|  |  |
| --- | --- |
| Inline Text Wrapping Picture | Inline Text Wrapping Picture |

硕士研究生学位论文阶段报告

学 号: 2015111624

姓 名: 赵成龙

学 院: 网络技术研究院

专业(领域): 信息与通信工程

研究方向: 通信网络服务与智能

导师姓名: 邹仕洪

北京邮电大学

2017年8月31日

|  |  |  |  |
| --- | --- | --- | --- |
| 论文题目 | 面向医院随诊应用的ETL工具研究和实现 | | |
| 论文类型 | 应用研究 | 选题来源 | 企、事业单位委托项目 |
| 开题日期 | 2016-12-14 | 是否开题题目 | 是 |
| 论文开始日期 | 2016-12-14 | 报告日期 |  |
| 报告地点 |  | 报告时间 |  |
| 研究内容简介  一、选题背景  数字化医院不仅是现代医院未来发展趋势，也是当前社会关注的热点，是对医院信息化建设的进一步深化。医院信息化建设是一项复杂而庞大的系统工程，还是一个整体和全面系统规划的过程，医院信息系统建设在医院的发展过程中越来越突显出其重要的作用。信息化系统需要运用计算机技术，对医院内各种信息进行采集、加工、分析和反馈。信息化建设已经成为加强医院现代化管理，走“优质，高效，低耗”发展道路的最有效途径。  由于医院信息化建设采取阶段性建设的方式和医院业务本身的特点给目前信息化建设推进带来了挑战。第一，医学数据用户术语种类多，不统一，信息系统元数据表示多样化，表示方式不唯一。目前部分病理名称或药品编码尚没有统一的编码标准，各个系统建立后存在术语不同的情况。存在编码不同，但是表述内容文字意义相同或相近的情况。第二，医院信息化应用仍处于操作层的事务数据处理，无法根据现有的数据预测未来的发展趋势，无法有效地利用信息和知识对诊断治疗进行指导和支持。大量数据相互之间的关系和规则等待去挖掘和发现。“信息孤岛”现象给现代信息化应用提出了挑战。第三，我国医院经过多年的信息化建设，很多都建立了自己的信息系统或电子病历系统。这些系统基本上是根据各医院当时的业务需求自主研发或引进的，由于缺乏信息标准和统一规划。目前，不同医院的系统自成体系，它们之间很难进行数据的交换和共享。所以在数据抽取应用过程中，如何完成信息整合，有效的进行元数据管理，提高数据质量是有待解决的问题。因此，如何将分散的独立的异构数据资源有效地整合到统一平台进行分析研究，实现数据共享；如何有效地将表达不一致的数据抽取转换整合到数据仓库中，保证数据分析的质量；如何从大量的现有数据资源中挖出有价值的知识，对医疗卫生事业发展的提供决策支持，显得越来越重要。  对于医院信息化建设过程中遇到的多源异构数据聚合、数据清理、数据处理等重要问题,可以采用数据仓库ETL工具的方式解决。ETL是数据仓库的主要部分，主要包含三个方面:数据的抽取、转换、加载。它在数据仓库和业务系统之间搭建了一座桥梁,使得业务数据能够进入到数据仓库中。本课题拟通过面向医院随诊数据仓库的ETL研究，改进传统的 ETL 工具不能很好满足医疗数据整合、处理和挖掘需求的问题，通过元数据管理，准确没有歧义的提取集成来自不同数据源的信息，从而提高 ETL 过程的数据质量。  二、研究内容  本课题以面向医院随诊数据仓库与数据挖掘分析为应用背景，主要研究ETL工具的设计，通过建立基于本体的元数据管理模块，从而更好的进行 ETL。设计随访数据仓库多维数据模型，为建立随访数据仓库分析系统提供了实际可行的方案。其主要内容包括三个部分：随访数据仓库ETL的研究、设计数据仓库 ETL 工具、ETL工具的实现和验证。  1、数据仓库ETL的研究  对数据仓库ETL的研究，包含三个部分：随访数据仓库的研究、随访数据仓库ETL工具的需求分析、ETL工具的研究。  数据仓库是为支持管理决策建立的、面向主题的、集成的、稳定的、随时间变化的数据集合。数据仓库主要由以下特征：面向主题的、集成的、非易失的、时变的。数据仓库系统从多种源数据经过抽取、转换、装载到数据仓库的存储部分,然后用户利用这些经过清洁处理,一致化处理的数据来建立各种应用。  要分析设计合理的ETL工具首先要充分理解医院信息系统中数据的含义和特点。对于医院信息系统的理解过程，主要针对面向医院随访门诊业务流程的理解和对信息系统数据表的理解。在充分的理解之后，更好的设计ETL工具的系统结构。  