

绿色网络 PDF 提取系统

龙 珑¹, 邓 伟², 覃 晓¹

(1. 广西师范学院 计算机与信息学院, 广西 南宁 530023;
2. 广西肿瘤防治研究所, 广西 南宁 530021)

摘 要:随着信息技术迅猛发展,很多不良信息与文化通过 PDF 文档传播,而传统的提取 PDF 内容的方法无法适应绿色网络提供优质内容并过滤不良 PDF 的社会需求。文中提出通过建立层次关键字自动机快速提取 PDF 内容并过滤不良 PDF 内容的方法。在提取准确性基本相同的情况下,文中方法提升了绿色网络系统提取 PDF 文档的速度,所用的时间仅为 PDFBox 方法的 16% ~ 36%,并能提供更好地过滤不良 PDF 的服务。

关键词:绿色网络;自动机;提取信息;不良内容 PDF;过滤

中图分类号:TP311

文献标识码:A

文章编号:1673-629X(2014)01-0204-04

doi:10.3969/j.issn.1673-629X.2014.01.052

PDF Extraction System of Green Network

LONG Long¹, DENG Wei², QIN Xiao¹

(1. College of Computer Science and Information Technology, Guangxi Teachers Education University,
Nanning 530023, China;
2. Guangxi Cancer Institute, Nanning 530021, China)

Abstract: With the rapid growth of Internet, a lot of unhealthy information and culture spread through the PDF file, traditional PDF extraction algorithm cannot adapt to the requirement of green network to provide quality content and filter undesirable PDF. A new method that extracts PDF content and filters undesirable PDF through establishing keyword automata is proposed. With the approximately equal extraction accuracy, the new method can enhance the speed of the green network system to extract the PDF document, the extraction time is only 16% to 36% of PDFBox, and provide better service to filter undesirable PDF file.

Key words: green network; automata; extracting information; undesirable PDF; filter

0 引 言

从现有资料没有发现具体定义“绿色网络”^[1],只能理解为预防人群感染上网瘾精神病的计算机网络系统。基于行为分析的绿色网络系统中提取 PDF^[2]文档内容子系统(下文称绿网 PDF 提取系统)能帮助青少年获得对他们身心有益的 PDF,对青少年有不良影响的 PDF 被过滤。系统把含不良内容的文档转换成 PDF 格式以逃避监控系统的监管,因此提取 PDF 内容对于绿色网络系统显得非常重要,绿网 PDF 提取系统使用基于自动机理论的方法提取 PDF 内容并过滤不良的 PDF 文档。

在如何快速准确提取 PDF 文档内容这个难题上,

不少学者进行了大量有价值的研究。William 等^[3-11]通过对 PDF 格式的分析,把 PDF 文档分拆为一些几何因素,依据几何元素之间的算法关系,再将几何元素组成一些逻辑块,最后分析提取整个 PDF 文档内容使用改进的 Blackboard 算法,该算法由于过程非常复杂,无法对高速网络中的 PDF 数据进行实时提取,可扩展性不强;宋艳娟等^[12-13]将 PDF 文档解析转成 XML 格式,然后利用文本特征、显示特征和位置特征对 XML 格式问题进行信息提取,该方法由于过分依赖文档的完整性,不能实时提取网络中部分到达的 PDF 文档信息,更加不能提取不完整的 PDF 文档的内容;李强等^[14-15]通过分析 PDF 文档的格式,使用对象方法设

收稿日期:2013-04-09

修回日期:2013-07-11

网络出版时间:2013-11-12

基金项目:国家创新基金项目(10C26224504901);国家自然科学基金资助项目(81260319);广西自然科学基金项目(2011GXNSFB0180825)

作者简介:龙 珑(1980-),男,硕士,高级工程师,研究方向为人工智能;邓 伟,通讯作者,博士,副主任医师,研究方向为流行病学、网瘾预防。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20131112.1627.004.html>

计和实现 PDF 阅读器,从而提取 PDF 文档的内容,该方法无法自适应不同版本的 PDF 格式,当 PDF 文档发生改变时无法识别,可移植性不好。

对提取 PDF 文档内容的研究尚处在初始阶段,结合绿网中 PDF 更新较快和多样性的特点,文中提出了建立层次关键字自动机快速提取 PDF 内容并过滤不良 PDF 内容的方法,实验表明该算法可有效地提取 PDF 文档内容,为绿色网络系统是否过滤该 PDF 文档提供依据。

