

论数字化图书馆的检索技术

[作者] 王甲佳, 马大川

[摘要] 武汉大学传播与信息管理学院

[摘要] 本文主要论述了数字图书馆的概念及其结构组成, 数字图书馆的检索技术和它的发展趋势。

[关键词] 数字图书馆, 检索技术

我们中的大多数人都有过在图书馆查阅资料、借阅图书, 在知识的海洋中遨游的经历。现在随着信息技术和互联网的迅速发展, 一个内容无限, 图文并茂、无边无际的“数字图书馆”呈现到了我们面前。

所谓的数字图书馆, 简而言之, 就是一种拥有多种媒体形式的、内容丰富的数字化信息资源, 是一种能为读者方便、快捷的提供信息的服务机制。这一概念最早出自欧美。在美国, 数字图书馆是克林顿政府所倡导的信息高速公路计划的一部分, 美国已经率先在这一领域开展了工程性研究, 目前, 一期工程已经结束, 二期工程正在实施, 我国政府也非常重视数字图书馆工程的建设, 1999年6月, 中国数字图书馆发展战略研究组等单位举办了99数字图书馆论坛, 同时, 被列入了国家科技创新重点工程之一的中国数字图书馆工程已经开始启动, 据悉, 由北京大学图书馆、信息科学中心和中国高等教育文献保障系统管理中心共同发起的北京大学数字图书馆研究所也已经正在筹建中, 这一切表明, 中国数字图书馆建设正在迅速地发展。

数字图书馆主要由简单的组件构成, 这些组件主要是数字对象。数字对象是结构化数字信息的一种方式, 其中包含元数据, 元数据中包含数字对象的唯一标识符句柄。为了表示数字图书馆中复杂的结构信息, 可以将若干数字对象组合起来, 形成数字对象集。所有的数字对象都具有相同的基本形式, 但是一个数字对象集的具体结构取决于它所表示的信息的复杂程度。数字图书馆中不同类习惯内的资料被分成不同的类目, 如 SGML 文本、WWW 对象、电脑程序等。每一类目都具有相应的规则和约定, 描述如何将该类目的信息组织为数字对象。

数字图书馆不同于传统的图书馆。传统的图书馆最主要的职能是收藏, 并在对图书资料保存、组织的基础上为读者提供各种服务, 而数字图书馆的收藏对象是数字化信息, 但数字化收藏加上各类信息处理工具并不能构成数字图书馆。他是一个将收藏, 服务和人集成在一起的一个环境, 他支持数字化数据、信息和知识的整个生命周期的活动, 包括生成、发布、传播、利用和生存。

数字图书馆是一个跨学科的综合研究课题, 涉及了近 30 个学科和分支, 主要包括了计算机科学、图书馆和信息科学、教育、生物信息、电子工程、新闻和传播、心理学、医学信息、环境科学、语言学、机器人等等。

实际上, 推动数字图书馆发展的动力是 Internet 的发展, 网络的互连使访问分散在各地的信息资源成为可能, 但这些各自独立的信息仓储具有各自独立的组织、检索和描述方式, 所收藏的信息的质量也良莠不齐, 网络环境下跨仓储的、统一的、高效的访问和利用工具, 以及高质量信息生成、组织和提取成为研究的重点, 而这正是研究的内容, 我们如果把 Internet 看成是一个巨大的无墙图书馆, 广义的数字图书馆的目标就是要把优化 Internet 的存储结构, 提供一致的检索接口, 使整个网络变成一个虚拟的、单一的、有组织、有结构的信息集合, 实现跨仓储的无缝查找。

数字图书馆的基本目标是创造一个良好的信息环境, 提供对分布式存储的信息知识化组织、智能化访问和服务。所谓智能化访问是指对信息的访问不是简单的对原始数据的查找(如

当前 Web 搜索引擎的关键词查找), 而是根据用户的信息需求进行知识的查找和内容提取。要实现这个目标有两个基本的问题:

