基于 XML 的 P2P 网络资源检索系统①

任文娟

(山东电子职业技术学院 计算机科学与技术系, 济南 250014)

摘 要: 结合 P2P 技术和 XML 技术在挖掘网络上广泛分布的异构资源的优势, 构建了一个基于 XML 的高效的 P2P 分布式网络资源检索系统, 采用强大的全文检索开源工具包 Lucene 进行了实现, 并对核心功能在基于 android 平台的移动客户端进行了延伸. 本系统的设计和解决方案对于解决网络上异构资源的共享具有重要的借鉴意义.

关键词: XML; P2P 技术; Web 资源检索; Lucene

P2P Network Resources Retrieval System Based on XML

REN Wen-Juan

(Department of Computer Science and Technology, Shandong College of Electronic Technology, Jinan 250200, China)

Abstract: By combining the advantages of P2P network technology and XML technology in mining the heterogeneous data distributed on the web, this paper built an efficient P2P network resources retrieval System. And it was implemented with the open source toolkit Lucene. The core function of this system is stretched on android platform of the mobile client. The design and solution of this system has its significant reference to solve the information sharing of hetero-geneous resources on the web.

Key words: XML; P2P; Web resoursces retrival; Lucene

目前, Web 搜索引擎已经成为人们从海量 Web 信息中快速找到所需信息的重要工具, 随着 Web 数据量的爆炸性增长, 传统的集中式搜索引擎已经越来越不能满足人们不断增长的信息获取需求. 随着对等网络(peer-to-peer, 简称 P2P)技术的快速发展, 人们提出了基于 P2P 的 Web 搜索技术并迅速成为研究热点[1]. 研究人员从不同的角度开始研究基于 P2P 的可扩展信息检索系统和数据管理系统, 构建基于 P2P 的搜索引擎, 如 Tang 等的 eSearch、Standford 大学的 SETS、复旦大学的 PeerIS 等试图建立基于 P2P 的全文检索系统, 而 ODISSEAt、Apoidea 以及 Loo 等提出基于有组织 P2P 网络构建基于 P2P 的分布式搜索引擎[2].

由于网络上存在着大量异构信息, 而 XML 可以有效解决 Web 数据集成和查询问题[3]. 为了充分的挖掘网络上存在的大量异构、动态和非结构化的数据, 本文采用 XML 格式对资源服务器的资源文件进行描述, 构建了一个基于 XML 的 P2P 的网络资源检索系统, 并采用当今比较流行的开源的全文检索工具包 Lucene 对该系统进行了实现, 为了便于用户随时随地搜索和下载资源的需要, 本系统还在基于 android 平台的移动客户端上进行了扩展.

1 基于P2P的分布式资源检索的系统架构

P2P 分布式资源检索是指从分布在各个独立节点上的网络信息资源中高效地索引、查找、检索出用户所需信息的过程[4]. 为便于说明, 本文定义: “拥有一定量的资源, 并且这些资源通过 web 方式可以供授权用户检索和访问的一个网络节点称为一个资源节点(或资源服务器)”. 在 P2P 网络中, 每个参与的节点既是服务器又是客户端, 既是信息的提供者又是信息的

① 基金项目:山东省高等学校青年骨干教师国内访问学者项目

收稿时间:2012-06-04;收到修改稿时间:2012-08-03

© 中国科学院软件研究所 http://www.c-s-a.org.cn

消费者. 本文结合应用实际, 采用基于 socket 通信的多线程并发机制, 设计了一种基于 P2P 的分布式网络架构. 如图 1 所示.

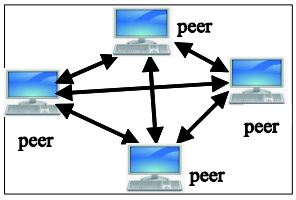

图 1 基于 P2P 的分布式网络架构

在每个资源节点上, 存储一定的各种类型的资源文件, 所有资源文件的信息通过一个统一格式的 XML 文档来存储, 每个资源节点通过开源工具包 Lucene 这一全文检索引擎对本地的 XML 资源文件建立一个索引文件, 这样就可以通过这个索引文件进行资源检索了.