ETL是三个单词首字母的缩写(Extract、Transform、Load),也就是抽取、转换、装载。ETL是构建数据仓库的重要一环，用户从数据源抽取出所需的数据，经过数据清洗，最终按照预先定义好的数据仓库模型，将数据加载到数据仓库中去。  2、设计数据仓库ETL工具  对基于本体的 ETL 工具设计，包含基于本体的元数据管理设计和随访数据仓库ETL过程设计两部分。  ETL 过程按照统一的规则，首先抽取数据源中的数据，然后根据一定的转化规则转化数据，最后将规范转化后的数据装载到数据仓库中去。通常，一个 ETL系统中包含三个过程：数据抽取过程、数据转化过程、数据加载过程。  元数据是 ETL 项目组面对的一个非常重要的主题，对于整个数据仓库项目也是非常重要的一部分。元数据(metadata)即“关于数据的数据”，是关于数据信息的描述，包括数据的定位信息、数据语义描述、数据结构描述等。不同数据源提供元数据的能力不同。关于数据库的元数据记录的是表、视图等信息；而关于表的元数据记录的则是字段、主键、外键是否可为空等约束条件。一些文档里关联有元数据，描述性元数据说明文档外在含义的字段，如作者、日期等；语义级元数据描述文档的内容。当语义级元数据被标准化后，可用于计算机自动从数据中抽取语义。  本体作为一种共享概念的形式规格描述，明确定义了概念以及概念之间的关系，并通过框架逻辑(frame logics)或描述逻辑(description logics，DL)等形式化系统提供推理。通过描述和表示特定领域中的概念以及概念之间的关系，本体能够精确定义该领域中各个概念的语义关系，为领域知识的描述提供术语，从而确定某一领域的基本知识体系，表达该领域中的公共知识。基于本体所具有的良好概念层次结构和对逻辑推理的支持的优势，实现从语义层面解决数据之间的异构性。并利用本体的共享、重用、包含语义等特点很好的解决补充可以有效地集成数据源，解决各数据源数据集成过程中的语义异构问题。最后建立医学随访领域使用的本体进行 ETL 过程。  3、ETL工具的实现和验证  在系统设计的基础上，完成基于本体的元数据管理模块的实现，用其来指导医学数据 ETL 过程，并且实现了医学数据 ETL 模块，并通过实验验证系统的可行性和性能。实验主要从三个方面进行测试：一方面验证数据仓库ETL过程的可行性；另一方面，通过与传统ETL工具进行对比，分析系统的性能；最后通过对系统进行调优，来分析系统的性能。  三、关键技术  1、数据仓库技术  数据仓库 是用于保存从一个或多个数据源获取的信息的中央存储库。数据通常从事务系统和其他关系数据库传输到数据仓库中，而且一般包括结构化、半结构化和非结构化的数据。这些数据将会定期处理、转换和提取。用户包括数科学家、业务分析师和决策者，他们通过 BI 工具、 SQL 客户端和电子表格来访问数据。  2、ETL技术  ETL（extract提取、transform转换、load加载）。ETL负责将分散的、异构数据源中的数据如关系数据、平面数据文件等抽取到临时中间层后，进行清洗、转换、集成，最后加载到数据仓库或数据集市中，成为联机分析处理、数据挖掘提供决策支持的数据。ETL是数据抽取（Extract）、清洗（Cleaning）、转换（Transform）、装载（Load）的过程。是构建数据仓库的重要一环，用户从数据源抽取出所需的数据，经过数据清洗,最终按照预先定义好的数据仓库模型，将数据加载到数据仓库中去。  四、论文计划  根据之前的研究内容及对系统的概要设计和详细设计，计划11月完成医院数据仓库的ETL过程的代码实现，使系统能正确运行，达到预期效果。12月开始撰写毕业论文。  五、论文进度和目标  目前已经完成了系统的需求分析、概要设计和详细设计，初步编码实现了ETL工具实现框架并对网络操作和数据库缓存进行了接口封装，后面可能需要完善系统框架，然后再继续系统开发。本系统目标是完成面向医院随诊应用的ETL工具研究和实现。解决多源异构数据聚合和ETL过程设计与实现两个关键问题。数据仓库系统的架构大体可分为三个部分:数据存储及计算引擎、数据展示界面、数据的抽取转换(ETL)。ETL又包含三个方面:抽取、转换、加载,它在数据仓库和业务系统之间搭建了一座桥梁,使得业务数据能够进入到数据仓库中,为计算引擎及展示界面体统数据基础。由于源数据提供者的业务逻辑及系统环境不同,源数据的格式也是大相径庭,因此本课题要实现面向医院随访服务的数据仓库和ETL的解决方案。 | | | |