1 数字对象的组织结构

数字对象是数字仓储中表示信息的基本逻辑单位, 如一篇文章、一张图片、一部音乐作品或是一段影像。数字对象的信息结构是数字图书馆的基本问题, 它决定着进一步的信息组织、处理和利用方式。

2 分布式仓储的组织结构

数字图书馆的收藏可以特指本地的信息仓储, 也可以是互连的信息仓储的集合。如何建立一个统一、互操作的、可伸缩的组织框架, 将分布互连的信息仓储组成为一个整体, 在此基础上提供高质量的信息服务, 如屏蔽各仓储的差异, 提供统一的服务接口、语义化检索、智能代理等。

在当前, Internet 上的信息检索模式是在交互的过程中进行浏览和自由词全文检索。自由词是指检索的关键词是由用户自由选择的, 不受任何限制。客户端的 WWW 浏览和全文查找分别是在服务器端的 HTTP 服务器和由 WEBCRAWLER 等自动搜索软件产生的索引表支持下完成的。面对网上巨大的信息量, 目前的浏览方法费时费力, 网络门户的分类索引并不能解决根本问题。全文检索的自由词, 也就是无序词, 可能来自于文献的标题、作者、文摘或全文; 而用户所选择的词又有很大的随意性。这样的全文查找, 其查准率是很低的, 更不用说查找图像、声频、视频等多媒体文档了。

造成这一问题的原因主要有以下几个方面:

(1) 自动搜索及索引软件只是进行关键词匹配, 而信息检索需要的是概念匹配。

(2) 网上电子文献的无结构性。当前网络上各种电子文献基本上是以 HTML 格式为主, HTML 基本上是无结构的, 其主要功能是提供资源的超级链接。

(3) 在传统的图书馆中, 用户的文献查找过程是在图书馆员的协助下完成的, 他们帮助用户确定准确的检索词, 选择查找的信息资源。而现在的网络检索机制几乎没有提供相应的支持。

我们为了克服上述检索方式的缺点, 就需要进行结构检索, 所谓的结构检索就是在服务器端对信息进行良好的组织和结构化, 将所有的信息文档按照统一的方式进行标识、存储和索引。并在此基础上, 利用文档中的结构化描述实现高精度的检索。具体的实现过程如下所示:

利用文档的细粒度结构。采用 SGML 来标识文档的结构, 包括全文、章节、图表、公式、文摘和参考引文。跨信息源的查询借助与一套规范的元数据和标签来实现, 将对 SGML 仓储的查找和目录、词表等其他图书馆的服务结合起来。

将文档对象化并保存在仓储中。仓储是有组织的对象集合并带有索引和视图, 索引支持查找, 视图支持显示。对分布式的仓储进行联邦化的操作, 在仓储中记录对象的结构。并利用这种结构导引跨仓储的联邦化查找。

按照查找的需求调整查找的界面。例如, 用户可以用布尔连接符来指定一个或多个词语, 用不同的邻近度来限定, 并使用 SGML 标签将查找限制在文档的某个指定部分。也可以从“词轮”列表中选择适合出现在收藏中的合适的词语, 使用事件选定的“典型”文档列表直接选择文档。

交互式术语提示。在用户的全文查询界面中进行交互式的术语提示,提供主题词表 and 同现词表。主题词表是由专业的图书馆员将某个专业领域内的重要术语按照语义层次结构排列而成的词汇表,其主题词都是专业内的规范词;同现词表是由自动索引程序对文档进行同现统计分析,根据词汇在文档中出现的频率排列而成的词汇表,其中包含有更广泛的词汇,能够反映新的词汇,也更加灵活。用户可以从任一词表中选取词汇进行全文查找。一般情况下,用户先参考主题词表,得到粗粒度提示,标识总的主题范围;然后参考同现词表,得到细粒度的提示,确定所需要的检索用词列表。最后,用这些词进行全文查询。

