基于内容图像检索的 Web 搜索器

李爱国，　白　冰

(西安科技大学计算机科学与技术学院　西安 710054)

摘要：构建图像 Web 搜索器是实现基于内容 Web 图像搜索引擎系统的关键，运行搜索器可为引擎系统提供 Internet 上的图像数据源. 提出了一种基于内容的 Web 图像搜索引擎的 Web 搜索器，它从初始 URL 网址集出发，根据广度优先遍历策略来获取新的 URL 网址以及互联网中图像文件信息，并建立 Web 图像特征库. 实验表明，当搜索器在设置较多父 URL 图像网址以及适合的遍历层数时，可获得图像信息检索的最佳性能，可获得较多数量网络图像信息.

关键词：图像 Web 搜索器；搜索引擎；URL 遍历；基于内容的图像检索

中图分类号：TP 391　　**文章编号**：1671－6841(2009)02－0060－03

0　引言

搜索引擎(Search Engine)是随着 Web 信息的迅速增加，从 1990 年开始逐渐发展起来的技术[1]. 搜索引擎以一定的策略在互联网中搜集、发现信息，对信息进行理解、提取、组织和处理，并为用户提供检索服务，从而起到信息导航的目的. 当前，搜索引擎主要是输入文字关键词，查寻文字资料和图像等信息；而当图像内容无法通过关键字描述，即图像文件本身的文件名描述不准确时，便很难进行检索. 图像 Web 搜索器在延续现有功能的基础上，结合使用示例图像进行搜索，从而找到与此图相似的互联网中的图像.

1　图像 Web 搜索器工作原理

搜索引擎执行搜索操作时并不真正搜索 Internet，而是搜索预先整理好的网页索引数据库. 引擎并非真正理解网页上的内容，它只是机械地匹配网页上的文字或图像信息[2-3]. 从用户使用角度，图像搜索引擎分为离线和在线两部分.

离线部分主要进行互联网中图像信息的采集. 首先图像 Web 搜索器创建信息采集任务，执行该任务到远程 Web 站点服务器端，在“本地”对下载的图像数据进行处理，如建立文本主题索引和图像内容索引等，最后将提取后的图像特征数据从远程服务器端传送到图像搜索引擎服务器端的 Web 图像特征库.

在线部分主要提供用户交互的接口. 用户通过 Internet 连接到本图像搜索引擎，采用用户接口模块对查询图像进行特征提取，然后提交给检索器，检索器将提取后的图像特征数据与 Web 图像特征库进行匹配，最后将排序后的结果返回给用户.

2　图像 Web 搜索器的开发

基于内容的 Web 图像搜索引擎在查找图像文件时，采用 IMG SRC 和 HREF 2 个 HTML 标签检测是否存在图像文件[4-5]. IMG SRC 表示“显示下面的图像文件”，导向的是嵌入式图像；HREF 则表示“下面是一个链接”，导向的是被链接的图像. 引擎通过检查文件扩展名判断是否为图像文件. 如果扩展名是“JPG” 或“BMP”等，即为图像文件. 图像 Web 搜索器是一个自动提取网页，解析下载图像文件，并提取图像特征信息的

收稿日期：2008－11－25
基金项目：陕西省科技攻关项目，编号 2008K01-58；陕西省教育厅自然科学专项计划项目，编号 07JK314.
作者简介：李爱国(1966－)，男，教授，博士，主要从事数据挖掘、信息融合、软件测试研究，E-mail：liag@xust.edu.cn.

程序.

2.1　图像搜索策略

图像 Web 搜索器的执行分 3 步:首先在 Internet 上攫取网页,建立 URL 网址索引库;然后将索引库中 URL 网址流文件中所包含的图像文件进行解析下载,并提取图像特征存入 Web 图像特征库;最后利用本地基于内容图像检索策略,从 Web 图像特征库中检索出 Internet 上的相似图像,并按相似度大小排序,返回给用户.图像搜索引擎系统的核心是构建 Web 图像特征库,构建过程如图 1 所示.构建了 Web 图像数据库后,便可进行图像的查找和相似匹配,完成基于内容的图像信息的检索.

