

基于 Internet 的图像检索

王 兵,王 汉,李 悦,王 瑞,黄 啸,汤 进

(安徽大学 计算机科学与技术学院,安徽 合肥 230601)

摘 要:随着 Internet 的迅速发展,在查找图像信息时因其信息资源量大而不能准确地找到所需的图像信息,所以一种基于 Internet 的图像搜索引擎也就应运而生了。提出利用爬虫技术实现可定位的图像搜索与获取,并分析和比较了基于内容和关键词的图像检索方法,讨论了上述方法在图像检索应用中的问题,进而提出将上述两种方法相结合的检索策略。与其它传统图像检索引擎相比,实验结果表明提出的策略能实现图像的定位检索,并有效提高图像检索的准确率。

关键词:URL;广度搜索;图像语义;图像特征;图像检索

中图分类号:TP31

文献标识码:A

文章编号:1673-629X(2011)05-0033-03

Internet-Based Image Retrieval

WANG Bing, WANG Han, LI Yue, WANG Rui, HUANG Xiao, TANG Jin

(School of Computer Science and Technology, Anhui Univ., Hefei 230601, China)

Abstract: Along with the rapid development of the internet, do not accurately retrieve the information due to the large amount of image information resources. So the internet based image search engines have emerged. The technology of WebSpider is proposed to locate, search and obtain images. As the foundation of this paper, the content and the keywords based image retrieval methods are analyzed and compared. In addition, the problems of their applications in image retrieval are discussed. Therefore, the combined strategy of these two methods is proposed for retrieval. The experimental results show that this image retrieval strategy can achieve image location and retrieval, and the retrieval accuracy outperforms other traditional image search engines.

Key words: URL; BFS; image semantics; image characteristics; image retrieval

0 引 言

图像检索^[1]是近年信息科学领域重要研究的热点,根据研究对象的不同形成多种检索技术。目前,在多媒体网络中,有两类图像检索技术^[2]正在研究和应用之中:①采用传统的基于关键词的图像检索技术;②采用基于内容的图像检索技术。

传统的基于关键词的图像检索是根据图像的文件名、图像所在网页的标题、图像附近的文本等一些与图像有关的因素来确定图像的内容^[3],而不是抽取图像本身的外部特征如颜色、纹理和形状,或者从更高的语义层次来获取图像的内容,由此确定的图像内容可能与图像的实际内容存在着偏差。为克服基于关键词的图像检索存在的问题,提出了基于内容的图像检索技术(Content-Based Image Retrieval, CBIR)^[4],其研究

内容主要是基于视觉特征的提取,基本方法是从图像中自动提取若干底层视觉特征(如颜色、纹理、形状、轮廓等),通过比较这些特征的相似度来获得检索结果。如经典系统有 Virage^[5]、VisualSEEK^[6]、QBIC^[7]等,这类方法在某些应用场合得到很好的效果,但仅仅从底层特征反映查询者的意图又容易产生语义上的歧义,因此,需要检索系统在此基础上再融合进关键词语义检索。

文中正是着眼于解决现有一般图像检索系统的缺点问题,提出了一种基于 Internet 的图像检索策略,该策略首先利用爬虫技术实现可定位的图像及其关键词语义的获取,然后再提取其图像特征,构建网络图库,进而通过采用基于关键词和内容相结合的图像检索策略——关键词和内容双重过滤的思想,实现基于 Internet 的图像检索。

1 总体设计和实现

本系统首先运用爬虫技术^[8~10]实现可定位的图像获取模块,从而为检索图像库提供了数据源。对于图像检索模块,本系统运用基于内容和关键词的两种

收稿日期:2010-09-13;修回日期:2010-12-21

基金项目:国家自然科学基金(60772122);国家大学生创新性实验项目(081035719);安徽省教育厅自然科学研究产学研重点项目(KJ2010A006)

作者简介:王 兵(1987-),男,河北衡水人,研究方向为图像处理和识别;汤 进,副教授,研究方向为图像处理与模式识别。

图像检索方法相结合的策略来实现。系统设有两个图像库:一个是网络图像库,此图像库来源于 Internet;另一个是标准图像库,此图像库用来进行二次图像检索过滤。本系统在用户检索时,提供相关反馈^[11,12]功能,扩充完善标准图像库,从而进一步提高用户检索准确率。其系统结构图如图 1 所示。

