文物信息获取系统关键技术分析

王永平
(北京联合大学 师范学院 电气信息系,北京 100011)

摘 要:为满足用户广泛、准确、快速获取文物信息的要求,设计了数字博物馆文物信息获取系统。在设计过程中采用了多线程、信息再过滤、信息重新分类等技术,对信息获取、信息分析、信息分类技术进行了改进,信息采集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。

关键词:信息获取;信息分析;信息分类;关键技术
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2010)07-0219-04

Heritage Information Retrieval System Analysis of Key Technologies

WANG Yong-ping
(Dept. of Electrical Information, Teachers' College, Beijing Union University, Beijing 100011, China)

Abstract: In order to meet the user a wide range, accurate, and quick access to cultural information requirements, design of digital museum objects information retrieval system. During the design process using multi-threading, information re-filtering, re-classification of information technology, information acquisition, information analysis, information classification technology to improve information collection, at the same time cultural information from multiple Web sites to achieve broad access to cultural information, to solve the current digital museum limitations of existing sources of information issues; Information filtering through the analysis of information collected to develop the corresponding information filtering rules, shielding the irrelevant information to solve a return to information overload, targeted poor problem to improve the accuracy of information; Information classification by analyzing the characteristics of cultural information, defines the use of classification methods, making a clear classification of cultural information to achieve fast retrieval of information.

Key words: information acquisition; information analysis; information classification; key technologies

0 引 言

在现有的数字博物馆中,提供的文物信息基本上都是馆内的文物信息,无论从广泛性还是针对性来说都不能满足用户在文物信息获取方面的需要[1]。通过设计数字博物馆文物信息获取系统,并采用广泛信息采集、信息再过滤和信息重新分类等关键技术加以实现,满足了用户广泛、准确、快速获取文物信息的要求。

1 信息获取技术

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,用户很难发现自己需要的信息,同时大量的信息中包含了许多与实际要求无关的信息。

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,根据自身的特点,做了进一步的改进,以适应数字博物馆的需要。

1.1 信息获取的基本方法

1.1.1 搜索引擎

搜索引擎一般由网络爬行机器人 Crawler、知识库 Repository、索引系统(包括索引器 indexer,文件索引等)、排序器 Sorter 和搜索器 Searcher 组成[2]。

搜索引擎首先用几个分布的网络爬行机器人 Crawler 同时从 Internet 下载信息,由 URL 系统负责向 Crawler 提供 URL 的列表,Crawler 所找到的信息被送到知识库(repository)中。每个信息都有一个关联 ID,称为 docID。当一个新的 URL 从一个信息中解析

收稿日期:2009-10-24;**修回日期**:2010-01-27
基金项目:北京数字博物馆平台—徐悲鸿纪念馆数字化建设项目(KM200611417012)
作者简介:王永平(1962-),男,北京人,硕士,研究方向为软件工程。

出来时，就被分配一个 docID。索引器(indexer) 和排序器(sorter) 负责建立索引，索引器从知识库中读取记录，将文档进行解析。每个文档被转换成一组词的出现状况，称为 hit。这些 hit 记录了词、词在文档中的位置、字号、大小写等。索引器把这些 hit 分配到一组桶"barrels"中，产生经过部分排序的索引。索引器同时分析信息中所有的链接，并将重要信息存在链接描述文件(Anchors)中，该文件保存了链接描述文字和其他一些信息，足以判断一个链接被链入或链出的情况。

1.1.2 垂直信息获取技术

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[3]。

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。

1.2 信息获取方法的改进

从上述分析可以看出，通常的信息搜索具有信息量大的优势，而垂直搜索的指向性较好，故从以下几方面进行改进。

1.2.1 不同搜索方法的结合

调用搜索工具进行文物信息搜索时，将广义搜索和垂直搜索相结合，广义搜索是不指定具体的网站，广义地获取信息，垂直搜索是到指定的网站(例如国家文物局、北京文物局、中国美术馆等)去进行搜索，这样可以获取针对性很强的信息。两种方法可以进行互补，当用户通过广义搜索方法找到了感兴趣的信息时，可以根据信息的来源通过垂直搜索方法进行进一步的搜索，保证了文物信息获取的广泛性。

