Exploration and Research on Archeology Information Technology

Tao ZHANG, Sihao FENG, Shuangyu ZHANG and Xiaohan LI

Department of Automation, Tsinghua University, Beijing 100084, China

考古信息化探索与研究

张涛，冯思豪，张双羽，李潇涵

清华大学自动化系，北京 100084，中国

Keywords: archeology, archeology information management system, data mining, database, B/S

Abstract

Nowadays, the integration of information technology and other subjects is quite significant. Archaeology, for example, is the study of human activity based on the recovery and analysis of the material culture and environmental data that have been left behind. Though Comprehensive data is gathered for the study, the massive data and the complicated relationships among them are far beyond the ability of manual work. A new method is needed by archaeology researchers, which not only can be used to record mass data, but also make rational use of the data to help the research. This paper mainly introduces the design and the implementation of the e-Arch archaeology information management system, which is a tool to assist researchers in the record, storage, searching, statistics, analysis and knowledge discovery of archaeological data. e-Arch archaeology information management system was developed under the cooperation of Tsinghua University and the Institute of Archaeology, Chinese Academy of Social Science. This paper also introduces the key issues and difficult points in the cooperative design and the implementation process.

摘要

信息化是指培养、发展以计算机为主的智能化工具，并使之造福于社会的过程。所以信息化的发展往往要涉及与其他学科的交叉融合。以研究古代社会的考古学为例，各种关于遗址、遗迹、遗物等的海量数据和复杂关联关系已经远远超过了个人的分析能力，考古研究学者们迫切需要一种新的方法用来记录保存大量的基础数据，并能够将这些数据合理利用起来，辅助其研究。本文详细介绍了考古学范畴内，以信息记录、存储、查询、统计分析和数据挖掘为目的的信息化工具的设计和实现方法，并以中国社会科学院考古所与清华大学共同设计开发的 e-Arch 考古信息管理系统为实际案例，介绍了其在设计和实现过程中的重点和难点。

关键字: 考古学, 考古信息管理系统, 数据挖掘, 数据库, B/S

1 综述

1.1 研究背景

考古学在了解古代信息，研究古代社会、文化等各个方面，重新认识历史经验教训、传承传统文化等方面意义和作用重大。最大限度的复原历史需要掌握最全面最细致的信息，而海量的信息既是考古研究的基础也是重大的挑战，时间紧迫的考古发掘、数量众多的遗物文物让考古工作者几乎无暇顾及聚落形态、文化结构和人类发展等历史观念[1]。为了发掘而发掘、沉迷于器物的分类排序等周而复始的研究道路，耗尽了许多人的宝贵时光。信息学的飞速发展，为满足考古对象和研究方法的变革提供了天然的优良工具。一种基于考古学本身的基本逻辑，能够分析各种情况下新的考古需求，利用新的信息技术，给出新时期考古学解决方案的信息系统，必将为探索考古学的发展方向产生积极影响，为其发展提供强大的动力。

之前的考古研究，更多地着眼于遗迹和遗物的技术和功能，或者通过遗物和遗迹将考古学文化作为单位来考察，较少研究遗迹和遗物之间的关系，即使研究他们之间的“总体关系”，也很少考察这种“关系”背后的文化含义和社会含义。而当前的考古研究也发展到了一个新的阶段。今天的考古研究已经从之前相对的“静态”变为“动态”，即考古研究将不止关注单体研究对象的单独意义，而是将其放在历史和当时社会的大背景下，来考虑研究对象的完整的考古学价值。比如希望通过与相同类型不同时期的研究对象的对比，发现历史发展的轨迹；希望通过同一时期不同类型、地域的研究对象之

间的联系，描绘当时的社会和文化[2]。考古学已经不是单纯的为历史学的发展提供证据和补充的学科了，而是具有人类学和社会学背景的一门综合性学科。在这种情况下，根据当前技术和考古学本身的发展趋势，将信息技术和数据挖掘技术引入到考古数据的处理之中，可以为实现上述目标创造条件。

在信息化方面，目前软件工程、WEB 技术、数据挖掘和数据仓库系统等各个领域都已经经过了很长时间的发展，在企业信息化、文献数据库、社交 web 服务等领域不乏大型综合性的信息系统，并且给人们的生活带来了巨大的变化。同样，在考古领域，综合性信息管理系统可以提供强大的信息获取、记录储存和统计分析、数据挖掘等功能，必然为考古信息记录和研究利用带来巨大影响和变革[3]。

1.2 研究内容

考古学是一门严谨细致的学科，同时具有与时俱进的研究方法。从挖掘探洞、采集探土的洛阳铲，到结合多媒体与虚拟现实技术的数字博物馆，科技的进步给考古学带来的不仅是劳动效率的提高与表现方法的革新，更是对研究者研究方法的改进和思维灵感的解放。目前通过信息化技术给考古研究带来的变革目前主要有三个方面：考古信息管理系统、文物保护与数字化展示系统、测绘与信息采集设备及软件。

考古信息管理系统是直接面向考古信息的，在三类系统中涉及考古领域工作流程范围最广，几乎涉及田野考古到发掘报告形成，考古研究和文物管理、收藏、展览的全过程；同时对硬件需求也是最少的，系统构成除了计算机之外几乎不需要其他额外的硬件设备。本文所述的考古信息系统设计与实现的项目研究成果 e-Arch 系统即属于此类。

e-Arch 考古信息管理系统是由清华大学与中国社会科学院考古研究所共同进行设计实现。主要解决了海量考古信息的储存、整理、统计分析和查询等问题，为考古学者的研究提供了有力工具。其设计实现过程也是信息学科与考古学交叉融合的典型案例，本文主要介绍了 e-Arch 考古信息系统的设计和实现方法，包括系统架构设计、数据库建模和实现、web 前端设计与实现、精确查询与全文检索以及非结构化数据检索的实现、数据挖掘的功能探索与研究等。

