

瀚海古籍全文检索平台

目录

1.	概述.....	4
1.1.	项目背景.....	4
1.2.	项目目标.....	4
2.	总体设计.....	4
2.1.	先进性.....	4
2.2.	安全性.....	4
2.3.	性能要求.....	5
2.4.	设计原则.....	5
2.5.	技术指标.....	6
2.6.	技术路线.....	6
3.	功能设计.....	7
3.1.	网络拓扑图.....	7
3.2.	模块架构图.....	7
3.3.	功能模块.....	8
3.3.1.	古籍数字内容管理.....	8
3.3.1.1.	古籍资源管理.....	8
3.3.1.1.1.	古籍分类管理.....	8
3.3.1.1.2.	古籍内容入库.....	8
3.3.1.1.3.	古籍内容维护.....	8
3.3.1.2.	应用机构管理.....	8
3.3.1.2.1.	机构客户管理.....	8
3.3.1.2.2.	机构资源权限.....	8
3.3.1.2.3.	机构 IP 管理.....	8
3.3.1.2.4.	机构外网用户.....	9
3.3.1.2.5.	注册用户管理.....	9
3.3.1.2.6.	机构区域管理.....	9
3.3.1.3.	应用系统管理.....	9
3.3.1.3.1.	角色管理.....	9
3.3.1.3.2.	用户管理.....	9
3.3.1.3.3.	权限管理.....	9
3.3.1.3.4.	参数管理.....	10
3.3.1.3.5.	热词管理.....	10
3.3.1.3.6.	日志管理.....	10
3.3.2.	古籍资源应用管理服务.....	10
3.3.2.1.	古籍资源服务管理.....	10
3.3.2.1.1.	应用栏目管理.....	10
3.3.2.1.2.	发布信息管理.....	10
3.3.2.1.3.	资源发布管理.....	10
3.3.2.1.4.	统计分析.....	11
3.3.2.2.	古籍资源应用服务.....	11
3.3.2.2.1.	信息服务.....	11
3.3.2.2.2.	全文检索.....	11

	3.3.2.2.3. 在线阅读.....	12
	3.3.2.2.4. 内容跟踪.....	12
	3.3.2.2.5. 内容收藏.....	12
	3.3.2.2.6. 热词记录.....	12
	3.3.2.2.7. 个人中心.....	12
4.	软硬件要求.....	13
5.	开发周期.....	14

1. 概述

1.1. 项目背景

项目中的数字内容都是比较珍贵的收藏价值、研究价值比较高的古籍图书，为了更好的展示古籍图书的内容，发挥图书最大价值，我们采用高质量扫描识别，形成可以在线检索、在线阅读的数字内容，根据资源类型和应用模式形成古籍主题数据产品，为机构提供科研学习服务。

1.2. 项目目标

- (1) 建立古籍图书管理系统，根据制定的标准对古籍资源进行批量入库；
- (2) 按照古籍图书的分类形成不同版本的估计专题产品；
- (3) 提供古籍图书的在线检索，双文对照阅读服务；
- (4) 提供资源及用户使用的统计分析；

2. 总体设计

2.1. 先进性

采用先进、成熟、实用的技术，既要实现出版平台的功能，又要确保在未来几年内其技术仍能满足应用发展的需求。

2.2. 安全性

保证出版平台和数据的高安全性，使整个系统在受到有意、无意的非法侵入时，被破坏的可能达到最小程度。制定出合理的安全策略，以有效保障系统安全。系统安全设计的主要目标和任务包括：

- (1) 禁止非法用户进入系统。非法用户是指那些没有通过认证的用户，他们没有合法的用户ID和密码，要进入系统必须通过非法手段。系统在安全设计时要重点防范这些非法用户的入侵；
- (2) 未授权用户不能越权使用系统和资源。未授权用户是指不具有对数据或资源进行某项操作的用户；
- (3) 数据加密机制，对于产品化的数据在网络传输和镜像分离时都是进行严格512位加密机制进行加密，并采用不同镜像服务器的唯一MARC地址作为加密

