基于结构一致和特征学习的网页信息标签提取

杜博远[1],王美清[1],陈长福[2],陈 飞[1]
DU Boyuan[1], WANG Meiqing[1], CHEN Changfu[2], CHEN Fei[1]

1.福州大学 数学与计算机科学学院,福州 350000
2.福建库易信息科技有限责任公司,福州 350000
1.College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350000, China
2.Fujian Ecallcen Information Technology Co., Ltd., Fuzhou 350000, China

DU Boyuan, WANG Meiqing, CHEN Changfu, et al. Tags extraction for Web information based on structure consistency and feature learning. Computer Engineering and Applications, 2017, 53(7):74-78.

Abstract: The Web information refers to the special contents of the Web pages which usually includes main body, title, release date and release media. Each content is put in the corresponding HTML tags. Extracting automatically such tags is able to obtain Web information under the same Web template. Such tags extraction for Web information is a great help for clawing contents from a large number of Web pages. Since Web structure consistency for the same template and the statistical features of Web information, this paper proposes tags extraction automatically for Web information based on structure consistency and feature learning. The algorithm consists of three steps: Web contrast, content identification and tags extraction. Experimental results on 51 Web templates from 1 620 Web pages show that the proposed algorithm achieves Web information extraction not only high-speed but also high-accuracy.
Key words: Website tags; information extraction; feature learning; structure consistency

摘 要: 网页信息指网页的正文、标题、发布时间、媒体等,每个信息都存在于HTML文档特定的标签中,自动获取这些标签可以实现在相同模板下的网页信息自动提取,对于大规模抓取网页内容有很大帮助。由于在相同模板下不同网页之间结构一致,网页信息有一定统计特征,提出了一种基于结构对比和特征学习的网页信息标签自动提取算法。该算法包含三个步骤:网页对比、内容识别和标签提取。在51个模块下对1 620个网页进行测试,实验结果表明,通过提取标签获取网页信息不仅速度快,而且抓取的内容更加准确。
关键词: 网页标签;信息提取;特征学习;结构一致
文献标志码: A **中图分类号:** TP391 doi:10.3778/j.issn.1002-8331.1509-0226

1 引言

随着网络的发展,网页的个数以惊人的速度增长,网页中包含用户所需的有用信息(网页的正文、标题、发布时间、媒体等)和无关的信息(如导航栏、广告等)。很多情况下无关信息会影响用户体验,因此如何提取网页中用户所需的信息成为研究的热点问题。由于网页格式多种多样,所以导致判断网页信息的位置很困难,同时多种类型的网页信息同时提取也很困难。传统的方法通常使用一定的手段对单个网页进行分块处理或者统计分析或者模板化处理,获取这个网页的正文信息。若需要对大量不同类型,不同格式的网页进行处理,传统方法信息提取由于方法复杂且针对一类网页的一种信息类型,导致速度慢,准确率低,提取信息类型单一。提出一种基于结构一致和特征学习自动获取信息标签的方法,用提取到的网页信息标签获取网页信息代替直接获取网页信息。获取标签后可以根据标签匹配准确的获得网页信息,所以相对单个网页信息提取速度快,而且在提取正文同时还可以提取时间、媒体、标题等。

基金项目: 国家自然科学基金(No.61401098);福州大学科研启动基金(No.022575);福州大学科技发展基金(No.2014-XY-21)。
作者简介: 杜博远(1991—),男,硕士研究生,研究领域为数据挖掘、机器学习;王美清,女,博士,教授,主要研究方向:数值算法、图像处理、数据挖掘;陈飞,男,博士,副教授,主要研究方向:机器学习、图像处理、智能计算,E-mail:chenfei314@fzu.edu.cn。
收稿日期: 2015-09-22 **修回日期:** 2015-11-30 **文章编号:** 1002-8331(2017)07-0074-05
CNKI **网络优先出版:** 2015-12-11, http://www.cnki.net/kcms/detail/11.2127.TP.20151211.1518.020.html