2 考古信息管理系统设计

2.1 系统需求分析

在系统设计和实现的过程中，需求分析是核心任务之一，决定着系统功能和方向；完善的需求分析会在系统设计和实施时减少很多不必要的资源浪费，大大提高系统实现效率。系统需求的辨识和确认也是不同专业背景的人员相互了解、达到信息共享和融合的过程，其主要实现方法是是通过访谈和调研，包括了会议、电话、电子邮件、小组讨论、演示等。最终形成《考古信息系统需求调研报告》，作为原始的用户需求提案，指导系统的设计和实施方向与范围。通过访谈、调研与分析，可以将考古信息系统的功能性需求归纳概括为：

- 针对田野考古发掘需要，可以提供科学、简便且稳定的数据记录和简单的信息统计分析、决策支持功能；
- 实现相对独立的信息记录和数据存储管理功能；针对考古研究需要，提供针对更大范围的已有电子数据资料的管理工具；
- 提供丰富、实用的数据浏览查询工具；
- 提供新颖的数据挖掘工具，帮助考古研究人员更好的认识数据及其数据背后的知识；
- 提供按照标准考古研究的要求输出固定格式的报告和信息表格；针对考古管理工作的特点，提供强大的中央信息汇总中心，并且可以关注区域的考古工作进展，为整体考古工作规划、长期考古发掘安排、协调不同区域工作进展和在更大范围内调配相关资源提供一个平台，更好的促进考古工作的科学管理和长远发展；
- 能够为对考古学和考古工作感兴趣的普通爱好者提供实时、权威和知识性较强的考古信息，和最新的考古研究成果和动态。

除以上功能性需求之外，还有如下非功能性需求：

- 系统必须满足多人在不同地点同时使用的需要；
- 提供详细的权限管理功能，不同的用户在系统中应可以设置不同的权限；
- 界面友好，操作简单方便，用户体验良好；
- 支持后续功能扩展等。

2.2 系统总体架构及相关技术介绍

在系统设计阶段，最大的困难在于没有类似的系统可以借鉴，所以需求并不完全清晰，研究的范围很容易扩散；且由于信息化专业人员对考古领域知识不了解，所以很难评估工作量和工作难点。所以在系统设计上采取了分层的设计，各层相对独立，对特定的一层的任何增添和扩展都比较容易实现，也便于与其他系统集成，并且不易造成混乱。

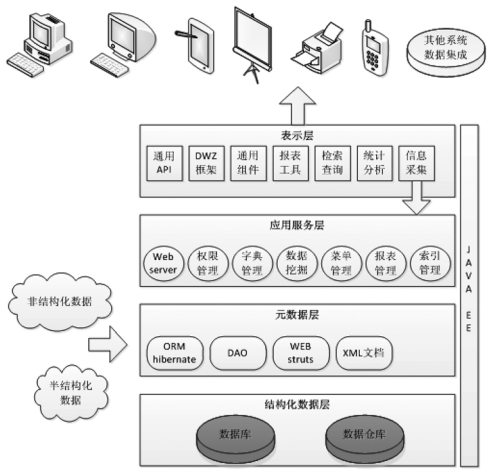

图1：考古信息系统整体架构

如上图 1 所示，e-Arch 考古信息管理系统共分为结构化数据层、元数据层、应用服务层和表示层。在持久化的数据层，数据库和数据仓库都采用 Oracle 数据库，不但性能优异而且数据库与数据仓库之间的数据处理几乎没有障碍[4]。

元数据层不光包含对数据的定义信息，在系统设计中，元数据层还主要包括将数据库中数据间的逻辑关系映射为前端可操作的对象的 pojo 类、数据访问接口、将前端的操作请求接收并请求，并根据该请求调用模型的业务逻辑方法进行处理，然后将处理结果返回给 jsp 页面显示的 action 类、丰富 jsp 的标签库、将非结构化数据进行索引建立得到的 XML 文档等。

应用服务层主要提供了数据的各种处理功能，与元数据层一起完成客户端与服务器端的数据交互。应用服务层因为与元数据和表现层都独立开来，所以对于新增功能和系统扩展来说，如果是数据的再利用，只需要在应用服务层增加应用即可。

表现层主要负责数据的录入和显示。e-Arch 考古信息管理系统采用 B/S 架构，可以在任何地方进行操作而不用安装任何专门的软件，客户端零维护使得系统的扩展、升级非常容易。同时，通过 ajax，javascript 等技术可大大减轻服务器负担，并优化用户体验。

2.3 数据库建模和实现

实现对考古信息的管理，首先需要收集归纳考古信息的各种属性特征，并理清各类型信息之间的逻辑关系。如果将纷杂的考古信息看作是细胞，那么分门别类的考古属性就是骨骼框架，而信息间错综复杂的关系则是神经脉络[5]。

对于考古信息的关系理解，通过不同的角度会得到不同的结论。考古学者非常熟悉考古信息的属性概念与相互之间的关系，所以往往会考虑的非常复杂全面，而忽略了抽象和简化；信息工程专家则几乎完全不了解考古概念所代表的含义和相互关系，却精于数据库设计和优化，掌握信息的组织和结构范式。只有学科间的融合，不同学科学生的交流和合作，才能完成考古信息属性概念逻辑关系的整理，并得到完善合理的考古数据编码。

