科技资源信息检索关键技术

张 辉 吴 辉 刘 瑞 周 刚

(北京航空航天大学 软件开发环境国家重点实验室, 北京 100083)

摘 要: 科技资源具有地理分布广、异构、复杂、海量的特点,为了高效查找和使用科技资源,提出了实现科技资源组织、表示、传输和访问的有效方法. 设计了数据层、数据连接层和表示层 3 层体系架构,使检索系统具有开放性和扩展性. 利用元数据技术实现科技资源的统一组织和表示,并基于轻量级目录访问协议 LDAP(Lightweight Directory Access Protocol)的目录服务机制实现了科技资源元数据的存储和访问. 提出了科技资源信息检索的多级缓冲机制和结果排序模型,对数据查询的性能进行了分析和优化,提高了查询速度,并可优先返回价值高的查询结果.

关 键 词: 科技资源; 信息检索; 资源组织; 元数据; 数据查询; 性能优化

中图分类号: TP 302.1

文献标识码: A **文 章 编 号**: 1001-5965(2006)09-1063-04

Key techniques of information retrieval from science and technology resources

Zhang Hui Wu Hui Liu Rui Zhou Gang

(Beijing University of Aeronautics & Astronautics, National Laboratory of Software Development Environment, Beijing 100083)

Abstract: In order to search and utilize science and technology resources, which are usually distributed, hybrid, complex and massive, with high efficiency, some measures were provided on how to organize, represent, transmit and access these resources. A three-layer architecture, which consisted of data layer, data-link layer and presentation layer, was designed to make the retrieval system more open and extensible. The metadata was utilized to organize and present different types of science and technology resources. Directory service mechanism based on LDAP(lightweight directory access protocol) was used for science and technology metadata storage and accessing. In order to analyze and optimize the performance of data querying, the multi-level buffer mechanism and the query result sorting model were introduced. In our proposed solution, the responding speed of the retrieval system is improved significantly, and the system is also able to return the query result sorted by the value.

Key words: science and technology resources; information retrieval; resources organizing; metadata; data query; performance optimization

"科技资源"是指科技领域的信息资源(科技文献、科技数据等)和实物资源(动植物标本、大型科学仪器设备等). 信息网络、数据、文献资料等科技资源的广泛社会共享,是充分利用知识积累和众人智慧,提高科研水平和效率的有效途径[1]. 经过多年的积累,我国拥有了丰富的科技资源,这些资源具有结构复杂、类型多样、数量巨大、地理分布广的特点,为发挥科技资源的作用,必须对其进行合理组织和管理,通过数字化、结构化、网络化形成有一定内在关系的、可共享的科技资源信息.

"信息检索"对信息进行有效组织,实现数据

收稿日期: 2005-11-10

基金项目: 国家科技基础条件平台建设资助项目(2003DKA5G015)

作者简介: 张 辉(1968-),男,浙江磐安人,副教授, hzhang@nlsde.buaa.edu.cn.

的高效查询. 信息组织是对被检索对象进行处理并建立索引,形成一个便于检索的信息集;数据查询则是通过输入关键词或查询条件,在已建立的信息集中找出和查询条件相匹配的信息并返回结果[2].

1 科技资源信息检索需求分析

2004年,科技部、发改委、教育部、财政部启动了"国家科技基础条件平台"(以下简称"科技平台")项目[3],其目的是实现我国科技资源的有效整合、共享和服务. 在科技平台建设中,科技资源的信息检索能够有机整合科技资源信息,实现逻辑上的统一定位和表示,同时能够为用户快速定位资源、发现资源提供帮助.

目前,我国科技资源的现状有如下特点:①数据量大. 我国科技平台存在动植物标本、科学数据、文献档案、大型科学仪器设备等多种科技资源,且数据量巨大;②资源类型复杂. 科技平台中既有数字化的图片、数据、科技文献等信息资源,又有矿藏标本、仪器等实物资源;③数据存贮结构复杂. 各种资源建库年代不同,存储方式及数据库类型千差万别;④数据存贮地理位置分散. 各类资源分布在各地、各单位,相互之间是信息孤岛.

相比google等基于Web网页的商业搜索引擎,针对我国科技资源特点建立起来的科技资源信息检索系统(以下简称"检索系统")主要对异构科技资源进行组织和处理,实现科技资源的统一表示、统一访问和统一传输,最后实现检索结果的统一展示,具有很强的领域性和专业性.

2 检索系统设计和实现

为了使检索系统具有开放性和可扩展性,系统体系架构设计采用了分层结构,如图1所示. 结构共分3层:数据层、数据连接层和表示层.

