ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.8, No.26, September 2012.

E-mail: info@cccc.net.cn
http://www.dnzs.net.cn
Tel:+86-551-5690963 5690964

基于版权信息的新闻网页去重策略研究

葛晓玢,刘杰,崔健

(铜陵职业技术学院 信息工程系,安徽 铜陵 244000)

摘要:随着网络技术的迅速发展和互联网络规模的不断扩大,人们能够获得的新闻信息资源也日益丰富。应用搜索引擎进行检索,经常会得到内容相同或相近的新闻网页,它们不但浪费了存储资源,而且加重了用户检索和阅读的负担。网页去重处理是提高搜索引擎的关键技术之一,因此,发现并去除重复网页信息的研究工作具有重要意义。文中提出了一种基于版权信息的新闻网页去重算法,其主要思想是:应用转载的新闻网页大多会标出其来源这一特征,并结合网页文本内容进行新闻网页去重。实验结果表明:该方法有效,对新闻网页实现较好的去重,能够得到较高的正确率及召回率,具有很好的应用价值。

关键词:网页去重;搜索引擎;版权;新闻网页;模糊匹配

中图分类号:TP311 文献标识码:A 文章编号:1009-3044(2012)26-6211-04

Detection and Elimination of Duplicated News Webpages Based on Copyright

GE Xiao-bin, LIU Jie, CUI Jian

(Department of Information,Tongling Vocational College, Tongling 244000, China)

Abstract: As the World Wide Web grows rapidly to become the largest and the most popular source of readily available information, it is increasingly abundant to access to information sources.Application of search engines, users often get the redundant news webpages with same content or similar news webpages, they will not only be a waste of storage resources, and increase users to retrieve and read the burden. Weeding out duplicated news webpages is one of the key technologies of search engine, Consequently,to detect and eliminate those pages in facsimile is of great significance. In this paper, a method based on copyright information is proposed to detect and eliminate the duplicated news webpages, This method basic thought is: reprint of most of the news webpages will be the source of its marked characteristics, combined with the text content of the page to re-page news. The experimental result indicates that,this method can complete in view of the news content duplicated news webpages,and can be a high accuracy rate and the rate of recall.

Key words: duplicated webpages; search engine; copyright; news webpages; fuzzy matching

随着互联网技术和规模的空前发展,应用搜索引擎已经成为人们从互联网上获取信息的主要渠道之一。搜索引擎以一定的策略在互联网中搜集、发现信息,对信息进行理解、提取、组织及处理,并为用户提供检索服务,从而起到信息导航目的[1]。但是,由于互联网上的信息经常被相互转载,因此检索出的网页大多是具有相同信息的重复网页。根据中国互联网络信息中心2005年7月发布的统计报告显示,用户在回答"检索信息时遇到的最大问题"这一提问时,选择" 重复信息太多" 选项的占44.6 %,排名第1位;据统计,目前Inte rnet 近似镜像网页数的比例约占全部页面的30%~45%[2];清华大学IT可用性实验室对Google、Baidu等中文搜索引擎的研究表明,重复网页占全部网页的比率,Google约占3.4%,Baidu约占2.1%[7]。国际上对转载文档去重方法的研究最初主要是针对大型文件系统,1993年,Arizona大学的Manber提出了一个SIF工具,用基于字符串匹配的方法来度量文件之间的相似性,用于在大规模文件系统中寻找内容相似的文件[8-9]。网页之间的大量转载使得网页冗余非常严重,大量重复网页的存在一方面加重了用户检索和阅读的负担;另一方面,也浪费了大量的存储资源,降低了索引效率,影响了准确率和召回率[3-7]。因此,准确、快速的发现并去除重复网页将是一项具有实际意义的工作。随着文本信息处理技术的发展,人们判断和处理近似网页的方法也日益丰富,如文本分类、聚类技术、特征码检索技术、特征串模糊匹配技术等等。