1 PDF 的结构

美国 Adobe 公司于 1993 年最先提出 PDF 文档格式结构。它由 PS(PostScript)页面描述语言发展而来,不但具有 PS 相似的页面描述能力,而且还具有交互功能、随机存取页面及字体仿真描述等特性,PDF 格式文档非常适合各种电子书与印刷出版。Adobe 公司对 PDF 格式给出了物理结构和逻辑结构两种方面的解析。PDF 的物理结构分为四部分:文件头部分、文件体部分、交叉引用表部分和文件尾部分。文件头部分出现在 PDF 的第一行,指明该 PDF 的版本号。文件体部分指 PDF 文档的具体内容,如图像、文字等。交叉引用表部分则是指为了随机存取间接对象而设立的一个间接交叉索引表。文件尾部分指明文件体,声明交叉引用表的地址,保存加密等安全信息,该部分相应地可以控制整个 PDF 文档。PDF 逻辑结构是一种树型结构。这种结构的树的根节点是 PDF 文档的根对象,根的节点下有四棵子树,分别为:页面树、书签树、线索树和名字树。

2 绿网 PDF 提取系统的设计

2.1 绿网 PDF 提取系统的功能设计

绿网 PDF 提取系统指对系统分析 PDF 格式文档,提取该 PDF 文字信息和图片信息,为系统分析不良信息提供基础数据。绿网 PDF 提取系统分析 PDF 格式文档过程如图 1 所示。系统按二进制文档格式规范对数据区域进行分割,可以分为分段文档数据,系统开始解压缩每段数据,形成普通文档数据,最后编码转换普通文档数据,形成正常编码的文档数据,是正常格式的文字和图片内容,从而绿网 PDF 提取系统可识别文档内容并提取该内容。

由于 PDF 文档的格式分析需要依据不同的文档规范并进行不同的实现,所以分析并分割 PDF 文档内容是绿网 PDF 提取系统提取方法的核心难点。加上各个厂家为了给自家产品增加新功能,系统同时也在不断修改同一个格式的文档的规范,为了更好地保护自家产品,半公开或完全不公开部分文档格式,所以根

据 PDF 格式去提取 PDF 内容的方法难以实现。文中采用通过建立层次关键字自动机快速提取 PDF 内容并过滤不良 PDF 内容的方法,解决这一难题。

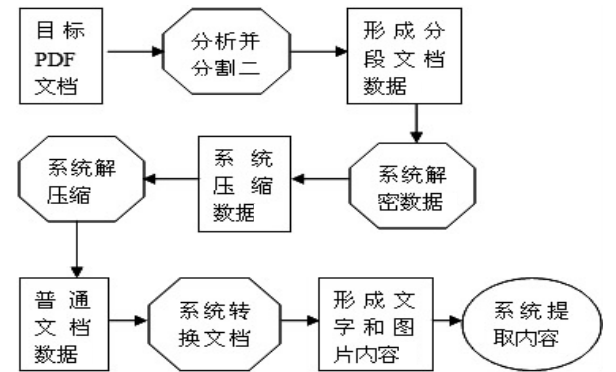


图 1 绿网 PDF 提取系统提取 PDF 文本内容数据流程

2.2 绿网 PDF 提取系统关键字树的建立

文字、图像、页面和表单等组成了 PDF 文档的主体内容部分,一个二元组表示绿网 PDF 提取系统抽象这些内容,二元组的格式系统规定为{关键字,操作},其中操作是系统处理关键字定义与其相关联的实体内容的动作。绿网 PDF 提取系统抽取文本内容时,需要找到与内容相关的目标关键字才可以依据关键字所定义的动作执行提取工作。

绿网 PDF 提取系统将 PDF 文档的关键字依据是否具有层次定位为两类,如图 2 所示系统把这些层次关键字组织成树状结构,进行提取使用了分层非确定的有穷自动机(Non-deterministic Finite Automation, NFA),以克服 PDF 文档过多的非结构化数据难以处理的问题。图 2 中,父节点要处理的内容分布在每一个子节点上,而最终系统提取的内容位于子叶子节点上。