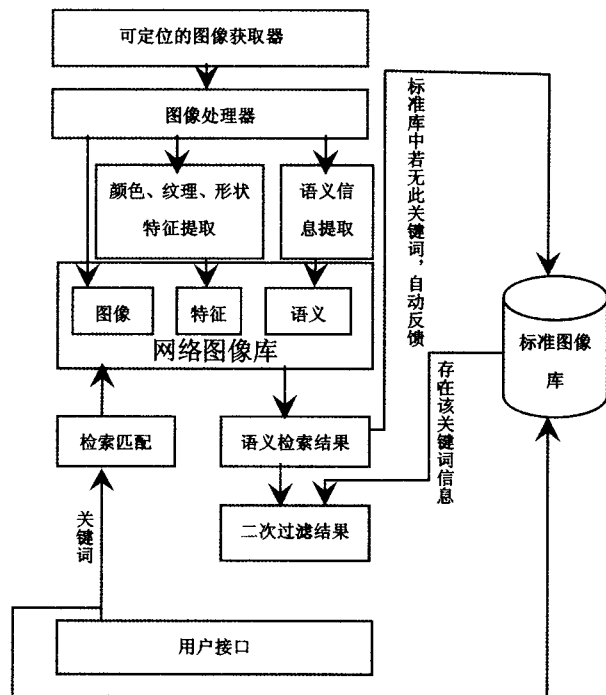


图 1 系统结构图

2 详细设计和实现

2.1 页面提取

本系统的页面提取模块是通过一个给定的初始 URL,把这个初始 URL 放到 URL 处理器中;网页读取器根据 URL 处理器提供的这个初始 URL,解析 URL 中标明的 Web 服务器地址、建立连接、发送请求和接收数据,采集到相应的网页;经过去重检测后,通过 URL 提取器从网页中提取出新的 URL 放入 URL 处理器;并将其存入数据库中;由标签信息获取器获取相应的标签,分两个方向分别将其存入 URL 处理器中和数据库中保存;如此反复采集、处理数据,直到网页读取器要求停止(URL 处理器中 URL 集合为空)为止。一般来说页面内容提取过程是模拟人浏览网站的过程进行对网页采集的。从网站某一个页面(通常是首页)开始,读取网页的内容,找到在网页中的其它链接地址,然后通过这些链接地址寻找下一个网页,这样一直循环下去,直到把这个网站所有的网页都爬取完为止。按这个原理来,如果把整个互联网当成一个网站,那么本系统就可以用这个原理把互联网上所有的网页都下载下来。其工作原理结构图如图 2 所示。

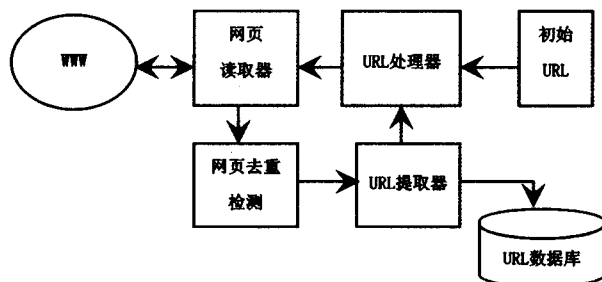


图 2 页面提取工作原理图

2.1.1 URL 处理器

URL 处理器是根据先进先出的策略向网页读取器分配 URL。URL 采用一个动态数组来实现存储,在 C#.NET 环境下可以实现对 URL 的快速处理。URL 处理器主要有两个数据来源:1) 初始的种子 URL,如上图所示;2) 从 URL 提取过来的 URL 集,它们是从已经读取到的页面中抽取出来并经过处理的。

2.1.2 网页读取器

通过各种 Web 协议来完成资料的采集。一般来说协议包括 http、ftp 等。但从主流上看,仍以 http 为主。根据分配的 URL 通过各种 Web 协议来爬取页面并读取页面内容。

