

文章编号:1001-9081(2007)06-1394-03

一种 Web 主题文本通用提取方法

蒲 强,李 鑫,刘启和,杨国伟

(电子科技大学 计算机科学与工程学院,四川 成都 610051)

(puqiang@uestc.edu.cn)

摘 要:为构建大规模中文文本语料库,提出了一种简单、有效、通用的中文 Web 主题文本提取方法。该方法巧妙地利用中文文本长度和标点符号序列,配合少量判别规则,便可准确地将主题文本从网页中提取出来。由于本方法不涉及具体的 HTML 标记分析,其通用性较强。实验结果表明该提取方法具有快速性和准确性,达到了构建大规模中文文本语料库的要求。

关键词:Web 文本;文本提取;文本语料库

中图分类号:TP311.52 **文献标识码:**A

Study on general extracting method of Web topic text

PU Qiang, LI Xin, LIU Qi-he, YANG Guo-wei

(College of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu Sichuan 610054, China)

Abstract: A simple and efficient method of generally extracting Chinese topic text from Web pages was proposed in this paper in order to build a large Chinese text corpus. This method just utilizes length of Chinese texts and series of punctuations, along with a few rules of discrimination, to extract needed text from Web pages accurately without analyzing HTML tags. The experiment shows the extraction is so fast and accurate that it can achieve the requirement of constructing a large Chinese text corpus.

Key words: Web text; text extracting; text corpus

0 引言

统计自然语言处理的首要工作是:基于语料库收集各种观察对象的出现次数,以此作为估计概率的基础^[1]。因此,语料库在研究机器可读的自然语言文本的采集、存储、检索、统计、语法标注、句法语义分析等问题时不可或缺。但是,从网络下载的语料库往往规模过小、内容更新慢,无法真正满足自然语言处理相关领域的研究需要。为此,本文试图利用 Web 网页丰富、及时的资源,建立一种通用的中文 Web 主题文本提取方法,来构建用于研究的大规模中文文本语料库。

Web 主题文本往往湮没在众多烦杂的 HTML 标记之间,加之 Web 网页缺乏相应的语义标记,使得我们必须利用一定的规则才能将“隐藏”的主题文本提取出来。目前有多种 Web 信息提取技术:文献[2]利用 DOM 树结构中的路径表达式来定位 HTML 文档中要提取的信息;文献[3]将 HTML 文档转换为一种含有位置信息的坐标树,结合位置特征和空间关系对网页进行分析和提取内容;文献[4]根据页面的组织结构将页面划分为若干个信息块来提取相应内容。

这些技术由于涉及网页标记的处理而比较复杂,提取规则也会随着网站模板格式的变化而变化。对构建文本语料库的任务而言,由于只需 Web 页面的主题文本,对链接、图像等

信息并不关心(相反,这些内容还会形成语料库中的“噪声”),因此上述 Web 信息提取技术在这里并不一定适用。

为此,本文提出一种简单、通用的提取方法,即:不在 HTML 标记上进行复杂分析,而利用简单的正则表达式快速滤掉所有的 HTML 标记及某些标记间的内容,然后通过简单利用中文文本长度和标点符号序列,配合少量判别规则,从网页中提取主题文本。由于不涉及分析具体的 HTML 标记,所以本方法通用性较强。

以上的处理方法基于以下假设:HTML 标记或脚本间的内容不是有效文本信息;Web 主题文本在网页内往往大块集聚在一起,并且文本句子的长度相对较长,标点符号序列较为完整,如“.,,.,。”。

1 Web 主题文本提取方法

我们提出一种不依赖网页 HTML 标记的提取方法,利用正则表达式简单强大的文本处理功能,以简洁的代码编程实现了 Web 主题文本提取工作。

1.1 文本提取的几点假设

假设 1 HTML 标记、脚本和链接信息对文本语料而言是“噪声”;

假设 2 主题文本一旦出现,就会大块集聚在一起;

收稿日期:2006-12-04;修订日期:2006-02-8

基金项目:国家自然科学基金资助项目(60471055);国家 863 计划项目(2005AA114030)

作者简介:蒲强(1971-),男,四川内江人,工程师,博士研究生,主要研究方向:统计自然语言处理、神经网络、模式识别;李鑫(1973-),男,湖北利川人,博士研究生,主要研究方向:语义 Web、逻辑计算、计算机网络;刘启和(1973-),男,重庆人,讲师,博士,主要研究方向:自然语言处理、人工智能;杨国伟(1939-),男,重庆人,教授,博士生导师,主要研究方向:自然语言处理、人工智能、计算机网络。

的平均情况。纵轴为文档各个操作步后剩余的平均字符数,横轴第 1 项 Original 为原始 Web 文档的平均长度,从第 2 项到第 5 项表示相应的去除 HTML 标记处理步骤,最后一项 ExtractedText 为实际提取的主题文本平均长度。曲线上数字表示文档经过相应标记去除操作后所剩平均字符数。可以看出,经过每步 HTML 标记处理,文档字符数的下降速率非常快。

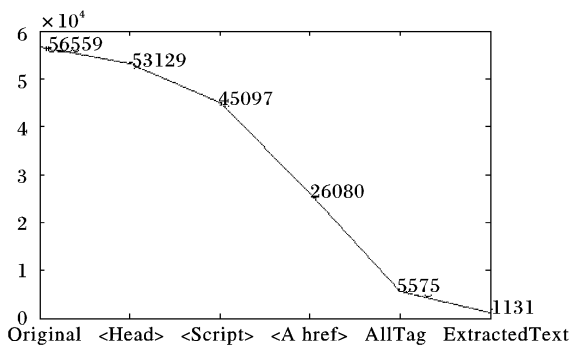


图1 去除 Web 冗余信息后文档字符变化情况

表 1 显示网页的主题文本内容只占到整个网页字符长度的 8%,可见,网页大部分内容都是 HTML 标记、脚本和链接文字,其总比例占到了整个网页的 90%。因此,快速地去掉这些冗余内容,可以提高提取可用文本的效率。

