Web 信息抽取技术综述*

陈 钊，张冬梅
（北京林业大学 信息学院，北京 100083）

摘 要： 快速高效地获取网页主题信息的需求使得 Web 信息抽取技术成为信息技术领域的研究热点。现有的 Web 信息抽取技术大致可以归纳为基于统计理论的、基于视觉特征的、基于 DOM 树结构的和基于模板的几类。由于网页文本本身具有树结构并且具有一定的相似性，基于 DOM 树结构和基于模板的抽取技术发展很快而且已经得到了广泛的应用。分别论述了上述几类技术在近几年来的研究进展，从自动化程度、适用范围和复杂性三个角度分析对比了几类技术的优缺点。

关键词： Web 信息抽取；网页噪声；URL 聚类；DSE 算法；RoadRunner 系统；MDR；视觉特征；模板
中图分类号： TP391 **文献标志码：** A **文章编号：** 1001-3695(2010)12-4401-05
doi:10.3969/j.issn.1001-3695.2010.12.001

Survey of Web information extraction technologies

CHEN Zhao, ZHANG Dong-mei
(*School of Information Science & Technology, Beijing Forestry University, Beijing* 100083, *China*)

Abstract: Web information extraction technology has been made the focus of the field of information technology by the needs of obtaining the topic contents of Web pages more efficiently. Existing technologies of this field could be classified into the following four categories, statistics based technology, vision based technology, DOM tree based technology and template based technology. The DOM tree based technology and template based technology had gained a rapid development and a wide employment because of the special structure and similarity owned by Web pages. This paper made a detailed survey and analysis of the above four technologies as well as the comparison of their advantages and disadvantages from points of automation, application filed and complexity.

Key words: Web information extraction; Web page noise; URL clustering; DSE algorithm; RoadRunner system; MDR algorithm; vision feature; template

0 引言

随着信息化进程的推进以及网络技术的发展，越来越多的人开始认识到互联网作为信息来源的重要性，而互联网也已经融入到了人们生活的方方面面。CNNIC(中国互联网络信息中心)在 2010 年 1 月 15 日公布的“第 25 次中国互联网络发展状况统计报告”显示，截至 2009 年 12 月，我国网民规模已达 3.84 亿，互联网普及率进一步提升，达到 28.9%。然而在发展的同时也带来了一些新的问题，如网页噪声几乎占据了主题型网页内容的一半[1]、许多由查询数据库自动生成的网页不能被搜索引擎检索从而成为了所谓的 hidden Web[2]。作为解决这些问题的有力工具之一的 Web 信息抽取技术就是在这样的背景下应运而生的。

根据 Proteus 工程的创建者，纽约大学计算机科学系的 Grishman[3] 对信息抽取的概念描述“信息抽取涉及到为从文本中选择出的信息创建一个结构化的表示形式”，Web 信息抽取可以引申为：从网页文本中抽取出指定的一类信息并将其形成结构化数据的过程。

信息抽取起源于文本理解。从自然文本中获取结构化信息的研究最早开始于 20 世纪 60 年代中期，这被看做是信息抽取技术的初始研究，它以两个长期的、研究性的自然语言处理项目为代表，并一直持续到 80 年代[4]。近几年，信息抽取技术的研究与应用更为活跃。在研究方面，主要侧重于以下几方面：利用机器学习技术增强系统的可移植能力；探索深层理解技术；篇章分析技术；多语言文本处理能力；Web 信息抽取以及对时间信息的处理等。

Web 信息抽取不同于对普通文本的抽取，这是由于网页本身在某种程度上具有一定的结构。大多数网页整体上都遵从 W3C 制定的 DOM 树型结构标准，从而降低了 Web 信息抽取工作的难度。这种结构化的形式在简化抽取的同时也带来了一定的缺点。在网页中，数据通常会被标签所分割，一个完整的句子中往往穿插着对句子本身不起任何作用的标签，从而无法表达句子原始的意义，这就使得传统的基于自然语言处理的文本信息抽取技术无法直接移植到 Web 信息抽取领域。本文将分别论述上述几类技术近几年来的研究进展，并着重探讨基于 DOM 树结构和基于模板的抽取技术，同时对比了几类技术的优缺点。