具体的资源搜索过程为: 当用户登录 P2P 分布式网络中任意一台资源服务器, 在用户查询界面中输入要查询的信息后, 服务器首先对输入的查询信息进行分词并经过查询过滤模块进行过滤, 然后通过基于 socket 的多线程并发通信将要搜索的关键词提交给 P2P 网络中的其他资源服务器, 各资源服务器根据采用 Lucene 建立的本地索引文件进行查询, 并将查询结果返回给提交查询请求的客户端服务器, 客户端服务器对查询结果根据相关性大小排序后以列表的形式展现给用户.

为实现方便, 本文实现的系统各资源节点上存储了一个统一的服务器地址列表, 服务器启动时, 会首先进行通信确认在线服务器, 以后每隔一段时间会确认一次. 这样在进行分布式通信时, 首先对在线的资源服务器进行筛选, 获取所有在线服务器的 ip, 然后采用多线程并发连接向每台服务器发出请求, 并建立计时线程进行计时, 若能在规定时间内返回处理结果, 则由回调函数接收每个线程返回的结果, 否则返回空, 作为失败处理. 在规定的时间内, 如果某一台服务器返回的错误数太多, 则认为此服务器出现问题, 分布式网络会主动剔掉该服务器, 并且会在管理员登陆的时候给予提示.

2 关键技术分析及实现

2.1 资源存储

由于网络中存在大量的异构资源, 本文使用基于 XML 的资源存储方式, 便于对网络中异构资源进行统一管理. XML(eXtensible Markup Language, 可扩展标记语言)作为互联网联盟制定的一种通用语言规范, 是全新的描述结构化数据的语言, 它开放标准, 简单, 易于使用, 支持国际化, 与平台、工具、数据库、协议、编程语言无关, 易读易写, 也易于在网络中传播, 因此在描述资源和信息传输有着一定的优势. 而且, XML 可以将各种结构化、半结构化和非结构化数据集成起来, 在 XML 平台上整合应用, 使杂乱无章的信息海洋得到根本改善, 每个数据节点将实现信息有序存储, 并能为对方接受.

本文对描述资源信息的 XML 文档进行统一规范. XML 文档以 allresource 作为根元素, 表示该元素下的子元素描述了该资源节点全部资源的信息. allresource 元素下包含至少一个 resourceitem 元素, 表明了该资源结节点下的一个资源项. 在 resourceitem 元素之下, 又分别定义了该资源的各种描述信息, 如标题、关键字、文件类型等. 这样, 在每一个资源节点上, 节点的管理程序会随着其所属的本地资源的变动维护一个 XML 文档, 通过这一个 XML 文档, 可以得到这个资源节点所管理的所有资源的信息, 包括资源的描述信息和资源的地址信息. 因此, 对资源的检索过程也就转化为对这个 XML 文档内容的检索过程.

2.2 Luncen 全文检索引擎

本文采用了 Apache Foundation 下的全文检索工具 Lucene 作为检索平台, Lucene 是一个高性能的 java 全文检索工具包, 它使用的是倒排文件索引结构, 可以在它的上面开发出各种全文搜索的应用. Lucene 将所有的源码分为七个模块, 图 2 展示了其中的主要部分及运行流程[5].

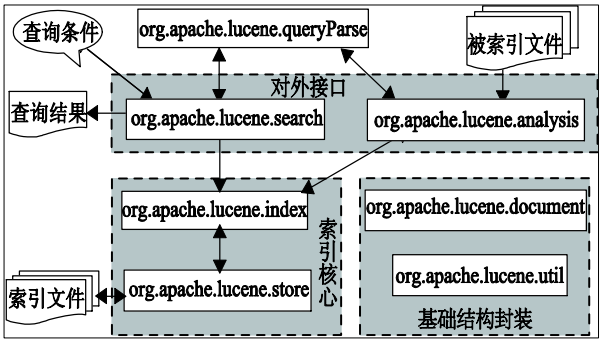

图 2 Lucene 逻辑结构图

© 中国科学院软件研究所 http://www.c-s-a.org.cn

2.3 中文分词技术

搜索引擎为了更好的辨别用户的需求，快速而准确的查找到用户需求信息，首先将输入的查询语句按一定的规则切分成有意义的词，即中文分词. 它关系到在搜索匹配索引时，词条是否能匹配上. 分词效果直接影响到查询效率. Lucene 自带的标准分析器 StandardAnalyzer 可以用于中文分词，但它是一元分词，机械地将一个汉字做为一个词元来切分的，尽管能保证查全率，但是速度慢而且连语义也会消失. 为此，本文设计的系统采用了 IKAnalyzer 分词器，它最初基于 Lucene2.0 版本 API 开发，新版本独立于 Lucene 项目，同时提供了对 Lucene 的默认优化实现. 采用了特有的“正向迭代最细粒度切分算法”，支持细粒度和最大词长两种切分模式，具有 83 万字/秒(1600KB/S)的高速处理能力、优化的词典存储和更小的内存占用. 针对 Lucene 全文检索优化的查询分析器 IKQueryParser，引入简单搜索表达式，采用歧义分析算法优化查询关键字的搜索排列组合，能极大的提高 Lucene 检索的命中率.

