基于 Web 的数字博物馆虚拟空间的构建与实现①

王永平，魏绍谦

(北京联合大学 师范学院，北京 100011)

摘 要： 为了解决数字博物馆信息孤立的问题，通过构建适合数字博物馆的虚拟空间的方法，将分散、孤立的文物信息实现共享。即这个虚拟的空间的作用是以一定的策略利用互联网进行文物信息的搜集、发现，然后对其进行理解、提取、组织和处理。设计过程中采用多线程、信息再过滤、信息重新分类等技术，对信息获取、分析、分类等技术进行改进。可解决目前数字博物馆信息来源局限性问题，同时提高了信息的准确性，使得文物信息分类清晰，实现文物信息的快速检索。

关键词： 数字博物馆；信息获取；虚拟空间；构建；实现

Web-Based Digital Museum Construction and Realization of Virtual Space

WANG Yong-Ping, WEI Shao-Qian

(Teachers' College, Beijing Union University, Beijing 100011, China)

Abstract: To solve the problem of isolated digital museum information, cultural information which is scattered and isolated will be shared by constructing a virtual space for digital museum. The role of the virtual space is to use some strategy to collect and find heritage information through the Internet, and then to complete its conduct, extraction, organization and processing. Using multi-thread, information filtering and information reclassification in the design process, techniques of information access, analysis and classification will be improved. The technology can solve the limitations of digital museum information, and can also increase the accuracy of information, which makes a clear classification of cultural information and realize fast retrieval of cultural information.

Keywords: digital museum; access to information; virtual space; construction; achieve

1 引言

由于博物馆中大量的文物信息缺乏统一的技术标准规范，使得各数字博物馆之间难以实现信息共享，不能满足用户在文物信息获取方面的需要，产生了信息孤岛现象，同时面对大量的文物信息，缺乏准确找到所需信息的手段。

2 虚拟空间的分析与构建

目前，获取信息常用的方法是通过搜索引擎来实现[1]，在现有的数字博物馆中，也提供了信息获取的手段，但提供的内容基本上是博物馆网站内部的文物信息，无论从信息的针对性还是广泛性来说都不能满足用户的需要。

2.1 数字博物馆虚拟空间功能需求分析

2.1.1 用户需求分析

用户在数字博物馆中可以进行文物欣赏、文物知识学习、文物信息获取等活动，其中用户在数字博物馆获取文物信息是用户目前的一个核心需求[2]。

用户文物信息获取是指用户通过登录数字博物馆，进入文物获取工作环境并提供需要获取的文物信息相对应的关键词，做为信息获取的依据，最终获得相应信息的过程。

2.1.2 性能需求

信息获取的性能需求包括信息的广泛性、准确性、

① 收稿时间:2010-05-06;收到修改稿时间:2010-05-27

快速性。其中，信息的广泛性是指除了能够优先搜索本数字博物馆的文物信息以外，还能够提供相关的外部的文物信息，以保证最大限度满足用户信息要求；信息的准确性是指系统根据用户要求，能够准确的提供相应文物信息，并应该能够屏蔽一些无关信息；信息的快速性是指信息提供要高速度，节省时间。

由此可见，用户希望数字博物馆提供一个获取文物信息更准确、更广泛的途径，满足用户在文物信息获取方面的需要。

2.2 虚拟空间的构建

2.2.1 虚拟空间的内容

文物信息的广泛性、准确性、快速性是数字博物馆的三个不同的层面，这三个层面的知识点有机地结合，形成了数字博物馆虚拟空间的内容。

为实现用户快速、准确、广泛的获取文物信息的需求，就要采取区别于一般信息获取的方法。数字博物馆虚拟空间中的文物信息不是简单的堆积，而是文物信息获取、过滤、分类、存储全方位的有机结合。在广泛获取信息的基础上，采用有效的方法提高信息的准确性和快速性，即以一定的策略利用互联网进行文物信息的搜集和发现，对其进行理解、提取、组织和处理。达到使现有的文物资源得到充分的利用，只有这样，才能够快速的提供给用户更准确、更广泛的文物信息，满足用户在文物信息获取方面的需要。