收稿日期：2010-06-28；修回日期：2010-08-12 基金项目：中央高校基本科研业务费专项资金资助项目(BLYX200928)

作者简介：陈钊(1971-)，男，甘肃天水人，副教授，博士，主要研究方向为信息推送及信息系统；张冬梅(1986-)，女，河北秦皇岛人，硕士研究生，主要研究方向为信息整合及信息推送(dongmei_761@126.com)。

1 Web信息抽取技术

1.1 基于统计理论的技术

基于统计的方法通过统计各个标签所包含的信息量或链接文本与普通文本的比值来获取网页的主题信息。这种方法克服了数据源的限制,并不只针对某一类网页,具有一定的普遍性。

Gupta等人[5]设计的Crunch系统利用区域中link/text(链接文本/普通文本)的比值与某个既定阈值的大小关系来确定网页的正文区域。认为在正文区域中,普通文本所占比例较大,相反,在广告区域或友情链接区域中,信息大部分以链接文本的形式出现。Gupta并没有给出具体的阈值,也没有提出阈值确定的方法,这种处理技术如果阈值确定不合理的话会大大影响最终的抽取准确率。

孙承杰等人[6]也提出了类似的方法,但是仅针对使用〈table〉标签来布局的网页。首先找出页面中所有的〈table〉,去掉这些区域内的所有HTML标签,得到不含标签的字符串,把长度大于某个阈值的字符串作为候选主题信息,对于所有的候选字符串按照长度进行排序,选取最长的字符串作为待抽取部分。由于〈table〉标签具有可嵌套性,最长的〈table〉区域可能会嵌套其他〈table〉标签,在选取最长区域的时候还要检查其中是否包含了其他的〈table〉。实验证明,采用这种方法抽取信息的准确率达到了90%以上。

曹冬林等人[7]改进了Gupta的方法,认为利用link/text的方法不能有效去除无用信息,只适合于对正文区域的定位。在结合了最大文本信息量的基础上,借鉴了香农[8]信息熵公式的理论,提出了如下的有效信息量公式:

$$EI = \log\left(1 + \frac{NW_c}{NW_a}\right)$$

其中:NW_c为网页中非链接文本数,NW_a为网页中的文本综述。该公式反映了有效文本在全体文本中所占的比例。在实际操作中,需要求出所有节点的有效信息量,选取信息量大于既定阈值的部分作为该网页的主题信息。

韩忠明等人[9]提出了基于行的判断方法,通过判断每行的link/text值是否小于某个阈值来确定该行是否是文本行,使用同样的方法对每行文本进行判断。他们在确定阈值的时候引入了生物学中经常被用来分析基因芯片数据的FDR(false discovery rate)算法[10],根据Benjamini的证明,设总共有n个候选输出,每个输出对应的P值从小到大排列分别是$P(1)$,$P(2)$,…,$P(m)$。若想控制FDR不能超过q,则只需找到最大的正整数i,使得$P(i) \leqslant (i \times q)/m$;然后,挑选对应$P(1)$,$P(2)$,…,$P(i)$作为输出结果。这样就能从统计学上保证FDR不超过$q$。将link/text的值当做$P(i)$来处理,从而可以确定阈值。

Song等人[11]不设定阈值,以〈table〉标签作为统计信息的最小容器,利用公式$WT = FC + 0.1 \times NC/HC$统计信息(其中:$FC$为中文句号个数,$HC$为超链接文字个数,$NC$为非超链接文字个数,$WT$为统计信息值),最终得到一棵$WT$值最大、仅有一个分支包含〈table〉节点的DOM树,遍历树得到正文。但是这种方法仅能处理正文内容在一个最小容器中的情况,不适合多正文体网页,而且在新的XHTML 1.0标准中,使用〈table〉标签作为页面布局的方式已经不再提倡,大多数符合设计规范的网页已经用〈div〉标签代替了〈table〉。