本文以新闻网页为研究对象,提出了基于版权信息的新闻网页去重策略,该方法应用转载的新闻网页大多会标出其来源这一特征,并结合网页文本内容进行新闻网页去重,提取主题信息。

收稿日期:2012-08-31

作者简介:葛晓玢(1983-),女,安徽淮北人,铜陵职业技术学院信息工程系,助教,淮北师范大学在读硕士,主要从事数据挖掘的研究;刘杰(1983-),男,安徽铜陵人,铜陵职业技术学院信息工程系,讲师,淮北师范大学在读硕士,主要从事网络安全、云计算研究;崔健(1980-),男,安徽安庆人,铜陵职业技术学院信息工程系,助教,合肥工业大学在读硕士,主要从事数据挖掘的研究。

1 基于版权信息的新闻网页去重策略研究

1.1 新闻网页特点分析

新闻网页是一种特殊的文体,是对最近发生的或正在发生的事情进行实时报道,体现了真实性、时间性和公开性等特点。对新闻网页的识别,人们通常从URL及结构特征两方面进行研究:1)新闻网页的URL特征主要表现在,绝大部分具有时间特征,不同网站的相同版块 URL 结构相似。URL的特征包括积极特征和消极特征。采用与内容无关的 URL属性可以使我们大大的提高网页分析的速度。2)新闻网页的结构特征主要表现在,网页含有丰富的结构信息,合理的利用这些信息,可以提高分类器的性能。在新闻网页中,其中<title> 和<H_n> 节点标注网页的标题信息, 节点划分网页的层次结构,提取并分析这些结构特征中对新闻网页去重有价值的网页特征[2,4,7]。

基于用户兴趣的驱动,网络信息流通中人们通过复制方式进行信息共享,一些重要的新闻网页,有时转载竟高达几十次之多。新闻网页在重复方面的特点,主要有:

1)重复率高。有时转载竟高达几十次之多,甚至有些网站为了提高点击率,进而大量转载热门的文章和新闻;

2)新闻网页的重复也来自对于同一事件的不同报道。尤其对于人们关注的热点新闻,对原网页进行再编辑或截取部分内容造成的新的网页,虽然此时信息量发生了变化,但内容和以前还是较为相似。重复网页的文本经过提取并去除噪声后,在内容和结构方面能够保持高度一致;

3)局部性明显。主要表现在转载内容的局部性和转载时间的局部性。前者是指其转载的新闻网页内容主要偏向于人们关注的热点新闻页面;而后者是指转载的时间比较集中,大都在一两天内进行转载,一段时间以后再转载则很少;

4)存在噪声。 新闻网页转载时一般都"原样照搬",保持文本内容和结构的一致,并尊重版权,在开头加入了引文信息。也就是说这些网页在去掉噪声以后,能够在内容和结构方面能够保持高度一致。本文也正是应用这个特点进行新闻网页的去重。

1.2 页面预处理与版权信息识别

Web新闻页面与普通文本类似,在新闻网页文本区上方通常是广告条、导航条、网站标志、版权信息等等,左右两侧是网页的超链接,下方是文本的相关超链接,多为半结构化的文本,有着大量的标签和嵌套的结构。大多数新闻网页通过明确的HTML标记来表示网页标题。应用新闻网页的结构特征,可以通过相关标记来抽取新闻标题、新闻页面转载时间、转载来源等信息,对页面进行预处理。文中提出的新闻网页去重算法主要是基于版权信息的,因此在网页去重前必须提取出新闻网页转载的版权信息,它直接决定着网页去重的效果。

文中主要借助HtmlParser来提取上述信息。HtmlParser主要靠Node、AbstractNode和Tag来表达HTML。 HtmlParser 为其提供了强大而灵活易用的开源类库来处理Internet上的网页,它提供了线性和嵌套两种方式来解析网页,主要用于HTML 网页的转换(Transformation) 以及网页内容的抽取 (Extraction)。HtmlParser 还有一些易于使用的特性,如过滤器 (Filters),访问者模式 (Visitors),处理自定义标签以及易于使用的 JavaBeans等供用户选择,实现对网页特定内容的提取和修改。