|  |
| --- |
| 论文进展情况  一、工作计划  2016.12 深入学习数据仓库技术，理解数据仓库体系结构的设计原理和联机事务处理 (OLTP)的实现过程  2017.1 学习ETL技术，理解ETL过程中如何实现多源异构数据的清洗、提取归一的实现过程  2017.2 学习数据库相关知识学习，学习并熟练掌握MySQL的使用  2017.3 学习现阶段医院中多用不同信息系统功能特点，如LIS、PACS等  2017.3 学习服务器接口协议，了解服务器接口格式，能与服务器进行交互  2017.4 系统基本需求调研，充分了解用户需求，进行需求分析  2017.5~6 进行系统概要设计和详细设计完成系统概要设计和详细设计  2017.7~9 系统开发实现，系统能运行，达到预期效果  2017.10~11 系统单元测试和集成测试，完成对系统的测试和修复  2017.12~2018.3 撰写毕业论文，完成毕设  二、实际进展情况：已完成对数据仓库主要技术、ETL工具核心技术的学习，掌握了现阶段主流的数据仓库和ETL实现工具的框架设计和工作原理。学习数据库知识和MySQL使用，能够基于数据库设计原理完成数据库表结构的设计，并能够熟练编写SQL语句。深入医院调研学习了，医院常用的几种信息系统：对医院信息系统（LIS）、医学影像存档与通讯系统（PACS）和电子病历（EMR）,对这些系统的结构设计和数据规范都有一定的了解。  根据医院现有信息系统的基础之上，深入分析数据仓库技术，并集合面向医院随诊服务的业务要求和特点，进行数据仓库系统的需求分析，确定数据仓库系统需要实现的主要功能。根据数据仓库的系统需求进行建模，并给出系统的总体架构和详细设计。根据整体的系统设计，已经完成数据提取、清洗脚本的开发，以及数据传输模块的开发。  实际的进展情况落后于开题时的工作计划，目前修改工作计划如下：11月完成系统开发实现及系统的测试和修复。12月开始撰写毕业论文。 |
| 工作成果  一、目前已完成的论文工作内容及取得的阶段性成果如下：  已完成了对医院数据仓库ETL的前期调研。数据仓库技术和ETL技术是本课题的核心技术，前期调研时对这两个技术进行了深入的学习和研究。深入分析数据仓库技术，并集合面向医院随诊服务的业务要求和特点，进行数据仓库系统的需求分析，确定数据仓库系统需要实现的主要功能。深入分析ETL技术，结合ETL过程相关思想和医院疾病数据自身特点，进行医院数据仓库中ETL过程的需求分析。  已完成了对医院数据仓库ETL的详细设计。根据数据仓库的系统需求进行建模，并给出系统的总体架构。最后给出了系统各个模块的详细设计过程。设计符合系统要求的ETL过程,提出了此ETL过程的设计方案。  已经完成医院数据仓库的ETL工具实现的框架设计和代码实现。根据医院信息系统特点和医院随访服务的实际需求，设计了完善了元数据标准，并且采用元数据方法，集中进行管理，保证接口、数据格式、传输有严格的规范；数据仓库的设计建模采用的雪花模型，以确保最大化地涵盖关键业务数据，把医院信息系统环境中杂乱无序的数据结构统一成为合理的、关联的、分析型的新结构，而ETL则会依照模型的定义去提取数据源，进行转换、清洗，并最终加载到目标数据仓库中。  二、本课题主要的创新点：本研究课题的可能创新之处在于多源异构医疗数据聚合及非结构化数据的提取和清洗。  （1）我国医院经过多年的信息化建设，很多都建立了自己的信息系统或电子病历系统。这些系统基本上是根据各医院当时的业务需求自主研发或引进的，由于缺乏信息标准和统一规划。而且信息系统元数据表示多样化，表示方式不唯一。目前，不同医院的系统自成体系，它们之间很难进行数据的交换和共享。所以在数据抽取应用过程中，如何完成信息整合，有效的进行元数据管理，提高数据质量是有待解决的问题。  聚合是指通过一定的规则,对多种异构的数据源提供统一的表示、存储和管理,.它能屏蔽多种异构数据源间的差异。数据聚合不是数据或数据载体的简单堆积,而是要实现数据的转换、数据源的统一、数据一致性的维护及异构环境下不同系统间数据的交换。  （2）通过对医院现有信息系统的了解，我们发现很多有价值的数据信息被存储在非结构化的文本信息中。例如出院小结中含有患者在院期间接受的治疗方案、服用的主要药物、主要体征参数和治疗效果等关键信息。由于其存在非结构化的文本数据中，在ETL提取、清洗、转化过程中就需要准确提取这些数据。这对于后续的患者随访、肿瘤临床专题研究和肿瘤临床流行病学研究有着巨大的价值。 |