如由考古学者最初整理的器物上文字与符号部分的数据编码为图 2 所示：

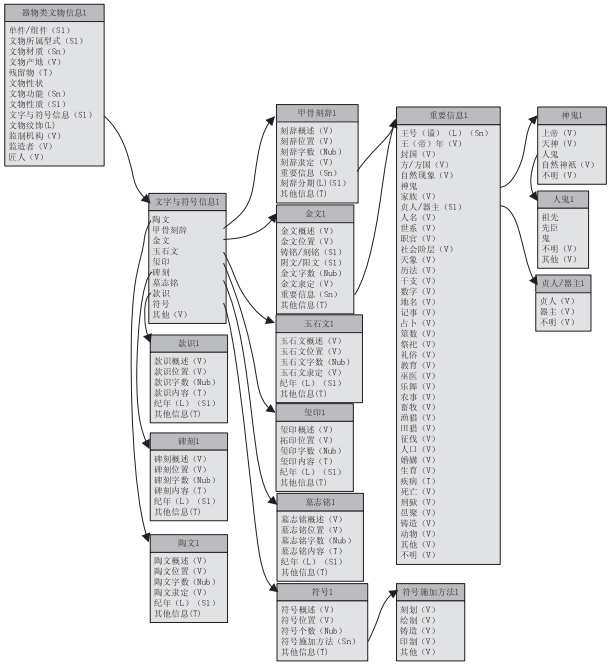

图2：器物上文字与符号部分的信息编码结构图

这样的信息编码是细致、正确、可行的，但是并不利于系统实现，依照此编码设计的系统，其信息结构复杂，数据库庞大，影响用户速度和操作简单的体验，且非常不利于信息的检索和统计分析。因为此信息编码结构复杂，分层过多，如：每种文字都有概述、位置、隶定、字数等信息，却未归纳其共性，而是分开为了不同的意义。其次，属性名称与属性值混淆，如：“刻辞概述”、“刻辞位置”等，属于属性名称，表示的是对象有此属性。而：“陶文”、“甲骨刻辞”等，属于文字种类的属性值；“刻画”、“绘制”、“印制”等，属于符号施加方法的属性值；“王号（谥）”、“王（帝）年”等，属于重要信息的属性值……。这些属性值并不需要在数据库结构中得到体现。或者说这里列出属性值的意义只是需要辅助用户进行数据记录。

所以，对图所设计的信息编码结构，可进行总结归纳，简化结构，以消除重复，如：

删除：“文字与符号信息”，此属性组合总称。此级无需在数据库结构中体现，可在界面上进行调整示意即可。

增加：“文字与符号信息种类”，其参考选项为：“陶文”、“甲骨刻辞”、“金文”等原第一级数据结构。

增加：“文字概述”替代原有各种文字概述。

增加：“文字位置”替代原有各种文字位置。

增加：“文字字数”替代原有各种文字字数。

增加：“隶定”替代原有各种隶定。

增加：“重要信息种类”替代原有各种重要信息种类。提供辅助输入，其参考选项为：“王号（谥）”、“王（帝）年”等。

增加：“重要信息”替代原有各种重要信息。

增加：“纪年”替代原有各种纪年。

增加：“其他信息”替代原有各种其他信息。

这样的设计可能会造成一些细部信息结构的简化，如在记录重要信息的“神鬼”重要信息时，无法在数据库结构上细分“上帝”、“神鬼”、“祖先”、“先臣”等属性（属性值），所以需要增加“重要信息备注”，用以记录。

依据数据库设计范式进行优化和修改后，器物上文字与符号部分的信息编码结构变更为：

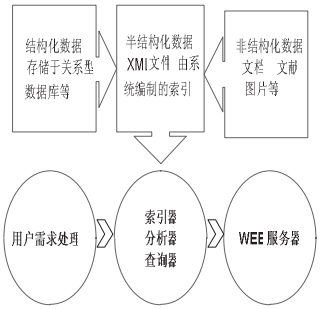

图 3：器物上文字与符号部分的信息编码结构图

依照此编码结构设计的系统界面如下：

图 4：器物上文字与符号部分系统界面

可以看到，内容集中在一个页面上，无需用户了解其背后的逻辑结构和过多操作即可看到信息的全貌，在后台数据库端，由原来的 10 个数据库表，减少到了 2 个数据库表(含通用的枚举值表)，大大降低了数据库的复杂度，并降低了信息的检索和统计分析难度。

在考古数据编码的整理收集以及数据库设计过程中，通过不同领域知识的融合，不同背景研究人员之间的相互学习、共同探讨和辛苦工作，目前已经建立了条理清晰、完整全面的考古信息网络，不但编码原则统一、信息属性完备，而且严格符合考古逻辑和考古习惯以及数据库设计范式。

通过对数据的编码规则定义和考古逻辑关系整理之后，就需要对这些编码和逻辑关系进行综合、归纳与抽象，把概念结构转换为数据模型，并不断调整优化，得到既符合数据库设计范式，又便于开发扩展的理想数据库模型。

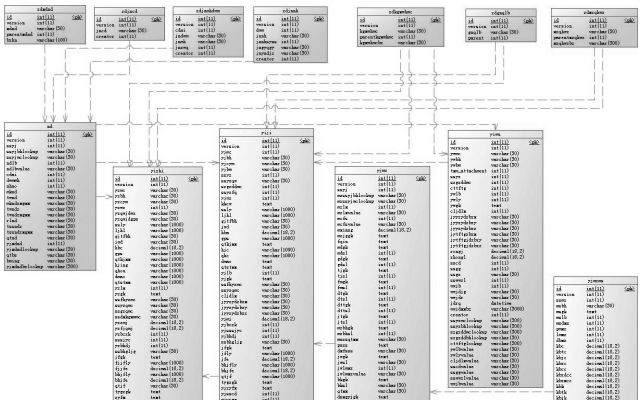

图 5： 数据库 E-R 图(部分)