1.3 去重算法设计

1.3.1 去重策略及系统原型设计

通过前面的工作,可以应用HtmlParser的visitors访问模式,提取新闻标题、发布时间以及新闻转载来源等信息过滤掉无关的一些超级链接信息。据统计,绝大部分新闻网页信息在转载时,有关版权信息的描述方式主要有:1)给出新闻网页转载出处信息,没有关键字"来源",如图1所示;2)直接标出"来源:×××",如图2所示。

http://www.sina.com.cn 2008年03月18日11:40 人民网

图1

www.hellobj.com.cn • 2008-1-4 9:49:52 • 来源: 新京报

图2

应用上述特征,可以结合新闻页面主题,解决新闻页面的去重问题。对于极少部分重复但又没有标明来源的新闻网页可以采用其它去重方法,如基于发布时间的去重方法、基于特征码的新闻网页去重方法等,以提高的去重的准确率和召回率。

系统原型设计主要包括三个模块,分别为页面预处理模块、特征码生成模块及网页去重模块。在页面预处理模块,一方面需要借助HtmlParser来提取有关版权信息;另一方面,去除包含导航、广告等一些噪声信息。在特征码生成模块,使用参考文献[3]中的方法来实现。在网页去重模块,考虑两方面。对于正确提取版权信息的新闻网页,一般基于新闻内容,可以比较版权信息,继而很方便地去除重复页面;针对转载时没有标记版权信息的页面,应用特征码生成模块,获取特征码信息,应用特征码,去除这小部分的重复页面。

1.3.2 算法流程

基于版权信息的新闻去重算法流程如下:

1)页面预处理,提取有关版权信息内容,并存入文本文件中,该过程主要代码描述如下:

```
Public static getKeyWordText(String url, keyword)
//提取网页的HTML文档,生成解析对象
```

```
{
StringBuffer document=new StringBuffer( );
      try {
 URL url=new URL(urlString);
url.SetEncoding("gb2312");
//设置网页编码
URLConnection conn = url.openConnection( );
//建立到URL的链接
BufferedReader reader = new BufferedReader(new InputStreamReader(conn.getInputStream( )));
String temp =" ";
while (temp = reader.readLine( ))!=null)
   { document.append(temp);
     document.append("\r\n");
//读取HTML文档
reader.close( );
}
catch {…}
return document.toString( )
}
//获得HTML文档代码
}
```

2)重复新闻网页判定。经过页面预处理以后,对提取出来的文本进行检索。

```
while("版权信息"在提取的文本中)
{
删除该文本;
返回其出现次数;
}
//对余下的极少部分网页应用基于特征码的去重算法进行分析
while (文本不为空)
   { //从网页文档中提取特征码信息
S1: 进行扫描;
 //根据特征码提取方法,假定在网页文档的某一个指定的"。"处提取特征码信息
S2: if (扫描到指定位置)
则提取它的前五个字,作为主码和后五个字,作为辅码;
            返回true;
         else
返回false;
S3: 用特征码将网页文档逐个索引构建一个检索系统;
S4: 将每个网页文档的特征码逐个投入检索系统,将检索到的全部网页聚成一类。
S5: 对特征码进行比较,去掉含有相同特征码的网页。
}
```

2 实验及结果分析

分析去重效果主要有两个重要指标:正确率及召回率[4]。正确率,它反映了应用该算法所能发现的近似镜像网页中有多少是正确的近似镜像网页结果,不妨设应用该算法检测到了S个重复网页,其中S_0个正确结果,则算法的正确率可定义为:

$$\text{正确率}=\frac{\text{正确去重文本数}}{\text{检测出的重复网页个数}}=\frac{S_0}{S}$$

召回率,它反映了应用该算法所能发现的正确的近似镜像网页数量占全部近似镜像网页数量的百分比。不妨设应用该算法检测到了S_0个正确的近似镜像网页,而数据集中实际存在S_N个近似镜像网页。则算法的查全率为:

$$\text{召回率}=\frac{\text{正确去重文本数}}{\text{实际存在的重复网页个数}}=\frac{S_0}{S_N}$$

为了验证文中提出的算法性能(包括算法的有效性,正确率、召回率以及处理每一类网页的去重平均时间),我们当时应用搜索引擎进行搜索"金融危机中各国竞争力调查,中国仍然高居榜首"新闻信息,人工收集了搜索返回的约8420篇网页,其中包括约1548

