中国科技资源导刊 ISSN 1674－1544
2009年11月 第41卷第6期 37－40
CHINA SCIENCE & TECHNOLOGY RESOURCES REVIEW
ISSN1674-1544 Vol 41 No 6 37-40,Nov. 2009

Web多媒体资源搜索与相关文本提取研究

于文超 刘 菲
(山东师范大学传播学院,山东济南 250014)

摘 要:Web多媒体网页中多媒体资源的相关文本对于描述Web多媒体资源具有重要意义,利用Web多媒体网页搜集器搜集网络中包含多媒体资源的网页,对网页进行区域分析。根据多媒体资源所在网页中的嵌入形式,设计Web多媒体资源相关文本信息提取系统,准确提取Web页面中多媒体资源的相关文本。实验结果表明,该系统提取Web多媒体资源的相关文本准确率较高,有助于提高多媒体信息检索系统的查全率与查准率。

关键词:网页搜集;区域识别;文本提取
中图分类号: TP391.4 **文献标识码**: A DOI: 10.3772/j.issn.1674－1544.2009.06.007

1 引 言

针对Web多媒体资源的检索主要有基于内容的多媒体检索与基于文本的多媒体信息检索两种方式[1],其中基于内容的多媒体信息检索是当前的研究热点,它通过分析图像、视频、音频等多媒体资源,提取能反映该多媒体资源的特征和语义信息,再利用这些内容建立索引库,但多媒体低层特征与高层语义之间存在鸿沟,其低层特征往往不能真正反映其语义信息,二者之间的映射与转换尚无统一标准,因此利用该方法检索的效果往往不理想。故在对多媒体资源的检索很多时候还要依靠基于文本的多媒体检索方式,或者两种方式相结合,而能否准确提取Web中多媒体资源的相关文本来描述相应的多媒体资源对检索的查准率和查全率具有重要影响[2]。

目前,国内外对多媒体相关文本自动提取技术的研究一般集中应用于Web图像领域,传统的效率低下的手工提取多媒体相关文本的方式已经不能满足人们的需求。新加坡国立大学计算机系的Heng Tao Shen教授针对Web中图像的检索提出了在Web文档中提取图像语义信息的一种新方法。他认为,网页中图像的图像名、图像替换文本、图像周围文本、图像所在网页标题与图像语义密切相关,并以此为依据建立了相应的提取模型来提取图像的相关信息,从而提高了Web图像检索的准确率[3]。国内外著名的搜索引擎如雅虎、百度、Google等,对分布广泛的多媒体网页进行自动标注,并在此基础上实现了对多媒体的标注。但是,通用的搜索引擎目前的查准率还远远不能满足人们的需要,主要由于描述多媒体资源的相关文本信息的准确性不高,并且真正用于音频、视频、动画的相关技术研究相对较少,因此针对各种媒体形式,开发相关技术,以准确提取多媒体资源的相关信息,对多媒体信息的检索具有重要意义。

Web多媒体相关文本提取是以利用多媒体搜集器广泛搜集Web中多媒体网页为基础,对搜集到的方式,设计多媒体信息提取算法,准确提取Web多媒体资源的相关文本,对所提取的多媒体网页进行预处理及区域分析,然后结合多媒体资源在网页中相关文本进行分词处理,得到反映该多媒体资源的相关信息[4]。系统结构图如图1所示。

第一作者简介:于文超(1982－),男,山东师范大学教育技术系研究生,研究方向是网络多媒体资源分析与数据挖掘。

收稿日期:2009年3月1日。

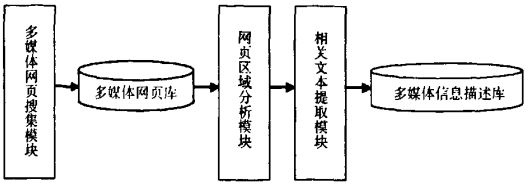

图1 Web多媒体资源搜索与相关文本提取系统结构图

2 Web多媒体网页的搜集

搜集Web中的多媒体网页需要通过专门的多媒体网页搜集器,图2为搜集器的体系结构。

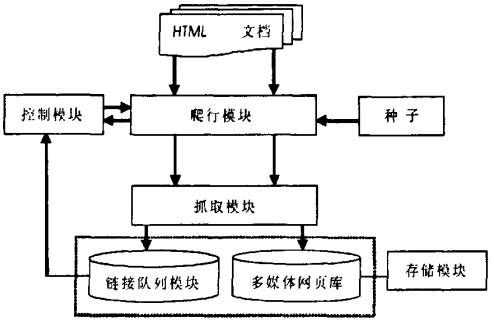

图2 多媒体网页搜集器体系结构图

在该系统中,控制模块既作为程序的入口,控制程序的运行,又能实现爬行模块与连接队列之间的通信;爬行模块的功能是通过控制模块向链接数据库请求一个合适的URL进行爬行,并分析链接;抓取模块用于提取链接并将链接存入链接数据库,将符合要求的多媒体网页的HTML代码存入多媒体网页库;存储模块用于存储链接和多媒体网页。