1.2.2 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，例如，用户提交的关键词是"陶瓷"，则在其后加上"文物" 或年代，实际上是以"陶瓷 文物 年代"进行信息搜索的，这样可以使搜索的指向更加明确。

1.2.3 多网站信息同时收集

采用垂直信息获取技术到指定网站获取信息时，利用选定的搜索工具，采用多线程技术，同时从多个网站搜索文物信息，再聚集到一起进行分析处理。

1.2.4 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。

1.2.5 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。本系统采用了多线程技术，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。

1.2.6 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[4]，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。本系统在两个地方用到了数据缓存技术，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。

2 基于权重的信息分析技术

2.1 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[5]，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。

2.2 信息分析方法的改进

数字博物馆文物信息获取系统对信息的分析方法分为两个步骤，第一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)。

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;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1.1(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。

拟定特征词时可以根据当时的实际情况进行设定,以达到较好的效果。比如,现在人们热衷于"收藏"等话题,则可以在特征词中加入与其相关的词,像"瓷器"、"木器"等。

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 信息分类技术

3.1 文物信息分类的选择

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[6]。

在文物分类中,同类相聚是一个重要原则。同类相聚的"同类",因标准不同其内容也不尽相同[7]。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[8]。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[9]。

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[10]。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像哪个分类;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	分类包括的关键词列表
分类A	AAA	A1,A2,A3,A4……
分类B	BBB	B1,B2,B3,B4……
……	……	……

3.2 文物信息分类方法的改进

在所搜索的文档中,是由一组有代表性的词(称为索引项)来表示的[11]。通常将要处理的文本集合中含有的所有的索引项抽取出来,索引项集合 T 可描述为 $(t_1, t_2, \cdots, t_M)$,其中 M 表示文本集合中含有的索引项的个数,实际使用中 M 都是随着文本集合的不断变化而增加的。然后通过预处理保留文档中最具有明显标识作用的索引项。对初始文档 d_j 描述为 $t_{j1}, t_{j2}, \cdots, t_{jN_j}$(其中 N_j 是文档 d_j 含有的索引项的数目)经过预处理以后得到 d_j',包括 $t_{j1} * t_{j2} * \cdots * t_{jN_j}'$,其中 $N_j' \leqslant N_j$,预处理可以很好地减小计算量,然后将文档表示成索引项权重的向量。

通过分配权重(weight)给文档中的索引项从而将文档表示为权重的向量 W,见计算公式(2)。W_j 可描述为 $\langle w_{1j}, w_{2j}, \cdots, w_{Mj} \rangle$,其中 w_{ij} 表示索引项 t_i 在文档 d_j 中的权重。

$$W_{ij} = (1 + \log(t_f(t_i, d_j))) * (\log(1 + N/d_f(t_i))) \quad (2)$$

其中 $t_f(t_i, d_j)$ 表示词 t_i 在文档 d_j 中出现的次数,N 表示要处理的文档的个数,$d_f(t_i)$ 表示包含有词 t_i 的文档个数,同时在检索时也需要将查询 Q 表示成权重

的向量以计算查询与文档的相似度[12]。

这种权重计算方式中 w_{ij} 的大小与 t_i 在文档 d_j 中出现的次数成正比，而与 t_i 在整个文本集合中出现的次数成反比。计算相似度公式见公式(3)，它是通过考察特征向量余弦夹角实现的。

$$\mathrm{Sim}(Q,d_j)=\frac{\sum_{k=1}^{M}W_{ki}\times W_{kj}}{\sqrt{\left(\sum_{k=1}^{M}W_{ki}^2\right)\left(\sum_{k=1}^{M}W_{kj}^2\right)}} \tag{3}$$