周佳颖等人[12]对Song等人的方法进行了改进,可以同时处理用〈table〉和〈div〉布局的网页,并且能够应用于多正文体网页。首先利用类似的统计方法识别出一条信息量最大的区域,再利用这个区域的特征来识别其他正文区域。通过实验证实,这种方法对单正文体和多正文体网页信息抽取的准确率分别达到了94%和91%。

1.2 基于视觉特征的技术

构成网页的HTML是一种语法较为松散且灵活的标记语言,不需要经过编译就可以直接解释执行。这种特性为网页制作人员带来了方便,一些语法错误也可以被浏览器的容错功能隐藏。HTML在很大程度上是用来展示数据而不是描述数据结构的,在HTML DOM树结构中拥有同一父节点的子节点并不一定在浏览器中呈现同样紧密的联系;同样在视觉上相关的两个对象在HTML中的结构有可能差距很远。对于设计并不规范的网页,仅仅从HTML代码的角度去分析其主题信息是不科学的,因而出现了结合页面的视觉特征来抽取信息的方法。

微软亚洲研究院的Cai等人[13]最早提出利用网页的视觉特征来抽取信息的VIPS(vision-based page segmentation)算法。设三元组$\Omega = (O, \Phi, \delta)$,其中,$O$表示页面内的区块,$\Phi$表示分隔符,$\delta$表示两相邻区块的分隔符。那么如图1(a)所示的一个页面可以用图1(b)来表示。算法首先要找出页面中所有类似〈table〉〈p〉〈hr〉〈ul〉等分隔符,把页面分成了各个视觉信息块。在分割区域时充分利用了字体大小、背景颜色、空白区域等视觉特征,并总结出以下几条规则:a)类似〈hr〉等标签通常用于分隔不同的主题,因此如果一个区域内包含〈hr〉标签,那么倾向于分割这个区域;b)如果一个区域的背景色与其内部子区域的背景色不同,则分割这个区域;c)如果一个区域内大部分节点都是文本类型,则不再分割这个区域。这能在一定程度上满足复杂页面对算法的要求,但由于视觉特征的复杂性和网页设计的多样性且存在许多不符合规范的页面,这种基于视觉的信息抽取技术在实施过程中依然存在许多的问题。

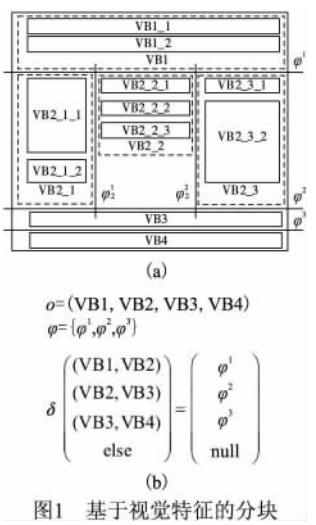

图1 基于视觉特征的分块

1.3 基于DOM树结构的技术

构成网页的HTML的标签具有可嵌套性,不同于普通无结构的文本,一个网页中所有标签组成的DOM模型通常呈现树状结构。在Web信息抽取中可以在网页默认的树结构的基础上通过一些常见的针对树的操作,从而总结归纳出待抽取部

分的特征。基于 DOM 树结构的技术克服了对网页数据源的限制,可以用来处理各种类型的单正文体和多正文体页面,其操作过程相对于基于视觉的方法更加易于实现。在基于 DOM 树结构的抽取技术领域有许多成型的系统和经典算法,使其成为 Web 信息抽取技术中发展极为迅速的一个分支。

1.3.1　DSE 算法

Wang 等人[14]提出了一种通过判断页面中 data-rich(数据密集)区域达到抽取页面主题信息目的的 DSE(data-rich section extraction)算法,并定义页面内包含了主题信息的区域为 data-rich 区域。这种算法主要针对通过查询数据库动态生成的网页,认为同一网站的页面通常具有相似结构的信息组织方式。

作者使用基于页面内 out-going link(出链接)的方式来查找要处理的网页。考虑到页面内存在着许多广告链接,因此并不是所有的链接都是指向其他页面,使用如下所示的 US(URL similarity)公式可以很好地作出判断:

$$US(i,j) = \alpha \times \frac{pre(i,j)}{len(i)} + \beta \times \frac{pre(i,j)}{len(j)}$$