状态网关要从实质上提高 Web 查找的性能,需要保留 Web 的交互性能,通过状态网关以提供会话历史。

我们知道,信息检索的目的就是在信息收藏中查找包含用户所需要信息内容的文档。这里有两个问题需要解决。一个是描述文档的信息内容,另外一个清楚的表达用户的信息需求。在传统的检索技术中,解决这两个问题的方法是受控词匹配。我们可以在统一的主题词表的控制下,让用户选择规范的主题词表达自己的信息需求,其优点是双方参照同一的词表选用相同的词语表达概念,但缺点是受制于词表。

在当前的网络信息检索系统中,主要采用的是自由词匹配。用户任意选用词语比哦啊大自己的检索需求,在文档的全文中进行描述和标引,其优点是灵活;缺点是有大量的误匹配和漏查。由于并不是文档中的任何一个词都可以表达文档的内容,因此,用户所选择的检索词也不一定是文档中的,尽管他们表达的是同一个概念。

未来的信息系统应该是概念,也就是语义检索。即自动抽取文档的概念,即主动抽取文档的概念,加以标引;用户在系统的辅助选择合适的词语表达自己的信息需求,然后在两者之间执行概念匹配——匹配在语义上相同、相近、相包含的词语,例如,用户在查找的是“操作系统”,那么,“UNIX”将是与之概念匹配的词语之一,人工智能和自然语言理解在这一领域进行了富有成效的研究,但是目前所构造的这类系统要求将文献资源限制在较窄的专业领域。

概念匹配还可以解决信息检索中的“词汇问题”。研究人员经常需要借鉴其他领域的研究成果,但是由于专业术语的隔阂,即使是在非常接近的领域也常常难以找到所需要的文献。例如,在山谷中架桥的工程师为了研究风力对桥梁结构的影响,希望能够参考在海底铺设隧道的工程师的研究水流对管道结构的影响的成果,解决词汇问题的方法是从多涉及的专业领域中在语义上进行转换,如前述的桥梁工程师可以直接利用已经熟悉的空气动力学术语,系统则自动将他转换为海洋流体方面的术语。

语义检索只有在相应的信息基础结构上才能实现,特别是在一个分布的、异构的信息仓储构成的多媒体网络信息环境中实现仓储的语义联邦和检索的概念匹配——语义互操作,这是数字图书馆面临的巨大挑战。

数字图书馆要求我们建设一个互联的信息空间,以实现跨仓储的语义联邦和语义检索,它为我们指出了本世纪网络信息环境的发展方向——信息分析环境,它的主要内容如下:

语义索引。首先识别并抽取表达文档内容的概念。方法是上下文同现统计分析(CO-Occurrence),分析那些词一同出现在同一句中,并统计其频率,构造同现词概念图。然后用抽取出来的这些概念词对文档自动标引。仓储中各文档概念图的集合形成了本仓储的概念空间,也就是该仓储所属的专业领域的概念空间。

语义互操作。即跨专业领域的词汇切换。在不同领域的同现概念图间交互互连,既在分属不同的概念空间,具有相应的术语间进行映射,实现跨仓储的语义联邦。由于这些概念空间常常来自不同的社区图书馆,这样,就提供了一条在不同的图书馆之间进行概念映射的途径,实现跨专业、跨图书馆的语义互操作。

语义检索。完全的语义检索有待于人工智能技术和自然语言理解技术的成熟,在词汇切

换和语义联邦的基础上,借助交互式的术语提示来实现语义检索的,在用户的检索的过程中,系统向用户提供概念图,并根据用户的输入词来定位相关的部分,供其选择候选的检索词,对于词汇切换问题,由用户在两个不同的领域指定相同的术语,系统根据此来检索在两个专业领域之间的概念图中交叉连接,并显示这两个领域中此术语周围的概念图。如此,用户就有了 2 个术语提示表,以比较那些分属于不同的专业领域却表达了同一概念的 2 套词汇。

信息组织结构是数字图书馆中组织信息的结构,研究如何有效、灵活的在数字图书馆中表示了丰富多样的数字化收藏信息,它是数字图书馆的一个关键的问题,直接影响着数字图书馆中数字资源的存储、管理和检索。

下面我们就简单的介绍一下数字图书馆信息组织结构的特点:

