# 启发式规则网页主题定位方法绿网系统的应用

龙 珑[1,2],宁德鹏[1,2],宁 葵[1,2]

(1.广西师范学院 计算机与信息管理学院,广西 南宁 530003;
2.广西大学 计算机学院,广西 南宁 530001)

**摘 要**:随着Internet的迅猛发展,我国网民的数量激增。为了能快速地过滤网上的不良信息的传播,绿色网络软件就必须能快速地抽取网页的信息才能完成这个项目设计任务。这绿色网络软件要求提取信息抽取速率非常快速。传统的绿色网络信息抽取的方法无法适应绿色网络网页信息抽取在速率上的需求。文中提出启发式规则网页主题提取方法,去解决绿色网络中网页信息快速抽取这一问题,从而有效地解决快速抽取任务。给出了结合绿色网络实际情况实现一个可以快速定位主题的模块系统。

**关键词**:绿色网络;启发式规则;信息定位

**中图分类号**:TP311 **文献标识码**:A **文章编号**:1673-629X(2011)10-0226-03

# Approach of Subject Information Location Based on Heuristic Rules Applying in Green Network

LONG Long[1,2], NING De-peng[1,2], NING Kui[1,2]

(1. Dept. of Computer Science and Information Technology, Guangxi Normal College, Nanning 530003, China;
2. College of Computer Science, Guangxi University, Nanning 530001, China)

**Abstract**: With the rapid growth of Internet, tradition forecasting methods in green network can not adopt to require of web services of green network. In Green Network, a new approach of subjection information location based on heuristic rules firstly proposed for Green Network Grid service. The new way can prove effective of the web services of green network, new approach of subjection information location based on heuristic rules is applied for green network, in the end the result of model is broposed.

**Key words**: green network; heuristic rules; subject information localization

## 0 引 言

"绿色网络"的定义尚未明确,一般性的理解为畅通、健康和安全,并可以预防人群感染上网瘾精神病的网络。基于行为分析的绿色网络系统软件是为了解除和预防青少年的网瘾而设计的,对绿色网络的建设具有非常重要的现实意义,并已经在实际应用中取得了较好的社会和经济价值。绿色网络要求快速过滤网络不良内容,这就需要系统具有信息抽取(information extraction, IE)能力[1,2]。信息抽取是结构化形式信息,从而可以直接在自然语言文字中抽取信息。绿色网络系统抽取信息进一步用于绿网系统对Web页面数据收集和不良内容Web文档对信息抽取相当困难,主要由于主要信息隐藏在没有关联的文字和结果中。而网页内容主要包含两部分信息内容:第一部分为网页的主题信息内容,如文章网页的文章标题、文章正文等;另外部分是与主题无关的内容,如广告信息等,也称为噪声信息。消除与主题无关的内容,抽取与主题有关的信息是信息抽取领域的主要课题之一,也是绿色网络系统网页过滤中一个难题。

目前在网页信息抽取方面,国内的相关研究有如文献[3,4]表述为:

(1)根据网页布局特征的抽取方法,区分主题内容和噪声内容;根据使用标记在布局方面的作用对网页进行结构分析,从而有效提取主题信息。

(2)模板的抽取方法,利用模板来直接提取网页主题信息根据采用机器学习方法来建立模板库。

国外的相关研究包括:

(1)可以通过计算网页文本域标记的比率将网页聚类成内容和非内容的区域,从而在HTML文档的文本内容与标记的比率特性从网页中提取信息[5,6];

**收稿日期**:2011-03-03;**修回日期**:2011-06-30

**基金项目**:科技部科技型中小企业创新基金(06C26224501689),广西科学研究与技术开发计划重大项目(桂科合0815007-1-15);2007年广西科技进步三等奖

**作者简介**:龙 珑(1980-),男,留英硕士,高级工程师,主要研究方向为计算机安全、人工智能。

(2)在网页中可以抽取信息块的 EIBA 方法,把网页划分为语义块,标注的块用来分类模型的训练数据集,再将信息块抽取出来得出结果[7]。

但是这些方法都有一个共同问题,只是针对主题信息块的提取,根本没有办法解决绿色网络系统网页过滤问题。系统设计者就提出基于启发式规则的主题信息精确定位方法,对提取的信息的主题信息块进行定位,分离出准确单独的主题信息,最后能用到绿色网络主题抽取信息系统中过滤不良信息[8~11]。

