基于主题网络爬虫的不良网页的发现与识别

方育柯， 傅 彦， 周俊临， 夏 虎

(电子科技大学 计算机科学与工程学院 四川 成都 610054)

摘要：针对互联网中出现的大量不良内容，分析出其主要特征，首次提出将不良网页的文本特征与搜索引擎中网络爬虫相结合的技术来主动寻找互联网中的不良网页及不良网站，并将结果分级别反馈到用户层以便对不良网页和网站进行处理，以达到净化网络环境的目的. 实验结果表明，所提出的算法能够有效检测不良网页，并且能够很好地应对不良网站的反关键字过滤策略.

关键词：主题网络爬虫；不良网页；文本特征

中图分类号：TP 391；TP 181 **文献标识码**：A **文章编号**：1671－6841(2010)02－0026－05

0 引言

互联网在造福经济社会的同时，也带来了新的问题，如涉及色情、欺诈、暴力等不良有害信息在网络上的传播、扩散，对此国家先后出台了一系列法律法规和政策措施[1]，以推动互联网健康有序发展. 然而在进行社会道德和法律管理互联网的同时还需要技术层面来做支撑，为了净化互联网内容，就需要在大量 Web 网页中发现不良网页和不良网站. 目前互联网上对不良网页的处理方法主要有以下 3 种：①强制在大规模的 pc 客户端上安装过滤软件；②依靠网民的举报；③网站管理人员的审核发现. 以上 3 种方法需耗费大量的人力和财力，而且很难快速有效地消除互联网中大量存在的不良网页. 为了有效解决这个问题，本文首次提出结合搜索引擎中的网络爬虫和文本内容识别技术来检测互联网中的不良网页. 结果表明，所提出的算法能有效检测不良网页，同时能够应对不良网站反关键字过滤策略.

1 不良网页特征分析与识别

1.1 不良网页的形式特征

不良网页就是指涉及色情、暴力、反动言论等不良有害信息的网页. 互联网中的非法文本的特征可以粗略地从其具有的内容结构和用词形式两方面来描述. 其中，非法文本内容结构特征主要表现为以下 2 个方面[2]：

1)非法文本中有些内容与合法文本内容一致，在某些文本中还占相当大比例. 如有关反动内容的文章，只有少数包含反动内容.

2)使用暗语或寓意表达隐含的意思. 由于非法文本的特殊身份，文中的词常常以非正常形式出现，以逃避基于关键词的屏蔽. 对非法文本用词的特征初步归纳如下：①用标点符号将一个多字词分割开，这个词往往是文中的关键词；②在一个词中用拼音或者同音字代替其中的某些汉字；③用符号、图标或空格代替一个词，用众所周知事件暗示其代表意思. 由上述特征可以看到，非法文本中包含大量合法内容，使非法内容与合法内容相似性增大，影响对文本内容的理解；在用词上，非法文本中用各种非正常形式表示一个词，从而影响了对文本中词的正确切分.

收稿日期：2009－12－01

基金项目：国家自然科学基金资助项目，编号 60973120，60903073；国家 863 计划项目，编号 2007AA01Z440；四川省科技攻关项目，编号 2008GZ0009.

作者简介：方育柯(1984－)，男，博士研究生，主要从事 Web 文本挖掘、数据流挖掘及异常检测研究，E-mail：fangyuke@uestc.edu.cn.

1.2　不良网页的内容特征与识别

从互联网上爬下的网页中提取有效的文本内容，因为不良网页除了一些关键词本身外，还有很大一部分是为了逃避监管而在形式上有所改变的关键字变种.