2.2.2 虚拟空间的结构

与数字博物馆虚拟空间的内容相对应，其结构可划分为三个部分，即信息获取、信息过滤、信息分类和存储，如图 1 所示。

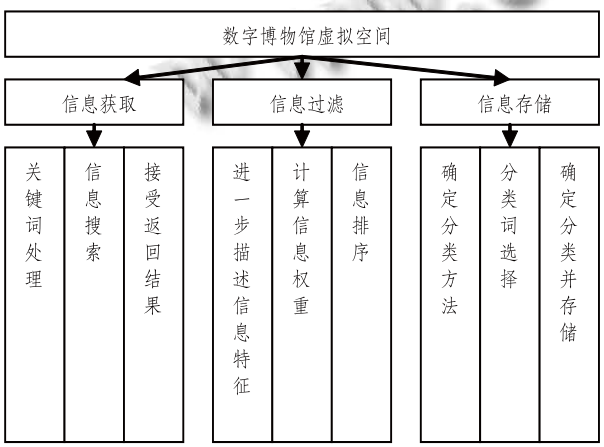

图 1　数字博物馆虚拟空间结构

信息获取的主要作用是保证信息来源的广泛性[3]，这是数字博物馆虚拟空间的基础，它包括三个步骤，即关键词处理、信息搜索、接受返回结果。

信息过滤的功能是在上述基础上进一步对得到的结果进行筛选，得到更加符合用户要求的信息，包括三个步骤，即进一步描述信息特征、计算信息权重、信息排序。进一步描述信息的特征是找出文物信息的自身特点，对其进行深入理解；计算信息权重是以一定的策略确定信息的得分；信息排序是根据信息得分排列顺序。

信息分类的目的是为了方便信息的查找，方法是充分利用信息表现出来的差异特性，构成合理的分类方法。

3　虚拟空间的实现

3.1 虚拟空间的设计

在实现的过程中，其整个功能实现通过两个部分体现，即搜索请求分流和搜索。包含有搜索请求分流服务器、搜索服务器、数据库服务器、web 服务器。它们在同一个局域网中，除 web 服务器必须有对外 IP 之外，其它服务均可以是内网 IP，其框架结构如图 2 所示。

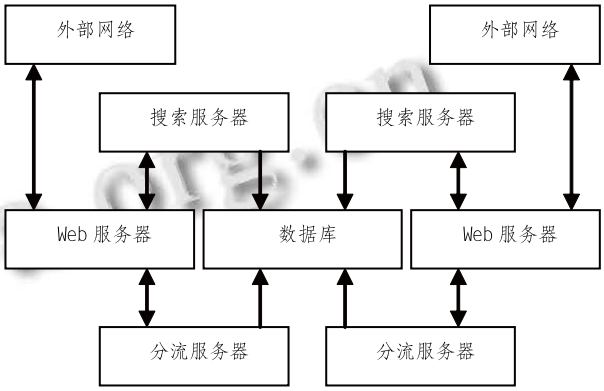

图 2　框架结构

3.1.1 分流服务器的实现

分流服务器(GUID)是一个常驻内存的服务程序，启动之后监听一个 TCP 端口等待客户端的连接请求，主要功能是决定一个搜索关键词应该由哪台搜索服务器来处理请求。同时实现一个 GUID 客户端接口，供 web 程序调用。分流服务器主要由事件分发模块、关键词分流计算模块和客户端模块构成。客户端向服务器发请求时，客户端根据多台 GUID 服务器的权重，

随机选取一台服务器与其通信。

3.1.2 搜索服务器的实现

搜索服务器(MSS)监听 TCP 端口，等待客户端的连接请求，主要完成的功能是接收用户的搜索请求，从互联网或数据库中搜索相关信息，对搜索到的信息进行过滤排序，返回给用户。与此同时，实现一个 MSS 客户端接口，供 web 程序调用。

搜索服务器主要由事件分发模块、网页搜索模块、信息过滤排序模块、文物自动分类模块和客户端模块构成。

3.2 虚拟空间构建技术及改进

3.2.1 信息获取技术

(1) 不同搜索方法的结合

目前，获取信息的常用方法是利用专用的信息搜索工具(搜索引擎)，搜索引擎的优点是能够按照关键词的指引获取数量很多的信息，缺点表现为返回的信息数量过大[4]。

垂直信息获取技术是针对某一个行业的专业信息搜索，是搜索引擎的细分和延伸，是对某类专门的信息进行整合、定向分字段抽取出需要的数据进行处理后再以某种形式返回给用户。垂直信息获取技术和普通信息获取的最大区别是对信息进行了结构化信息抽取，也就是将信息中的非结构化数据抽取成特定的结构化信息数据。

数字博物馆文物信息获取系统的信息获取技术将通用的信息搜索工具(搜索引擎)作为信息获取的基本手段，同时选择固定的有代表性的文物网站进行搜索，充分发挥各自的优势，解决了搜索信息的数量和质量的矛盾，达到了较好的效果。

(2) 关键词预处理

在进行广义搜索时，对用户提交的关键词进行搜索前的相应处理，即在用户关键词后加上相应的附加词，这样可以使搜索的指向更加明确。

(3) 信息搜索请求的分布式部署

信息搜索请求的分布式部署主要是为了满足大流量搜索请求的需要，合理分配当前的信息搜索请求应该由哪个搜索服务器处理，即建立关键词和服务器编号的对应关系，方法是对关键词计算 hash 值，根据搜索服务器权重比例，选择一台搜索服务器并将此关键词和搜索服务器编号记入数据库。