1) 数据层:检索系统的数据源,包含各种类型的数据资源,既有Oracle,Sybase等异构数据库,又有目录服务、文本、流媒体等其它形式的信息,这些数据资源在结构、存储、访问模式等方面千差万别.

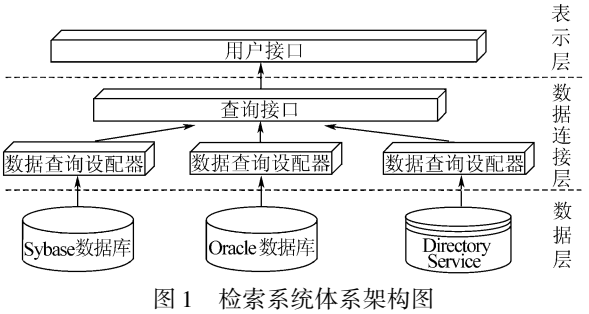


图1 检索系统体系架构图

2) 数据连接层:对底层数据源在逻辑和物理上实现屏蔽,消除数据源的异构性,达到整合数据的目的. 逻辑屏蔽就是通过统一查询接口来消除各数据源在结构上的差异,本系统采用数据查询适配器实现逻辑屏蔽. 数据查询适配器是针对每一种数据源(如Oracle数据库、html文本等)实现统一封装的程序实体,并向上层提供统一的查询服务. 物理屏蔽就是通过一种查询接口将各类型的数据资源整合起来,消除它们在存储位置上的差异. 查询接口能够根据具体数据源对数据查询适配器进行选择调度,并将各个数据查询适配器的返回结果整合起来,提交给表示层.

3) 表示层:检索系统的用户接口,实现系统与用户的交互工作,负责接受用户的查询请求并展现检索结果.

在检索系统的具体实现中,资源的组织和数据查询是系统设计的关键,本文重点讨论这两个关键问题.

2.1 资源组织

目前,科技平台拟接入6类科技资源:自然科技资源、科学数据、网络科技资源、大型科学仪器设备、科学试验基地和科技文献[4]. 各类资源地理分布、结构不一致,如果直接对这些资源进行基于内容的检索实现起来比较困难,因此需要建立一种独立于具体资源表现形式的数据表示方式,即元数据. 元数据是关于数据的组织、数据域以及关系的信息,也就是"关于数据的数据"[5]. 不同类型的资源表现特性存在较大差别,通过定义和规范元数据可实现统一表示多种类型资源的目的. 目前国际上通常采用都柏林核心集来描述各种数字化对象[6],本系统在借鉴都柏林核心集和对象元数据的基础上设计了平台中的元数据,对于各类资源均给出相应的元数据格式.

平台中的元数据格式,即:专业元数据=基本元数据元素+专业元数据扩展+相关字段映射. 其中,基本元数据元素包含了6类资源的共有元素属性;专业元数据扩展是在基本元数据不足以表示某一资源时,通过专业扩展来表示该资源信息;相关字段映射是为了充分利用基本元数据,当专业元数据中的元素可以在基本元数据中找到与之相对应的元数据元素时,通过映射来表达而不必扩展,从而减少元数据的字段个数. 元数据设计如图2所示.

由于元数据具有层次型的分类体系,因此元

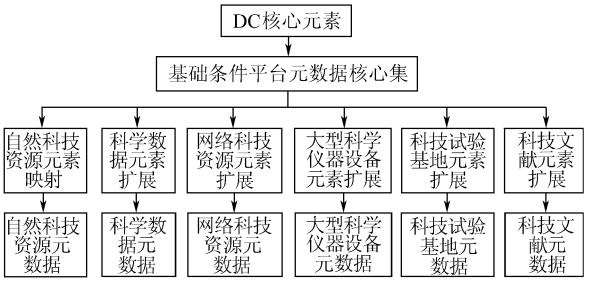


图 2　元数据设计

数据存储和访问采用轻量级目录访问协议 LDAP (Lightweight Directory Access Protocol) 的目录服务机制. 相对于关系型数据库,科技资源元数据管理采用目录服务机制,在存贮、扩展、查询等方面具有较大优势.

2.2　数据查询

数据查询的性能是检索系统设计的关键,好的检索系统应该具有响应速度快且价值高的查询结果优先返回的特点,即及时性和序列性.

2.2.1　及 时 性

及时性是指检索系统能够在短时间内反馈查询结果. 提高及时性的常用办法是“缓冲技术”,即在内存中保存一段时间的查询信息. 本系统设计了 3 种缓冲技术:

1) 缓冲一段时间内的查询结果. 将一段时间内频繁访问的查询请求和相应结果缓存起来,减少查找数据源的频率;

2) 缓冲单个用户的当前查询结果. 用户提出查询请求后,将该请求的结果全部缓存到服务器内存中,供该用户翻页查询时使用;

3) 缓冲数据源连接,即连接池. 事先与数据源建立多个连接等待调用,从而减少实时连接数据源的频率.