密钥，保证产品数据的安全。

- (4) 系统要保证全天候不间断运行，各种服务不能停止。非灾难原因，系统要保证7*24小时不间断运行，各种系统服务不能停止或服务质量不能下降。对于出现的灾难，系统具有符合国家规定的相应灾难处理能力，及时恢复运行；
- (5) 保证业务连续性，利于今后的扩展，避免重复建设。

2.3. 性能要求

- (1) 响应性能：应具备快速的系统响应能力，响应速度在用户心理所能忍受的范围内；响应速度不致影响业务工作、造成业务工作的低效率。
- (2) 可扩展性：在设计上必须具有适应业务变化的能力，当系统新增业务功能或现有业务功能改变时，应尽可能的保证业务变化造成的影响局部化，新增功能时不应需要改造原软件系统。
- (3) 易用性：所有的业务功能界面风格和操作流程一致，应具有良好的中文操作界面、详细的帮助信息，系统参数的维护与管理通过操作界面完成。
- (4) 可靠性：应保证在正常情况下和极端情况下业务逻辑的正确性，并能保证系统可靠连续7×24小时的无故障运行。
- (5) 准确性：保证系统数据处理的准确性，提供核查、审计手段。
- (6) 开放性：系统应采用主流的、开放的技术，以保证系统对各种数据业务的服务，以及与相关系统的互连能力。
- (7) 实时性：实时完成大容量数据处理，对业务提供并发处理支持。

2.4. 设计原则

- (1) 资源与应用分离。本项目的资源要能与应用进行分离。
- (2) 项目可以按自身设计的标准进行存储数据资源。
- (3) 内容资源通过分类进行关联，以提供更人性化的服务。
- (5) 结合目前市场的情况，项目数据库产品形态可以自由制作。

2.5. 技术指标

选型关键性能指标主要针对其数据简单查询和复杂查询的响应时间：

事务类型	最大响应时间	百分比	响应时间的定义
数据简单查询	2 秒	95 %	从用户按下“Enter”键开始，到显示下一个页面的第一个汉字之间的时间间隔
数据复杂查询	5 秒	95 %	从用户按下“Enter”键开始，到显示下一个页面的第一个汉字之间的时间间隔