2.1.3 网页去重检测

网络上的资源,网页中的内容经常被其他网站、网页引用,本系统模块找到的网页有很多重复的,如不进行网页重复内容的检测过滤,将极大地浪费了网络带宽和系统的运行效率。因此,重复内容检测是网络蜘蛛中的重要组成部分。本系统以 Hashtable 类建立一个全局对象,保存已经爬行的 URL 地址,分析获得的网页,提取出其中的 URL 与 Hashtable 里的 URL 进行比较,若检测出 URL 还没有爬取,则再对此 URL 进行爬取。简而言之,即本系统采用了哈希去重法。

2.2 图像获取

本模块直接调用上一模块所得到的 URL 数据库中的 URL 集合,把这些 URL 存储在一个动态数组中,按顺序对这些 URL 进行解析,运用正则表达式匹配得到每个 URL 中所包含的有效的图片 URL 链接,并提取每个图片 URL 链接的标签信息,对其进行处理,得到图片的语义信息,并将图片 URL 链接和其语义信息对应着存入数据库中,方便对其进行索引。本模块的基本结构如图 3 所示。

2.3 图像语义信息、内容特征双重过滤检索

本系统的核心模块就是该部分。首先在本地有一个标准的图像库,每个常用的语义都对应着多张标准图片,例如“大海”,在本地库对应着 20 张图片,然后运用图像的 RGB 模型理论提取每张图片的颜色特征值。

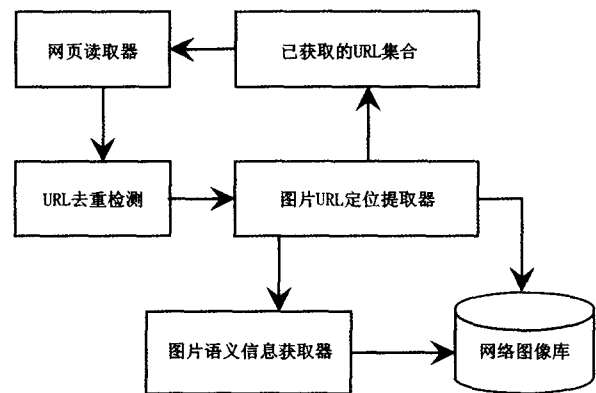


图3 图像获取原理结构图

由此可获得“大海”这一语义对应的颜色特征值的范围。

运用同样的流程来获得其纹理和形状的特征值范围。

当用户提交“大海”这一语义信息来检索图片时,本系统先通过图片对应的语义信息进行语义模糊查询,得到符合语义条件的图片后,再获取这些图片的颜色、纹理、形状特征值,然后再通过由本地标准库产生的“大海”这一语义对应的颜色、纹理、形状特征值范围进行二次过滤,最后把经过双重过滤检索的图片结果反馈给用户。

本模块的原理结构如图4所示。

3 实验及结果分析

本系统运行环境:Windows XP 及以上版本操作系统,需要.NET Compact Framework3.5及以上版本和SQL Server2008 服务器组件的支持;开发工具为 Visual Studio 2008。