本搜集器的爬行过程是目标驱动的,根据用户定义的目标主题,从一些种子URL出发,沿着Web页面上的超链接在线遍历Web,搜集包含多媒体资源的页面,利用改进的Fish-Search算法,搜集Web中包含多媒体资源的网页。

本系统采用的后台数据库系统为Access 2003,包含4个数据表,即Image、Audio、Flash、Video,分别用来存放搜集到的图像、视频、动画、音频等多媒体类型的网页信息,以Flash为例,其数据表的逻辑内容如表1所示。

表1 Flash数据表逻辑结构图

字段名称	数据类型	字段意义
ID	自动编号	用于记录包含Flash数据的网页数量
URL	备注	网页的URL地址
URLCode	备注	网页的HTML代码
VisitedTime	日期/时间	网页的访问时间
FlashCount	数字	Flash动画的个数

3 多媒体网页区域分析

绝大多数的多媒体Web页面都有区域分布的特性,如重要的信息一般放在网页的上部和中部等显著位置,下方一般放置如网页的版权等信息。一般情况下,多媒体资源所在的区域周围的文本与该多媒体资源的主题显著相关,在多媒体网页中,除网页的主题信息外,还存在很多与多媒体信息无关的文本内容。大量的导航条、广告信息、版权信息等内容也会干扰多媒体相关文本信息的提取。准确提取多媒体资源的相关文本信息、判断多媒体资源所在网页的区域是非常重要的[5]。本文利用网页的HTML语言的语法特点与网页布局的外部特征相结合,完成对网页区域的分析。

HTML是超文本标记语言的英文缩写,由许多元素按照HTML语法组成的,而大部分元素由开始标记(Start Tag)、元素内容和结束标记(End Tag)组成,还有一部分元素为空标记,如
。每个HTML文档都应至少有如下标记<HTML>、<HEAD>和<BODY>。一个HTML文档分为头部和主体两部分。头部由成对标记<HEAD>定义,主要说明文档的类型、性质及与其他文档的关系等。主体由成对标记<BODY>定义,用来描述文档的内容。标题在文档头部中,由成对标记<TITLE>定义,是在浏览HTML文档时浏览器顶部标题栏中描述文档的文字。在设计网页时大多是用表格对网页布局。与表格有关的标记有<TABLE>、<TR>、<TD>等。利用表格布局页面元素往往采用树形结构,即表格中又嵌套了表格。

根据HTML页面的布局特点,对HTML文档顺序扫描,跳过脚本中<SCRIPT>、样式表<STYLE>、表单<FORM>等无用标记,然后按文本条在HTML

文档中先后位置进行统一编号，对其视觉、语义属性进行提取，同时利用堆栈生成文本条的结构。文本条的属性与内容分别存储，利用内容、属性标识号建立文本条内容与属性一一对应关系，再顺序扫描映射表中文本条属性映射部分，将标识号连续、结构相同的文本条聚类，形成具有不同结构的多个区域，从而完成多媒体网页的区域分析。

4 多媒体资源相关文本提取

要提取网页中多媒体资源的相关文本，首先分析与多媒体嵌入相关的 HTML 标记，相关标记如下：

(1)字符集属性标记

字符集属性标明了网页所采用的字符集，可由“<meta> …… </meta>”标记内的 charset 属性值来获得。根据其属性值是否为 gb2312 或 gb2312-80，即可判断相应网页是否为简体中文网页。

(2)网页标题标记

网页标题是对网页内容的概括。网页“<title> …… </title>”标记中的内容即为网页的标题。

(3)网页内容描述关键词

在网页头部有两个很重要的关键词 keywords 和 description，标记为<meta name="keywords" content="网页的关键词">，<meta name="description" content="网页的简述">，这两个标记中的信息往往对文档的内容有很高的概括性，利用这些标记信息可以提高特征提取精度。

(4)超链接标记

超链接标记主要用来实现网页之间或者网页与媒体文件之间的导航。其语法格式为超链接热点(注："#"代指某一段字符串)，其中 href="#"表示一个超文本引用，可以指向任意一个目标地址；name="#"表示该超链接标记在当前网页中的名字；超链接热点是在网页上显示以便浏览者点击的文字或图像，其中往往包含了重要的语义信息。

(5)图像嵌入标记

图像嵌入标记可以将图像嵌入在网页中指定的位置。其语法格式为<img src="#" alt="#" width="#" height="#">，其中 src="#"表示嵌入的图像文件的路径和文件名；alt="#"表示在浏览器不能或者尚未完全读入图像时，在图像位置显示的替换文字，又称为图像的标签；width="#" height="#"分别表示显示图像的宽度和高度。图像嵌入标记中的 src 属性和 alt 属性对于提取图像的语义信息具有重要意义。