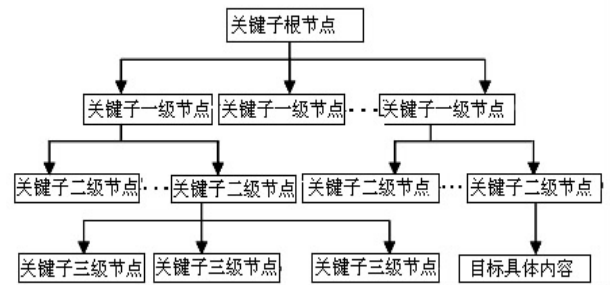


图 2 绿网 PDF 提取系统关键字组织图

绿网 PDF 提取系统主要是提取页面信息、编码解码信息、内容信息和字体信息等几种基本类型信息。页面信息的关键字是“Page”及该字相近词“Pages”等一组字,这一组关键字定义了 PDF 文档的网页结构,并且通过“Page”这一组关键字建立起了页面树。编码解码信息的关键字是“Filter”及该字相近词“Filters”等组字,用哪种方法进行解码需要找到该组关键字。因为一个内容流一般情况下使用了多种编码,绿网 PDF

提取系统按照编码的逆序技术方式依次解码相应的内容流,而系统内容信息的关键字是“Content”及该字相近词“Contents”等组字,绿网 PDF 提取系统一旦发现该组关键字才提取文本内容,否则直接忽略该页。字体信息的关键字是“Font”及该字相近词“Fonts”等一组字,绿网 PDF 提取系统通过该组关键字确定内容的文本对象应该要做哪些操作之后才能将内容存储在用户可以理解的文本文件中。

2.3 绿网 PDF 提取系统 NFA 的分层构造

由于 PDF 文档有层次结构,绿网 PDF 提取系统采用基于 NFA 的方法提取 PDF 文档中文本信息,绿网 PDF 提取系统依据 2.2 部分内容建立关键字,然后通过分层建立 NFA 的方式(见图 3),最终提取出 PDF 文档中的文本信息。

绿网 PDF 提取系统提取 PDF 文档信息的顶层 NFA 如图 3 所示。顶层 NFA 完成了系统识别目标 PDF 逻辑结构每部分的开始与结束,检测到一部分开始时,绿网 PDF 提取系统调用下层 NFA 来具体实现抽取不同信息内容的操作,当操作完毕后,通过返回一个程序消息告知上层 NFA 操作已经完成,收到程序消息后上层 NFA 的操作会继续进行。

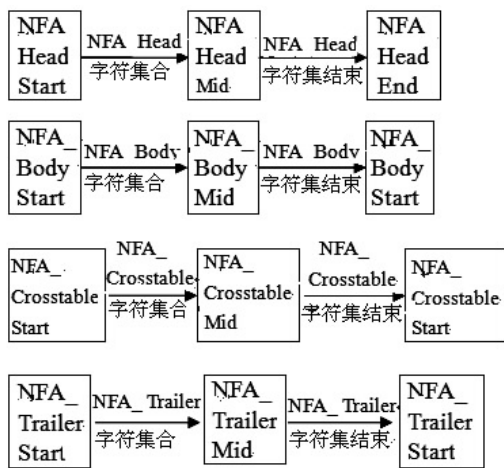


图 3 绿网 PDF 提取系统提取 PDF 文档信息的顶层 NFA 示意图

不同的操作要在逻辑结构每部分完成,具体必须定义下层 NFA 中每一部分的操作。识别文件不同部分的 NFA 需要强制被绿网 PDF 提取系统识别。NFA 文档的版本号根据文件头部分被系统识别,根据版本不同,绿网 PDF 提取系统解析文本信息时所要做的有些操作也会不相同。文本体中包含了全部解析的所有内容,间接对象组成了这些内容。每个间接对象的格式相同,但内容不同。遇到间接对象需要通过查找 {关键字,内容} 的二元组具体执行的操作。交叉索引表存放 PDF 文件中所有间接对象的位置信息,绿网 PDF 提取系统通过查找交叉索引表,快速找到每一个

间接对象。交叉索引表的开始位置以及加密信息在文件尾 NFA 找到。

2.4 绿网 PDF 提取系统提取文本内容

绿网 PDF 提取系统提取 PDF 文档内容的步骤如下:

- (1) 系统先配置一个 {关键字,操作} 的二元组文件;
- (2) 系统将二元组中的关键词构建成关键字树;
- (3) 系统分层提取关键字数,按 PDF 文件的逻辑结构勾画每层关键字,用 AC 算法建立不同的 NFA;
- (4) 建立好自动机后,开始监控和扫描绿色网络系统中的数据流或者完整的 PDF 文档,从而提取 PDF 文件中文本信息;
- (5) 把 PDF 文本信息内容与绿色网络云关键词进行实时对比,过滤不良内容的 PDF 文档。

图 4 为绿网 PDF 提取系统提取 PDF 文档中的文

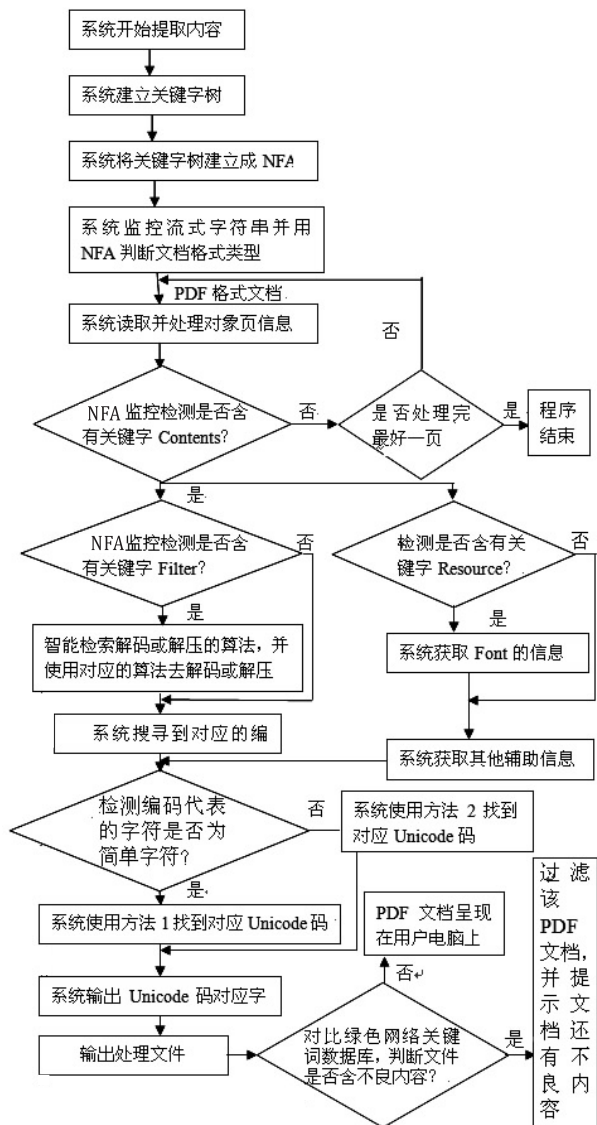


图 4 绿网 PDF 提取系统提取 PDF 文档中的文本信息流程图

本信息流程图,其中方法 1 为非 type() 的字符提取 Unicode 码的方法,方法 2 为 type() 的字符提取 Unicode 码的方法。详细步骤参考文献[2]。

3 实验及分析

3.1 实验测试条件

文中选取中文 PDF 文档和英文 PDF 文档两类数据进行测试。中文 PDF 文档数据从不同的中文电子书中选取文件大小从 153 kB 到 53 350 kB 不等;英文 PDF 文档数据从不同的英文电子书中选取从 241 kB 到 1 375 kB 不等。

测试的硬软件环境如下:
CPU:i5-3210M,2.5 GHz;内存:4.0 GB;操作系统:Windows 7 Professional 32 位。

3.2 实验结果与分析

绿色网络 PDF 提取系统与现在最流行的 PDFBox 的方法相比,提取 PDF 文档内容的效果如表 1 和表 2 所示。

表 1 中文 PDF 文档提取结果对比表

文档编号	文档大小/kb	提取内容时间/ms	
		PDFBox 方法	文中方法
1	241	497	86
2	355	584	141
3	515	909	361
4	1 122	1 738	309
5	1 331	887	244
6	1 375	961	343