2.4 索引的创建

进行资源检索，首先要对存储在本地资源服务器中的 xml 资源文件建立索引. 建立索引时，本文首先采用 dom4j 技术对 xml 文件进行解析，解析 xml 依次得到文件中存储的每个资源文件的标题、关键字、文件类型、描述信息和资源的 URL 等信息；然后对利用 Lucene 提供的类 Document, Field, IndexWriter, Analyzer, Directory 等分别对上述信息依次进行建立索引，建立索引时按资源类别建立，大大提高检索效率. 索引创建核心代码如图 3 所示.

2.5 资源检索

当建立完索引后，就可以对索引进行搜索了. Lucene 的搜索过程原理如图 4 所示.

在得到用户输入的关键词后首先进行分词，对分词后的词条封装成 Lucene 能够识别的 Query，然后通过检索索引文件，查找匹配的词条. Lucene 提供了几个基础的类来完成这个过程，它们分别是 IndexSearcher, Term, Query, TermQuery, Hits. 可以利用 Query 类的子类 TermQuery, BooleanQuery, PrefixQuery 等把用户输入的查询字符串封装成 Lucene 能够识别的 Query. IndexSearcher 是用来在建立好的索引上进行搜索的. Hits 是用来保存搜索的结果的.

```
a.采用 dom4j 技术对 xml 进行解析
  //建立xml数据流，对xml进行遍历
SAXReader reader = new SAXReader();
File file = new File(RESOURCE_XML_PATH);
if(!file.exists()){……//报告错误}
  org.dom4j.Document document =
        reader.read(file);
  Element root =
document.getRootElement();
   // 迭代根元素下面的所有子元素
 for (Iterator i = root.elementIterator();
    i.hasNext();) {
      Element element = (Element)
i.next();
        ……}
b.根据分类创建不同的索引文件系统，将音乐，图片，视
频，文档，单独分离出来
//如果是音乐格式，将从XML中解析出来的标题关键字等
信息存储到类别为音乐的索引文件中
if(Music.indexOf(element.elementTextTrim(
"kind").toLowerCase())>0){
 Document musicDoc = new Document();
 musicDoc.add(new Field("title",
 element.elementTextTrim("title"),
 Field.Store.YES, Field.Index.TOKENIZED));
 …… }
//如果是其他类型的格式，同样的创建其他类型的索引文
件，便于只搜索该类型资源
 ……
//最后将所有类型文档资源创建到一个混合类型的索引
```

图 3　索引创建核心代码

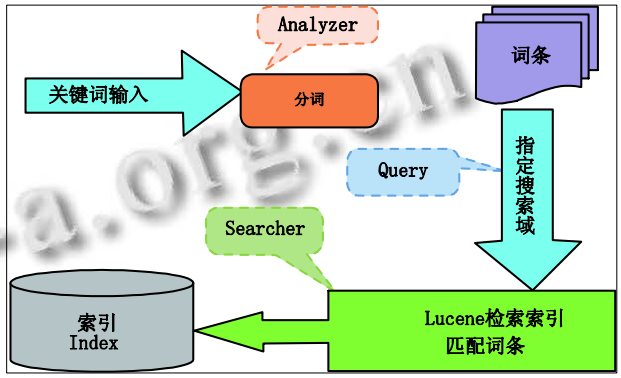

图 4　lucene 搜索原理

2.6 基于 android 客户端的搜索

为便于用户随时随地下载和预览搜索到的资源，本文实现的系统在基于 android 平台的移动客户端上进行了延伸，移动客户端搜索过程如图 5 所示.

