基地导师可根据需要发送；
3. 校内实习学生每2周要书面报呈周报给导师，当面指导并签字。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **本科生实习&毕业设计（论文）工作周报** | | | | | | | | |
| 学生姓名 | 李天宝 | 联系电话 | 15704600640 | | Email | turingmac@hotmail.com | | |
| 实习基地  (或实验室) | 上海骇咕赛信息科技有限公司 | | 基地导师 | 丁盛豪 | | | 校内导师 | 王宏志 |
| 报告日期 | 2017-03-12 | | 实习周次 | 第34周 | | | | |
| 报告区间 | 2017-03-06 到 2017-03-12 | | | | | | | |
| 工作重点 | 准备中期答辩 | | | | | | | |
| 工作内容及进度 | 1. 整理现有工作和代码 2. 撰写中期报告 | | | | | | | |
| 遇到问题 |  | | | | | | | |
| 自我小结 | 本周主要为中期答辩做准备 | | | | | | | |
| 下周计划 | 准备中期答辩 | | | | | | | |
| 导师意见 | 签字： | | | | | | | |

**说明：**

1. 上述表格大小可以根据文字多少调整；
2. 每周要写工作周报并发送给校内导师，实习基地导师可根据需要发送；
3. 校内实习学生每2周要书面报呈周报给导师，当面指导并签字。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **本科生实习&毕业设计（论文）工作周报** | | | | | | | | |
| 学生姓名 | 李天宝 | 联系电话 | 15704600640 | | Email | turingmac@hotmail.com | | |
| 实习基地  (或实验室) | 上海骇咕赛信息科技有限公司 | | 基地导师 | 丁盛豪 | | | 校内导师 | 王宏志 |
| 报告日期 | 2017-03-19 | | 实习周次 | 第35周 | | | | |
| 报告区间 | 2017-03-13 到 2017-03-19 | | | | | | | |
| 工作重点 | 准备中期答辩 | | | | | | | |
| 工作内容及进度 | 1. 完善开题报告 2. 做开题答辩PPT | | | | | | | |
| 遇到问题 |  | | | | | | | |
| 自我小结 | 本周主要准备中期答辩，未进行更多的工作 | | | | | | | |
| 下周计划 |  | | | | | | | |
| 导师意见 | 签字： | | | | | | | |

**说明：**

1. 上述表格大小可以根据文字多少调整；
2. 每周要写工作周报并发送给校内导师，实习基地导师可根据需要发送；
3. 校内实习学生每2周要书面报呈周报给导师，当面指导并签字。