表 2 英文 PDF 文档结果提取对比表

文档编号	文档大小/kb	提取内容时间/ms	
		PDFBox 方法	文中方法
1	153	497	155
2	508	584	206
3	999	600	223
4	1 430	1 650	515
5	13 750	10 605	2 281
6	53 350	64 346	9 593

从以上实验结果表明:在提取准确性基本相同情况下,文中方法所用的时间仅为 PDFBox 方法的 16%~36%,绿色网络系统提取 PDF 文档速度得到提升。

4 结束语

PDF 格式已经成为网络上一种非常流行的文本格式,部分网站为了逃避网络监管,把不良内容存储成为

PDF 格式文档,为了过滤不法 PDF 文档,必须快速有效地提取 PDF 文档内容。基于以上需求,文中提出通过建立层次关键字自动机快速提取 PDF 内容并过滤不良 PDF 内容的方法。测试结果表明该方法提取 PDF 内容速度比 PDFBox 高。由于基于行为分析的绿色网络系统的云数据采集库的关键字更新,使用云技术实时更新关键字将成为今后研究重点。

参考文献:

[1] 宁 葵,龙 珑,覃 晓,等.绿色网络不良内容语义分析方法研究[J]. 计算机应用研究,2010,27(12):4643-4645.

[2] Adobe system incorporated;PDF reference:sixth edition[EB/OL].[2010-10-23]. http://www.adobe.com/content/dam/Adobe/en/devent/acrobat/pdfs/pdf_refernce_1-7.pdf.

[3] William S L, David F B. Document analysis of PDF files: Methods,results and implications[J]. Electronic publishing origination dissemination and design,1995,8(2/3):207-220.

[4] 杨 洁,季 铎,蔡东风,等.基于联合权重的多文档关键词抽取技术[J]. 中文信息学报,2008,22(6):75-79.

[5] 李 珍,田学东.PDF 文件的信息的抽取与分析[J]. 计算机应用,2003,23(12):145-147.

[6] 李贵林,李建中,杨 艳.用 Plug-in 实现对 PDF 文件的信息提取[J]. 计算机应用,2003,23(2):110-112.

[7] 张秀秀,张立峰.PDF 文件文本内容提取研究[J]. 科技情报开发与经济,2008,18(36):118-120.

[8] 王晓娟,谭建龙,刘燕兵,等.基于自动机理论的 PDF 文本内容抽取[J]. 计算机应用,2012,32(9):2491-2495.

[9] Yuan Fang, Liu Bo, Yu Ge. A study on information extraction from PDF file[C]//Proceedings of the 4th international conference on advance in machine learning and cybernetics. Berlin:Springer-Verlag,2005:258-267.

[10] Chao Hui, Fan Jian. Layout content extraction for PDF documents[C]//Proceedings of document analysis systems. Berlin:Springer-Verlag,2004:213-224.

[11] 郑皎凌,唐常杰,姜 玥,等.基于伪属性语义匹配的 Deep web 信息抽取[J]. 四川大学学报(工程科学版),2009,41(2):173-178.

[12] 宋艳娟,张文德.基于 XML 的 PDF 文档信息抽取系统的研究[J]. 现代图书情报技术,2005,21(9):10-13.

[13] 张晓李,王西锋.基于概念图的汉语语义计算的研究与实现[J]. 计算机工程与应用,2011,47(10):120-123.

[14] 李 强,刘时进.PDF 阅读器的设计与实现[J]. 计算机工程与设计,2010,31(7):1635-1638.

[15] 杨道良.面向对象的中文 PDF 阅读器的设计与实现[J]. 计算机应用,1999,19(6):1-4.

作者:

龙琬, 邓伟, 覃晓, LONG Long, DENG Wei, QIN Xiao

作者单位:

龙琬, 覃晓, LONG Long, QIN Xiao(广西师范学院 计算机与信息学院, 广西 南宁, 530023),
邓伟, DENG Wei(广西肿瘤防治研究所, 广西 南宁, 530021)

刊名:

计算机技术与发展

ISTIC

英文刊名:

Computer Technology and Development

年, 卷(期):

2014(1)

本文链接: http://d.wanfangdata.com.cn/Periodical_wjz201401052.aspx