|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 学生姓名 | 李天宝 | | 专业 | 物联网工程 | 学号 | | 1133730206 | |
| 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 |
| 主要实习工内容汇报（实习开始到目前为止）（包括培训、学习、项目开发等内容）  （不少于1000字） | | | | | | | | |
| **1. 实习公司/实验室的情况介绍（包括主营业务、软件产品、研究/开发方向等）**  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及数据挖掘领域，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、分布式计算服务提供热缓存数据交换。该系统特点为支持极大数量的机器同时运行、同时访问、同时修改，系统依据 PAXOS 算法等业内领先的分布式计算论文设计了具有极高容错能力的一致性控制系统。支持随时热插拔机器，同时允许任何几台运行系统的机器遇到故障，能够被检出、重启和修复。该系统为大数据处理、机器学习等热数据交换频繁、计算结构并行且复杂的计算框架提供了极大的性能支持。目前该项目有多名系统底层和网络底层方向的工程师参与开发，具体细节公司保密。  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。相比于Apache Hadoop等分布式解决方案，公司希望能够实现功能更为完善的基础接口，甚至包括一些具有科研创新性的内容。公司力图在科研人员和工程师的协作下，能够将创新点及时有效的加入到系统中，使得该分布式系统实用性和技术领先性兼备。  作为以大数据为核心内容的初创公司，骇咕赛主要业务都是依据现有分布式平台，一方面为该平台完善功能、增添业务逻辑，另一方面旨在基于该平台为企业提供定型化服务，诸如数据挖掘/机器学习服务器集群、用户画像提取系统，数据可视化展示平台，数据采集/众包服务（APP）等等。  **2. 到目前为止学习、工作情况介绍**  入职到现在为止的工作主要以学习为主，对相关知识进行积累。  一方面，由于公司所依靠的分布式集群系统过于庞大，实习至今很大一部分时间用于熟悉该系统以及该系统相关的知识，诸如 PAXOS 算法、并行程序开发等，并对整个系统功能和逻辑有一个简单的了解。同时，为了保证下一步开发能够良好地进行，研读了公司的代码规范，研究了C++网络通讯相关内容，并温习了学校内所学的网络通讯相关模型。  另一方面，为了更好地满足后续研究和开发的需求，还对海量数据相关内容进行了学习，包括map-reduce模型、Apache Hadoop、Pig、R语言以及一些常用的机器学习和数据挖掘的算法，阅读了《Hadoop权威指南》《Pig编程指南》《R语言入门与实践》等书籍。学习过程中，利用公司提供的计算集群，对所学内容进行了简单的练习。  对于接下来的工作，个人主要承担底层集群系统之上的数据API接口开发和复杂功能研发、设计工作。基于在计算机学院海量数据研究中心的学习内容和研究经历，试图在该平台上设计出高效且实用的数据挖掘、分析、集成的相关功能。 | | | | | | | | |
| **毕业设计（论文）草拟题目：** | | 基于知识库的海量异构数据集成系统的设计与实现 | | | | | | |
| 毕业设计（论文）相关项目简要介绍（1000字以上）：  **1. 项目开发的目的和意义**  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。信息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。作为数据收集和分析的基础，数据集成在诸多领域具有重要的意义，包括数据清洗、模式识别、生物信息等等。  在现有的学术界和工业界，关系型数据库仍占有着无法忽视的地位。为了将异构、异源的关系型数据库进行集成，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合并。对于数据库集成，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得到预定的全局模式，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，我们希望能够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。  **2. 项目需求描述**  整个系统给定一个数据库模式集合Σ，其中每组模式s（来自于同一个数据表）包含多个属性a，希望通过探求各个属性之间的关系，生成一个全局的数据模式Sg，其中包含的属性能够涵盖Σ中所有的属性内容，同时Σ中相近的属性在Sg­合并成一条体现。  根据对已有数据的分析，属性名之间的相近关系可大体分为以下几类：   * 全等的属性名，即多个属性名完全相同 * 缩写，例如属性名num和number * 同义词，例如output和 productivity * 近义词，例如pie和puff可合并成属性dessert * 形近词，常常因录入过程的拼写错误导致   本系统希望能够在满足以上几种情形的属性名之间建立联系，将相近的属性归并成具有代表性一条属性，并将全局模式Sg中合并的属性与原有的每一条建立对应的联系。  此外，由于处理的目标数据量很大，整个系统包括设计的算法都将是基于外存的，如何高效的利用磁盘读取能力、降低磁盘读取次数和时间也是提升系统性能的一个挑战。  **3. 项目开发环境、工具、相关技术介绍**  目前阶段，本系统设计为在高性能的服务器上单机运行，通过C++实现相关数据结构和算法，并使用web前端对系统的结果进行简单的展示。  对于相似性判断，在算法实现中分为两个阶段。形近属性采用基于q-gram的编辑距离技术，通过使用倒排表提升处理效率。语义相近的属性我们采用知识库来判断之间的关系，目前可使用的有 Freebase、 WordNet、YAGO等等，将在具体的实现过程中进一步选择。由于数据库属性和知识库的数据量预计可达千万条级别或以上，本系统实现过程中仅将内存作为数据缓冲区，而所有的数据读取、写入都将在外存上进行，并采用hash、B树等数据结构予以优化并加速。 | | | | | | | | |
| **实习单位导师意见**（校外实习者可以用邮件截图，P2不需填写）    签字： 年 月 日 | | | | | | | | |
| **校内导师意见**（若导师出差可以用邮件截图，否则必须签字）    签字： 年 月 日 | | | | | | | | |
| **选题检查小组意见：**  **◎通过 ◎不通过**  签字：  年 月 日 | | | | | | | | |

**说明：**

1. 上述表格大小可以根据文字多少调整；
2. 若用邮件截图，则信息必须完整，可以放到表格中，也可以以单页附在后面。