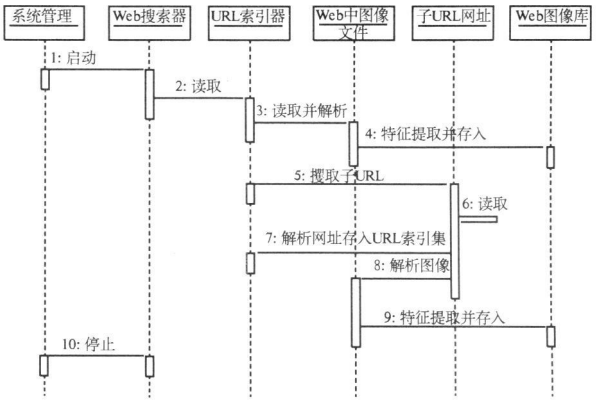

图 1　Web 图像特征库时序图

Fig.1　Sequence diagram of Web image feature DB

系统管理程序启动图像 Web 搜索器,开始 Internet 中图像搜索操作.首先读取 URL 索引集中网址,以流文件形式处理 URL 源文件,并将源代码中的图像文件内容进行解析、过滤并下载,特征提取与量化之后,将特征存入 Web 图像特征库.在读取 URL 源文件信息过程中,同时可解析该 URL 网址中的子 URL 网址信息.子 URL 网址的操作同父 URL 网址一样,一方面可将 URL 网址存入到 URL 索引集中,另一方面解析相关的 Web 图像信息,并在处理后存入 Web 图像特征库.

2.2　搜索器开发步骤

(1)URL 网址攫取策略

互联网中 URL 网址采集步骤为:从一个初始的 URL 集出发,将这些 URL 全部放入一个有序的待处理队列中[6].而 URL 采集器从这个队列里按顺序取出 URL,通过 Web 上 HTTP 协议获取 URL 所指的页面,将其以文件流的形式进行下载,然后读取这些已获取的页面并按照一定的策略提取新的 URL,将获取的 URL 信息入库.如此反复进行,直到人工停止或依据某种策略停止.

攫取网页的时候,搜索器有 2 种策略:广度优先和深度优先[7-8].广度优先指 Web 搜索器首先攫取起始网页中链接的所有网页,再选择其中的一个链接网页,继续攫取在此网页中链接的所有网页.这个方法可让 Web 搜索器并行处理,提高其攫取速度.深度优先是指 Web 搜索器会从起始页开始,链接相互跟踪下去,处理完这条线路之后再转入下一个起始页,继续跟踪链接.由于 Internet 中的页面是海量数据集的,因此,图像 Web 搜索器在对图像库网站遍历时设置了访问层数,采用了广度优先遍历策略进行 URL 网址搜索.

将 URL 定义为一个对象,按照对象的属性建立 URL 索引库.以 URL 字段作为插入关键字,以此来限定 URL 的遍历层数.当前网页大多采用半结构化的 HTML 超文本语言进行定义,由于 HTML 语言本身不规则以及不同用户开发时编写 HTML 语言代码的不规范性,造成对其 URL 解析时需要不同的策略.攫取网页中的 URL 采用 Java 语言所提供的正则表达式匹配.如 String regEx＝“href＝\”([^\“]＊)\”.

(2)生成 Web 图像特征库

建立 Web 图像特征库,首先读取 URL 索引库中的网址,然后从网址流文件中解析出图像信息,根据检索规则对图像文件进行过滤,然后提取过滤后的图像特征,最后将提取的图像特征插入到 Web 图像数据库.在图像 Web 搜索器的开发过程中,图像文件的下载与特征提取实际上是和 URL 攫取同步的.在建立 Web 图像特征库时需要指定 Internet 中的图像地址,因此,存入 Web 图像特征库中的 URL 网址应该是该图像信息的绝对地址.

由于 HTML 本身的半结构化,造成了图像路径描述也是多种多样的.因此,图像文件的路径需要转化为绝对路径.如:URL 地址 http://www.wallcoo.com/nature/index.htm 中的＜IMG SRC＝“图像名称”/＞标签;指定了一幅图像的相对路径.设 URL＝http://www.wallcoo.com/nature/index.htm;A＝http://www.wallcoo.com/;B＝http://www.wallcoo.com/nature/,在解析图像绝对路径时采用的策略如表 1 进行转换,可得到 Internet 中图像文件的绝对路径.

3 搜索器性能评测

实验目的是测试图像 Web 搜索器在运行时的稳定性以及在图像搜集方面的最优性能.所开发的图像 Web 搜索器在奔腾 CPU 2.4 单独主机、Windows 操作系统以及 512 M 内存的环境中经过数小时运行后,状态良好,没有内存溢出.实验采用"相同个数父 URL 网址,不同遍历层次"进行了搜索器检索性能评测.其中,父 URL 均采用了图像网站的网址.以"相同个数父 URL 网址,不同遍历层次"进行搜索器性能的评测.启动搜索器后,通过锁定 3 个父图像网站 URL 网址,设定不同遍历层数,测试一个小时后,检索到 Internet 图像个数以及遍历的子 URL 个数,结果如表 2 所示.