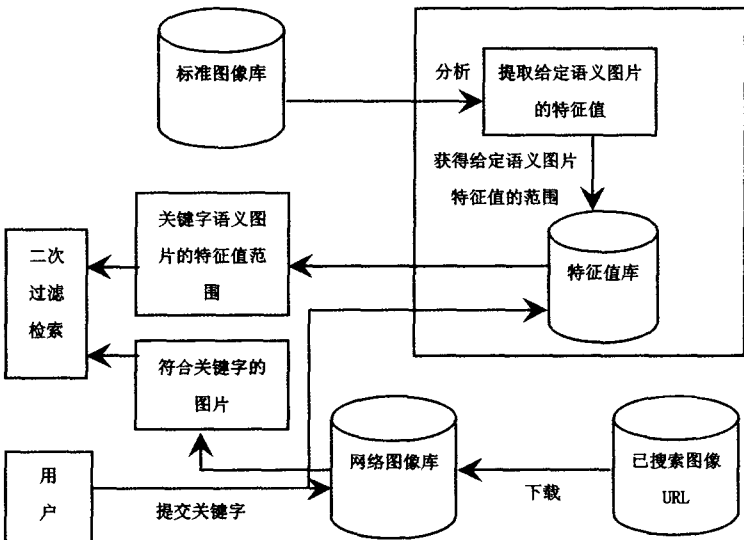


图4 双重过滤检索原理结构图

(1)本系统先分别对搜狐、新浪、网易、中华网、安徽大学首页各检索1分钟,然后对中华网分别检索1,3,5,10分钟,获得的有效连接数和含语义的图片数如图5所示。

(2)分别以关键词“战斗机”、“中国人”和“龙”为例进行双重过滤检索的结果如表1所示。此时网络图像库为3000张图片,标准图像库为100张图片。

表1 双重过滤检索结果

关键词	战斗机	中国人	龙
检索出的图像数(张)	11	31	19
图像内容、语义错误率	0	3.2%	0

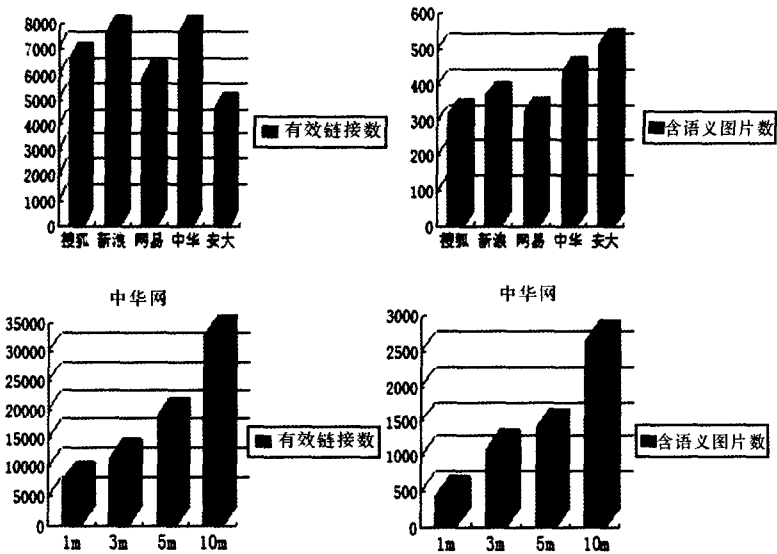


图5 定位检索测试效果图

通过上述仿真测试表明:本系统与传统的百度等图片搜索引擎相比,具有可定位图像搜索功能,从而很好的方便用户自定义图像源;本系统的基于内容和关键词相结合的图像检索策略得到了很高的图像检索准确率。

4 结束语

文中提出了一种基于 Internet 的图像检索系统实现策略,详细介绍了整个基于 Internet 的图像检索系统的运行原理,并进一步给出了这个系统的实现方法。

提出了一种新的综合图像语义和图像内容特征的图像检索方法。这种新方法简单易行,能同时体现图像的内容特征和语义信息,提高了图像检索的准确度。

虽然本系统在提高图像检索精度方

(下转第39页)

基本稳定在0.8左右,当启动DDoS攻击时,Hurst指数从0.8迅速下降到0.3。也就是说,当发生DDoS攻击时,Hurst指数将有较为明显的变化。从Hurst值的变化,就可以检测到是否发生了DDoS攻击。

4 结束语

选取Daubechies3小波对网络的流量数据(时间序列)进行小波分解,可以得到较为平稳的时间序列——近似分量和细节分量,进而求得自相似指数Hurst值,通过Hurst值的变化可以将繁忙时的正常业务流量同DDoS攻击时的异常流量区别开来,这为网络流量异常检测提供了新的思路。

参考文献:

- [1] Leland W, Taqqu M, Willinger W, et al. On the Self-Similar Nature of Ethernet Traffic [J]. IEEE/ACM Transactions on Networking, 1994, 2(1):1-15.
- [2] Paxson V, Floyd S. Wide area traffic: the failure of Poisson modeling [J]. IEEE/ACM Transactions on Networking, 1995, 4(1):226-244.
- [3] Paxson V, Veitch D. Wavelet Analysis of Long-Range-Dependent Traffic [J]. IEEE Transactions on Information Theory, 1998, 44(3):2-15.
- [4] Crovella M E, Bestavros A. Self-similarity in World Wide Web traffic: evidence and possible causes [J]. IEEE/ACM Transactions on Networking, 1997, 5(6):835-846.
- [5] Willinger W, Taqqu M S, Sherman R, et al. Self-similarity

through high-variability: statistical analysis Ethernet LAN traffic at the source level [J]. IEEE/ACM Transactions on Networking, 1997, 5(1):71-86.