篇部分重复新闻网页，在PC机器CPU为T5750，2.00GHz，内存为1024M，操作系统为Windows XP的实验环境下进行实验，应用正确率及召回率对算法进行评价，实验结果如表1所示：

表1 实验结果

	实际存在的重复网页	检测出的重复网页	检测出正确重复网页	正确率	召回率
全部重复	6872	6769	6553	96.81%	95.36%
部分重复	1548	1522	1416	93.05%	91.42%

另应用基于特征码的全程去重算法以及基于版权信息的去重算法，在上述相同环境下进行对比实验，实验数据取上述人工收集的约6872篇全部重复的新闻网页，实验结果如表2、图3、图4所示：

表2 实验结果

算 法	正确率%	召回率%
基于特征码	97.15%	92.71%
基于版权信息	96.81%	95.36%

根据表2，我们可以得出，两种算法在实现新闻网页去重，其正确率方面较为接近；但在召回率方面，基于版权信息的去重算法比单纯应用基于特征码的去重算法较为优越。究其原因分析，基于版权信息的去重算法，它在分析转载网页时，应用版权信息，已经去掉了大量重复网页，对于极小部分网页，没能正确提取版权信息，我们采用特征码方法进一步去重，从而达到了较高的召回率。而基于特征码算法去重，并没有考虑相似文本，所以它的召回率相对来说就不是很理想。

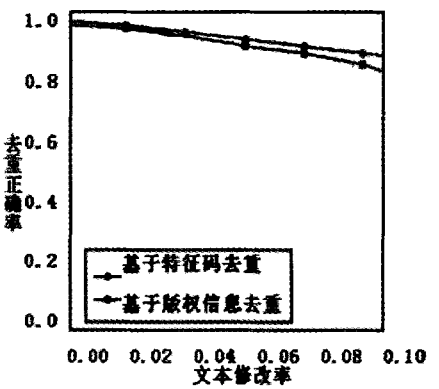

图3 文本修改率与去重正确率的关系

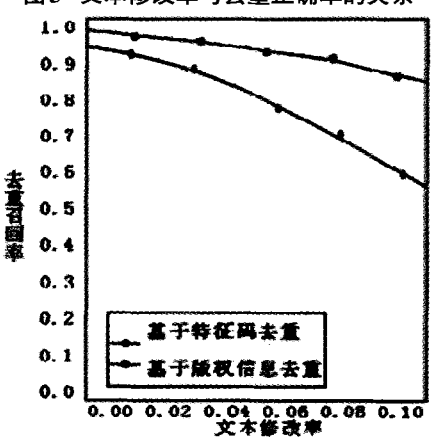

图4 文本修改率与去重召回率的关系

根据图3，基于版权信息去重算法，其去重正确率稍低于基于特征码去重算法，究其原因分析，有部分新闻网页在转载时，没有尊重版权，缺少版权信息的文字描述；根据图4，当新闻网页部分内容修改时，基于版权信息的去重算法明显优于基于特征码的去重算法，究其原因分析，主要是因为基于特征码的去重算法对文本修改非常敏感，微小的文本内容修改均有可能导致新闻网页特征码的变化。

