第26卷 第3期
2011年6月

郑州轻工业学院学报(自然科学版)
JOURNAL OF ZHENGZHOU UNIVERSITY OF LIGHT INDUSTRY(Natural Science)

Vol. 26 No. 3
Jun. 2011

文章编号:1004-1478(2011)03-0085-03

一种新的用于数据挖掘工具的网页净化算法

孙楠, 张华伟

(河南财经政法大学 计算机与信息工程学院, 河南 郑州 450002)

摘要:为了更好地消除网页噪声,有效地提取网页的主题内容,提出了一种新的网页净化算法.该算法认为网页的主题内容主要包含在<table>标记和<p>标记里面,并据此对网页噪声进行预处理,然后与相关网页进行内容匹配,通过计算节点重要度,获取网页的主题内容.对门户网站的6 318个网页的检测表明,该算法可以有效地提取网页的主题内容,准确率达到98.2%以上.用于数据挖掘工具时,该算法优于其他同类算法,可以有效地去除网页噪声.

关键词:网页净化;网页噪声;文档对象模型;阈值

中图分类号:TP393.08 **文献标志码**:A

An new algorithm of Web page purification for data mining tools

SUN Nan, ZHANG Hua-wei

(*College of Comp. and Infor. Eng., He'nan Univ. of Econ. and Law, Zhengzhou* 450002, *China*)

Abstract: In order to eliminate noise preferably and extract topic content from Web pages efficiently, an algorithm of Web page purification is presented. This algorithm argues that topic content of Web page is mainly contained in <table> and <p>, hereby Web noise can be preprocessed. Then with the content match of relevant Web page, the topic content of Web page can be acquired by way of calculating the importance of node. This algorithm has achieved very precise results, correctly extracting 98.2% of the pages in a set of 6 318 pages in portal sites. When used for data mining tools, this algorithm is better than the other similar algorithms. It can eliminate noise efficiently.

Key words: Web page purification; Web noise; DOM; threshold

0 引言

随着科技的迅猛发展,互联网已经成为用户获取信息的重要途径.面对海量的数据,用户一般采用数据挖掘工具来获得自己需要的信息,其中最流行的是搜索引擎(如Google、百度等).可是,大量的网页噪声会导致主题漂移[1],使得同一网页中存在多个主题.传统的Web数据挖掘工具以整个网页为粒度进行搜索,难以准确快速地获取所需要的信息.因此,如何消除网页噪声,获取网页的主题内容(又称网页净化),成为数据挖掘领域的研究热点之一.

目前,国内外对网页净化算法展开了诸多研究.Lan Yi 等[2]认为所有的噪音块具有相似的风

收稿日期:2010-12-29

作者简介:孙楠(1983—),男,河南省新野县人,河南财经政法大学助教,硕士,主要研究方向为数据挖掘、图像处理、优化算法.

格,基于此提出了 SST(site style tree)的思想. 通过对网站网页的取样建立 SST,根据 SST 的测量确定网页的噪音部分. 但是,由于需要对网页结构进行多次递归处理,效率相对较低,面对现在海量的网页数据,其实用性欠缺. 张恒等[3]提出网页的主题信息存在于 <p> 标记中,对网页中的所有 <p> 标记进行处理,以提取网页的文本信息. 这种方法在获得结构简单的网页的主题内容时简洁有效,但面对复杂的网页结构尤其是图片信息时效果不佳. 邱江涛等[4]提出通过一次遍历把网页的主题内容和噪声划分至不同的块中,根据块的分布,使用合适的分类方法来判别网页中是否存在主题内容,并使用孤立点分析的方法提取网页的主题内容. 但仅仅依靠单个网页进行主题内容提取,准确度不高. S. H. Lin 等[5]利用 <table> 标记的布局特性进行解析,每个 <table> 标记在网页中所占的长宽一般各不相同,越是重要的 <table> 标记在网页中所占的比例越大. 通过统计每个 <table> 标记所占的比例,寻找网页的主题内容,显然该算法对没有使用 <table> 标记布局的网页不适用.

为了更好地消除网页噪声,有效提取网页的主题信息,本文认为网页的主题内容主要包含在 <table> 标记和 <p> 标记里面,并据此对网页噪声进行预处理,有效降低算法的时间复杂度. 然后,使用内容匹配的方法进行节点重要度的计算,获取网页的主题内容.

1 网页噪声分析

由于 HTML 页面是一种半结构化数据,用于显示的代码和网页的主题信息混在一起,而且网页设计者可以随意加入其他无关的信息,因而网页中包含了大量的噪声信息. 根据其粒度的不同,网页噪声可以分为总体噪声和本地噪声.