上图为数据库的部分实体-联系图，图中表示了遗址、遗迹、遗物、墓葬及年代、行政区划等字典实体之间的关联关系。通过对考古概念逻辑关系的梳理和多次的迭代优化，最终完成了数据库的设计和建模，最终设计和创建了 187 个考古实体表以及 29 个系统基础表。

2.4 数据挖掘在考古领域的应用探索

数据挖掘是指找寻隐藏在数据中的有用信息，如趋势、特征及相关性的过程[6]。换言之，数据挖掘是一个知识发现系统，它能发掘特定模式的知识，比如规则、分类、关联等。常用的数据挖掘方法有分类预测方法、聚类方法和关联分析等。

分类预测方法是对数据进行有指导的学习，输入的训练数据中必须包含已知的类别信息，它们指导着模型的学习，使模型能够最大限度地揭示现有数据所包含的分类规律，并用这样的规律去对未知类别的数据进行预测，获得尽可能高的准确率。聚类方法是一种无指导的对实体进行分类的方法，训练样本中不包含分类信息，甚至不知道它们可以被分为多少类，而仅仅根据样本点之间的远近关系来分类。关联规则是另一种重要的数据挖掘方法，属于无指导型，能够探索事物的内在结构，解释大量复杂数据中隐含的关联特征。

考古学者的研究任务，是根据时代和社会环境产物——人类，来试图复原人类世界的形成过程。他们所要搜集的考古学资料是人类行为改变物质世界的结果，简而言之即包括人类行动的所有痕迹。在此基础上，考古学者寻找考古学记录，即被发现的有意义型式的共存关系，因而型式与共存是考古学最注重的“知识”。数据挖掘的常用方法非常适合用来挖掘这类知识。

所谓型式即为很多考古发现的某种共有的标准。一件单独发现的石器是完全没有意义的，除非它同发现于有意义状态中的石器非常相似，即从技术上同已知的型式一致。因而考古学同植物学、地质学一样，是一门分类科学。当一件考古发现横空出世的时候，考古人员首先想到的是能否将它归入已知的某种标准型式，使其具有考古价值（分类预测方法）。否则只能等到更多的相似发现到来后一起形成新的型式（聚类方法）。

共存关系是指考古学资料在一起发现，并表明它们同时被使用的状态，这样的组合具有很大的社会意义。

比如欧洲古代异教徒的墓葬，一个武士带着他的装备、徽记，桌上摆着食物和饮料，躺在用橡树干挖空的棺材里，上面覆盖土冢（坟丘），为我们展现的当时社会的一种风貌。因而如果我们找到一些共同被发现的考古资料之间的联系，就能够揭示和还原当时社会的一些场景与习俗（关联方法）。

3 考古信息管理系统实现与验证

在归纳分析了用户需求的基础上，目前 e-Arch 考古信息管理系统重点实现了以下功能：数据收集与展示、信息检索、统计分析与报表、以及固定规则的数据挖掘应用。这些也是作为考古信息管理系统，最为核心的功能应用。

3.1 信息检索

信息的收集、整理、记录的意义在于便捷高效的利用信息，而一个可以对大量信息进行检索并可生成统计报表的工具，无疑是使信息得到最大限度利用的最佳手段。对于数据库中符合考古逻辑的结构化数据，e-Arch 考古信息管理系统提供了精确的组合查询工具，可根据各种条件组合，精确查询到符合要求的结果。而对于未进行信息录入的非结构化考古数据，如 word 文档、excel 表格、pdf 文件、图、照片、拓片、影像等，e-Arch 系统同样提供了强大的索引和检索功能，可快速查询得到相关信息和文件。对于系统中的数据或查询得到的结果数据集，e-Arch 系统还提供了统计分析工具，可快速生成统计图或报表。

3.1.1 结构化数据的高级检索

结构化数据的高级检索是指对于第 3 章中介绍的数据库中存储的结构化数据进行的检索查询。由第 3 章介绍可知，由于考古信息的相关信息类别多，考古逻辑复杂，所以数据库的设计相当复杂庞大，这就为考古信息的组合精确检索带来了障碍。

对于结构化考古信息的查询主要用于两个方面，在信息编辑时快速定位到要编辑的信息；在统计分析时能组合过滤，得到符合条件的所有数据集以用于研究。所以，在系统中结构化数据的查询有两种表现形式：

信息浏览页面的数据查询快速定位：

图 6：List 页面的结构化数据查询

结构化数据标准组合查询页面：

图 7： 标准查询页面的的结构化数据查询

对于结构化考古信息的查询由如下的方法实现：

考古数据的特点是种类庞杂，记录详细全面。针系统中不同的数据类型，需要不同的查询方式，而对于不同的查询方式，不但其缺省的操作符不同，而且标签的属性有差异，页面渲染的效果不同，更重要的是查询结果也相应不同。为了为考古学者提供智能、精确的检索体验，e-Arch 系统的提供的查询方式有：

表 1: 结构化数据查询逻辑

数据类型	查询方式	缺省操作符	描述
字符型	遗址名称:	LIKE	对应数据库的 like 操作。
数值型	第几层:	=	
日期型	更新日期: 2012 三月 日 一 二 三 四 五 六 … 清空 确定	>= <=	提供两个文本框，以制定时间区间。
枚举型	遗迹保护等级: 全部 全部 全国重点文物保护单位 省级文物保护单位 市级文物保护单位 县级文物保护单位 区级文物保护单位 文物保护点 未定级	=	用枚举值的 id 进行准确比较，枚举值仅用于展示。
查询带回型	所属遗址编号: 所属遗址分区	LIKE	与枚举不同，并未对比 id，而是采用了 like 对比带回值，扩大了查询范围。