1.2.1　*特殊符号间隔敏感词的处理*　汉语的书面用法通常是以逗号、句号等符号断句，文本分词也是以断句符号标志进行分词处理的.识别文本的合法性首先从文中是否使用诸如“^”、“&”、“[ ]”等特殊符号间隔敏感词，直接识别很困难，最简单的方法是建立正则表达式并利用循环语句查找所有字串间的特殊符号并删除，使得敏感信息恢复自然的组合状态.当然文中的这些特殊符号很可能是其他用途，比如计算某商品总价 = 单价 * 数量，像这样的情况将特殊符号删除是不会影响文本性质的，然而可能会影响正确分词的效果，比如将关键词分成单个汉字的情况.这种情况可以利用词库对已分词后的单个汉字进行左右相邻字间的词匹配，从而较准确地反应文中出现该词的概率及特征值.

1.2.2　*同音字或拼音替换的关键词处理*　一些非法不良文本中常常出现的敏感词中，经常会使用一些同音字或者拼音替代，这使得文本处理或者文本分词更加困难，因为现有的分词都是基于精确匹配的分词，如果不对其特殊处理，最终将不能提取任何有效的敏感词特征.针对此种情况，本系统设计了 2 个映射表来解决这个问题：拼音汉字映射表(表 1)，这是个多对一的映射；另一个是敏感词拼音映射表(表 2).通过这 2 个映射表就可以间接地实现敏感词到其多种同音字替代词的映射，显然是一对多的映射.

表 1　拼音汉字映射表

Tab.1　Phonetic transcriptions of Chinese characters

拼音	拼音对应的汉字			
pinyin(W_0)	$W_{0,0}$	$W_{0,1}$	…	$W_{0,M0}$
pinyin(W_1)	$W_{1,0}$	$W_{1,1}$	…	$W_{1,M1}$
⋮	⋮	⋮	⋮	⋮
pinyin(W_N)	$W_{N,0}$	$W_{N,1}$	…	$W_{N,MN}$

表 2　敏感词拼音映射表

Tab.2　Phonetic transcriptions of sensitive Chinese characters

敏感词	拼音
S_0	pinyin(W_0)
S_1	pinyin(W_1)
⋮	⋮
S_M	pinyin(W_M)

$S_n=(W_{n1},W_{n2},W_{n3},\cdots,W_{nk})$，$nk$ 表示字符串 S_n 中所包含的汉字的个数.因此在具备上述映射库的情况下，系统使用敏感词集的首字触发机制来设计敏感词识别算法.

输入：一段文本 $S=(s_1,s_2,s_3,\cdots,s_i\cdots s_k,\cdots,s_n)$，这里假定 S 已经使用拼音汉字对照表替换过；

输出：识别出的敏感词.

1)对 S 中当前处理的拼音 s_i，如果 s_i 出现在敏感词集的首字拼音中，则转到 3).

2)指示符向后移一位，如果 S 处理完毕则转 5)，否则转 1).

3)在 s_i 所对应的映射库进行类正向最大匹配，如果匹配到的拼音对应的词($s_i\cdots s_k$)，就说明($s_i\cdots s_k$)为敏感词并记录该词.

4)指示符向后移 $k-i+1$ 个字的拼音，转 1).

5)识别结束，返回识别出的敏感词.

1.2.3　*敏感字拆分成偏旁部首的处理*　识别不良信息的第 3 种情况就是要查询文中是否有偏旁部首及非单字出现的情况，此种情况还需要借助字典进行匹配和识别.因为汉字都是由一些简单的偏旁部首构成的，并且这些偏旁部首的总数也比较少，因此可以考虑将偏旁部首和相对应的汉字做一个映射表.同时由于文本自身的输入特征，即绝大部分的敏感字能够拆分的都是左右或者左中右拆分，因此只需对左中右结构的偏旁部首及其相应的汉字进行映射即可.

输入：一段未知的汉字串 $S=(s_1,s_2,s_3,\cdots,s_i\cdots s_k,\cdots,s_n)$，$P$ 暂存当前遇到的偏旁部首集，初始 P 为空；

输出：识别出的敏感词.

1)对当前处理的汉字 s_i，判断能否找到 s_i 对应的偏旁部首，如果不存在相应的偏旁部首，则认为 s_i 是普通汉字，清空 P，转 3)；否则记录 s_i 对应的偏旁部首到 P 中去.