1)总体噪声是指整个 Web 页面在互联网上的副本,主要包括历史网页、镜像站点和非法链接等.

历史网页主要用于搜索引擎,如百度快照、Google 的网页快照等[6],当原始网页所在的服务器出现故障或链接失效时,发挥替补作用. 镜像站点作为主站点的备份,一般不与主体站点在互联网上同时出现. 综合来说,总体噪声在网页噪声中所占的比例较小.

2)本地噪声是指单一网页中与主题无关的信息,主要包括广告、导航条、交互类元素和装饰类信息等.

广告是网站以营利为目的的宣传手段. 导航条用于保持网页间的链接关系,使用户可以方便地访问所需的信息. 交互类元素是网站为确认用户身份、提取用户意见等而设置的元素. 装饰类信息是为了增加用户兴趣、提高点击率而使用的图片和动画等. 本地噪声在网页噪声中所占的比例很大,是网页净化的重点.

2 网页净化算法描述

2.1 基础启发性规则

1)同一网站的网页之间,主题内容各不相同,而噪声内容却大致相同.

2)网页的正文内容主要存在于 <p> 标记和 <table> 标记之中.

以上的启发性规则是合理的. 同一网站的不同网页之间主题内容是截然不同的,然而导航条、广告植入以及版权类信息等却是大致相同,不可能完全单独维护. 网页的主题内容包括文字类主题内容和图片类主题内容,前者一般包含在 <p> 标记和 <table> 标记里面,后者一般由 <image> 标记控制显示并包含在 <table> 标记里面,当然 <p> 标记和 <table> 标记里面也包含了大量的网页噪声.

2.2 算法流程

输入:页面集合(DOM 树结构)

for 对 DOM 树进行后序遍历中遇到的节点 E

switch(E):

case E 是 <title> then 保留节点 E

case E 是 <p> then 用公式①计算 E 的节点重要度 *NodeIMP*(E)

if *NodeIMP*(E) > β then 保留节点 E//β 为最佳阈值

else 不保留节点 E

case E 是 <table>

if E 包含未被访问的 <table> then 继续遍历

else 用公式①计算 E 的节点重要度 *NodeIMP*(E)

if *NodeIMP*(E) > β then 保留节点 E

else 不保留节点 E

输出:网页净化结果(即所有保留节点内容)

算法说明:

1)算法开始时使用开源工具 HTML Parser 对页面集合进行解析[7],获得网页的 DOM 树结构.

2)网页中的文本内容多数存在于 <p> 标记中,它的显示控制代码对 Web 数据挖掘工具来说完

全没有意义. 而且, 标记之间不存在嵌套关系. 因此,只要提取 标记和 标记之间的内容就可以了.

3)单个网页的结构都十分复杂,网页中一般有多个<table>标记. 由于<table>标记之间存在复杂的嵌套关系,所以要对<table>标记进行递归分离,消除<table>标记之间的嵌套关系.

4)通过计算节点的重要度,以确定该节点是网页的主题内容还是噪声内容. 节点重要度定义为

$$NodeIMP(E) = 1 - \log_m P \qquad ①$$

其中,m 为网页的总数,P 是节点 E 中的内容出现在网页集合中的次数. 实验表明,$\beta = 0.75$ 是算法的最佳阈值(实验在第3部分给出),即当 $NodeIMP(E) > 0.75$ 时认为节点 E 属于网页的主题内容,当 $NodeIMP(E) \leqslant 0.75$ 时,节点 E 属于网页的噪声内容.

5)把属于网页主题内容的节点集中起来,并单独提取网页的标题,获得网页净化的结果.

2.3 方法实现及净化结果

本方法在 Windows XP 系统下,使用 Eclipse 3.4 平台开发,内嵌 JDK 为 1.6 版本. 网页净化前后的结果对比如图1所示.

图1　网页净化前后结果对比

3　净化算法最佳阈值确定及实验结果评价

网页净化的效果依赖于2个方面:一是网页主题内容的获取率,其值由公式②确定;二是网页噪声内容的消除率,其值由公式③确定.

$$P_1 = \sum_{i=1}^{m} u'_i \Big/ \sum_{i=1}^{m} u_i \qquad ②$$

其中,u' 是净化后单个网页主题内容的字符数(u 是 useful information 的缩写),u 是单个网页原有主题内容的字符数,m 是选取网页的总数.

$$P_2 = \sum_{i=1}^{m} n'_i \Big/ \sum_{i=1}^{m} n_i \qquad ③$$

其中,n' 是净化后单个网页被消除噪声内容的字符数(n 是 noise 的缩写),n 是单个网页原有噪声内容的字符数,m 是选取网页的总数.