所有的查询项之间默认通过 AND 关系符进行连接，所以查出来的结果是组合条件的交集。[7]

查询结果以列表方式呈现，同时进行分页，提高了查询速度，同时优化了用户体验。

3.1.2 非结构化数据的索引和检索

信息的爆炸性增长使得其内容的存储和表现形式日益多样化，今天的信息已经不仅仅局限于文本，而越来越多的是文本、多媒体和元数据的混合。传统关系数据库管理系统处理的数据信息主要为结构化数据，往往是最为核心并经过精简的关键信息。而用纸质记载，电子记录，图片、报告、音频等形式记录的信息数据，即非结构化数据，在数据量上往往远大于结构化数据。[8]Forrester Research 的统计表明，非结构化“内容”量正在以每年 200 %的速度增长。在考古学界非结构化数据的增长更为明显，考古工作者动辄出版几千页的典籍，却只是个目录索引。如何有效的管理和利用这些非结构化的信息，已经成为学术界越来越关注的问题。

同样，仅在考古研究领域就已经存在并仍在不断产生着大量的非结构化的考古数据，其中包含但不限于：文档方面：发掘报告、记录表、出土物统计表、墓葬、房址、灰坑等遗迹登记表、考古文献等；图像资料方面：遗迹位置分布图、墓葬、房址、灰坑等遗迹平面图平面图、层位关系图、照片、拓片、影像资料等。大量非结构化数据蕴含着重要的信息，而对其进行结构化改造，不但费时费力几乎无法实现，而且过程中还有可能丢失大量细节信息。e-Arch 系统通过提供强大的数据索引和检索功能，为非结构化数据的研究和利用提供了便利工具。

在用户操作上，只需在检索页面输入想要检索的信息，之后点击搜索即可得到相关的结果。

图 8：智能检索功能界面

图 9：高级检索功能界面

在图 9 检索结果中可以看到：结果集中不仅有服务器上建立了索引的文件信息，而且还有数据库中结构化的数据，通过对结构化和非结构化数据的浏览对比，可对检索项的相关资料有大致了解。对于非结构化的数据，还可以点击下载，得到服务器上的原文件。

对于非结构化数据的检索在实现方法上，首先对非结构化数据和结构化数据一起编制索引，即生成 XML 文件，作为半结构化的数据 XML 具有强大的灵活性，不仅可以用来置标无结构的文本信息，而且还可以置标高度结构化的规则数据。利用 XML 文档结构索引的方法、特点和优势对设定的非结构化数据以及结构化数据进行建立索引之后，即可通过 XQuery XML 查询语言进行查询。[9]最后通过 WEB 服务器经查询结果进行展示。

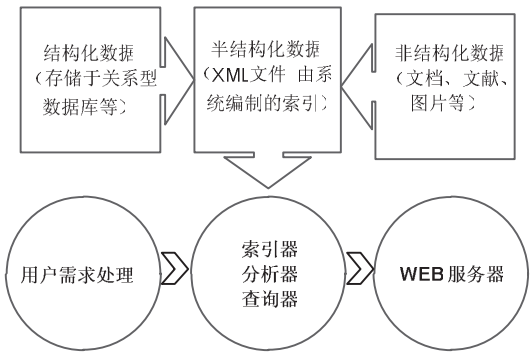

图 10：非结构化数据搜索模型原理图

e-Arch 系统与 R3 企业搜索系统进行集成，构建于 Solr 和 Lucene 之上，集成了 POI、PDFBox 和 Apache Tika 等第三方开源项目，建立了强大的数据搜索系统，通过 WEB 服务为系统中授权用户提供结构化数据和非结构化数据的搜索功能。目前以实现支持包括 TEXT、HTML、PDF、OFFICE 文档(Word/Excel/PPT 等)、iWork(Pages /Keynote)等 40 多种格式文件的存储、索引和检索。支持多媒体数据的存储管理。支持多语种、多编码管理。实现了高效的编制索引压缩，超低空间膨胀。编制索引速度可达到索引 10G 文档仅需 0.5 个小时。

3.2 统计分析与报表

在考古研究中，对信息的利用需求除了管理和查询之外，同样重要的是能够快速的根据已有规则制定报表或通过自定义规则进行统计分析。这也是 e-Arch 系统为考古研究者提供的对考古信息进行管理和利用的重要工具之一。

e-Arch 系统中的统计分析和报表工具分为定制表格和自定义规则统计分析两个部分。定制表格是指将系统中的数据按照国家文物局制定的标准进行报表输出，目前可输出的报表包括：墓葬登记表、灰坑登记表、房址登记表、遗址出土物统计总表、动物遗骸出土状况总表、遗址出土植物遗存统计总表、发掘资料记录登记表等考古工作常规记录表和档案管理表。

在 e-Arch 系统中，定制表格输出的实现方法是与 Finereport 系统集成，通过配置应用服务器解析设计好的 cpt 模板文件，从而将数据库中的数据查询出来按照

报表模板方式进行展现。可在展示报表 web 页面上通过 Flash 报表打印技术实现浏览器零客户端打印，同时也支持将报表下载为 PDF 文件或直接发送到电子信箱。