其中: α 和 β 是标准化因子,$0 \leqslant \alpha \leqslant 1$,$0 \leqslant \beta \leqslant 1$ 且 $\alpha + \beta = 1$; $pre(i,j)$ 是指 URL_i 和 URL_j 之间相同的前缀字符数(即出现第一个不相同的字符之前字符串的长度); $len(i)$ 和 $len(j)$ 分别表示 URL_i 和 URL_j 各自的长度。当 $US(i,j)$ 的值大于某个阈值时,说明这两个页面是要找的同源网页。

去掉一些内容无关的标签后,使用迭代算法自顶向下地深度优先遍历两棵网页 DOM 树,同时对比其相应的节点。如果两个节点相同,则算法继续比较它们的子节点,如果两个子节点相同且是叶子节点,则将两个子节点移除,否则回到它们的父节点,按照相同的规则继续比较其他的孩子节点。如图 2 所示,如果一个节点的所有子节点都被比较过了且均被移除,那么这个节点也要被移除。

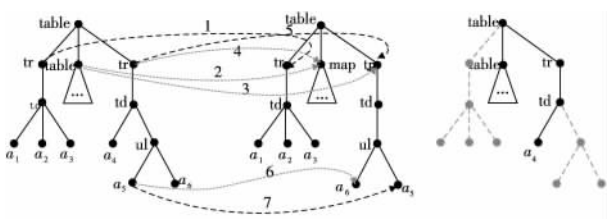

图2　DSE算法对比树的过程

DSE 算法对网页主题区域判断的正确率很高,但仍然存在以下几方面的问题: a) 针对每个网页都必须进行重复的处理,作者并没有很好地利用已经处理过的页面结构,如果可以总结学习得出一个通用的网页信息抽取模板,必将会提高抽取效率; b) 作者认为对于同源网页来说,只有与页面结构有关的标签才具有相似性,但事实上如果两个页面只有内容(即 text 节点)不同的话,那么按照算法的描述,页面中所有的节点都将被去除,这样的结果显然不是所想得到的。

1.3.2　RoadRunner 系统

RoadRunner[15]系统中的信息抽取算法通过比较两棵同源网页的 DOM 树之间的匹配与不匹配部分来得到一个网页主题信息抽取程序。与 DSE 算法类似,该算法也是针对结构相似的网页,并且对数据密集型网页的抽取效果较好。RoadRunner 系统中的算法建立在两个假定已经完成的工作的基础上。为了避免由于网页的不规范而造成的错误,假定所要处理的网页都是遵守 XHTML 规则的;假设网页内的全部内容已经被处理成标签和字符串的形式。

整个算法的中心思想就是处理两棵 DOM 树之间的 mismatch(不匹配)。将 mismatch 的情况分为字符串不匹配和标签不匹配两种,标签不匹配又可以处理为 repeat(重复项)和 optional(可选项)两个子类。对于同源网页来说,字符串的不匹配在很大程度上是由于读取的数据库信息的不同,也就是说字符串发生不匹配的区域正是要抽取的页面主题部分,将这部分节点用字符串"#PCDATA"代替,用来标志待抽取区域。对于标签的不匹配,其处理方法分为两种。如图 3 所示,sample 页面的第 20 ~ 25 行发生了标签的重复,首先需要找出发生 mismatch 的位置的最后一个节点,在本例中为〈/LI〉,这个节点被称为 terminal tag(终点标签),从第 20 ~ 25 行中找出该 terminal tag,从而确定了一个可能的重复区域,即从第 20 ~ 25 行的部分。为了确定这个部分确实是重复,需要从后向前进行比较,首先比较第 25 行和第 19 行,然后向上移动比较第 24 行和第 18 行,依此类推,直到整个区域都找到与之匹配的部分,从而证明这个区域确实是重复,并用"+"标志出来; 对于图中 wrapper 和 sample 页面,在第 6 行发生了 mismatch,首先检查这个 mismatch 是否是重复,显然本例中不是重复,那么就认为这个 mismatch 是由可选项引起的,只需要跳过这个位置,继续比较后面的节点,并在这个位置用"?"来标志以供后续处理使用。