系统在稳定性和数据管理能力等方面需求如下：

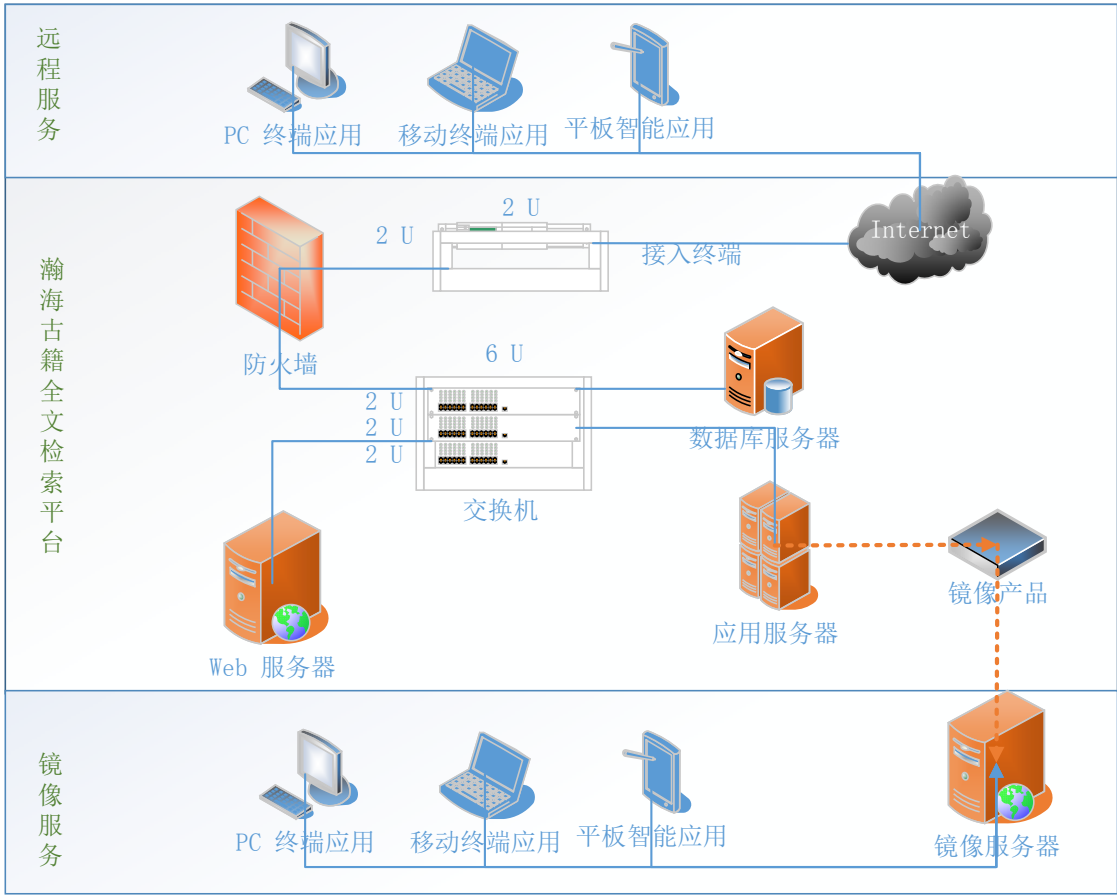
事务类型	性能要求	描述
稳定性	能够 7×24 小时连续稳定工作，并具备故障无缝迁移能力。	7×24 小时指系统具有长期稳定运行的性能，故障无缝迁移能力反映了软件在底层系统现故障时的可靠性。
数据管理能力	能够有效存储与快速索引海量网络数据，数据量达到 TB 级。	系统具有超大的数据管理能力
中文文本切词速度	全切分达到 1MB/S，最大匹配达到 5MB/S	高切词速度是保障中文内容分析实时性的必要前提。
中文文本切词准确率	达到 95%以上。	高切词准确率是保障中文内容分析准确性的必要前提。
全文检索速度	达到 500GB/S 以上的数量级。	全文检索速度定义为：系统每秒钟扫描的全文数据的速度。

2.6. 技术路线

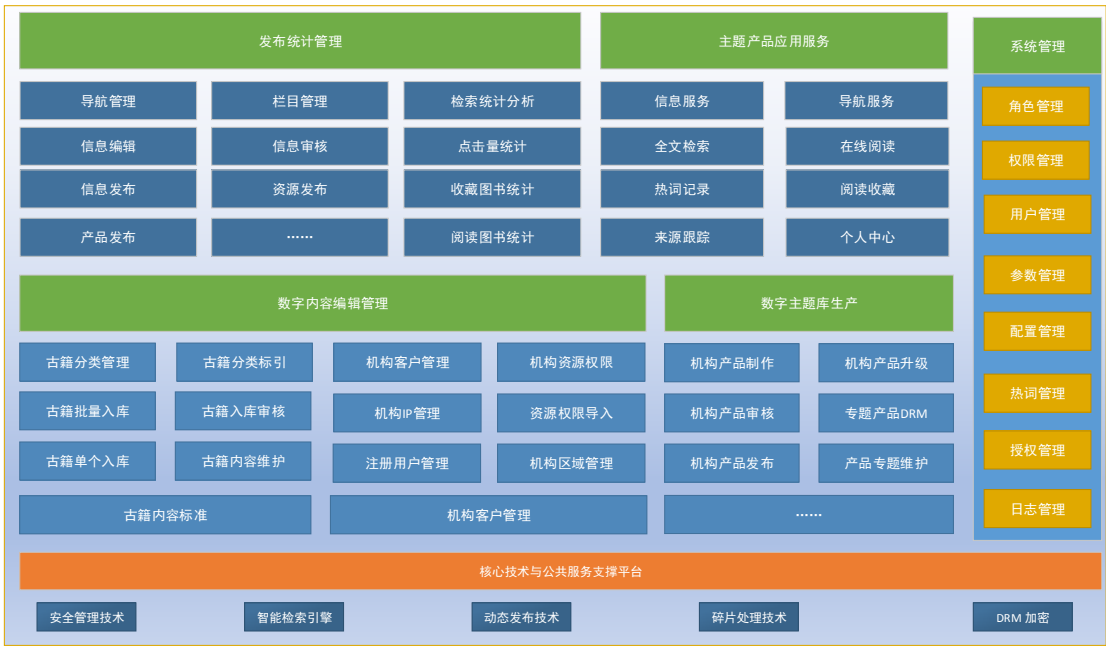
方案采用分层的 J2EE 架构设计，保证系统扩展性和运行效率，数据库支持跨平台。