万方数据

因此这种方法适合于舆情分析、搜索引擎等的前期网络信息大规模获取工作。

2　相关工作

对于网页的信息提取，已经有了大量的研究工作。文献[1]提出里一种计算网页节点权值的方法来获取正文的方法。文献[2]微软亚洲研究院提出基于视觉的网页分块算法VIPS，对不同类型的网站的正文信息提取都有很好的效果。文献[3]提出基于统计估计的朴素文本行选择算法，对于网页有关文本信息的提取有很好的效果。文献[4]提出基于视觉分块后在考虑块的取舍的方法，在微软VIPS方法上进一步提高算法的效率。文献[5]提出了基于模版的方法，先经过建立dom树，再在多个网页中寻找网页的模版，对于新发现的网页，根据模版的相似度，判断网页所属的模版类，最后再提取网页信息。现有这些方法可以分为三种：基于统计的方法、基于模版的方法和基于分块的方法。

基于统计的方法[6-9]是将网页源码分解成dom树，根据对节点内容的统计特征的观察，如文字数量、汉字的个数、标点率、是否有句号，构造判断节点内容是否为正文的判断函数。再通过判断函数判断每个节点的内容是否属于正文，从而获取正文。这种方法简单易行，对网页的样式没有要求，但是正确率不能保证。在判断正文时，会把标题，版权信息等不属于正文的信息当作正文获取。并且，这种方法设计的判断函数基本是基于正文的设计，没有对网页的标题、时间等信息进行判断提取。

基于模版的方法[10-12]是先把网页分解成dom树，再把多个相同或者相似的网页的dom树进行比对，提取出相同的部分，称之为模版。多种模版的集合成为模版类。对于新进的网页通过计算相似，把新网页所属的模版从模版类中提取出来，根据模版提取网页信息。由于这种方法仅仅使用HTML文档的标签信息，也就是只考虑了页面的布局信息，忽略了内容信息。一方面难以准确的确定网页信息的位置，另一方面也无法对多种信息同时分类提取。

基于视觉信息的方法[13-15]是将网页分成几个块，再根据算法进行取舍，获取正文所在的块。这种方法对于网页结构特殊的网站效果不好，块的取舍决定的因素较为模糊，不同网站需要人工调整（即就是没有良好的判断阈值）。同样的，这种方法只能获取网页的正文信息，对时间、标题、媒体无法提取。

针对传统方法的问题，提出一种基于结构一致和特征学习的网页信息标签提取方法。综合利用HTML文档中的标签信息和内容信息，弥补了基于统计的方法仅使用内容信息和基于模版的方法仅使用页面信息的缺点。同时，使得信息提取的位置准确，不需要基于视觉信息的方法需要人工干预的问题。另外，此种方法可以做到信息提取和信息分类同时进行，也就是可以同时提取多种信息，解决了传统方法提取信息单一的问题。

3　信息标签自动提取

由于HTML的书写格式很灵活，所以网页的格式多种多样，直接获取网页信息是比较困难的。目前绝大部分网站的网页是动态生成，所以网页设计具有一致性，即在相同网站相同板块下的网页是有相同的模版。这里相同的模版是指在同一个栏目下，它们各个类型的信息都放置在固定的标签中。例如，新浪福建新闻下的全部网页具有相同的模板，它们的标题都存储在标签$<h1>/</h1>$中。只要获得了这些标签和标签对应的类型，那么在该网页栏目中的全部网页的信息就可以根据标签信息获得。显然提取网页信息的问题就简化为提取其对应的标签。

因此提出一种可以针对大量网站的信息标签提取方法。首先通过建立dom树将网页源码分解成节点信息和节点信息对应的标签路径，然后对比同一模板下m篇网页，把m篇中网页信息互相重复的去除，接着判断剩余网页信息内容的类型，最后进行标签提取。为了便于算法介绍和分析，引入如下几个定义。

定义1　类型：正文、标题、时间、媒体、点击率、来源地等的集合。这里主要讨论网页信息中的正文、标题、时间、媒体，其他网页信息的类型可以通过相似的方法推广。

定义2　模板结构：类型和标签的对应关系集合，两个网页如果全部类型和标签的对应关系都相同，则称这两个网页有相同的模版结构。

定义3　节点路径：网页源码经过构建dom树后，同层节点用1，2，3，…表示，节点父子关系用“-”表示。根据深度遍历后，各个节点可以用类似X_1-X_2-X_3-…表示，称为节点路径。