连接池的应用比较普遍,本文着重介绍前两种缓冲技术. 这两种缓冲共同构成了系统的结果缓冲区,即核心缓冲区和个人缓冲区. 个人缓冲区存贮的是单个用户当前查询关键词的全部结果;核心缓冲区存贮近一段时间内频繁被查询的关键词及其结果. 当用户提出查询请求时,检索系统首先根据查询关键词和该关键词显示的页面参数,在结果缓冲区中寻找,如果找到,则反馈给用户;否则,按照该关键词信息在资源库中查找,并将其调入缓冲区. 其流程如图 3 所示.

如果用户的查询请求已经在个人缓冲区中,那么就直接读取;否则,调用核心缓冲区中的内容. 个人缓冲区在服务器内存中开辟,核心缓冲区在服务器数据库中设置. 当用户点击搜索结果时,

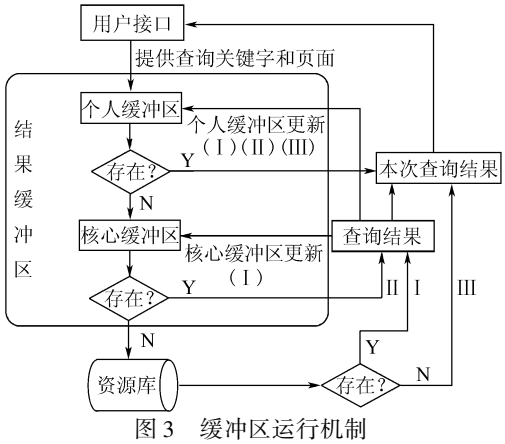


图 3　缓冲区运行机制

服务器先传递第一页的结果(15 个结果项);当用户对同一关键词进行翻页选择时,由于该关键词的所有结果已经在服务器的内存中,所以结果的页面转换较快.

本系统试运行情况表明,增加个人缓冲,检索时间明显缩短;增加核心缓冲区,对于在缓冲区中已经存在的检索关键词,查询速度明显提高. 表 1 为一组测试数据,检索系统运行环境为:CPU:Intel(R) Xeon(TMP) MP CPU 3 GHz,内存:3 G,操作系统:Windows 2003.

表 1　测试数据、返回结果及返回时间

关键词	结果项数/项	分页数/页	无缓冲时间/ms	有缓冲时间/ms
激光器	175	12	359	<1
质谱仪	268	18	297	<1
显微镜	758	51	360	<1
航空航天	76	6	140	<1
色谱仪	555	37	266	<1
分析仪	986	66	422	<1

其中,关键词为用户输入的科技名词;结果项数为检索系统返回的结果个数;分页数为按每页显示 15 条结果生成的页面数;有无缓冲时间为采用缓冲技术前后的响应时间,单位为 ms.

从测试数据可以看出,采用缓冲技术减少了一次传输的信息量,减少了磁盘 I/O 操作次数以及数据库连接时间,查询速度得到明显改善.

2.2.2　序 列 性

序列性是对查询结果进行合理的排序,依据检索结果的内容和用户查询请求的相关性程度,确保高相关性资源优先返回. 检索系统从结果集的内容出发,采用集合运算的方法实现粗排序,利用资源重要性权值再进行细排序,从而实现从堆排序到堆内排序的过程. 堆排序对结果集进行粗粒度的顺序排列,而堆内排序则实现了细粒度的顺序排列.

堆排序算法是依据关键词在资源中出现的完

整性,将包含查询关键词较多的资源排在前面. 算法的基本思想是将用户的查询请求分割为关键词的集合,假设用户提出的关键词共有 $K_{(1)}$,$K_{(2)}$ 和 $K_{(3)}$ 共3个. 该算法先显示 $K_{(1)}$,$K_{(2)}$ 和 $K_{(3)}$ 都包含的结果堆,再显示只包含两个关键词的结果堆($K_{(1)}$ 和 $K_{(2)}$,$K_{(1)}$ 和 $K_{(3)}$ 或 $K_{(2)}$ 和 $K_{(3)}$),最后显示只含有一个关键词的结果堆($K_{(1)}$,$K_{(2)}$ 或 $K_{(3)}$),且堆与堆之间相互独立,不存在重复的元素. 结果堆排序原理如图4所示,图中"&"对应逻辑"与";"||"对应逻辑"或";"!"对应逻辑"非".