3. 功能设计

3.1. 网络拓扑图



3.2. 模块架构图



3.3. 功能模块

3.3.1. 古籍数字内容管理

3.3.1.1. 古籍资源管理

3.3.1.1.1. 古籍分类管理

建立古籍的内容的分类结构，用来对古籍内容的分类进行管理。提供古籍分类结构的增加、修改、删除、关联、导入、导出等维护管理功能。

3.3.1.1.2. 古籍内容入库

系统采用磁盘监控的方式对不同的数据资源进行自动入库，建立不同的入库资源的模型，存放到指定的磁盘目录下，系统会自动扫描监控目录，根据模型标准进行资源的匹配入库。

3.3.1.1.3. 古籍内容维护

支持对所有入库的古籍资源进行统一管理，可以对于入库的资源可以根据资源分类按不同的条件进行内容的检索，对检索出来的资源在权限允许的范围内可以进行资源分类的重组、内容的分类标引、或资源内容的删除等操作。

3.3.1.2. 应用机构管理

3.3.1.2.1. 机构客户管理

机构客户管理主要是对客户的基础信息管理进行管理，机构客户分为远程客户和镜像客户，远程客户可以设置客户的访问期限，并关联机构资源权限对远程客户可以访问的资源进行权限设置。为了方便对机构用户的管理，增加了机构客户相关附件上传功能，例如合同扫描件等。

3.3.1.2.2. 机构资源权限

对远程应用的机构客户，可以根据服务费用或选择的资源内容进行资源使用的权限分配，资源权限的分配可以按主分类、子分类、具体资源本身进行操作，方便资源权限的灵活设置。

3.3.1.2.3. 机构 IP 管理

对于机构用户可以设置他们的 IP 使用范围，根据访问的 IP 地址来确定用户

来源，根据用户来源进行资源使用权限的分类，管理用户可以设置不同的 IP 段，也可以对机构用户的 IP 段进行删除或修改。

3.3.1.2.4. 机构外网用户

对不在机构 IP 范围内容的用户，可以采用外网用户的方式来登录系统，登录后系统会自动归类到对应的机构，拥有和机构用户相同的操作权限，管理用户可以对外网用户进行添加、删除和修改的操作。

3.3.1.2.5. 注册用户管理

注册用户主要是指自由注册的使用用户，注册用户登录后可以收藏阅读的估计内容，系统也会自动记录每次阅读的内容，采用注册用户登录个人中心可以看到自己收藏的和阅读的古籍内容。

3.3.1.2.6. 机构区域管理

按照中国省市级划分形成树状区域机构，在建立机构用户时直接和区域机构数进行绑定，方便后期根据区域机构数对用户进行统计分析，管理员可以自由维护区域机构树，对区域结构树的节点进行添加、删除和修改操作。