定义4　节点内容：构建dom树后节点中的文本信息。

3.1　网页对比

因为相同网站相同板块的模板结构是一样的，所以在同一个板块下的多个网页有很多相同的节点，而且这些节点大多是系统自动生成的。显然，有效地识别这些相同的节点，可以获得大致的模板结构。但是，由于网页噪声的影响，相同节点的识别并不那么容易。通过网页对比可以有效识别相同节点，从而减少后面识别和标签提取的节点个数，提高系统的速度和准确率。

3.1.1　网页预处理和建立dom树

网页预处理和建立dom树可以使得对比更加容易准确。网页预处理就是对下载好的HTML文档直接处理，把<script>、<head>等不需要的节点，含有超级连接的节点，注释等删除。预处理后，由于网页中每个元素

万方数据

或者域都在一对标签中,所以根据一个开始标签(如<$h1$>)和一个结束标签(如</$h1$>),以及其相对应的属性和属性值,可以把HTML的文档分解成树形结构,称为dom树。

3.1.2 节点对比

为了分析一个板块的模版结构,可以选择该板块下n篇网页对比获得相同节点。由于网页信息主要存在于dom树中的叶子节点,那么网页对比就体现在节点间的对比。

假设集合A_i为网页i的全部叶子节点的节点内容,$A_i=\{a_{i1},a_{i2},\cdots,a_{in}\}$,其中,$a_{ij}$为第$i$个网页下第$j$个节点的内容。显然,这些节点包含需要提取的网页信息和无效信息。为了便于分析,假设W_i为网页i需要提取的网页信息的节点内容集合(由于网页信息互不相同,所以W_i间互不相同,即若$i\neq j$,$W_i\cap W_j=\varnothing$),G_i为包含无效信息的集合;那么,$A_i=W_i\cap G_i$。所需要提取的网页信息都存在于W_i,所以应尽可能的减少G_i中的元素,使后面的识别量减少,同时提高准确率。设G为该板块下全体无效信息的集合,若已知集合G那么,$W_i=A_i-A_i\cup G$,但是全集G无法直接获得,所以获取$n-1$篇该模块下的网页A_2,A_3,…,A_n。用A_1与A_2,A_3,…做对比,剔除相同的元素。设A_1裁剪后剩余集合为S,那么:

$$S=A_1-A_1\cap A_2-A_1\cap A_3-\cdots A_1\cap A_n=$$
$$A_1-A_1\cap(A_2\cup A_3\cup\cdots\cup A_n)=$$
$$A_1-A_1\cap(G_2\cup G_3\cup\cdots\cup G_n)$$

可以看出当n越大(即就是用于对比的网页越多),$G_2\cup G_3\cup\cdots\cup G_n$就越趋向于$G$。显然,$n$越大越精确,但是会增加网页对比的计算量。实际工作中,选取m个网页,对这些网页都做3.1.1小节的处理,获取每个网页的dom树。再用网页1叶子节点节点内容集合A_1与其余的网页节点内容集合$A_2,A_3,\cdots,A_n$做交集。

下面以对比网页1和网页2为例介绍对比过程。

网页1和网页2建立dom树后如图1、图2所示。

(1)可知网页1的叶子节点为集合为$\{N_4,N_5,N_6,N_7,N_8,N_9\}$,其中$N_4$=(1-1-1-1,标题1),$N_5$=(1-1-1-2,正文1),同理可得网页1的其他节点和网页2的叶子节点表示。

(2)获取网页1节点内容集合A_1={标题1,正文1,媒体1,广告1,广告2,广告3},网页2节点内容集合A_2={标题2,正文2,媒体2,广告2,广告3}。删除A_1、A_2中相同的元素“广告2”、“广告3”,得到S={标题1,正文1,媒体1,广告1}。

对网页1和其余网页做相同的处理,这样就完成了网页对比。

3.2 节点内容识别

经过3.1节的网页对比后,剩余叶子节点的节点内容包含需要提取的网页信息和少量残留的无用信息。若已知剩余节点的的节点内容属于何种类型,那么就可以根据节点内容和节点路径的一一对应关系以及节点路径和标签的一一对应关系获得各个类型的标签了。利用节点内容的统计特征,采用机器学习的方法识别节点的类型。

3.2.1 特征的选取

在分类过程中,特征是将节点内容分类的的可分性判据,所以好的特征能使得分类效果更好。特征的选取原则是:选择的特征可以明显表现出一类和其他类的差异(例如:正文的字符数很多,其他类较少,所以字符数