2)如果 P 不为空，假定 $P=\{p_1,p_2,p_3,\cdots,p_i\cdots p_j,\cdots,p_n\}$，假定 T 暂存当前处理的偏旁部首，初始为 $T=\{p_1,p_2\}$，

a.如果 T 中的偏旁部首 $p_i\cdots p_j$ 能够构成汉字 w，则获得这个 w，转 b；

b. 如果 P 中还有未处理的部首,将指示符后移2个位置,设置 $T=p_{j+1}p_{j+2}$,转a;

c. 如果 P 中还有未处理的部首,指示符向后移1个位置,将 p_{j+1} 添加到 T 中去,转a;

d. 算法结束,清空 P 并返回识别出的词.

3)如果 S 中还有未处理的汉字,指示符向后移一位,转1).

4)算法结束,返回识别出的词.

1.3 不良网页的识别

通过对不良网页提取特征词以后,得到网页中的敏感词以及相应的词频,然后依据敏感词库中词的相应权值来计算这个网页是不良网页的概率.其中敏感词的权重使用基于互信息的特征选择算法[3]生成,并且用户可以添加指定自定义的敏感词.

假设网页 P 经过上面处理以后得到的特征词 $FW=(w_1,n_1;w_2,n_2;\cdots;w_N,n_N)$,$w_i$ 在敏感词库中相应的权值为 v_i,则网页 P 是不良网页的概率为

$$P(FW)=\sum_{i=1}^{N}\frac{v_i n_i}{\text{sum } v * \text{sum } N}, \tag{1}$$

其中,$\text{sum } v=\sum_{i=1}^{N}v_i$,$\text{sum } N=\sum_{i=1}^{N}n_i$. 以上描述的只是对已经下载下来的不良网页的识别方法,但是系统的目的是识别互联网上所有的不良网页,这就用到了搜索引擎中的网络爬虫技术.

2 搜索不良网页的主题网络爬虫设计

搜索不良网页的主要任务是尽可能多地探测不良网页及其相关网页,尽可能少地下载无关网页,提高网络资源的覆盖度,因此主要解决有效资源发现和不良网页识别两个问题.

2.1 基于超链结构的资源发现技术

基于链接结构评价的搜索策略,是通过对Web页面之间相互引用关系的分析来确定链接的重要性,进而决定链接访问顺序的方法. 通常认为有较多入链或出链的页面具有较高的价值,PageRank算法[4-5]是最具有代表性的算法,本文也采用PageRank技术.

PageRank算法中,页面的价值通常用页面的PageRank值表示,若设页面 p 的PageRank值为 $PR(p)$,则 $PR(p)$ 采用如下迭代公式计算:

$$PR(p)=\gamma\frac{1}{T}+(1-\gamma)\sum_{c\in\text{in}(p)}\frac{PR(p)}{\text{out}(c)}, \tag{2}$$

其中,T 为计算中的页面总量,$\gamma<1$ 是阻尼常数因子,$\text{in}(p)$ 为所有指向 p 的页面的集合,$\text{out}(c)$ 为页面 c 出链的集合.基于PageRank算法的网络爬虫在搜索过程中,通过计算每个已访问页面的PageRank值来确定页面的价值,并优先选择PageRank值大的页面中的链接进行访问.

虽然基于链接结构评价的搜索考虑了链接的结构和页面之间的引用关系,但忽略了页面与主题内容的相关性,在某些情况下,会出现搜索偏离主题的问题[6].另外,搜索过程中需要重复计算PageRank值,计算复杂度随页面和链接数量的增长呈指数级增长.

2.2 基于启发式规则的搜索策略

针对上面使用超链结构来进行搜索所出现的问题,加上不良网页在整个互联网里所占比例小的特点,如果单纯使用超链结构的搜索算法往往事倍功半,因此需要依据不良信息的特征来给网络爬虫制定一些启发式规则,以便给出网络爬虫在资源搜索时的指导策略,以提高资源搜索的准确性和有效性.