3.3.1.3. 应用系统管理

3.3.1.3.1. 角色管理

对系统后台管理员划分不同的角色类型，可以分为超级用户、机构管理用户等，机构管理用户可以直接关联到对应的机构，在登录时候自动按机构权限进行资源和功能的展示

3.3.1.3.2. 用户管理

后台系统的管理人员，超级用户可以添加不同角色下的用户信息，并为他们分配不同的权限。

3.3.1.3.3. 权限管理

对用户角色进行权限设置，可以控制不同角色的用户拥有不同的操作权限和操作界面。

3.3.1.3.4. 参数管理

为了增加系统的灵活性，根据系统和业务的特点抽象出一些功能的模型形成配置项或数据字典，要对系统进行修改时候用户可以直接通过参数进行调整，方便系统后期的维护。

3.3.1.3.5. 热词管理

对前台用户输入的检索关键字进行管理，对不与健康的检索热词可以进行删除，修改等操作。

3.3.1.3.6. 日志管理

日志要能记录每位用户在平台的登录、检索、浏览、等的使用情况，实现对用户操作日志的检索、统计功能，日志同时记录用户操作的成功和失败，要保证日志不可篡改，以及数据的完备性和安全性。

3.3.2. 古籍资源应用管理服务

3.3.2.1. 古籍资源服务管理

3.3.2.1.1. 应用栏目管理

对前台应用中不同的展示栏目进行动态配置，可以动态的添加门户中的栏目，对栏目的内容进行静态或动态的发布，是门户页面展示起来更加灵活方便。

3.3.2.1.2. 发布信息管理

数字资源应用服务不但有资源应用还包含信息服务，为了为用户提供更好的服务，系统集成新闻内容发布模块，可以制定不同新闻栏目，发布对应的新闻内容，为前应用服务提供强大的发布功能。

3.3.2.1.3. 资源发布管理

资源发布是对系统采集整合的全媒体数字内容进行在线发布，根据资源分类和资源形态的不同，可以创建不同的资源栏目，同系统中选择适合栏目的资源内容进行动态发布，也可以根据资源的使用量、资源的入库时间、资源的喜爱程度进行资源内容的自动发布。

3.3.2.1.4. 统计分析

统计分析分为超级用户统计和机构用户统计，系统根据用户角色自动判断用户拥有的统计权限，超级用户可以统计所有机构用户的使用数据，机构用户只能统计机构内容的使用数据信息。统计分为五种类型：