## 1　绿色网络系统启发式规则及定位算法

### 1.1　绿色网络中启发式规则定义

(1)绿色网络主题抽取信息系统中文本节点集合被定义:TM = {TM | TM 是文本节点};

(2)绿色网络主题抽取信息系统中标题节点集合的方法被定义:LM = {lm | lm ∈ TM ∧ lm 的文本是标题内容};

(3)绿色网络主题抽取信息系统中正文节点集合的方法被定义:SM = {SM | SM ∈ SM ∧ SM 的文本是正文内容};

(4)绿色网络主题抽取信息系统中发布时间节点集合的方法被定义:TMN = {tmn | tmn ∈ TM ∧ tmn 的文本是正文内容};

(5)绿色网络主题抽取信息系统中来源节点集合方法被定义:绿色网络主题抽取信息系统 CM = {CM | CM ∈ TM ∧ CM 的文本是来源内容};

(6)绿色网络主题抽取信息系统中发布时间特征词汇集合的方法被定义:Timeset = {times | times 具有时间含义的词},而绿色网络主题抽取信息系统 Timeset 中元素为“时间”“发布时间”“更新日期”“日期”等;

(7)绿色网络主题抽取信息系统来源特征词汇集合的的方法:Sset = {ss | ss 具有含义的词},而绿色网络主题抽取信息系统 Sset 中元素为“来源”“转自”等。

1)绿色网络主题抽取信息系统中节点偏序关系的方法。绿色网络主题抽取信息系统对 DOM 树进行先序遍历,绿色网络主题抽取信息系统可以得到的节点序列为 $a_1\ a_2\ a_3 \cdots a_n$,绿色网络主题抽取信息系统 $i \in \{1,2,\cdots,n\}$,$j \in \{1,2,\cdots,n\}$,绿色网络主题抽取信息系统且 $i < j$,则 $a_i$ 与 $a_j$ 满足偏序关系,记为 $a_i < a_j$。绿色网络主题抽取信息系统节点直接先后的方法:对 DOM 树进行先序遍历,绿色网络主题抽取信息系统取得文本节点序列为 $TM_1\ TM_2\ TM_3 \cdots TM_m$,对于 $i \in \{1,2,\cdots,m-1\}$,称绿色网络主题抽取信息系统 $TM_i$ 直接先于 $TM_{n+1}$,绿色网络主题抽取信息系统记为 $TM_i < TM_{n+1}$。

2)绿色网络主题抽取信息系统中 Content 规则的方法:

a. 假设绿色网络主题抽取信息系统 $\lambda_1 \leqslant$ length(TM),绿色网络主题抽取信息系统 TM<SM。其中绿色网络主题抽取信息系统 length( $n$ )是求节点 n 文本长度,绿色网络主题抽取信息系统 $\lambda_1$ 为正文文本长度的阈值。

b. 绿色网络主题抽取信息系统。设绿色网络主题抽取信息系统 pn = parent(TM),绿色网络主题抽取信息系统中集 dbm ∈ { $n$ | $n$ = <b>N ∈ sibling(bn)},如果 bn = <b>,而且 count(dbm) ≥ $\varepsilon_1$,绿色网络主题抽取信息系统得到 TM ∈ SM。绿色网络主题抽取信息系统 parent( $n$ ) 为节点 $n$ 的父亲节点,绿色网络主题抽取信息系统 sibing( $n$ ) 为节点 $n$ 在 DOM 树中兄弟节点集合;绿色网络主题抽取信息系统 count( $n$ ) 为统计 $n$ 节点个数,$\varepsilon_1$ 为 dbm 节点个数阈值。

c. 绿色网络主题抽取信息系统。设 ps = previosSibing(TM),绿色网络主题抽取信息系统集 dbm ∈ { $n$ | $n$ = <b>N ∈ sibling(bn)},bs = <br>,绿色网络主题抽取信息系统就可以 count(sbs) ≥ $\varepsilon_2$,绿色网络主题抽取信息系统就可以 TM ∈ SM。previosSibing( $n$ ) 为节点 $n$ 前一个兄弟节点绿色网络主题抽取信息系统 $\varepsilon_2$ 则为 sbs 节点个数阈值。

d. 绿色网络主题抽取信息系统对集 tmn<TMN,设 tmn<TM,TM ∈ SM。

### 1.2　绿色网络系统中启发式规则的定位

绿色网络的启发式定位算法可以描述如下:

(1)绿色网络主题抽取信息系统中对一棵已经去掉重复内容的精简过的 DOM 树,可以先序遍历得 DOM 树中所有文本节点集合 TM_list,| TM_list | = $n_3$;绿色网络主题抽取信息系统再初始化主题集合 topic_list,| topic_list | = $n_2$。绿色网络主题抽取信息系统可运用 IWMA 获得权值矩阵 $\boldsymbol{A}$。

(2)绿色网络主题抽取信息系统中对于主题,先获得该主题的可能度向量。绿色网络主题抽取信息系统则可能度是用来衡量该节点能够成为该主题的可能性。绿色网络主题抽取信息系统中可能度表示如下:

$$p_{kj} = \begin{cases} p_{kj} + A_{ij} \\ p_{kj} \end{cases}$$

绿色网络主题抽取信息系统中 $B_{kj} + A_{ij}$ 节点 $k$ 满足主题 $j$ 规则 $i$,否则其他表示为 $B_{kj}$。

(3)绿色网络主题抽取信息系统判断 topic_list 中所有的主题是否遍历完毕,成立则转绿色网络主题抽取信息系统中(4),其他转绿色网络主题抽取信息系统中(2)。

(4)绿色网络主题抽取信息系统定位各个主题的节点后,可获得绿色网络主题抽取信息系统中各个主

题节点在 DOM 树中路径,则作为该主题的抽取规则。

## 2 绿色网络系统阈值选取和权值生成

(1)绿色网络系统中 1WMA 中权值选择根据经验人工合理制定,固定的权重很难处理不同类型和不同风格的网页。但是对于某些网页它们符合标题启发式规则。

(2)绿色网络系统可以用标题启发式规则。

(3)绿色网络系统中权值比较高,绿色网络系统标题提取效果会比较好一些;一些网页它们符合标题启发式规则(4),则标题启发式规则(4)权值比较高,绿色网络系统标题提取效果会比较好。绿色网络系统权值根据网页特征自动进行调整,绿色网络系统得到更好的信息提取效果。

## 3 绿色网络系统中算法测试结果和分析

绿色网络系统中算法包括两个重要的部分:绿网系统的长度阈值估计方法可用的准确性;绿网系统的测试系统的 HRBA 效果。绿网系统的标题启发式规则(1)的抽取标题信息,检测绿网系统的阈值估计方法的准确性;后者将 HRBA 应用到绿色抽取的系统中,评测在绿网系统中该算法的有效性。

绿色网络主题抽取信息系统为了考查算法在绿网系统应用效果,第一要避免单种风格带来的影响,这样达到测试目标,就选取绿网系统最具四种代表性网站 8000 个页面进行了对比测试。绿网系统标题启发式规则(1)提取绿网系统抽取标题信息,在绿网系统中根据长度阈值 $\lambda_2$ 取值变化,检测结果如表 1 所示。

表 1 绿色网络主题抽取信息系统中 $\lambda_2$ 误判率影响值

| $\lambda_2$ 值 | B1 值 | B2 值 | B3 值 |
|---|---|---|---|
| 8 | 0.078 | 0.2519 | 0.4329 |
| 9 | 0.0958 | 0.1592 | 0.3549 |
| 10 | 0.1415 | 0.0472 | 0.2884 |
| 11 | 0.1695 | 0.0225 | 0.0667 |
| 12 | 0.2049 | 0.0083 | 0.0673 |
| 13 | 0.2458 | 0.0043 | 0.0769 |
| 14 | 0.2968 | 0.0026 | 0.0908 |

其中在绿色网络主题抽取系统定义:绿色网络主题抽取信息系统中 B1 定义为绿色网络主题抽取系统标题误判为噪声信息的概率;把 B2 定义为噪声信息误判为绿色网络主题抽取系统标题的概率;把绿色网络主题抽取信息系统中 E 定义为绿色网络主题抽取系统总的误判概率,绿色网络主题抽取信息系统 B = B2 B1 + B1B2。