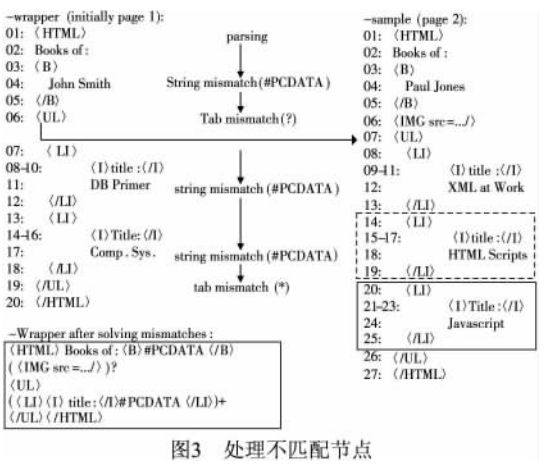

图3　处理不匹配节点

RoadRunner 系统的算法并没有像 DSE 算法一样直接认为标签不匹配的部分是主题信息,而是将其细化为两种情况。但是算法中并没有指出如何寻找同源网页,执行过程是建立在假设网页设计符合 XHTML 标准的情况下。事实上网页中或多或少都会存在一些不规范的部分,如果直接应用这个算法的话可能会出现问题甚至影响信息抽取的准确率。作者并没有提出对网页进行预先处理去掉一些噪声信息,这样直接对网页树进行对比会降低算法的执行效率。

1.3.3　MDR 算法

Liu 等人[16]在 2003 年提出了 MDR(mining data records)算法,并在 2005 年对该算法进行了改进[17]。该算法基于以下两条通过观察得到的规则: a) 结构相似的数据块往往固定在页面的某个区域中,并且以相似的结构进行组织,这样的一个数据块被称为 data region(数据区域); b) 一条完整的数据描述极少会在一棵子树中开始却在另外一棵子树中收尾。MDR 算法的关键在于判断 data region 的位置,首先定义了一个 generalized node 的概念,指的是具有相同的父节点并且相邻的一系列的节点。一个 data region 就是两个或两个以上具有如下特

征的 generalized node 的组合: 所有 generalized node 的长度相同、具有相同的父节点且都是相邻的,并且它们之间的树编辑距离[18]小于某个固定的阈值。

MDR 算法提出了一种独特的判断 data region 的方法,不但适合单正文体网页,对于多正文体网页的处理也达到了很好的效果。但是在基于视觉特征的技术中提到过,HTML 相对于数据结构的组织来说,更加注重数据的表现形式,甚至有时候为了达到视觉效果而将毫无关系的内容添加到相邻的标签中,而一整条完整的数据又可能被分配到不相邻的节点当中。如图 4 所示,〈table〉的第 1 行是两个商品名,第 2 行分别是两个商品的描述,如果按照 MDR 来判断 data region 的话会将一条完整的数据分开,这显然是不合理的。

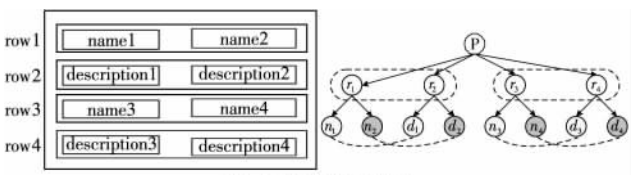

图4 同行数据不相关

1.3.4 几种改进的基于 DOM 的处理技术

刘亚东等人[19]提出了一种基于 MDR 的改进算法,引入了 URL 聚类的概念,将待抽取的页面按照 URL 进行分类,针对每类页面生成一种抽取方法。在处理页面之前,首先对页面进行了预处理,直接去掉了〈script〉〈b〉〈font〉〈img〉等标签及其内部包含的内容,并有选择地去掉普通文本数/链接文本数小于某个固定阈值的分块。该算法不但可以抽取信息,还可以根据现有的关键词集和关键词权重对抽取到的信息根据领域进行分类。这种算法直接去掉了某些标签及其内部的内容,这样有可能会去掉关键内容,使得待抽取信息变得不完整;而且,作者并没有给出阈值 T 的确定方法,阈值的合适与否会对算法的效果产生很大的影响。