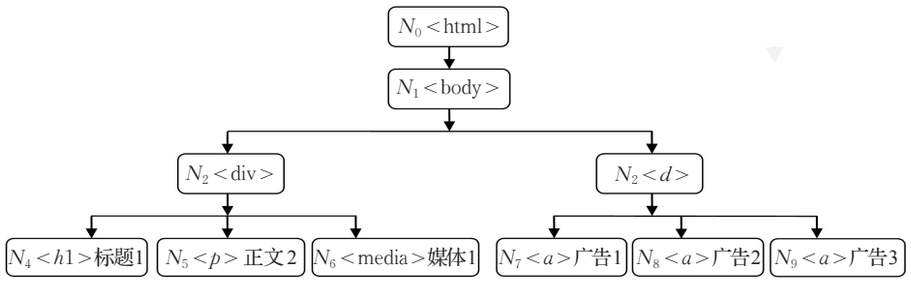

图1 网页1对应的dom树

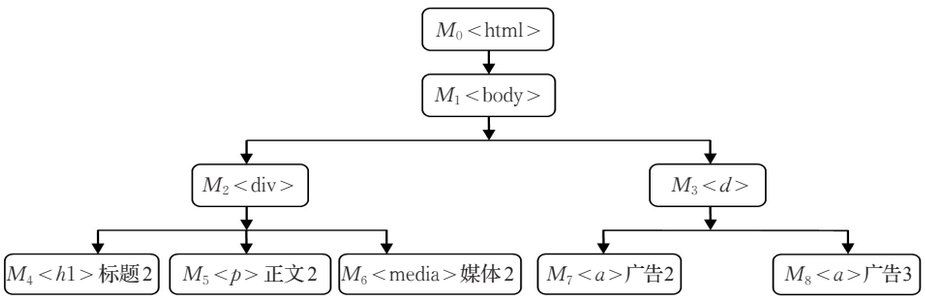

图2 网页2对应的dom树

万方数据

就可以把正文和其他类型分别开)。依据特征的选取原则，对大量已知类型的样本可选择的特征进行统计分析，观察特征对于不同类是否有明显的差异，若有则保留，否则舍弃。这样既可以保证无用的特征不会干扰分类结果，同时又尽可能多地保留了可用的特征，提高分类的准确率。最终选取的特征如表1第一列所示。

表1　特征转化方法

类型	特征描述	选取方法一	选取方法二
字符数	统计字符个数	简单统计	分段
是否有句号	内容是否有句号	有句号为1否为0	同方法一
数字个数	统计数字个数	简单统计	分段
汉字个数	统计汉字个数	简单统计	分段
汉字连续性	5个汉字连续	是为1否则为0	同方法一
数字率	数字字符百分比	数字个数/字符数	同方法一
汉字率	汉字字符百分比	汉字个数/字符数	同方法一
标点	统计标点个数	简单统计	分段
标点率	标点百分比	标点个数/字符数	同方法一

特征选取后还需把特征转化成数字形式。首先把各个类型用数字表示，也就是建立类型和数字的一一对应关系。采取标题对应1、时间对应2、媒体对应3、正文对应4。节点内容的特征转化成数字的方法是：对原始特征进行直接统计和变换统计相结合。具体来说使用了两种方法：方法一是对统计性数据(如字符个数)采用整数连续型进行统计，是非型(如是否有句号)数据使用布尔型进行统计，变化型(如数字率)采用0-1连续型统计；方法二是对统计性数据使用分段布尔统计(如把字符个数分成0~10，10~50，50~200的区间，统计结果属于哪个区间，哪个区间对应的维数为1)，是非型和变化型采用与方法一相同的方法，具体如表1。

3.2.2　基于KNN的节点识别

在进行叶子节点内容分类时，由于正文多，其他类型少，而 K 近邻法(KNN)能较好地避免样本不平衡的问题。同时，不同类型的节点内容可能存在交叉重叠(例如正文信息和标题信息可能相似导致交叉重叠)，而KNN方法主要靠周围有限的邻近的样本，而不是靠判别类域的方法来确定所属类别的，因此对于节点内容的样本集来说，KNN方法较其他方法更为适合。