(4) 多线程并发技术

多线程并发是为了使得多个线程并行的工作以完成多项任务，提高处理器和内存等系统资源的利用率。采用多线程技术，在启动时创建一批工作线程，避免运行过程中创建和销毁线程带来的 CPU 和内存消耗，这些线程平常处于空闲状态，运行过程中从工作线程池中找出一个空闲的线程，向其发一个消息，该线程即开始工作，由空闲状态转为繁忙状态，任务运行之后即转入空闲状态。

(5) 数据缓存技术

数据缓存是 web 开发中常用的一种性能优化方法[5]，使用数据缓存技术可以将内存中常用的数据提前放到缓存中，利用缓存速度快、容量小的特点提高工作速度。可在两个地方使用数据缓存技术，一是将最近访问的关键词与服务器编号的对照表存储在内存中，当搜索服务启动时，将近期最多访问次数的关键词和搜索服务器编号的对应关系调入至缓存中，可以加快常用关键词的请求响应速度；二是将最近的关键词搜索结果存放在内存中，下次有同样关键词请求时，直接从内存中返回给客户端。每次从内存中查找数据时，更新最后访问时间，当内存中缓存的数据超过一定的时间没有访问时，将该内存块释放，有效地节约了存储空间。

3.2.2 基于权重的信息分析技术

(1) 信息分析的基本方法

通用搜索引擎的信息分析技术主要从召回率和准确率两方面考察[6]，召回率是一次搜索结果中符合用户要求的文档数与所有符合要求的文档总数之比,衡量的标准是搜索引擎的查全率，由于很难统计文档库中含有的相关文档的数目, 所以召回率在 Web 搜索系统中使用很少；准确率指一次搜索结果中符合用户要求的数目与该次搜索结果总数之比。

(2) 信息分析方法的改进

虚拟空间对信息的分析方法分为两个步骤，第一个步骤是通过关键词的指引利用通用搜索引擎对相关信息进行有序的分析过滤；第二个步骤是首先制定出文物信息的权重和特色，然后在此基础上进一步分析出文物信息的标题、链接、内容简介中的关键信息，结合对信息的权重和特色描述，得到一个新排列顺序，信息的排列顺序是每个信息的得分降序排列，信息得分的依据是信息得分计算公式，其形式如公式(1)所示。

信息得分=(100－返回信息的顺序号)×信息来源系数＋标题偏爱得分＋内容偏爱得分－标题非法扣分－内容非法扣分　　(1)

返回信息的顺序号是指利用通用搜索工具得到的返回信息的排列顺序号，按照重要程度升序排列，排在前面的为重要信息[7]；信息来源系数是描述信息所在网站重要程度的指标。对于信息来源系数可以对一些有影响的权威的文物网站的信息数量和质量进行对比分析，确定合理的来源系数。例如，从国家文物局和北京文物局网站的来源系数设置为1.5(来源系数默认为 1)，也就是在众多的信息中比较看重从国家文物局和北京文物局网站返回的文物信息。

“标题偏爱”和“内容偏爱”、“标题非法”和“内容非法”可以称之为“偏爱词”和“禁用词”，统称为“特征词”，事先可以拟定好相应的特征词，如果搜索到的信息标题或内容简介中含有偏爱词，增加该网页的得分，含有禁用词的信息，减少得分。

确定返回信息得分的原则有两个，一个是充分考虑返回信息已有的排列顺序，排在结果集前面的信息，给予较高的权重。这样做的原因是由于返回信息是通过搜索工具按照用户提交的关键词得到的，所以其排列顺序是值得参考的；二是进一步描述信息的特征，即设置不同信息的来源权重和信息的内容权重，通过这些权重再次计算返回信息的得分，并以此为根据进行再排序。

信息分析方法改进的优点是即保持了原有信息的特点，充分尊重了原有信息的排列顺序又加入了新的得分元素，得到的结果是比较满意的。

3.2.3 信息分类技术

(1) 文物信息分类方法的选择

文物信息的分类是一个重要环节，由于文物本身具有很强的学术性，所以在进行划分时应充分了解和尊重现有的文物分类方法，文物的分类方法从不同的角度可分为时代分类法、区域分类法、存在形态分类法、质地分类法、功用分类法、属性分类法、来源分类法、价值分类法等。