移动客户端的实现原理同 PC 客户端，基于 android 客户端的搜索过程为：(1) 用户启动移动客户端，进行网络设置，设置登录的服务器的域名或 IP 地址及端口号，此时便可以通过 Internet 访问所登录的服

© 中国科学院软件研究所 http://www.c-s-a.org.cn

务器. (2) 当用户在资源界面中输入关键字进行查询时, 系统首先在登录的服务器中进行查询, 然后将查询信息通过多线程并发通信提交给 P2P 网络中的其他资源节点进行检索, 其他资源节点接收到查询信息后, 在本地检索并将结果返回给请求的客户端, 移动客户端登录的服务器根据某种排序策略将返回的结果显示在移动客户端.

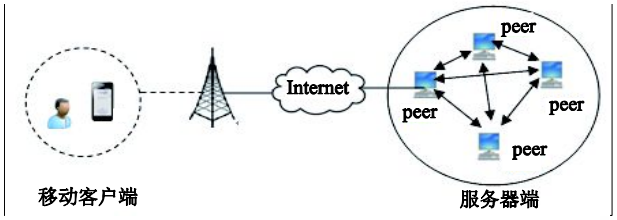

图 5 移动客户端搜索

在 android 客户端进行搜索, 很关键的一点是要考虑查询结果的处理和可视化, 本文移动客户端搜索界面采用了选项卡的形式实现资源的分类检索, 检索结果对不同的资源采用不同的图标, 并采用自动积累分页模式, 滚动到底部自动加载下一页. 界面底部菜单友好, 分别定向于下载管理页面, 设置页面和退出. 下载界面采用 android 优良的组件合理布局, 用户体验性强. 图 6 展示了 android 客户端的相关界面.

图 6 android 客户端的相关界面

3 系统功能及性能评价

3.1 功能评价

本系统设计时充分考虑了用户的体验及系统安全, 在 PC 客户端提供了用户注册模块, 用户注册为会员登录后可以进行界面的个性化设置和用户搜索偏好设置, 以实现个性化搜索功能. 为了便于系统管理, 设置了后台管理模块, 管理员登录后可以对用户、资源、词库和服务器等进行管理. 在 android 客户端开发了系统设置模块, 提供用户主服务器设置和备用服务器设置, 支持用户皮肤设置和自动更新版本功能.

在 android 客户端, 由于分布式网络可能会出现某台服务器不稳定的情况, 所以我们的客户端支持主服务器配置和辅服务器配置, 当主服务器因为出现故障, 客户端会自动的去连接辅服务器. 降低了客户端因为分布式网络中的服务器故障而导致客户端不稳定的概率. 同时自主开发的多任务多线程下载功能, 不卡机, 充分利用手持设备带宽. 友好的网络连接方式提示, 告诉用户当前连接方式是 3G, 还是 wifi, 还是 2G, 让用户合理选择, 避免失误导致超额流量费用.

3.2 性能评价

分布式通信作为系统的核心, 采用了多线程并发+回调函数机制+socket 通信等技术, 如果不考虑通信的网速延迟, 理论可以达到 10 毫秒完成并发 10 台服务器, 30 毫秒处理所有返回结果, 每台服务器平均占用内存 200K 左右(粗略测试), 并采用超时机制, 可以过滤掉不同服务器的不同缺陷导致的结果返回延迟. 而且分布式通信包含了复杂的负载均衡机制, 当一台服务器规定的用户并发数超过限制时, 就会将以后连接此服务器的用户定向于其他并发数较少的服务器. 解决了服务器因访问量不均衡造成的分布式网络不稳定现象.

文档解析采用 dom4j 技术, dom4j 对 xml 的操作速度很快, 可以在很短的时间内读取并解析 xml 文档, 取得 root 节点, 遍历 xml 树,字符串与 xml 之间的转换等. 分词采用了 IK Analyzer 分词器, 具有较高的分词效率和性能, 检索模块采用了 Lucene 全文检索引擎工具包, 并且对程序进行了极大的优化, 极大加快了索引建立速度, 为 200MB 左右的资源文件建立索引最快速度可以在 7 秒左右完成. 为了提高检索效率, 根据文件类型的不同, 建立不同的索引目录, 根据分类进行检索比对所有类型的文档进行检索在搜索效率上有很大提高. 而且利用索引进行检索时, 查全、查准率非常高, 速度也非常快, 可以达到毫秒级别, 前 100 条记录可以满足几乎所有用户的要求.