识别的流程：先把学习样本与类型进行对应(在实验中对已获取的正文、标题、时间、媒体各10 000条数据作为学习数据)，类型对应关系如3.2.1小节中所述。再对样本根据3.2.1小节的方法进行特征选取和统计。例如："新华社"根据方法3.2.1小节中选取方法一结果为：3 0 0 3 0 0 1 0 0，选取方法二结果为：100 0 000 100 0 0 1 000 0(其中字符个数分成0~10，10~50，50~200的区间，数字个数分成0~3，3~10，>10的区间、汉字个数分成0~10，10~50，50~200的区间、标点个数分成0~3，3~10，>10的区间。不同类型间为了清晰已用空格区分，实际操作中不需要)。然后把数字化后的学习数据放入KNN学习器进行学习，获取分类标准。最后，根据分类标准判断剩余叶子节点的节点内容的类型。

3.3　标签提取

经过上述步骤后，已经获得了节点内容属于的类型，如表2所示。

对于中文网页来说，类型间存在一定的位置关系：标题、时间、媒体类型都应在正文的上方。因此首先提取正文内容，正文存在于多个节点中，所以将正文部分的内容应进行合并，方法是获取两个判断为正文的节点

表2　节点识别后结果

节点id	节点路径	节点内容	节点内容类型
1	1-1-6-1-2-1-1-	永暑岛造陆进度曝光 完工图似巨型航母	标题
2	1-1-6-1-2-3-1-1-1-	2014/11/19 7:22:22	时间
3	1-1-6-1-2-3-1-2-1-2-1-1-	重庆华龙网	媒体
4	1-1-6-1-2-5-1-1-1-1-1-	第1页：永暑岛造陆进度曝光	正文
5	1-1-6-1-2-5-1-2-1-1-2-1-1-	随着2013年底开始的填海作业不断加速，我国…	正文
6	1-1-6-1-2-5-1-3-2-2-1-	—	时间
7	1-1-6-1-2-5-1-5-2-	11月17日，一组最新的永暑礁扩建进度图曝…	正文
8	1-1-6-1-2-5-1-6-1-	延伸阅读：	正文
9	1-1-6-1-2-5-1-7-1-	‘’	时间
10	1-1-6-1-2-5-1-8-1-	‘’	时间
11	1-1-6-1-2-5-1-10-1-1-	1	时间
12	1-1-6-1-2-5-2-1-1-	2	时间
13	1-1-6-1-2-5-2-4-1-1-	重庆华龙网	媒体
14	1-1-6-1-2-5-2-8-1-	9398	时间
15	1-1-6-1-2-15-2-1-1-2-1-1-	11/19	时间
16	1-1-6-1-2-15-2-1-1-3-1-1-	11/19	时间
17	1-1-6-1-2-15-2-1-1-4-1-1-	11/18	时间
18	1-1-6-1-2-15-2-1-1-5-1-1-	11/18	时间
19	1-1-6-1-2-15-2-1-1-6-1-1-	11/17	时间
20	1-1-6-1-2-15-2-1-1-7-1-1-	11/18	时间

万方数据

(节点4、节点5)的公共节点路径(1-1-6-1-2-5-1)并将拥有此路径的节点合并为正文。然后根据类型的位置关系获得唯一的标题路径(1-1-6-1-2-1-1)、时间路径(1-1-6-1-2-3-1-1-1)、媒体路径(1-1-6-1-2-3-1-2-1-2-1-1)。最后,再根据建立dom树时路径和标签的对应关系,就获得了各个类型的标签。

4　实验与分析

4.1　节点识别与比较

内容识别作为标签提取过程中最重要的部分,识别的准确性直接决定了标签提取的准确性,所以选择良好的数字化方法和类型判断器就很重要。因此做了KNN与目前主流分类算法逻辑回归、集成学习器、判别分类器和随机森林的对比实验(见表3)。

表3　内容识别对比实验结果

分类器	数字化方法	错误个数	错误率/%	耗时/s
逻辑回归	1	245	6.12	60.25
判别分类器	1	83	2.08	0.53
集成学习器	1	66	1.65	10.34
随机森林	1	52	1.30	25.44
K近邻(k=10)	1	52	1.30	1.41
K近邻(k=10)	2	58	1.45	8.30

在实验中对正文、标题、时间、媒体各获取10 000条数据作为学习数据,同时各获取与学习不同的各类型数据4 000条(每类型1 000条)作为检验数据做为实验数据。