顾韵华等人[20]对 DOM 树节点进行了扩展,将节点分为标题类、正文类、视觉类、分块类、链接类和其他类,并定义了每类节点的影响因子。节点的影响因子是指节点对内容影响的相对程度,取值为 [0,1],值越大表示影响力越大。同时节点的影响因子具有传递性,父节点的影响因子可以传递给子节点,一个叶子节点的影响因子是该叶节点所有附件点影响因子之和。

李朝等人[21]根据关键词在 DOM 树中的位置来确定待抽取信息,将文本信息也加入到 DOM 树中,文本通常以叶子节点的形式出现。给定关键词集$\{k_1,k_2,\cdots,k_n\}$,首先找到每个关键词对应的文本节点,记录该节点相对于 DOM 根节点的路径,从而形成 n 个路径集$\{p_1,p_2,\cdots,p_i\}$,寻找 n 个路径集中最长的公共路径,这个路径对应的子树就是最终要抽取的主题信息。

1.4 基于模板的技术

互联网上存在着大量的通过读取数据库数据然后填充到统一模板的方式自动生成的网页,针对这类具有模板的网页产生了一种基于模板的抽取技术。通过对产生于同一模板的网页的对比分析总结出一个通用的抽取模板,从而免去了对众多网页进行重复处理的繁琐。

欧建文等人[22]针对由模板生成的网页提出了一种生成网页抽取模板的方法,并且已经应用到某搜索引擎当中。同一网站内的网页地址(URL)往往被组织成一个很好的树状层次结构,尤其是在动态网页技术出现之后,根据不同的参数动态地从数据库中取出数据生成的动态网页,其层次结构更加明显。根据上述特性,作者提出以下三个假设: URL 树中在同一个目录节点下存在大量由同一模板生成的网页; 由模板生成的网页布局基本一致; 网页中链接文本是对目标网页主题内容的描述。作者将抽取模板定义为 $T=\{$urlfix, tags prefix, tags suffix, label point$\}$。其中: urlfix 指的是在 URL 树中,从根节点到目录节点的路径,以"/"分隔; tags prefix 是指从根节点到待抽取信息节点(不含待抽取信息节点)的标签序列; tags suffix 是指从待抽取节点(不含待抽取节点)到树的末尾的所有标签序列。它们是确定网页模板的关键特征,即如果两个网页具有相同的 tags prefix 和 tags suffix,就可以认为这两个网页是由同一个模板生成的。Label point 是网页待抽取信息块在网页 DOM 树中的路径。在模板生成阶段,通过对一定数量的网页的对比操作总结出模板 T,将 T 存放到这些页面的目录节点中,并认为这个目录节点下的所有网页都可以按照这个模板来处理。这种观点只适合于严格按照模板生成的网页,如果网页由于主题信息的不同(过长、过短或包含图片)使得 tags prefix 或 tags suffix 有所区别的话,就会被认为不是来自同一模板。根据作者对模板 T 的定义,显然这种方法只针对一个待抽取点,不能处理多正文体的网页。

张彦超等人[23]的方法不但生成了页面模板,还引入了 URL 模板匹配的概念。URL 模板匹配是根据待抽取页面的 URL 与 URL 模板匹配库进行模板匹配、识别该页面是否可解析及确定该页面所用的解析模板。

郑长松、陈治昂和杨少华等人[24~26]也提出了类似的能够产生信息抽取模板的方法,只是对模板的定义形式略有不同。时达明等人[27]提出了一种针对博客页面的抽取方法,认为大多数信息都嵌套在 class 属性等于某个值的〈div〉标签内。这种方法具有局限性,仅能针对某个具体网站进行抽取。