对考古数据自定义规则的统计分析类似于商业智能（BI）的查询统计报表，可由用户自定义组合查询条件对数据仓库中的信息进行筛选过滤、分组排序、报表输出和统计图绘制。

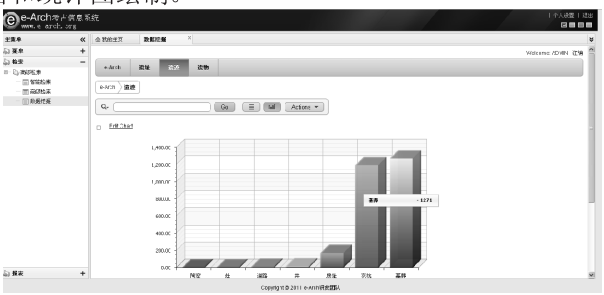

图 11：自定义统计图表

自定义规则的统计分析目前是与 oracle apex 集成实现的。对于 oracle 数据仓库中预处理后的考古数据用 oracle apex 进行统计和分析是比较合适和成熟的方案。

3.3 数据挖掘工具在考古信息处理上的应用方法与效果

3.3.1 商代古瓷片归类

我们对商代 5 个古瓷器厂遗址出土的 53 片瓷片的化学组成数据（含量百分比）进行数据挖掘，化学成分的字段包括铝（AL）、钡（BA）、铯（CE）等总共 19 项。我们假设相同地方出产的瓷器会有相似的化学组成，而不同窑口之间则不相同，因而可以根据某瓷片的化学组成推测判断其出产地[10]。使用 Clementine 12.0 软件的人工神经网络方法构建数据流来生成分类器。

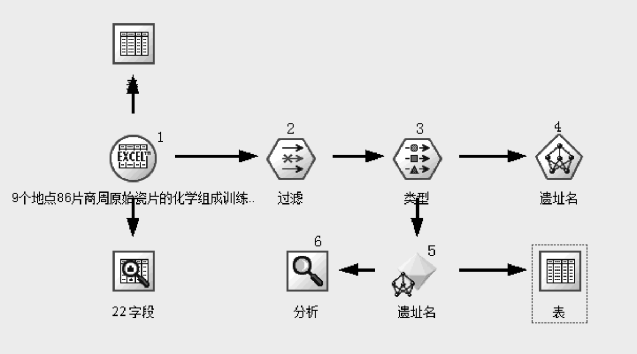

图 12：商代古瓷片归类模型

针对本方法构建的 Clementine 数据流，模块 1 读取 excel 文件，从而获得数据源；模块 2 过滤掉没有用的字段，本例可以将编号和先验组（与遗址名只需保留一个）过滤掉（也可以保留）；模块 3 是选择各字段代入人工神经网络模型的类型，本例中 19 个化学成分的字段是输入类型，遗址名是输出类型，如果在模块 2 中没有过滤掉编号与先验组，可以在此将它们设置为无类型。模块 4 是人工神经网络模型，双击可以对它进行设置，本例选择快速算法，并将停止的条件设为准确率 90%以上，为防止过度训练，在训练样本中随机抽取 50%的样本来训练模型，再用剩余样本集来计算模型误差；模块 5 是运行模块 4 后生成的人工神经网络分类器，将模块 4 的数据输入到模块 5，即可得到分类器对输入字段的分类结果，可以用表模块观察；模块 6 是对分类结果的分析，可以给出分类的正确率。

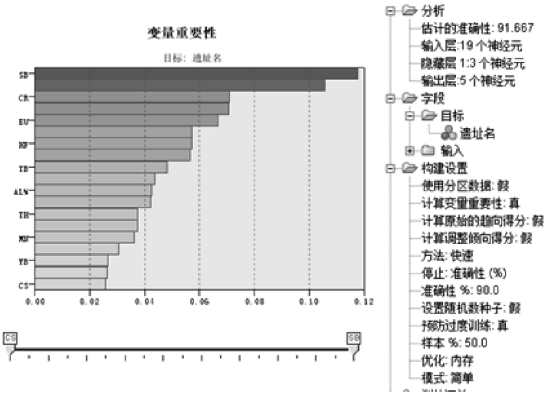

图 13：商代古瓷片归类数据挖掘参数设置

模块 4 的训练结果如图 2 所示：生成的神经网络是的三层神经网络模型，即有 19 个输入层神经元，3 个隐藏层神经元和 5 个输出层神经元，将样本点在有指导的情况下划分成了 5 类（因而是 5 个输出层）。训练的正确率达到 91.667%，意味着所生成的神经网络应用于训练数据（训练数据是从训练集中随机选出的 50%的数据，不用全部数据是为了防止过度训练）的时候，将会有 8%左右的错误率。

	遗址名	$N-遗址名			
1	吴城	吴城	27	黄梅山	黄梅山
2	吴城	吴城	28	黄梅山	黄梅山
3	吴城	吴城	29	牯牛山	牯牛山
4	吴城	吴城	30	牯牛山	牯牛山
5	吴城	吴城	31	牯牛山	牯牛山
6	吴城	吴城	32	牯牛山	吴城
7	吴城	吴城	33	牯牛山	牯牛山
8	吴城	吴城	34	牯牛山	牯牛山
9	吴城	吴城	35	苍圆垮	苍圆垮
10	吴城	吴城	36	苍圆垮	苍圆垮
11	吴城	吴城	37	苍圆垮	苍圆垮
12	吴城	吴城	38	苍圆垮	苍圆垮
13	吴城	吴城	39	苍圆垮	苍圆垮
14	吴城	吴城	40	苍圆垮	苍圆垮
15	吴城	博罗	41	苍圆垮	苍圆垮
16	吴城	吴城	42	苍圆垮	苍圆垮
17	吴城	吴城	43	博罗	博罗
18	吴城	吴城	44	博罗	博罗
19	吴城	吴城	45	博罗	博罗
20	吴城	吴城	46	博罗	博罗
21	黄梅山	黄梅山	47	博罗	博罗
22	黄梅山	黄梅山	48	博罗	博罗
23	黄梅山	黄梅山	49	博罗	博罗
24	黄梅山	黄梅山	50	博罗	博罗
25	黄梅山	黄梅山	51	博罗	博罗
26	黄梅山	黄梅山	52	博罗	博罗
			53	博罗	博罗