|  |
| --- |
| 计划及进度安排  1、完成ETL工具的开发  工作内容：通过编码完成ETL工具的开发，实现预期功能及目标  预计完成时间：9月下旬  预计工作量：该部分工作主要是数据抽取模块、数据转换模块、数据装载模块开发以及实现ETL处理的自动化。在源数据、中间数据定义好以后,能够实现源数据的自动处理来生成系统的输出数据。  2、系统测试及修复  工作内容：进行单元测试和集成测试，并对问题进行修复，重复测试直至系统能正确运行。  预计完成时间：11月中旬  预计工作量：该部分工作主要在于对问题进行修复，修复过程中会出现某部分需要重新编码实现的情况。 |
| 问题及整改方案  论文后期的工作是继续进行系统开发、测试及测试修复，并撰写论文。  在ETL工具开发过程中，由于医院内有多种不同的信息系统，这些信息系统之间数据结构、信息维度、数据组织方式各不相同，为了能够把这些不同数据源的数据整合在一起，需要在数据仓库模型和元数据设计上进行良好的设计，并不断优化以获得更好的效果。同时由于医院内有多个信息系统，当一个患者来院时，会在多个信息系统中产生该患者数据，当把这些信息系统的数据整合时，会存在大量重复冗余的数据，对这些冗余数据进行整合时，需要进行数据去重，减小对数据存储资源的占用，进行去重时，要更好的设计去重规则，在保证数据完整性、一致性的基础上，尽量降低数据冗余度。 |

|  |
| --- |
| **参考文献**  [1] 耽美.电子病历建设研究[J].科技档案.2011,3(27).  [2] Wang Yong，Liu Like，Lu Hao.Thinking on the development of computer-based patient record system.West China Hospital of Si Chuan University, Chengdu610041.  [3] 张仁辉，王晓明.电子病历系统的设计与实现。《为计算机信息》.2009.3．  [4] Zhang Q，Matamoras Y，Takeda H. Making clinical data warehouse as an audit tool to detect wrong drug errors. Med info.2004,9:123-156  [5] Jose-Norberto, Juan Trujillo. An MDA approach for the development of data warehouses,Decision Support Systems, 2008, 45(1):41-58  [6] 袁淑丹.浅谈电子病历系统开发的意义.电脑知识与技术.2010.8.26.  [7] 陈金雄,刘雄飞,王庆森. 医院数据仓库的设计与实现.医疗卫生装备.2004,9(8):37-39  [8] 黄昊,柯新华,曾凡. 数据仓库技术在医院信息管理下的应用. 解放军医院管理杂志,  2002,(2):174-175  [9] William H. Inmon. Building the data warehouse [M], 2005 - Wiley India Pvt. Ltd  [10] 王克龙,王玲,王平立,宋斌. 数据仓库中 ETL 技术的探讨与实践.计算机应用与软件,  2005,11(11):30-31  [11] Ralph Kimball. The data warehouse lifecycle toolkit: expert methods for designing,developing, and deploying data warehouses[M].2004 |
| |  |  |  |  | | --- | --- | --- | --- | | 姓 名 | 职 称 | 职务 | 工 作 单 位 | |  |  |  |  | |  |  |  |  | | 邹仕洪 | 副教授 | 成员 | 北京邮电大学 |   **评审小组** |

|  |
| --- |
| **导师评语** |
| 导师：  日期： 年 月 日 |
| **阶段报告小组意见：** |
| 负责人：  日期： 年 月 日 |
| **学院意见：** |
| 负责人：  日期： 年 月 日 （签章） |