通过对多个超链接的研究发现,互相链接的网页之间内容相关的链接仅仅占全部链接的7成以上,因此一般的链接并不意味着网页间存在内容上的相关性.为了找到主题相关页面,需要从各个方面捕捉有用信息,将其融入网络爬虫的搜索策略中.本系统中考虑以下几种Web信息[6-7],它们对于某个URL所指向页面的主题相关性的判断提供了重要的帮助,示例如图1.

URL信息:就网页制作者来说,他们一般比较习惯于在自己制作的页面所对应的URL旁边加入与该页面主题有关的一些信息来反映页面的主题.图1中的title信息"http://www.epochtimes.com/gb/8/3/

21/n2053083.htm"，由于"www.epochtimes.com"是一个典型的反动网站，因此可以判定后面紧跟的就是一些不良网页的链接.

```
<A title=血腥镇压西藏告急 href="http://www.epochtimes.com/gb/8/3/21/n2053083.htm"><SPAN class=style9>
【热点互动】又见血腥镇压西藏告急(1)<FONT color=#ff0099 size=-1>图
</FONT></SPAN></A>
```

图 1　不良网页内容源文件示例

Fig.1　Example of source file in harmful Webpage

锚文本信息：超链接中的锚文本，即标记文本对该链接所指向的页面也起到了概括描述的作用，这种概括在一定程度上可能会比该页面的作者所做的概括(页面的标题)更为客观、准确.比如图 1 例子中的"又见血腥镇压西藏告急"信息，所代表的页面主题就很可能是不良网页.

父亲页面信息：如果页面 u 中包含页面 v 的链接，那么页面 u 叫做页面 v 的父亲页面.一般情况下，若父亲页面的内容是不良文本的话，那么父亲页面所包含的链接也是不良文本的可能性也较高.

兄弟页面信息：兄弟页面是指位于同一父亲页面的链接所指向的页面.同样根据前文的讨论，对于链接到某一主题页面的那些页面，它所包含的其他链接也趋向于链接到该主题.即若某个页面有较多的关于某个主题的兄弟页面，那么该页面很可能是与该主题有关的.

参照上面 4 个信息制定启发式规则，即在获取链接的 4 个信息以后，综合分析上面的几个信息来优先考虑与主题内容相关的网页链接，从而进一步提高了不良网页检测的速度.

3　实验与分析

基于主题网络爬虫的互联网不良网页检测分为 3 个阶段：①系统初始化阶段.主要是设置系统运行的初始参数，如网络爬虫所使用的最大线程数量、初始种子网站、网络爬虫一个网站内部链接爬下的最大深度大小、系统所使用的各种词库的路径、爬下网页内容分析时所使用的各种参数等.②网络爬虫获取网页阶段.在链接队列按照 PageRank 值和所属网页是不良网页的概率值进行优先级重排以后，网络爬虫使用多线程来从优先队列中取出网页链接，获取网页.③网页分析阶段.针对爬虫爬下来的网页分析的过程，然后按照前面的算法来计算该网页是不良网页的概率，并依据该值来判断该网页是否为不良网页.如果是不良网页，则保留该网页，登记这个不良网页.最后将上面得出的概率值累加到所属的网站上面去，以便能够识别不良网站.

为了对本文设计的算法作出准确客观的评价，构造如下训练集：人工标注 40 个网站的 400 个网页，每个网站包含 10 个网页，共 400 个网页(即 400 个 URL 信息).其中 10 个网站不含不良信息，20 个网站包含部分不良网页，剩余 10 个网站全部为不良网页.