图 14：数据挖掘分类结果

模块 5 对训练集数据的分类情况如上图（定制表模块，使其仅显示需要对比的两个字段）所示，可以看到 53 个训练数据中仅有 2 个出现了分类错误。

3.3.2 殷墟墓葬青铜器随葬品关联分析

殷墟 80 年以后出土了一大批的青铜器，种类繁多，共有日用器、兵器和工具 39 种。考古学家将 1398 个墓葬出土随葬青铜器的情况做成了 Excel 统计表格。大多数墓葬中都或多或少地有一些青铜器随葬，但是出现的规律难以把握。另外整个样本达到 1398 个数据，规模已经超过了人工分析的极限，因而这样的数据在考古队仅仅作为一个记录表格存在，这是非常可惜的。

利用数据挖掘的关联分析方法，我们可以研究一下这些随葬器物间的内在关联关系，即它们在一个墓葬中怎样共同出现的规律。

使用 Clementine 12.0 的简单关联算法构建数据流来生成关联规则（图 3）

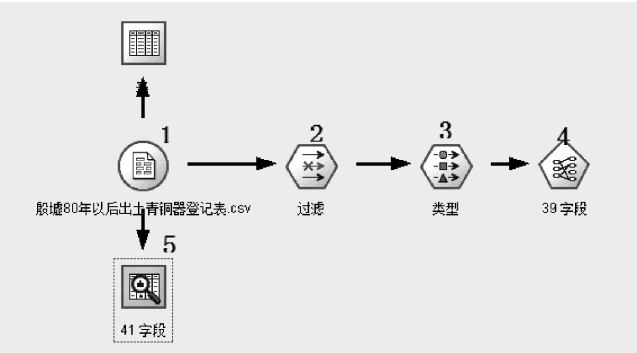

图 15：随葬品关联性数据挖掘模型

此问题的数据流建立较为简单，如上图所示，将模块 4—Apriori 算法模型的挖掘条件改为：最低支持度 10%，最小规则置信度 50%。最低支持度 10%可以保证每一条规则至少会有 140 个样本点印证，而置信度 50%的规则在考古学中已经是非常有意义的贡献规律了。

在此最低要求下，我们得到了 12 条规则，对其中的一些进行分析如下：

觚→爵[27.897,89.487]：当一个墓葬中有觚随葬时，同时也有爵的概率为 89.487 %。这个规则得到了 27.897%的样本点的支持。

爵→觚[26.18,95.355]：当一个墓葬中有爵随葬时，同时也有爵的概率为 95.355%。这个规则得到了 26.18%的样本点的支持。

通过这两条规则我们得出的结论是觚与爵作为随葬品经常成对出现。

我们的挖掘结果得到了经验丰富的考古学者的肯定，爵（图 4 左）在出土的青铜酒器中较常见，特别是商代，是古代饮酒必备之物。觚（图 4 右），形状像两个喇叭组合，中间细小，也是当时流行的一种酒器。觚与爵在商代一般是组合使用的，因而在作为随葬品的时候也成对下葬。可以说我们的数据挖掘方法与考古学、历史学的方法在这个问题上是殊途同归的。

图 16 商代酒器爵与爵

鼎→簋[12.16,50.588]：当一个墓葬中有鼎随葬时，同时也有爵的概率为 50.588%。而没有这个先验条件时，簋的出现概率仅为 7.22%，因而规则提升度达到了 7.002。

7.002 的提升度说明鼎的出现对簋显然有非常大的促进作用，尽管这条规则置信度并不是太高。但是说明了在商代墓葬中，常常用鼎与簋的组合来显示墓主的身份地位，如帝王九鼎八簋，诸侯七鼎六簋，因而两者常常同时出现在一些规格较高的墓葬中。我们将这条结果反馈给考古学者之后，得到了同样的肯定。

4 总结

4.1 设计与实现情况总结

e-Arch 考古信息管理系统项目是由中国社会科学院考古研究所与清华大学共同进行设计实现的，自 2009 年立项至今已经经过了 3 年的设计开发和运行。项目实现过程中，中国社科院考古所与清华大学的老师、学者和同学们反复的讨论、交流、沟通、演示，使各自擅长的知识领域得到了充分的共享，极大的促进了系统的快速优质实现和投入运行。

目前，e-Arch 考古信息管理系统已实现的功能有：

信息的记录和存储功能。为考古发掘人员提供了便捷标准的信息记录工具。其逻辑结构清晰严谨、具有完善标准的数据字典、操作简单方便。采用 B/S 结构设计，兼容各种移动终端，使数据记录快捷、高效、准确。

考古信息浏览。针对系统设计的记录标准全面，字段丰富；而具体考古信息又相对稀疏的情况，设计实现了有针对性的信息浏览功能，结合检索功能，使用户能够在海量数据中迅速得到想要的信息。

数据检索。提供了快速精准的检索功能，不但使用户可以通过多个条件进行组合查询，而且还能够全文检索文档、图片、资料等非结构化数据。使考古数据在考古研究中能够并易于得到充分的利用。