3 结束语

本文以新闻网页为研究对象，提出了基于版权信息的新闻网页去重算法，与传统的一些新闻网页去重算法相比，本文提出的算

（下转第6227页）

络的消亡时间去换取网络的生存时间是值得的。

4 结束语

论文基于距离的概念，对LEACH协议的路由算法进行了改进，提出了LEACH-DB路由算法。该路由算法考虑了无线传感网络中各个节点与基站之间的相对位置，并通过这种相对位置的关系来有意识的影响各个节点成为簇头的概率，从而影响了簇头的总体地理位置分布，使它们更加靠近基站，有效的减小了簇头与基站之间的数据通信开销，延长了网络的生存时间，提高了网络性能。从仿真时间的结果可知，LEACH-DB路由算法对于网络生存时间的提升，相对于LEACH协议，延长了大约25%。这是一个比较可观的提高，说明LEACH-DB路由算法是行之有效的。

论文中LEACH-DB算法并没有考虑各个节点剩余能量情况，而根据节点剩余能量的概念来均衡整个网络的能量消耗也是一个延长网络生存时间的有效手段。因此，今后的研究工作会围绕这个问题继续深入下去，以期将距离和剩余能量这两个概念结合起来，更加有效的提高整个网络的工作性能。

参考文献：

[1] 廖明华,张华,王东.基于LEACH协议的簇头选举改进算法[J].计算机工程,2011(7):112-114.
[2] Wendi Rabiner Heinzelman,Anantha Ch, Hari Balakrishnan.Energy-Efficient communication protocol for wireless microsensor networks[C].Proceedings of the Hawaii International Conference on System Science,January 4-7,2000,Maui,Hawaii.[S.l.]:IEEE Computer Society,2000:3005-3014.
[3] 张伟华,李腊元,张留敏,等.无线传感器网络LEACH协议能耗均衡改进[J].传感技术学报,2008(11):1918-1922.
[4] 路纲,周明天,佘堃,等.无线传感网络路由协议的寿命分析[J].软件学报,2009(2):375-393.
[5] 吕涛,朱清新.一种基于LEACH协议的改进算法[J].电子学报,2011(6):1405-1409.
[6] 胡钢,谢东梅,吴元忠.无线传感器网络路由协议LEACH的研究与改进[J].传感技术学报,2007(6):1391-1396.
[7] 谢丽惠,汤碧玉,施海彬.基于NS3的LEACH协议仿真与改进[J].厦门大学学报:自然科学版,2010(2):193-197.

(上接第6214页)

法具有速度快，检测率高，算法容易实现等特点，能够有效地去除检索结果集合中内容相同或相近的新闻网页，能够在网页发生修改时兼顾查全率和查准率，更适合网页的去重。文中对新闻网页的结构特征进行了深入分析，但结构特征的选择具有一定局限性。下一步研究的重点是对文中提出的算法进一步进行完善，使得聚集的结果更合理，更符合网页本身的特征，提高检索质量，指导工程实践。

参考文献：

[1] 李晓明,闫宏飞. 搜索引擎原理、技术与系统[M].北京:科学出版社,2004.
[2] 高凯,王永成,肖君.网页去重策略[J].上海交通大学学报,2006,40(5):775-777.
[3] 陈基漓,牛秦洲.基于特征码的网页去重[J].微计算机信息,2006,22(3-3):113-115.
[4] 罗永莲,罗永秀,张永奎.突发事件新闻网页的去重方法研究[J].计算机应用与软件,2008,25(8):24-26.
[5] 魏丽霞,郑家恒.基于网页文本结构的网页去重[J].计算机应用,2007,27(11):2854:2856.
[6] 王建勇,谢正茂,雷鸣,等. 近似镜像网页检测算法的研究与评价[J].电子学报,2000,28(11):130-132.
[7] 阎亚杰.网页去重方法研究[J].电脑开发与应用,2008,21(8):60-62.
[8] Cho J H,Shivakumar N,Garcia-Molina H.Finding ACM International Conference on Management of the Data[M]. USA:ACM Press,2000(2):355-366.
[9] Liu C J, Wechsler H. A shape and texture based en- hanced Fisher classifier for face recognition[J]. IEEE Transactions on Image Processing, 2001,10(4):598-608.