表 3　不同特征词数量的实验结果

Tab.3　Experimental results of different feature words

特征词数量	网页(400)					网站(40)				
	实际不良网页数	检测不良网页数	准确率/%	误报率/%	漏报率/%	实际不良网站数	检测不良网站数	准确率/%	误报率/%	漏报率/%
500	200	118	79.5	0	41	30	16	76.6	0	46.6
1 000	200	134	83.5	0	33	30	20	83.3	0	33.3
2 000	200	183	94.7	1	8.8	30	27	88.3	10	13.3
3 000	200	189	95.2	2	5.5	30	29	91.6	10	6.66

从系统的运行结果可以看出，特征词数量选择从 500 到 3 000 的过程中，准确率逐渐上升，误报率也在上升，这基本上与实际情况相一致，因为选择的特征词太多，一些不太明显区分正常信息和不良信息的词语会对分类器产生干扰，影响分类效果.之所以探测不良网站的准确率没有不良网页高，并且网站的误报率比网页误报率也高，是因为准确率计算方式的影响，在不良网站的准确率计算中，健康网站和不良网站的比率为 1∶3，而网页的比率为 1∶1.总体来看，本文所提出的基于主题网络爬虫的不良网页的识别还是比较令人

满意的.另外,用户可以在准确率和误报率之间进行折衷,以达到满意的效果.

4 结束语

本文在分析了当前互联网上关于不良网页的发现技术及缺陷的基础上,提出了一种从互联网上检测和识别不良网页的算法和适用于不良网页发现的网络爬虫的爬行策略,并实现了一个原型系统.实验结果表明,本文提出的方法能够切实有效地对不良网页和网站进行发现,同时能够应对不良网站反关键字过滤策略.下一步的研究工作是提高网络爬虫的爬行效率和不良网页识别的准确度和识别效率,同时融合图像识别技术来增加对色情网站的发现和识别.

参考文献:

[1] 姜帆,张霁雪. 我国政府对互联网的管制[J].财经界:下半月, 2006(12):75-81.

[2] 张永奎,李东艳.互联网中非法文本特征分析及其属性预选取新方法[J].计算机应用,2004,24(4):114-115.

[3] 陈平, 刘晓霞,李亚军.文本分类中改进型互信息特征选择的研究[J].微电子学与计算机,2008,25(6):194-196.

[4] Page L, Brin S. The PageRank citation ranking: bringing order to the Web[EB/OL].[2009-11-01]. http://www.db.stanford.edu/~backup/PageRanksub.ps.

[5] Arasu A, Novak J, Tomkins A, et al. PageRank computation and the structure of the Web: experiments and algorithms [EB/OL]. [2010-03-01]. http://citeseerx.ist.psu.edu/viewdoc/summary? doi=? doi=10.1.1.18.5264.

[6] Ester M, Gross M, Kriegel H P. Focused Web crawling: a generic framework for specifying the use interest and for adaptive crawling strategies [EB/OL]. [2010-01-11]. http://www.dbs.informatik.uni-muenchen.de/~ester/papers/VLDB2001.submitted.pdf.

[7] Arasu A, Cho J, Garcia-Molina H, et al. Searching the Web[J].ACM Transactions on Internet Technology, 2002, 1(1): 1-42.

Unhealthy Webpage Detection Based on Topic-focused Web Crawler

FANG Yu-ke, FU Yan, ZHOU Jun-lin, XIA Hu

(*School of Computer Science and Engineering, University of Electronic Science and Technology, Chengdu* 610054, *China*)

Abstract: Internet is making massive amounts of harmful information, and it is very important to remove as much harmful information as possible to purify the internet. After the analysis of a large amount of harmful information on the internet, the key text features of harmful contents are presented. The novel approach is to find harmful Webpage and site by embedding the harmful text features into the Web spider of the search engine, and generate multi-level results to the users so that they can deal with the harmful Webpage and site to purify the internet environment. The experiments show that the proposed algorithm is capable to detect unhealthy Webpage effectively, and cope with the strategy of anti-keywords filtering from the unhealthy Website.

Key words: topic-focused Web crawler; unhealthy Webpage; text feature