(6)视频嵌入标记

视频嵌入标记是<object>、<embed>。两者的区别是<object>标签只支持 IE 系列的浏览器或者其他支持 Activex 控件的浏览器，<embed>标签支持<mozilla>系列的浏览器或其他支持 Netscape 插件的浏览器(mozilla family of browsers)，Netscape 和 Mozilla 系列的浏览器只读取<embed>标签而不会识别<object>标签。

通过对多媒体的相关文本进行分析，将能够反映多媒体信息的文本提取出来，用以标识多媒体的信息，如果多媒体文件存在周围文本，则根据前述的区域分析后的结果提取多媒体资源所在区域内与多媒体资源相关的文本。

5 实验结果

我们对多媒体网页搜集器中搜集的各种类型的网页进行测试，实验结果如表 2 所示。

实验结果表明该方法对图像的相关文本提取效果较好，音频的相关文本提取效果次之，视频、动画的相关文本提取效果相对较差。

经分析，以下原因导致了上述实验结果的产生：

(1) 网页中图片的嵌入方式相对简单并且单一，图像利用其单一的嵌入标记“<img>”进行嵌入，结构和形式相对稳定，并且主题网页中的图片

表 2 实验结果

多媒体类型	搜集到网页总数	提取正确的网页数	正确率
图像	843	783	92.88%
视频	182	135	74.18%
动画	465	343	73.76%
音频	176	143	81.25%

一般都有相关的文字进行描述，从而使图像类网页的相关文本提取准确率较高。

(2) 音频类多媒体网页中音频的嵌入方式也相对简单，一般采用 <embed> 标记嵌入，形式相对固定，但很多音频类网页中缺乏对音频的直接描述，导致音频类网页的相关文本提取准确率不如图像类网页。

(3) 视频、动画类多媒体网页相似，这两类多媒体嵌入网页的形式多种多样，结构过于复杂，并且很多网页都采用了代码隐藏技术，在客户端很难分析并提取其相关文本信息，从而导致提取效果不如图像类和音频类多媒体网页。

6 结 语

随着 Internet 的迅速发展及信息化程度的日渐提高，Web 中的网页数目呈几何级数爆炸性增长，网页内容呈现出海量性、多样性和动态变化的特点。据中国互联网络信息中心 (CNNIC) 发布的《第 23 次中国互联网络发展状况统计报告》，自 2002 年开始，中国的网页规模一直保持高位增长，截至 2008 年底，中国网页总数超过 160 亿个，较 2007 年增长 90%。如此众多的网页中有很多是多媒体网页，包含了大量的多媒体资源。一方面，互联网技术的发展使得这些多媒体资源的发布与共享不再受时间、空间的限制，成为我们获取这些资源的一个重要途径；另一方面，网页数目的激增及其无序性使网民查找所需多媒体资源变得非常困难。因此，进行 Web 多媒体资源搜索及提取其相关文本，对描述网上多媒体资源的信息具有重要意义。与基于内容的多媒体检索相结合，可提高多媒体搜索引擎的查准率和查全率，并且该方法可用于中文网页的自动分类、搜索引擎的个性化查询服务和面向主题的信息搜集等。

参考文献

[1] 张波．两种图像检索技术的比较研究 [J]．情报杂志，2005(2)：103－104.

[2] 万钧，钟亦平，傅维明，等．启发式相关文本提取技术研究 [J]．小型微型计算机系统，2004，25(4)：582－586.

[3] Heng Tao Shen，Beng－Chin Ooi. Giving Meanings to WWW Images [R] // Proc of ACMMM2000. Los Angeles，California：ACMPress，2000：39－47.

[4] 孟祥增，钟义信．基于语义的 WWW 图像检索 [J]．现代图书情报技术，2004(3)：35－37.

[5] 吴鹏飞．面向 Web 的多媒体信息提取及其教育应用 [D]．济南：山东师范大学，2007：21－32.

Research on Web Multimedia Resources Search and Related Text Extraction

Yu Wenchao，Liu Fei

(School of Dissemination，Shandong Normal University，Jinan 250014)

Abstract: The related text of multimedia resources in multimedia webpages has great significance in the description of multimedia resources in web. Webpages which contain the multimedia resources are collected by Web－oriented multimedia searcher in the web，then analysing the regions of the page. Based on the work，according to the embedded form of multimedia resources in the web pages，we design web information extraction system of multimedia resources related text to extract related text of Web pages of multimedia resources accurately. The experimental results show that the system has a higher accuracy in extracting related text of multimedia resources，the method is helpful in increasing recall ratio and precision ratio in multimedia retrieval system.

Keywords: webpage collection，region recognition，text extraction