从实验数据可以看出使用的数化方法一、K 近邻分类是较优的方法。

4.2　标签自动提取实验

为了验证所提算法的有效性和可行性,对7个网站中51个模块做了大量实验。对比网页数 m 的选取是根据实际情况动态选择,对于复杂网站 m 应取较大值,对于网站简单的 m 取小值。在获取各个类型的标签信息后,再在该板块下随机抽取一些网页,字符串匹配寻找各个类型的信息。

本实验 $m=3$,结果如表4所示。正文正确是要求把正文的全部内容全部提取,即就是查准率和查全率都为100%。

从实验中可以发现,介绍的方法对正文、时间、标题都有很好的效果,对于媒体效果一般。主要原因是用于提取标签的网页中媒体名称过长,判断时判断为标题,或者用于提取标签的网页中媒体都一样在清洗过程中被清洗掉,导致未寻找到媒体标签。若更换获取标签的网页或进一步后处理。可以解决上述问题。

对比文献[1]、文献[2]在相同网站下的实验结果,可以看出提出的方法对于抓取正文的准确率高于传统方法,并且获取了标题、时间、媒体信息。由于采用先根据网页信息获取各个类型网页信息的标签,再根据标签获取相同板块下的网页信息,所以具有唯一性好,查全率高的特点。

4.3　算法复杂度分析

对于相关工作中所属的三种方法,方法一中需要遍历网页的全部节点,计算每个节点的函数值;方法二需要对新进网页建立dom树,也就是遍历了每个节点;方法三的分块过程和块的取舍过程,都需要对新进网页的节点进行处理。综上,可以看出,无论何种方法,都需要对新进单个网页的HTML文档先进行处理,然后再提取网页信息。而基于标签的算法,只在获取标签时,需要对网页HTML文档进行处理,在提取新进网页内容时,不需要对HTML文档进行任何处理,也就是不需要遍历每个节点,只需要对HTML文档进行简单的检索工作。对与海量网页的信息提取,时间效率明显高。

5　结束语

针对相同模版下的网页,提出了基于结构一致和特征学习自动获取信息标签的算法。首先对下载好的网页源码转换成dom树,再根据网页对比去掉相同的部分,然后根据 K 近邻分类器对剩余叶节点进行类型判断,最后获取该板块各个类型的信息标签。本文算法可以提取网页的多种信息,准确率高。由于获取的标签在相同板块下都适用,所以在获取其余网页信息时只需要字符串匹配,效率高,适合大规模网页信息获取。

表4　实验结果

模版来源	模版个数	测试网页个数	正确正文个数	正文正确率/%	正确时间个数	时间正确率/%	正确标题个数	标题正确率/%	正确媒体个数	媒体正确率/%
新浪	10	300	300	100.0	300	100.0	300	100.0	300	100.0
搜狐	10	300	298	99.3	298	99.3	299	99.7	278	92.7
人民	5	200	188	94.0	188	94.0	188	94.0	160	80.0
新华	10	300	300	100.0	300	100.0	300	100.0	300	100.0
凤凰	5	200	199	99.5	199	99.5	199	99.5	100	50.0
福大	1	20	20	100.0	20	100.0	20	100.0	17	85.0
网易	10	300	300	100.0	300	100.0	300	100.0	300	100.0

(下转120页)

万方数据

本文算法简单,易于实现,计算速度也快,具有一定的应用前景。今后的工作,将会把如何恢复图像篡改区域作为研究重点。

参考文献:

[1] 霍耀冉,和红杰,陈帆.基于邻域比较的JPEG脆弱水印算法及性能[J].软件学报,2012,23(9):2510-2521.

[2] Tong Xiaojun,Liu Yang,Zhang Miao,et al.A novel chaos-based fragile watermarking for image tampering detection and self-recovery[J].Signal Processing:Image Communication,2013,28(3):301-308.

[3] 张玉梅,和红杰,陈帆.浏览器端定位篡改的网页脆弱水印算法[J].计算机研究与发展,2014,51(12):2604-2613.

[4] 刘东彦,刘文波,张弓.图像内容可恢复的半脆弱水印技术研究[J].中国图象图形学报,2010,15(1):20-25.

[5] 赵春晖,刘巍.基于分块压缩感知的图像半脆弱零水印算法[J].自动化学报,2012,38(4):609-617.

[6] 杨晋霞,鞠杰,邵峰.基于超混沌加密的半脆弱音频水印算法[J].计算机应用与软件,2014,31(11):295-298.