2 Web 信息抽取技术对比

1) 自动化程度

基于统计原理的技术和基于视觉特征的技术在多数情况下都涉及到对待抽取页面本身进行区域划分等处理,需要对页面进行人工干预,因此操作人员的主观行为可能会造成区域划分不合理从而直接影响信息抽取的效果。基于模板的技术需要依赖于表示待抽取位置的节点串,通常需要针对某一类待抽取页面进行分析和标记,总结出一个统一的模板节点串。尽管利用模板来抽取信息较为便捷,但生成模板的过程却需要大量的人工操作。基于 DOM 树结构的技术针对网页本身的结构优势,通过对网页树进行对比操作就可以确定页面内主题信息的位置进而实现信息的抽取,极少受到操作者主观因素的影响。

2) 适用范围

基于统计原理的技术适用于以文字为主题并且文字部分相对于其他部分来讲具有明显数量优势的一类页面,针对不同的页面要应用不同的阈值。基于视觉特征的技术过多地依赖页面的组织结构,因此比较适用于结构清晰、符合一般设计规则并且没有过多标签错误的页面。基于 DOM 树的技术对网页的类型没有限制,对于出自同一个网站并且具有相似结构的

页面都能进行处理。基于模板的技术适用于相似度较大的页面,如通过动态查询数据库生成的页面,并且只能针对单正文体网页。

3) 复杂性

基于统计原理的技术在理论上易于实现,但其难点在于确定一个合理的阈值。阈值的确定方法会对主题区域的确定产生直接的影响,并且对于不同种类的页面必须分别讨论阈值。基于视觉特征的技术对网页的分块更加注重可视化信息的组织形式,比单纯考虑网页标签嵌套结构的方法更合理,但网页本身的一些标签错误、结构不规范以及数据分块与视觉效果分块的不统一等多种因素使得这种技术的实现过程非常繁琐。基于 DOM 树结构的技术不需要再对待抽取页面进行分块处理,可以直接通过对比得出页面的主题信息区域,但却需要对每个页面都进行同样的处理,没有充分利用已有的结果总结出针对同类相似页面进行处理的统一方法。基于模板的技术免去了对同类网页的重复操作,针对相似页面总结出统一的抽取模板,但在模板的生成方法和模板通用性方面还有待于改善。

3　结束语

Web 信息抽取是网络信息挖掘和信息检索的一个非常重要的前处理步骤,在实际的工程项目和信息获取中也存在着明确的需求。本文详细探讨了基于统计理论的、基于视觉特征的、基于 DOM 树结构的和基于模板的几类常用的 Web 信息抽取技术及其发展现状。几类技术各有其优缺点,在实际应用中,只有将几类技术结合起来,取长补短,才能更准确地在页面中抽取所需要的内容。

参考文献:

[1] 张志刚,陈静,李晓明. 一种 HTML 网页净化方法 [J]. 情报学报,2004,23(4):387-393.

[2] BRIN S, PAGE S. The anatomy of a large-scale hyper-textual Web search engine [J]. Computer Networks and ISDN Systems, 1998,30(1-7):107-111.

[3] GRISHMAN R. Information extraction: techniques and challenges [EB/OL]. (1997) [2010-06-01]. http://cs.nyu.edu/cs/faculty/grishman/proteus.html.

[4] GAIZAUSKAS R, WILKS Y. Information extraction: beyond document retrieval [J]. Journal of Documentation, 1997,54(1):70-105.

[5] GUPTA S, KAISER G. DOM-based content extraction of HTML documents [C]//Proc of the 12th World Wide Web Conference. New York: ACM Presss, 2003:207-214.

[6] 孙承杰,关毅. 基于统计的网页正文信息抽取方法的研究 [J]. 中文信息学报,2004,18(5):17-22.

[7] 曹冬林,廖祥文,许洪波,等. 基于网页格式信息量的博客文章和评论抽取模型 [J]. 软件学报,2009,20(5):1282-1291.

[8] SHANNON C E. A mathematical theory of communication [J]. ACM SIGMOBILE Mobile Computing and Communications Review, 2001,5(1):3-55.

[9] 韩忠明,李文正,莫倩. 有效 HTML 文本信息抽取方法的研究 [J]. 计算机应用研究,2008,25(12):3568-3574.