关联关系。数字化资料经常以部分/整体、序列等关系互相关联。例如一份由页、章、索引、插图等构成的数字化文本,一个包含多页文本、若干嵌入图片和许多链接的 WWW 对象。

多种存储格式。同样的项目带有多种数字存储格式,有些是可以互相转化的,如一幅未压缩的原始图像和它的无损压缩版本,有些则含有不同的信息,如一页文本的 SGML 格式和 Postscript 格式。

不同的版本。如扫描图片的高质量存档版本和它的缩微版。

不同的权限和许可。构成一个信息项的每个元素都可能有不同的权限和许可,例如页的文字和图片可以分属不同的作者。

不同的工作模式。用户获取资料的方式受限与其所处的计算机系统、网络环境及资料的多少。例如,拨号上网的用户和通过专线上网的用户的工作模式可能完全不同,尽管他们所进行的工作一样。

2.1 信息组织结构设计要求

- (1)能够提供用户及其应用以相当的灵活性;
- (2)使收藏易于管理;
- (3)能够及时反映信息基础结构在经济、社会和法律等方面的发展。

2.2 结构元数据和元对象

信息组织结构有三个基本概念:

数据类型。描述数据的技术属性,如格式和处理方法。

结构元数据。是描述数字资料的类型、版本、关系等特性的元数据。

元对象。提供对数字对象集的引用。最简单的元对象是一个指向其他数字对象的句柄的列表。如列出某物理项的所有数字化版本的数字对象就是一个元对象。

2.3 数字图书馆中的计算机系统组件

在数字图书馆的系统结构中,主要的计算机系统组件如下所示,他们可以运行在网络中不同的计算机系统上。

用户界面。用户界面有两种,一种为用户服务,一种用于图书馆员和系统管理员对收藏进行管理,每种用户界面都有两个互联的部分,即浏览器和客户服务,浏览器负责与用户交互,客户服务在浏览器与系统的其它部分之间提供中介功能,允许用户自己决定到哪里查、

怎么查,以及解释结构化数字对象的信息,协商条款、管理数字对象间的关系,记忆交互状态和转换协议等。

仓储。存储和管理数字对象及其它信息。大型数字图书馆有许多各种类型的仓储,包括传统数据库、Web 服务器等。仓储的界面通常实现为仓储访问协议(RAP),RAP 显式的识别在用户访问一个数字对象前必须满足的权限和许可,支持数字对象的传播、开放结构和定义良好的界面。

句柄系统。为 Internet 资源提供句柄的分布式目录服务。与仓储一起应用时,句柄系统收到输入的数字对象的句柄,返回保存的数字对象的仓库的标识符。

查询系统。用户的查询过程分为三步,现在我们假设用户要查找一张周恩来和尼克松的合影照片。让我们看一看具体的查找过程。

第一步 查找符合要求的数字照片,用户在浏览器中输入查询要求,将查询传给客户服务,客户服务按查找系统的协议和格式对查询进行转换,最后返回按由句柄组成的查询结果列表。

第二步 用户在结果集中选取要观看的一幅数字照片。

第三步 检索该数字照片,客户服务将选取的照片的句柄传给句柄系统,句柄系统将返回仓储地址,客户服务用 RAP 将句柄交给仓储,所有的 RAP 交易都必须通过明确的条款验证,这可能需要在客户服务和仓储之间协商,或直接与用户协商。最后,所需要的数字照片从仓储中通过客户服务传到用户的浏览器中显示出来。

参考文献

- 1、韩得志. 数字图书馆技术探讨. 华中理工大学学报, 1999(1): 30 ~ 32
- 2、孙承鉴. 走向数字图书馆. 计算机世界, 1998(2)
- 3、王卓杰. 论数字图书馆的含义及其功能. 图书馆学报, 1998(4): 14 ~ 17
- 4、张滨声, 于爱香. 各国数字图书馆方兴未艾. 图书馆建设, 1999(5): 52 ~ 53
- 5、中国数字图书馆工程扫描: 大事记. 光明日报, 2000-3-8