从表 1 绿色网络主题抽取系统测试结果可以看出绿色网络主题抽取信息系统中 B1 与 绿色网络主题抽取信息系统中 B2 成反比,而绿色网络主题抽取信息系统中 B2 随 $\lambda_2$ 增大而逐渐减少。证明绿色网络主题抽取信息系统中该方法准确性。

绿色网络主题抽取系统的抽取结果如表 2 所示。

表 2 绿色网络主题抽取信息系统抽取结果表

| 网站类型 | 抽取网页 | tbe/% | tmbe/% | ebe/% | Obe/% | be/% |
|---|---|---|---|---|---|---|
| 小说网站 | 2000 | 95 | 98 | 94 | 99 | 91 |
| 新闻网站 | 2000 | 92 | 94 | 91 | 96 | 90 |
| 机关网站 | 2000 | 93 | 96 | 90 | 95 | 87 |
| 学术网站 | 2000 | 96 | 99 | 95 | 98 | 93 |

其中:把绿色网络主题抽取信息系统 tbe 定义为绿色网络主题抽取;绿色网络主题抽取信息系统 tmbe 定义为绿色网络主题发布时间抽取正确率;绿色网络主题抽取信息系统 ebe 定义为绿色网络主题正文抽取正确率;绿色网络主题抽取信息系统 Obe 定义为绿色网络主题来源抽取正确率;be 为绿色网络主题总体抽取正确率。

(1)绿色网络主题抽取系统中正文抽取的正确率分析。绿色网络抽取系统的 Ebe 平均能达到 91.5%。导致不够准确原因有两类:绿色网络主题抽取系统中部分噪声信息夹杂在不同段落文字中,错误认为<b>和<br>等是段落文字标签中,软件系统导致误把噪声信息当正文抽取,造成软件的抽取的过度;另外一些正文内容以超链接形式出现,部分信息没有被软件系统抽取出来,信息造成抽取的错误。

(2)绿色网络主题抽取系统中发布时间的抽取正确率分析。绿色网络主题抽取系统中 Tmbe 则相对较高,系统可以平均达到 97%。绿色网络主题抽取系统是发布时间特征,定位很准确,并不是所有网站都达绿色网络主题抽取系统全部准确率。绿色网络主题抽取系统影响发布时间定位的因素有:网页当前的时间、网页正文时间、网页相关链接时间三类。绿色网络主题抽取系统通过对出错页面的分析,发现影响发布时间主要是当前时间、正文时间,绿色网络主题抽取系统相关链接时间由于其位置特性,没有对系统抽取效果造成影响。

(3)绿色网络主题抽取系统中总体抽取的正确率分析。绿色网络主题抽取系统理论上总体正确率计算如下:绿色网络主题抽取系统中 WS 为某一网站测试网页集,B 越接近 minB 时,绿色网络主题抽取系统总体抽取效果越好。

从绿色网络主题抽取系统应用效果可以看出,小说网站、机关网站类 be 相对低一些,新闻类网站和学术网站 be 相对高一些。对具体页面进行分析,小说类网站夹杂噪声信息比较多,机关网站建设不规范,新闻网站和学术的网站建站水平就高不少。总体来说,实际应用效果还是不错。

(下转第 236 页)

小的，与单纯地判断字符编码的奇偶性信息隐藏算法相比，隐藏信息的容量具有很大的增强。

### 3.4 安全性

文章中算法采用零水印的信息隐藏方法，由于载体文档中不包含隐藏信息，文档本身是安全的。隐藏信息的安全主要取决于标记文本的安全，而标记文本可以通过隐蔽信道或安全信道在通信双方之间传递。标记文本体现不出信息的内容，只有将两者结合才能得到正确的隐藏信息，因此该方法具有较强安全性。

## 4 结束语

文章采用的信息隐藏算法是基于通用的 Unicode 编码方式的二进制字符串形式，适用于多种语言的文本。该方法采用零水印的信息隐藏思想，在信息隐藏的过程中不对载体文本做任何形式上的修改，不依赖于字符特征和文本格式，具有抗干扰能力强的特点。应用过程中，对于隐藏的信息可以采用较强的加密算法方法来增强信息的安全性。同时，该方法实现简单，但是却具有较好的信息隐藏容量和鲁棒性，在信息完整性认证、隐蔽通信与对抗等方面具有很好的研究和应用前景。

**参考文献：**


[1] Podilchuk C L, Delp E J. Digital Watermarking Algorithms and Applications [J]. IEEE Signal Processing Magazine, 2001 (1):33-46.