[10] PAWITAN Y, MICHEALES S. False discovery rate, sensitivity and sample size for mieroarray studies [J]. Bioinformatics, 2005, 21(13):3017-3024.

[11] SONG Ming-qiu, WU Xin-tao. Content extraction from Web pages based on Chinese punctuation number [C]//Proc of International Conference on Wireless Communications, Networking and Mobile Computing. 2007:5573-5575.

[12] 周佳颖,朱真民. 基于统计与正文特征的中文网页正文抽取研究 [J]. 中文信息学报,2009,23(5):80-85.

[13] CAI Deng, YU Shi-peng, WEN Ji-rong, *et al*. VIPS: a vision based page segmentation algorithm [R/OL]. (2003-11). http://research.microsoft.com/apps/pubs/defanlt.aspx? id=70027.pdf.

[14] WANG Ji-ying, LOCHOVSKY F H. Data-rich section extraction from HTML pages [C]//Proc of the 3rd International Conference on Web Informations Systems Engineering. Washington DC: IEEE Computer Society, 2002:2313-2322.

[15] CRESCENZI V, MECCA G. RoadRunner: towards automatic data extraction from large Web sites [C]//Proc of the 27th VLDB Conference. San Francisco: Morgan Kaufmann Publishers, 2001:109-118.

[16] LIU Bing, GROSSMAN R, ZHAI Yan-hong. Mining data records in Web pages [C]//Proc of the 9th International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2003:601-606.

[17] ZHAI Yan-hong, LIU Bing. Web data extraction based on partial tree alignment [C]//Proc of the 14th International Conference on World Wide Web. New York: ACM Press, 2005:76-85.

[18] ZHANG K, SHASHA D. Tree pattern matching [M]//Pattern Matching Algorithms. New York: Oxford University Press,1997.

[19] 刘亚东,彭舰,张达平. 基于智能的网页信息提取系统的研究与设计 [J]. 四川大学学报: 自然科学版,2009,46(4):957-962.

[20] 顾韵华,田伟. 基于 DOM 模型扩展的 Web 信息提取 [J]. 计算机科学,2009,36(11):235-238,289.

[21] 李朝,彭宏,叶苏南,等. 基于 DOM 树的可适应性 Web 信息抽取 [J]. 计算机科学,2009,36(7):202-210.

[22] 欧建文,董首斌,蔡斌. 模板化网页主题信息提取方法 [J]. 清华大学学报: 自然科学版,2005,45(9):1743-1747.

[23] 张彦超,刘方,李勇,等. 基于自动生成模板的 Web 信息提取技术 [J]. 北京交通大学学报: 自然科学版,2009,33(5):40-45.

[24] 郑长松,傅彦,余莉. 基于模板的 Web 信息自动抽取方法 [J]. 计算机应用研究,2009,26(2):570-582.

[25] 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]. 计算机应用研究,2009,26(7):2646-2649.

[26] 杨少华,林海略,韩燕波. 针对模板生成网页的一种数据自动抽取方法 [J]. 软件学报,2008,19(2):209-223.

[27] 时达明,林鸿飞,赵晶. 基于模板化的 blog 信息抽取 [J]. 计算机工程与应用,2008,44(9):156-162.

[28] EIKVIL L. Information extraction from World Wide Web [EB/OL]. http://citeseerx.ist.psu.edu/viewdoc/download? doi=10.1.1.41.4905&rep=rep1&type=pdf.

[29] 赵欣欣,常江光,刘玉树. 基于标记窗的网页正文信息提取方法 [J]. 计算机应用研究,2007,24(3):144-146.

[30] GUPTA S, KAISER G E, GRIMM P, *et al*. Automating content extraction of HTML documents [J]. World Wide Web,2005,8(2):179-224.

[31] SHIH L K, KARGER D R. Using, and table layout for Web classification tasks [C]//Proc of the 13th International Conference on World Wide Web. New York: ACM Press, 2004:193-202.

[32] 杨桢,赵艳萍,朱东华. 基于正则表达式的信息抽取系统在国防技术监测中的应用 [J]. 北京理工大学学报,2006,26(6):74-78.