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，可以得到文物分类如表 2。

表 2 文物分类

文物分类名称	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 宋武帝 西夏
……	……
清代文物	清朝 清初 清末 清代 康熙 乾隆
其他	……

4 结束语

通过对数字博物馆文物信息获取系统所采用的各种主要技术进行分析，阐明了这些技术的实现方法及采用这些技术所带来的好处。信息获取方法的改进主要体现在不同搜索技术的组合、关键词预处理、多网站信息同时采集等方面，提高了信息获取的准确性；信息分析方法的改进是在充分尊重原有信息排列顺序的基础上，加入了新的得分元素，得到的结果是比较满意的；信息分类方法选择了文物信息的时代特征为基本依据，通过进一步设置分类关键词得到不同的分类权重，最终通过权重实现信息分类。

参考文献：

[1] 鲍 泓.基于 Web Services 的虚拟文物博物馆架构[J].系统仿真学报,2005(6):1412-1417.

[2] 张佳强,周锦程,王士同.基于领域模型的信息系统分析与应用[J].微计算机信息,2009(3-3):195-196.

[3] 王郁新.Web 服务在数字博物馆中的应用[J]. 计算机科学,2007(10):58-60.

[4] 黎 文.数字博物馆关键技术[J].北京科协,2005(5):40-43.

[5] 陆宜梅.Web 搜索技术现状分析[J].沈阳大学学报,2006(4):34-36.

[6] 张宏斌.智能化搜索引擎技术的研究进展[J].信息与控制,2003(12):526-530.

[7] 姚全珠.基于数据挖掘的搜索引擎技术[J].计算机应用研究,2006(11):29-30.

[8] 龚正伟.数字博物馆的建设与发展[J]. 北京科协,2005(5):17-19.

[9] 王永平.基于 Web 的数字博物馆虚拟空间分类索引研究[J].计算机科学,2007(10):58-60.

[10] 何淑庆.URL 分级散列在分布式搜索引擎中的应用[J].电子技术应用,2006(7):25-27.

[11] 张绚丽.基于搜索技术的科技期刊网站建设要点研究[J].武汉科技大学学报,2006(10):76-78.

(上接第 218 页)

[3] 荚 伟,谢 军,奚红宇,等.遗传算法在软件测试数据生成中的应用[J].北京航空航天大学学报,1998,24(4):434-437.

[4] Jullier E M. Tunneling between ferromagnetic film[J]. Phys Lett,1975,54(3):225-226.

[5] 常先英,李荣钧.改进粒子群优化算法及其在 CVaR 模型中的应用[J]. 统计与决策,2009(8):144-146.

[6] 王溪波,马 春,杜晓舟.面向路径的测试数据自动生成工具设计与实现[J].沈阳航空工业学院学报,2009,26(6):54-59.

[7] 李爱国,张艳丽.基于 PSO 的软件结构测试数据自动生成方法[J].计算机工程,2008,34(6):93-94.

[8] 虞 凡,覃 征,贾晓琳.基于 XYZ/E 规范的软件测试用例自动生成方法[J].计算机工程,2005,31(19):76-78.

[9] 夏 芸,刘 锋.基于免疫遗传算法的软件测试数据自动生成[J]. 计算机应用,2008,28(3):723-725.

[10] Khurshid S, Suen Yuk Lai. Generalizing Symbolic Execution to Library Classes[J]. ACM SIGSOFT Sotfware Engineering Notes,2006,31(1):103-110.

[11] 常瑞花,张 力,慕晓冬,等.基于遗传算法的结构测试数据自动生成[J]. 火力与指挥控制,2009,7(3):76-78.

[12] 高海昌,冯博琴,朱 利,等. 改进的遗传算法在测试数据自动生成中的应用[J]. 系统工程与电子技术,2006,28(5):1077-1081.