- 点击量统计：

在一定的条件内容，按机构、作者、分类对所有使用的数据进行统计，形成列表和决策图，并能导出 Excel、PDF、JPG 格式的文件下载；

- 阅读量统计：

在一定的条件内容，按机构、作者、分类对所有阅读的数据进行统计，形成列表和决策图，并能导出 Excel、PDF、JPG 格式的文件下载；

- 收藏量统计

在一定的条件内容，按机构、作者、分类对所有使用的数据进行统计，形成列表和决策图，并能导出 Excel、PDF、JPG 格式的文件下载；

- 关键词统计

在一定的条件内容，按机构、按词频对所有检索的关键字的数据进行统计，形成列表和决策图，并能导出 Excel、PDF、JPG 格式的文件下载；

3.3.2.2. 古籍资源应用服务

3.3.2.2.1. 信息服务

根据后台的特色服务，展示不同的信息栏目内容，提供信息栏目的在线阅读。

3.3.2.2.2. 全文检索

3.3.2.2.2.1. 快速检索

提供类似 Google 等快速检索模式，用户直接输入关键内容，系统对关键内容进行智能分词然后采用全文索引机制进行相关内容的匹配，以列表的方式进行结果内容的展示。

3.3.2.2.2.2. 高级检索

根据后台配置可以自定义检索字段，采用与或非的多重逻辑条件对不同的属

性字段进行综合全文检索检索。检索方式是采用对属性内容进行智能分词然后采用全文索引机制进行相关内容的匹配，以列表的方式进行结果内容的展示。

3.3.2.2.2.3. 二次检索

对检索的结果列表进行二次过滤，在检索列表中重新按制定的条件进行过滤，以便更加精确的定位所需要的资源内容。

3.3.2.2.2.4. 列表排序

对检索出来的列表内容可以按不同的字段进行排序，排序的选择字段可以有后台进行灵活配置，根据具体资源具体对象进行配置。

3.3.2.2.3. 在线阅读

通过导航或检索功能进入图书列表页面，在列表页面直接点击对应的图书就可以打开图书进行在线阅读了，图书的阅读分为两种模式，在图书阅读页面左侧为图书的板式页面，右面为图书的流式阅读页面，翻页时左右内容自动关联，并且在右侧的流式页面中可以进行笔记添加操作。

3.3.2.2.4. 内容跟踪

通过检索或导航阅读的每一篇文档都可以映射到对应的出处，来源的图书、图书的分类等信息。

3.3.2.2.5. 内容收藏

对阅读的图书内容进行收藏，收藏后阅读的图书会进入个人中心，以后可以在个人中心对收藏的图书内容进行阅读。

3.3.2.2.6. 热词记录

对读者用户检索的关键字以及关键字的词频进行管理，对于在前台高频显示的关键词进行审核，也可以对不健康的关键词进行删除和修改等操作。

3.3.2.2.7. 个人中心

进入用户可以收藏检索的数字内容，对阅读的图书、听看的音视频进行标签，建立自己的收藏或上传分类，把本地的资源上传到个人中心，方便以后的应用与管理，上传个人资源在经过审核后可以作为系统资源为其他用户提供服务，让固定的数字资源随着用户应用而丰富起来。

1.1.1.1.1. 个人信息修改

对个人的账户密码、安全信息、个人头像、基础信息等进行修改。

1.1.1.1.2. 内容分类管理

对个人中心的各类资源进行分类管理，内容分类采用树状结构不限层级，可以根据自己的需求进行资源分类的无限制创建，对创建的分类可以修改、删除等操作。

1.1.1.1.3. 个人收藏管理

对自己收藏的数字资源进行管理，可以按不同的条件对收藏的资源进行查询、标引、删除、查看等操作。

1.1.1.1.4. 笔记标签管理

在阅读过程中，用户可以对流失阅读中的内容添加笔记，在个人中对添加的笔记进行修改，关联查看等操作。

1.1.1.1.5. 阅读历史管理

个人中自动记录登录用户阅读过的网络资源，可以通过阅读历史快速的跳转到上次阅读的位置，也可以对个人中心的阅读历史内容进行删除操作。

4. 软硬件要求

编号	类型	型号及版本
0	服务器	CPU：2.0GHz，16 核及以上；内存：32GB；硬盘：4TB 及以上； 网卡：1000MB 及以上；独立 IP
1	操作系统	Windows 2008 Server 及以上
2	数据库	MySql 6.5

3	JDK	1.6
4	Tomcat	6.5

5. 开发周期

● 开发周期 2.5 个月，约为 55 工作日；

主要开发远程服务模式，包括古籍内容管理、古籍资源应用管理服务；

分为五个阶段

第一阶段：5 工作日，需要确定与平面设计

- (1) 确定需求细节
- (2) 确定美工设计页面
- (3) 确定具体开发计划

第二阶段：5 工作日，数据库设计与功能架构设计

- (1) 进行系统数据设计
- (2) 进行功能模块设计
- (3) 进行系统架构设计

第三阶段：30 工作日，功能模块开发

- (1) 进行功能模块开发
- (2) 进行开发模块预演
- (3) 进行开发模块测试

第四阶段：10 工作日，功能模块开发

- (1) 进行实际问题反馈修改
- (2) 进行数据批量入库整合
- (3) 进行用户测试

第五阶段：5 工作日，部署试运行

- (1) 进行项目部署
- (2) 进行项目试运行
- (3) 进行项目正式部署