[2] 数字水印技术：概念、应用及现状[EB/OL]. 2010-03-26. http://www.chinaai.org/ip/image-hiding/watermarking.html.

[3] 张 浩，钟尚平. 基于字符编码的信息隐藏算法[C]//第九届全国信息隐藏暨多媒体信息安全技术大会会议论文集(Proceedings of CIHW2010). 出版地不详：出版者不详，2010:62-67.

[4] 白 剑，杨 榆，徐迎辉，等. 基于文本的信息隐藏算法[J]. 计算机系统应用，2005(4):32-35.

[5] 付 兵. 基于字符 Unicode 编码奇偶性的文本信息隐藏算法研究[J]. 福建电脑，2008(12):66-66.

[6] 于晨裴. 基于二次余数的 Word 文档数字水印[J]. 计算机仿真，2007，24(11):324-326.

[7] 刘显德，唐国维，富 宇，等. 一种基于 Word 文档的信息隐藏方法[J]. 电子技术应用，2005，31(15):129-131.

[8] 王海春，邱寄帆，邱敦国. 一种基于 Word 文档的数字密写设计与实现[J]. 微计算机信息，2003，22(10):47-48.

[9] Atallah M J, Raskin V, Christian F, et al. Natural Language Watermarking and Tamperproofing [C]//Proceedings of the 5th International Workshop on Information Hiding. [s. l.]: [s. n.], 2003:196-212.

[10] 温 泉，王锬锋，王树勋. 零水印的概念与应用[J]. 电子学报，2003，31(2):214-216.

[11] 邱发林，李 伟，周邵景. Unicode 以及中文到 Unicode 转换[J]. 科技信息，2006(3):21-22.

[12] 陆 绿，方 勇. 基于字符 Unicode 奇偶性的数字水印设计与实现[J]. 计算机技术与发展，2010，20(8):176-179.


(上接第 228 页)

## 4 结束语

由于现在对于网页信息抽取过滤研究大多集中在主题信块的抽取方面，还很少发现文献针对各个主题精确定位并过滤的系统，更没有在绿色网络建设的应用系统，所以还没有办法进行针对性比较。但是，文中提到的基于启发式规则的网页主题信息在绿色网络应用还是不错的。不过发现很多要进一步研究的内容，其中包括：(1)算法在绿色网络系统中自适应性的提高；(2)启发式对有部分主题提取通用性也要提高。

**参考文献：**


[1] Massimo B, Takahiro K, Bayne T R, et al. Importing the Semantic Web in UDBE[C]//Proceedings of Web Service, E-business and Semantic Web Workshop. Toronto, Canada: [s. n.], 2002:225-226.

[2] Sivashanmugam K, Verma K. Speed-R: Semantic B2B Environment for diverse Web Service Registries[D]. Georgia: Department of Computer Science, University of Georgia, 2002.

[3] Antonio J, Silva C. Quality of Service and Semantic Composition Workflowers[D]. Georgia: Department of Computer Science, University of Georgia, 2002.

[4] 胡金柱，周 星，舒江波，等. 基于启发式规则的网页主题信息精确定位方法[J]. 计算机应用研究，2010，27(2):494-497.

[5] 刘传昌，陈俊亮. 目标 Web 服务描述本体和服务发现模型[J]. 计算机工程，2007，33(9):187-189.

[6] 邱莉榕，史忠植，林 芬. 基于本体的语境信息模型与推理[J]. 计算机工程，2007，33(22):37-39.

[7] 石 倩，陈 容，鲁明羽. 基于规则归纳的信息抽取系统实现[J]. 计算机工程与应用，2008，44(21):166-170.

[8] 邓集波，洪 帆. 基于任务的访问控制模型[J]. 软件学报，2003，14(1):76-81.

[9] 张 欢. 基因表达式编程中的转基因关键技术研究[D]. 成都：四川大学，2006:106-108.

[10] 陆昕为. 一种改进的 GEB 方法及其在演化建模预测中的应用[J]. 计算机应用，2005，25(12):172-174.

[11] 谢方军，唐常杰，元昌安，等. 基于基因表达式的演化硬件进化和优化算法[J]. 计算机辅助设计与图形学学报，2005，17(7):1415-1420.