表 1 Web 页面中各种内容所占比例情况(长度单位:字符)

	原始	头	脚本	超链	剩余	实际
长度	56 559	3 430	8 032	19 089	20 505	4 444
比例	100%	6%	14%	34%	36%	8%

2.2 主题文本提取速度测试

我们的任务是对输入的几万到几十万个 Web 网页的每一个页面进行主题文本提取,因而要求文本提取速度达到实用的要求。下面给出系统的测试结果。

我们的文本提取系统使用支持正则表达式的 jdk1.4 包,用 Java 编程实现,在奔腾 3-800MHz,320M 内存的 PC 机上测试运行。

表 2 显示系统处理速度能够满足对大规模 Web 网页中提取文本的速度要求。

表 2 主题文本提取速度测试表

Web 页面 数量(个)	页面大小 总和/MB	平均页面 大小/KB	提取时间 /s	提取速度 页面数/s	提取速度 MB/s
2776	121	44	89	31	1.36

2.3 文本提取的准确性判断

在 Web 主题文本提取准确性判断上,先以人工方式判断提取的内容确实是主题文本内容,然后将提出的文本同网页中人工判断的准确内容进行文本长度比较,通过长度差异来反映提取的准确性。为了说明提取字符差异问题,我们使用了来自同一报刊网站的 50 个 Web 页面进行准确性判断。

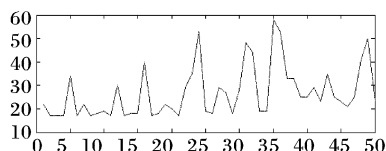


图2 提取文本和标准文本的差异波形

图 2 显示 50 个 Web 页面提取后的文档字符较标准内容

文档会多出 20~60 个字符。其中有 60% 的文档多出 20 个字符左右,例如文档 1,多出的字符为:“《中国汽车报》(2005 年 6 月 27 日 A2 版)” ;30% 的文档多出 30 个字符左右,例如文档 5,它多出了“(相关报道见 B6~B7 版)《中国汽车报》(2005 年 6 月 27 日 A1 版)” ;10% 的文档会多出 40~60 个字符,例如文档 24,多出了:“(作者是经济学博士、管理学博士后,现任职于清华大学汽车发展研究中心)《中国汽车报》(2005 年 6 月 27 日 A7 版)” 。可以看出,多出的字符往往是文末的作者介绍和引用出处。如果不去掉这些内容,实验显示系统提取平均准确率为 93.15%;这里为了说明问题,实际上我们的系统已经将这些内容去掉了,最后文本提取的准确性大于 98.63%。

2.4 实验讨论

实验表明了我们的系统在提取 Web 网页主题文本时,对各种网站的网页提取通用性较强,并且编程实现方便,处理速度很快。但这里仍然存在几个需要注意的问题:

1) 去除超链文字时,会发现正文中少量超链文字也会被去掉,这会影响正文的连续性。我们的处理办法是:将去掉的超链文字及其在页面中的位置存储起来,根据位置信息判断是否恢复相关超链文字成为所需正文。

2) 提取存在于 Web 标记之内的正文,形如:“<Tag 正文...>”,本文的方法则不适用,需要重新编写程序加以处理。幸运的是,在大型网站上,这类网页数量并不多。

3) 依照 1.2.3 节的句子权重计算,2.3 节中提取正文后多出字符形成的句子,其权重本该很低,例如:“《中国汽车报》(2005 年 6 月 27 日 A2 版)” 的权重为 $0.025 < 0.5$,本该为“噪声”文本,但由于其在程序中依附前面完整的句子一起出现,如“还需要清醒认识、理智应对、不断磨合。《中国汽车报》(2005 年 6 月 27 日 A2 版)”,其标点序列为“、。” ,计算权重为 $4.532 > 0.5$,所以判断其为主题句子。测试发现,这些“噪声”文本的内容形式和出现位置都很固定,而且字符变动范围也很稳定,因而很容易从主题文本中去掉。

3 结语

我们给出了一种简单、有效、通用的中文 Web 主题文本提取方法及具体实现细节。由于方法充分利用了中文文本长度和标点符号序列,配合少量判别规则,无需分析 HTML 标记。另外,利用正则表达式简单强大的文本处理功能,在系统编程实现上也简洁了许多。实验显示,我们的方法能够迅速自动地从 Web 网页中提取出主题文本,不依赖于网页结构,具有很高的处理通用性。只要有了大量的文本素材,就可以进行后续的分词处理、标注及统计相关信息,从而构建研究所需的大规模中文文本语料库。

参考文献:

- [1] MANNING CD, SCHÜTZE H. Foundations of statistical natural language processing[M]. Cambridge, MA: MIT Press, 1999.
- [2] 李效东,顾毓清.基于 DOM 的 Web 信息提取[J].计算机学报,2002,25(5):526-533.
- [3] 封化民,刘飏,刘艳敏,等.含有位置坐标树的 Web 页面分析和内容提取框架[J].清华大学学报,2005,45(S1):1767-1771.
- [4] 周源远,王继成,郑刚,等. Web 页面清洗技术的研究与实现[J].计算机工程,2002,28(9):48-50.