为了达到较好的净化效果,需要确定一个合理的阈值,通过对搜狐2 312个网页的实验分析,认为0.75是最佳阈值,实验结果如表1所示.

表1　最佳阈值实验表　　%

阈值	主题内容获取率	噪声内容消除率
0.65	99.71	88.24
0.75	98.38	98.16
0.85	92.23	99.64

网页主题内容的获取率,随着阈值的增加而减少,噪声内容的消除率随着阈值的增加而增加,在阈值为0.75时,2个参数达到最优.

确定最佳阈值之后,对网易、搜狐和新浪3个门户网站进行实验,计算网页主题内容的获取率与噪声内容的消除率,实验结果如表2所示.

表2　净化算法结果　　%

网站	主题内容获取率	噪声内容消除率
网易2 016个网页	98.19	98.23
新浪1 990个网页	98.51	97.88
搜狐2 312个网页	98.38	98.16

从表2可以看到,净化后网页主题内容的获取率与噪声内容的消除率都比较高,表明本文的网页净化算法是可行的.

4　结语

随着互联网技术的迅猛发展,网页噪声也日渐增多,因此,消除网页噪声对有效获取网页的主题信息有着重要的意义. 本文针对数据挖掘工具,提

(下转第91页)

4 结语

通过计算机仿真,本文分别得出几种典型的最大匹配调度算法在不同的流量模型下的性能:PIM和RRM类算法由于算法设计上的局限性而性能较差;iSLIP算法在均匀流量下能很好地达到100%的吞吐量;DRRM算法在性能负载较低时表现优于iSLIP算法.仿真结果表明,通过针对输入输出端口设置轮询指针并适当地变换更新轮询指针的方法,能够增加每个时隙内端口匹配的数目,有效地避免输出端同步问题发生,降低信元的平均时延.因此,在以后设计改进调度算法时,可以通过增加每个时隙中的端口匹配数目,使极大匹配更加接近最大数量匹配,来改善交换机的性能.

参考文献:

[1] Nong G, Hamdi M. On the provision of quality-of-service guarantees for input queued switches[J]. IEEE Com Magazine, 2000, 38(12): 62.

[2] 庞斌,贺思敏,高文.高速IP路由器中输入排队调度算法综述[J].软件学报,2003,14(5):1011.

[3] Anderson T, Owicki S, Saxes J, et al. High speed Switch Scheduling for local area networks[J]. ACM Trans on Comp Syst, 1993, 11(4): 319.

[4] Marsan M A, Bianco A, Leonardi E, et al. RPA: A flexible scheduling algorithm for input buffered switches[J]. IEEE Trans on Com, 1999, 47(12): 1921.

[5] McKeown N. The iSLIP scheduling algorithm for input-queued switches[J]. IEEE/ACM Trans on Networking, 1999, 7(2): 188.

[6] Chao H J. Saturn: A terabit packet switch using dual round robin[J]. IEEE Com Magazine, 2000, 38(12): 78.

(上接第87页)

出了一种新的有效的网页净化算法.利用网页的DOM树,避开了对半结构化数据的直接处理,使用合理的启发性规则,提高了算法的效率,通过对节点内容的比较,保证了算法的准确率.面对日益复杂的网页结构,如何提出更加合理的启发性规则,进一步提高网页主题内容的获取率和噪声内容的消除率是今后需要努力的方向.

参考文献:

[1] 陈治昂,周知予,李大学.一种基于模板的快速网页文本自动抽取算法[J].计算机研究应用,2009,26(7):2646.

[2] Yi Lan, Liu Bing. Eliminating noisy information in Web pages for data mining[C]//Proce of Int Conf on Knowledge Discovery and Data Mining, New York: ACM Press, 2003: 296-305.

[3] 张恒,屈景辉,张亮.网页文本信息提取及结果评价[J].微计算机应用,2007,28(9):921.

[4] 邱江涛,唐常杰,李川,等.基于块分布的新闻网页内容提取[J].吉林大学学报:工学版,2009,39(5):1326.

[5] Lin S H, Ho J M. Discovering informative content blocks from Web documents[C]//Proc of the 8th ACM SIGKDD Int Conf, New York: ACM Press, 2002: 588-593.

[6] 王琦,唐世渭,杨冬青,等.基于DOM的网页主题信息自动提取[J].计算机研究与发展,2004,41(10):1787.

[7] 王实,高文,李锦涛.Web数据挖掘[J].计算机科学,2000,27(4):28.