表 1 图像文件映射绝对路径

Tab.1 Image file path mapping

图像名称	Web 中图像绝对路径
b.jpg	*B*+b.jpg
a/b.jpg	*A*+a/b.jpg
../b.jpg	*A*+b.jpg
URL.jpg	全路径 URL.jpg

表 2 相同个数父 URL 网址不同遍历层次下搜索测试结果

Tab.2 Results of the same number URLs and different invite level

父 URL 个数	遍历层数	图像个数	遍历子 URL 个数
3	不设层	3 579	1 890
3	10	6 678	793
3	3	11 157	512

由表 2 可知,当父 URL 网址个数设置为 3 个时而设置遍历层数较少时,搜索器所检索到图像数目较多,而遍历到的子 URL 网址个数则较少.如:遍历层次为"不设层"时,可获得 3 579 个图像文件,1 890 个 URL 网址;而当遍历层次为"3"时,可获得 11 157 个图像文件,512 个 URL 网址.

通过实验可知,网址遍历层数设置较多时,当搜索器对指定父 URL 图像网址遍历完全后,搜索器会自动迁移到其他的非图像网站中.由于非图像网站图像信息资源不丰富,相对 3 层遍历,层次越深反而检索到的图像文件个数相对较少.由以上分析可知,为了在较短的时间内检索到较多的 Internet 图像信息,图像 Web 搜索器需要设置较多父图像 URL 检索源网址以及适合的遍历层数.

4 结论

通过搜索引擎核心技术网络爬虫的研究,结合图像文件 Internet 中解析与 URL 网址遍历技术,开发出了图像搜索爬虫——图像 Web 搜索器.通过实验分析可知,搜索器运行时具有良好的稳定性,并在设置较多父图像 URL 网址以及适合的遍历层数的情况下,可获得图像检索的最佳性能.选取 10 个父图像网址,3 级广度优先遍历层次进行了一个小时搜索实验后,为引擎系统提供了 11 000 多条 Internet 图像数据源.

参考文献:

[1] 张卫丰, 徐宝文, 周晓宇,等.Web 搜索引擎综述[J].计算机科学,2001,28(9):24-28.

[2] Liu T Y, Yang Y, Wan H, et al. An experimental study on large-scale Web categorization[C]//Proceedings of the 14th International World Wide Web Conference. Chiba, Japan,2005:1106-1107.

[3] Shen X, Dumais S, Horvitz E. Analysis of topic dynamics in web search[C]//Proceedings of the 14th International World Wide Web Conference. Chiba, Japan, 2005:1102-1103.

[4] 章勤,余洋,陶文兵.图像搜索中基于网页分块的图像分类研究[J].计算机工程与科学,2007,29(6):42-45.

[5] 陈福集,杨善林.一种基于 KDD 的 Web 搜索引擎框架[J].情报学报,2002, 21(3):264-268.

[6] 王继成,潘金贵,张福炎.Web 文本挖掘技术研究[J].计算机研究与发展,2000,37(5):1051-1057.

[7] 潘春华, 冯太明, 武港山. 基于移动爬虫的专用 Web 信息收集系统的设计[J].计算机工程与应用,2003,36:153-156.

[8] 徐远超,刘江华,刘丽珍,等.基于 Web 的网络爬虫的设计与实现[J].微计算机信息, 2007,23(7):119-121.

(下转第 72 页)

(GGA), and the lowest energy structures as well as metastable isomers are determined. The calculated results indicate that $AgAu_n$ clusters adopt planar stuructures as their ground state geometries for $2 \leqslant n \leqslant 7$, and the stable structures can generally be obtained by edge-capping an atom on the structures of smaller clusters. The onset of three dimensional lowest-energy structures starting from $AgAu_8$ indicates a 2D-3D transition around the size of $n=8$. The calculated results on the averaged binding energy, second-order difference of cluster energies, the vertical ionization potentials and HOMO-LUMO gaps indicate that the $AgAu_n$ clusters with odd n exhibit stronger stability than those with even n, and moreover, and the stability is more prominent for $AgAu_5$.
Key words: $AgAu_n$ clusters; density functinoal theory; geometrical structure and electronic property; stability

(上接第62页)

Web Searcher for Content-Based Image Search

LI Ai-guo, BAI Bing
(*School of Computer Science and Technology, Xi'an University of Science and Technology, Xi'an 710054, China*)

Abstract: Image Web searchers play a key role in content-based image search engine systems. Using image Web searchers, images data-sources on internet can be provided for content-based image search engines. The design of proposed image Web searcher begins with initialized URL sites aggregate. According to the breadth-first strategy, the searcher can get new URL sites and image files on internet, and construct Web image feature database. Experimental results indicate that the Web searcher can get optimal performance and get more image files data on Internet, when the searcher sets more father URL image sites and adapted depth of visitation.
Key words: image Web searcher; search engine; URL traversal; content-based image search engine