- [6] Duffy D E, McIntosh A A, Rosenstein M, et al. Statistical analysis of CCSN/SS7 traffic data from working CCS subnetworks [J]. Selected Areas in Communications, 1994, 12(3):544-551.
- [7] Garrett M W, Willinger W. Analysis, modeling and generation of self-similar VBR video traffic [C] // In: Proc. of the ACM SIGCOMM 94. London: ACM Press, 1994:269-280.
- [8] 王西峰,高岭,张晓李. 自相似网络流量预测的分析和研究 [J]. 计算机技术与发展, 2007, 17(11):42-45.
- [9] 淮文军,王明芳,汪梅. 基于小波分析的电缆故障特征提取方法研究 [J]. 计算机技术与发展, 2007, 17(11):209-211.
- [10] Beran J, Sherman R, Taqqu M S, et al. Long-range dependence in variable-bit-rate video traffic [J]. IEEE Transactions on Communications, 1995, 43(234):1566-1579.
- [11] 李永利,刘贵忠,王海军,等. 自相似数据流的Hurst系数小波求解法分析 [J]. 电子与信息学报, 2003, 25(1):100-105.
- [12] Boggess A, Narcowich F J. 小波与傅里叶分析基础 [M]. 北京:电子工业出版社, 2004.
- [13] Ethernet Traces of LAN and WAN Traffic [EB/OL]. 2008. <http://ita.ee.lbl.gov/html/contrib/BC.html>.
- [14] 罗关春,林夏,卢显良,等. 一种新型的基于网络流量自相似的DDoS入侵检测方法 [J]. 计算机科学, 2003, 30(12):54-58.

(上接第35页)

面有着不错的效果,但怎样进一步提升图像检索的准确度和检索速率,有待进一步研究和探索。

参考文献:

- [1] Datta R, Li J, Wang J Z. Content-based image retrieval—approaches and trends of the new age [C] // In Proceedings of the Seventh International Workshop on Multimedia Information Retrievals. Singapore: [s. n.], 2005:253-262.
- [2] 王学松,周明全,耿国华. 互联网www图像搜索引擎的研究与设计 [J]. 小型微型计算机系统, 2003, 24(7):1161-1164.
- [3] 侯东京,侯英梅. 在互联网上检索图像信息的方法 [J]. 现代情报, 2005(8):77-81.
- [4] 黄祥林,沈兰利. 基于内容的图像检索技术研究 [J]. 电子学报, 2002, 30(7):1065-1071.
- [5] Gupta A, Jain R. Visual information retrieval [J]. Communications of the ACM, 1997, 40(5):71-79.
- [6] Smith J, Chang S. VisualSEEK: a fully automated content-based image query system [C] // In ACM Multimedia. Boston, Massachusetts: [s. n.], 1996:87-98.

- [7] Flickner M, Sawhney H, Niblack W, et al. Query by image and video content: the QBIC system [J]. Computer, 1995, 28(9):23-32.
- [8] 杨学明,刘柏嵩. 基于本体的网络爬虫技术研究 [J]. 情报学报, 2007, 26(5):723-727.
- [9] Davulcu H, Koduri S, Nagarajan S. Datarover: a taxonomy based crawler for automated data extraction from data-intensive websites [C] // Proceedings of the 5th ACM international workshop on Web information and data management. [s. l.]: [s. n.], 2003.
- [10] Aggarwal C C. Collaborative crawling: mining user experiences for topical resource discovery [C] // Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining. [s. l.]: [s. n.], 2002.
- [11] 张毅,赵捧未,刘怀亮,等. 基于语义的图像检索相关反馈技术 [J]. 情报杂志, 2006(10):43-44.
- [12] 罗焯,汤进,罗斌. 一种基于检索结果集的图像检索相关反馈算法 [J]. 计算机工程与应用, 2007, 43(11):69-71.