统计分析。在系统内提供了简单方便而又功能强大的统计分析工具，使用户可以方便的对考古信息进行统计、分析等数据操作，其分析结果可以保存输出。结合数据挖掘规则，使系统成为了强大的考古研究辅助工具。为考古研究者提供了极大的便利和丰富的研究思路

报表输出。对于系统内信息的常规利用，提供了一键生成报表功能。目前支持的考古工作报表有：遗址出土物统计总表、遗址出土植物遗存统计总表、含墓葬发掘记录表、灰坑发掘记录表在内的各类遗迹发掘记录表、各类遗迹调查记录表、各类遗迹钻探记录表、各类遗迹常规记录表、遗迹统计表、动物遗骸出土状况总表等。通过报表输出，可以大大减少信息的重复记录和整理工作，并极大的提高了工作效率。

权限及系统管理。作为成熟完善的信息管理系统，e-Arch 还提供了严格细致的权限管理，权限设置粒度最小为针对单个用户的浏览、编辑、删除和特殊操作。同时在系统中海能够详细记录每一个用户的访问和操作记录，为系统稳定良好运行提供了保证。

目前，e-Arch 考古信息管理系统已经在河南安阳考古队实验运行，并在殷墟布局探索与研究中发挥了重要作用。拟进一步在全国考古工作站推广使用。同时软件注册权申请工作也在进行之中。

4.2 需进一步开展的工作介绍

考古信息化在考古研究领域仍然属于在时代新环境下产生的新事物，随着对考古领域的不断深入了解，项目进展也越来越迅速和顺利并小有成效，可是在项目初期，与其他其他新的研究领域一样，项目进程也经历过很多次反复、重构、需求和设计的重大变更及技术实现路线的更改，在 e-Arch 考古信息系统的设计和实现过程中，有成功的部分，也有不足的地方，他们都是宝贵的经验。目前项目运行中出现的问题和不足仍然存在，有待完善；研究进一步深入的方向和课题也仍同样丰富，并充满诱惑。

可以进一步开展的研究主要方向有：

GIS 在考古系统中的应用。考古信息中位置信息毫无疑问是至关重要的，目前 e-Arch 系统中提供了对于位置信息的详细记录功能，如遗址位置、遗迹位置、探方探沟位置、工作地点、工作地点编号、遗物出土位置、层位关系图示等，但目前还仅限于文字和图片描述，要达到考古信息在地图上的展示、查询、对比和浏览，还需要与 GIS 系统的集成和整合。目前 GIS 系统的相关功能正在由清华大学建筑学院的老师和同学设计开发进行中，进一步的整合工作将在下一个阶段共同完成。

便携式移动终端的系统移植和应用。移动终端以不受使用时间、环境限制，获取信息的同时能够显著提高分享与沟通效率等优势，迅速达到了一个相当大的不容忽视的规模[11]。目前 e-Arch 系统在移动终端仍然以 web 服务的方式提供支持，且所有控件和功能兼容几乎所有的移动设备和系统，但仅仅这样是不够的，e-Arch 系统目前尚未为移动设备或系统开发适应移动终端操作习惯或可充分发挥移动终端操作优势的功能或 app，对于 e-Arch 系统，在移动终端上仍然具有可以有大量可以研究和开发的提升空间，如除前述的桌面 web 服务和功能之外，还可以利用移动终端的定位功能、拍照功能做信息获取和数据采集，或有针对性的数据推送和获取等等。

工作流。田野考古工作存在一个从考古调查、行政审批、探方发掘、资料记录与整理，直至发掘报告发表与后续研究的一个完整的工作流程[12]。田野考古地理系统的研究与开发，需要具备工作流机制，以体现这种工作流程。目前 e-Arch 系统的工作流程还比较简单直接，包括了信息的记录、审核、生成考古发掘报告及后续的信息的统计分析。由于考虑到行政审批环节超过考古信息系统管理范围之外，而调查钻探发掘过程中，目前尚不具备使信息记录与发掘过程达到同步的条件，所以，完全符合考古发掘流程的工作流机制尚未在系统中建立。随着考古工作站和文物管理部门及田野工作人员的信息化应用水平不断提高，部门间信息系统集成度会得到增强，工作流程标准化程度提高，e-Arch 系统中的工作流机制也需要得到进一步提升。

参考文献

[1] 毛延辉. 应用于考古数据处理的决策树算法研究[J]. 2009.《殷都学刊》; 2010.

[2] 霍巍. 评欧美“新考古学派”——兼论我国史前考古学传统模式的变革[J]. 《四川文物》, 1992 年 01 期.

[3] Abajian, V. E-Archaeology+: An integrated expert system dedicated to Archaeology[C]. Information and Communication Technologies: From Theory to Applications, ICTTA 3rd International Conference on, Damascus, Apr. 7-11, 2008.

[4] 何明, 何茜颖. Oracle 快速 Web 应用开发[M]. 北京：清华大学出版社, 2010.

[5] 张鹏程.关于建立文物考古数据库的几个问题 考古与文物 2008.2

[6] 张兴会等. 数据仓库与数据挖掘技术[M]. 北京：清华大学出版社, 2011.

[7] 张昊, 冯思豪等. J-Hi 操作手册 http://www.j-hi.net/courses!ke_cheng.action?project.id=1

[8] 王金生.处理非结构化信息的桌面搜索模式的发展和竞争现状.现代情报.2005.9 第 9 期

[9] 夏立新等. XML 文档全文检索的理论与方法.2011

[10] 陈铁梅. 定量考古学[M]. 北京：北京大学出版社, 2005.

[11] 中国互联网络信息中心(CNNIC).《中国移动互联网发展状况调查报告》

[12] 孙懿清,黄家柱等.工作流机制在田野考古地理信息系统中的应用。