4 结语

本文实现的基于 xml 的 P2P 资源搜索原型系统, 可以充分挖掘网络上存在的大量的异构资源数据, 有效的消除网络上存在的“信息孤岛”. 为了能够方便用户随时随地下载和预览资源, 本文实现的系统在基于 android 的移动客户端上进行了扩展, 并具有良好的用

(下转第 82 页)

© 中国科学院软件研究所 http://www.c-s-a.org.cn

左到右，t 代表从上到下. 纹理图的左下角像素对应纹理图的原点(0,0),左上角对应点(0,1)，右上角对应点(1,1)，右下角对应点(1,0).

几何体的纹理坐标系和几何体的空间坐标系不相同，几何体的空间坐标系在三维空间确定几何体位置坐标系 x、y 和 z，一旦几何体移动，它的坐标系就会发生相应的变化，缩放、旋转时也是如此. 而纹理坐标测量的是纹理的重复，无论怎样放大、缩小，原点(0,0)总在左下角，右上角的点总是(1,1). 点(0,0)和点(1,1)之间是纹理的一个重复，若使用纹理的多个重复，就是相当于把纹理的许多拷贝逐个拼起来，每一个拷贝使 s 轴或 t 轴做标加 1，若在反方向则减 1.

对于复杂的几何模型映射纹理，直接映射纹理很难到达到设计要求. 例如窗帘的制作，窗帘的形状是波浪形，如果直接映射纹理效果很差. 使用 VRML 中的节点 TextureCoordinate 可以很好的完成对窗帘的纹理映射.

TextureCoordinate 节点的语法如下.

```
TextureCoordinate {
        ExposedField  MFVec2f  point []
}
```

TextureCoordinate 节点仅含一个 point 域，通过它标记纹理图中关键的点，使用时将纹理图的每一个点指定到要映射的模型的相应的顶点上. Texture Coordinate 的一个应用是对一个物体的一个面进行纹理映射.

本模型中对壁画的设计运用了节点 Texture Coordinate，这样就仅仅在壁画框的正面映射纹理贴图，不至于将该贴图纹理同时映射到壁画框上，从而保持了壁画框的木质材质.

1.3 卧室模型的结果

通过基础设计和在材质和纹理上的改善,得到了如下的卧室模型.

图 8 卧室效果图

2 结论

虚拟现实技术的沉浸感(Immersion)、想象性(Imagination)和交互性(Interaction)具有重要的现实意义.

本文基于虚拟现实的 VRML 技术设计了一个虚拟卧室模型系统，其中包括对各种虚拟物体的设计，并着重论述了纹理和材质技术在增强虚拟物体真实感方面的应用. 针对该模型，进一步的工作将是在优化其代码的基础上改进光照效果及动态交互，以增强其真实效果，更接近现实生活.

参考文献

1 吴小华,等.构建个性化网络虚拟世界—RML 从入门到精通.北京:国防工业出版社,2002.1–5.
2 数虎图像网. http://www.cgtiger.com.2012
3 王琦.3D STUDIO MAX 三维动画大制作(第一部).北京:宇航出版社,1997.416–420.

(上接第 61 页)

户体验. 实践证明，本文提出的解决方案不仅解决了分散信息检索的难题，而且可大大提高信息检索的工作效率，具有极好的应用价值和广阔的应用前景，如对于整合某个城市或社区的图书资源信息，建立统一的图书资源搜索平台有重要的借鉴意义.

参考文献

1 方启明,杨广文,武永卫,郑纬民.基于 P2P 的 web 搜索技术.软件学报,2008,19(10):2706–2719.
2 傅向华,明仲.基于 P2P 的个性化 Web 搜索系统的设计与实现.计算机工程与应用,2007,43(7):111–113.
3 唐永平.利用 XML 技术解决 Web 数据挖掘中数据异构的问题.计算机时代,2010(9):4–6.
4 韩毅.P2P 网络信息检索的研究进展.现代图书情报技术,2007(7):36–40.
5 李颖,李志蜀,邓欢.基于 Lucene 的中文分词方法设计与实现.四川大学学报(自然科学版),2008,45(5):1095–1099.
6 钱春江.Lucene 全文搜索引擎的应用.上海:上海理工大学,2009.
7 索红光,孙鑫.基于 Lucene 的中文全文检索系统的研究与设计.计算机工程设计,2008,29(19):5083–5086.
8 Manning CD, Raghavan P, Schutze H.信息检索导论.王斌译.北京:人民邮电出版社,2010.

© 中国科学院软件研究所 http://www.c-s-a.org.cn