[7] 蔡键,叶萍,刘涛.基于小波变换的用于医学图像的半脆弱水印算法[J].计算机应用与软,2011,28(6):278-230.

[8] 陈帆,王宏霞.定位像素篡改的安全脆弱水印算法[J].铁道学报,2011,33(1):63-68.

[9] Rawat S,Raman B.A chaotic system based fragile watermarking scheme for image tamper detection[J].AEU-International Journal of Electronics and Communications,2011,65(10):840-847.

[10] 刘敏,陈志刚,邓小鸿.基于混沌和脆弱水印的图像篡改检测算法[J].计算机应用,2013,33(5):1371-1373.

[11] Tareef A,Al-Ani A,Hung N,et al.A novel tamper detection-recovery and watermarking system for medical image authentication and EPR hiding[C]//Proceedings of the 36th Annual International Conference on Engineering in Medicine and Biology Society(EMBC),2014:5554-5557.

[12] 温泉,孙锬锋,王树勋.零水印的概念与应用[J].电子学报,2003,31(2):214-216.

[13] 隋淼,李京兵.一种基于Arnold置乱变换和DCT的医学图像鲁棒水印算法[J].计算机应用研究,2013,30(8):2552-2556.

[14] 吴伟民,丁冉,林志毅,等.基于混沌的医学图像篡改定位零水印[J].计算机应用研究,2014,31(12):3685-3688.

[15] Benrhouma O,Hermassi H,Ahmed A,et al.Chaotic watermark for blind forgery detection in images[DB/OL].[2015-09-30].http://link.springer.com.

[16] 刘瑶利,李京兵.一种基于DCT和Logistic Map的医学图像鲁棒多水印方法[J].计算机应用研究,2013,30(11):3430-3433.

[17] 周武杰,郁梅,禹思敏,等.一种基于超混沌系统的立体图像零水印算法[J].物理学报,2012,61(8):117-126.

(上接78页)

参考文献:

[1] 殷彬,杨会志.灵活结构网页的正文提取[J].计算机技术与发展,2011,21(9):111-113.

[2] Cai Deng,Yu Shipeng,Wen Jirong,et al.VIPS:A vision based on page segmentation algorithm,Microsoft Co,Tech Rep:MSR-TR-2003-79[R].2003.

[3] 韩忠明,李文正,莫倩,等.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):3568-3571.

[4] 安增文,徐杰锋.基于视觉特征的网页正文提取方法研究[J].微型机与应用,2010,9(3):38-41.

[5] Ji Xiangwen,Zeng Jianping,Zhang Shiyong,et al.Tag tree template for Web information and schema extraction[J].Expert Systems with Applications,2010,37(12):8492-8498.

[6] 王少康,董科军,阎保平.使用特征文本密度的网页正文提取[J].计算机工程与应用,2010,46(20):1-3.

[7] 刘军,张净.基于DOM的网页主题信息抽取[J].计算机应用与软件,2010,27(5):188-190.

[8] Mantratzis G C,Orgun M A,Cassidy S.Separating XHTML content from navigation clutter using dom-structure block analysis[C]//Proceedings of Conference on Hypertext,2005:145-147.

[9] 杨钦,杨沐昀.一种基于标点密度的网页正文提取方法[J].智能计算机与应用,2015,5(4):42-44.

[10] 高屹.基于树先剪枝的网页正文抽取方法研究[J].科技创新与应用,2013(36):63-64.

[11] 张瑞雪,宋明秋,公衍磊.逆序解析DOM树及网页正文信息提取[J].计算机科学,2011,38(4):213-215.

[12] 朱逢春.基于DOM树的网页去噪技术[J].电子制作,2015(8).

[13] Liu W,Huang G,Liu X.Detection of pubishing Web pages based on visual similarity[C]//Proceedings of the 16th Intl Conf on World Wide Web,2007:61-70.

[14] Kai S,Lausen G.VIPER:Augmenting automatic information extraction with visual perceptions[C]//Proceedings of the ACM CIKM Int'l CorIf on Information and Knowledge Management.[S.l.]:ACM Press,2005:381-388.

[15] Chibane I,Doan B L.A Web page topic segmentation algorithm based on visual criteria and content layout[C]//Proceedings of SIGIR Coference,2007.

万方数据