在文物分类中，同类相聚是一个重要原则[8]。同类相聚的“同类”，因标准不同其内容也不尽相同。例如，按质地聚类，铁器类中只有铁制的器物，不会有其他质地的文物；按功用聚类，炊器类中的鼎，就有陶鼎、铜鼎、铁鼎，分属于三种材料制成，是三种不同质地的器物。但不论用哪一种标准聚类，同类文物都有内在的联系。这种联系由聚类标准决定，同时又要受到聚类标准的制约。在文物分类或归类的时候，首先要确定对具体的文物对象以什么做为分类的标准，凡是符合标准的文物，就可以归纳到一起，取舍均从标准出发。在分类标准确定之后，用它去衡量复杂的文物，把符合该标准的文物筛选出来，集合成类，以达到归类的目的。

在实现分类的技术手段上，一般表现为树形的层次目录结构,要将上述分类方法进行合理整合[9]。例如，可将文物按年代分类，按存在形态分类，在存在形态分类下又按文物的类型进行分类。

如果一条文物信息的标题中含有分类 A 的关键词，就把这个关键词归为分类 A；如果既包括分类 A 中的关键词，也包括分类 B 中的关键词，那就看包含哪个分类的关键词多，也是说决定这条文物信息更象哪个分类；如果不能完成匹配某个分类的关键词，再依次模糊匹配，如果一个都匹配不上，就归为“其他类”。与之对应的抽象分类方法描述如表 1 所示。

表 1　分类描述

分类	名称	关键词列表
分类A	AAA	A1,A2,A3,……
分类B	BBB	B1,B2,B3,……
……	……	……

(2) 文物信息分类方法的改进

文物信息获取系统的信息分类方法是以文物信息的时代特征为依据，通过设置分类关键词得到不同的分类权重，最终通过权重进行分类。事先制定好每个文物分类包含哪些关键词，如果文物信息标题或介绍中包括哪个分类的某些关键词，则认为该条文物信息属于该分类。每一个文物信息可以按照多个方式分类，事先定义好使用什么分类方法以及应包含哪些关键词，然后用文物信息的标题和内容与每个分类的关键词比较，与哪个分类的符合度高，就属于哪个分类。举例来说，如按时代分，可以得到下面的文物分类，

如表 2 所示。

表 2 文物分类

文物分类	包括的关键词
夏代文物	夏初 夏朝 夏代 夏末…
商代文物	商朝 商初 商末 商代…
秦代文物	秦初 秦始皇 秦末 秦朝 秦代…
……	……
宋代文物	宋朝 宋代 北宋 南宋 宋末…
……	……
其他	……

4 小结

在构建和应用数字博物馆虚拟空间的过程中，通过将文物信息获取、过滤、分类几个过程有机的结合，并以一定的策略利用互联网进行文物信息的搜集、发现、理解、提取、组织和处理，综合利用多种技术实现了数字博物馆提供信息服务的功能，这其中包括信息采集、信息过滤和信息分类等关键技术，并在其基础上针对实际问题进行了深入的探讨和改进。

信息采集能同时从多个网站获取文物信息，实现了广泛的获取文物信息，解决了目前数字博物馆存在的信息来源局限性的问题；信息过滤通过分析采集到的信息，制定出相应的信息过滤规则，屏蔽了无关信息，解决了返回信息过多、针对性差的问题，提高了信息的准确性；信息分类通过对文物信息特点的分析，定义了使用分类的方法，使得文物信息的分类清晰，实现了信息的快速检索。

在构建数字博物馆虚拟空间的过程中，由于明确了构建的内容并对其结构进行了合理的规划，在实际应用中较好的体现了信息来源的广泛性，信息的准确性和快速性得到了提高，说明在数字博物馆虚拟空间的构建及应用方面所作的工作是有意义的，能够满足用户在数字博物馆中进一步获取信息的需要。

参考文献

1 徐浩,欧阳松.一种高效的无结构对等网络搜索机制.计算机系统应用,2009,18(9):41－44.
2 鲍泓,刘宏哲.基于 Web Services 的虚拟文物博物馆架构.系统仿真学报,2005,17(6):1411－1416.
3 马文涛,刘虹.基于 Web Services 工作流管理系统新模型.计算机系统应用,2009,18(8):27－31.
4 尹焕亮,孙四明,张峰.基于本体的 Web 智能检索研究.计算机工程,2009,35(23):44－46.
5 柯剑,那文武,朱旭东,许鲁. 性能虚拟化存储系统的设计与实现.计算机工程,2009,35(23):4－6,9.
6 马晓宁,冯志勇,徐超.Web 服务中基于信任的访问控制.计算机工程,2010,36(3):10－12.
7 李辉,王瑞波.多条件分页查询优化的设计方法.计算机工程,2010,36(2):51－52.
8 王琼,张量,刘闯.基于关联规则的检索结果聚类优化.计算机工程,2010,36(3):47－50.
9 任刚,曹三顺.水利系统异构数据动态集成的设计和实现.计算机系统应用,2009,18(7):11－14.