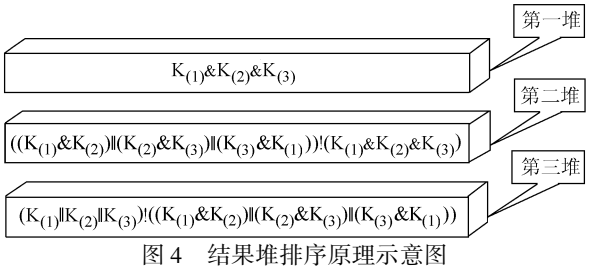


图4 结果堆排序原理示意图

推广到关键词总数为 N 的情况,其堆排序的原理为:第一堆为包含所有 N 个关键词的结果集合,第二堆为包含 $N-1$ 个关键词的结果集合. 依此类推,第 i 堆为包含 $N+1-i$ 个关键词的结果集合,最后一堆为只包含一个关键词的结果集合. 假设关键词为 $Key_{(1)}$,$Key_{(2)}\cdots Key_{(N)}$,共 N 堆,采用过滤控制条件表达式 $Filter_{(i)}$ 可生成第 i 堆结果集合. 式(1)和式(2)分别给出 $Filter_{(1)}$ 和 $Filter_{(i)}$ 的表达式.

$$Filter_{(1)} = Key_{(1)}\&Key_{(2)}\&\cdots\&Key_{(N)} \quad (1)$$

$$Filter_{(i)} = \cup \sum_{j=1}^{C_N^{N+1-i}} (Key_{(j_1)}\&Key_{(j_2)}\&\cdots \&Key_{(j_N+1-i)}) - \cup \sum_{j=1}^{i-1} Filter_{(j)} \qquad i > 1 \quad (2)$$

因为第 i 堆中的每个结果项只包含 $N+1-i$ 个关键词,那么第 i 堆中的关键词有 C_N^{N+1-i} 种组合,需要将 C_N^{N+1-i} 种含有 $N+1-i$ 个关键词的结果集合相加,式(2)中 $Key_{(j_m)}$ 表示第 j 种情况的第 m 个关键词. 此外,因为第 i 堆只能含有 $N+1-i$ 个关键词,所以在第 i 堆结果集合中应除去关键词数超过 $N+1-i$ 的所有结果项,即式(2)中的 $\cup \sum_{j=1}^{i-1} Filter_{(j)}$,最后利用式(2)生成第 i 堆结果集合.

堆内排序将依照每条记录(堆内的每个元素)的权值大小进行,该权值通过系统的权值计算模型来生成. 权值计算模型综合多个对记录产生影响的因素,通过加权函数计算出记录的权值. 根据科技平台资源特点,各类资源的参考因素各不相同,通过权值公式计算后形成资源的优先级,权值高的资源排序优先. 以科技文献类为例,式(3)给出了相应的权值计算公式.

$$权值 = a \cdot A + b \cdot B + c \cdot C + d \cdot D + e \cdot E \quad (3)$$

其中,a,b,c,d,e 为加权因子,$a+b+c+d+e=1$;A,B,C,D,E 是各种参考因素的计算公式,下面介绍各计算公式的形成过程:

公式 A:考虑关联信息,即关联类型和数目. 目前定义关联相关的元素项共有13个,可用式(4)表示:

A = 版本继承+版本关联+被替代+替代+被需求+需求+部分于+部分为+被参照+参照+格式转换于+格式转换为+遵循 (4)

根据这些元素的重要性分别给每个因素赋予一个数值,当资源元数据具有某个属性时便得到该赋值分,否则没有(即"0-1"变量函数).

公式 B:考察资源权威性,主要通过资源所属单位或级别来衡量,如刊物的等级越高,论文的权威性也越高.

公式 C:考虑资源本身的技术属性,如在同等条件下,篇幅长的文献其价值相对较大.

公式 D:考虑时间属性,排序时将最新的资源排在前面. 该时间因素采用资源发表时间与查询时间的差值来表示,差值越小则价值量越大. 公式 D 是这个差值的递减函数.

公式 E:考虑算法扩充性,以便于引进新的控制因素,同时使整个权值可由系统进行调节和控制.

堆内结果将通过该计算模型算出重要性权值,并按照权值进行排序. 该排序算法考虑了多个资源因素,在一定程度上保证了排序的公平性. 查询结果排序算法的思想是先粗分再细分,粗分算法尽量简单,细分算法尽量全面、公平.

3 结束语

本文结合国家科技基础条件平台中科技资源信息检索的需求,设计了具有3层结构的系统总体架构,使多种异构资源能够方便接入,屏蔽了科技资源在地理位置和数据模式上的差异,实现了科技资源信息的统一表示、统一访问. 同时,由于采用了多级缓冲机制和有效的结果排序模型,提高了科技资源的查询性能,并在试运行中取得较好的效果.

(下转第1103页)