网页正文提取方法研究

赵明明[1]，陶华[2]，伏虎[2]，李昕[1]
（1. 北京邮电大学网络与交换国家重点实验室，北京 100876;
2. 河南省电力公司朝阳供电公司）

摘要：网络成为人们获取信息的重要途径。而网页上的内容除了主题内容外，还有如广告、版权信息、欢迎信息等与主题无关的内容，如何将网页中的正文内容提取出来已经成为机器学习和数据挖掘界的一个研究热点。本文将对网页正文提取方法的研究现状做一个简要介绍，并对未来的研究工作进行展望。

关键词：网页正文提取; DOM 树; VIPS 算法
中图分类号：TP301.6

Research on webpage content extraction algorithm

Zhao MingMing[1], Tao Hua[2], Fu Hu[2], Li Xin[1]
(1. State Key Laboratory of Networking and Switching, Beijing University of Posts and Telecommunications, Beijing 100876;
2. HeNan Electronic Power Company, Xinyang Power Supply Company)

Abstract: Network has become an important way for people to obtain information. The web pages contents include subject matter, in addition, there also including advertising, copyright information, welcome message and other topics unrelated with the contents, how to extract the contents of Web pages out of the body has become a research focus for machine learning and data mining sector. This articleal will make a brief introduction of the gorithm research extracting the body of the page, and make prospects for future research work.

Keywords: webpage content extraction; DOM tree; VIPS algorithm

0 引言

随着 Internet 的不断发展和日益扩大，Web 已经成为世界上最大的信息来源之一。Web[1]作为信息技术的载体，已成为学习、生活、工作、娱乐，经济，政治必不可少的工具之一。Web 的发展给人类生活带来了巨大的方便，人们可以跨越空间和时间限制获取大量信息。但是网页的内容除了主题内容外，还有如版权信息、广告、导航栏，装饰信息等与主题内容无关的内容，称为"噪音"消息。如何屏蔽噪音信息，将网页中的正文内容提取出来，在互联网技术迅猛发展的今天具有重要意义。

目前在这个领域已经发表了很多的研究成果[2]，分类方法也各有不同，目前主流有三大类的研究方法，一种是基于模板的网页正文提取的方法[3][4][5][6][7]，一种为基于语义信息的网页内容提取方法，另一种为基于视觉的网页内容提取算法。但目前还没有一种方法能适用于所有的网页内容提取，从而没有达到人们期望的程度，因此还需要不断地研究和探索。本文接下来对目前的网页内容提取方法进行简要介绍，然后对该研究领域进一步的研究工作进行展望。

1 基于模版的网页正文提取算法

基于模板的网页正文提取算法依赖于 HMTL 内部结构特征。该方法设定 WEB 的同类网页中有着相似的结构特征或者相似的 Dom 树(Document Object Model)结构。可以通过制定模板[3]获取同类网页的正文内容。基于模板的网页正文提取方法一般使用分装器(wrapper)

作者简介：赵明明，(1985-)，女，硕士研究生，主要研究方向：宽带通信网. E-mail: mingmingbupt@gmail.com
通信联系人：李昕，北京邮电大学宽带网研究中心副教授、硕士生导师，研究。E-mail:cplalx@gmail.com

[4][5][6][7]来抽取网页中正文数据。分装器是一个程序,该程序根据网页的页局特征，制定模板，编写分析器，解析出正文在页面中的位置，即它根据特定的信息模式从信息源中抽取出需要匹配的内容,并以某些形式展示出来。

基于模板的网页正文提取方法的重点和难点是如何确定以及维护模板，分装器如何生成。分装器的生成和维护都是费时费力的。目前研究人员人仍在研究如何高效的构建分装器。目前较为流行的有 TSIMMI 系统中的分装器[8]，Ontology 系统中的包装器[9]，XWRA 系统中的包装器[10]。下面简要介绍。

1.1 TSIMMIS 工具简要介绍

TSIMMIS 工具[8]通过手工提取规则来构建分装器。所提取出的规则以[variables , source , pattern] 的形式表示，其中 variables 存储提取出的网页正文内容；source 存储输入的 html 文件；pattern 存储 source 中文件的模式信息。提取出的规则被放到专门的文件中存储。

TSIMMIS 工具实现了规则与数据的分离，减轻了维护的负担。该方法的瓶颈在于需要手工编写规则，对工具的使用者的要求很高，需要专家级别的使用人员，限制了工具的使用范围。因此使用该工具是费时费力的需要高智商的工作。

1.2 Ontology 工具简要介绍

使用 Ontology 工具[9] 构建分装器需要首先构造一个完备的数据库，数据库中元素为网页的一个个包含数据的记录块的抽取模式以及记录块之间的联系。Ontology 工具抽取的模式没有使用依赖于自然语言技术代表性的特定模式，比如特定文档的特定的分隔符或者词性等，而是主要使用的是词法模式，比如系统中判定电子邮件的词法模式为"^\w+[-+.]\w+)*@\w+([-.]\w+)*\.\w+([-.]\w+)*$"。

Ontology 工具方法的优点很明显，不依赖于结构和表现形式。该工具使用 Ontology 的数据库来定位关键信息并使用关键信息与网页进行匹配，匹配成功的部分即为网页正文区。但是构建一个完备的 Ontology 数据库是一件困难的事情，因为有些信息无法给出对应的 Ontology。同时在很多情况下一个简单的任务并不需要构建一个巨大的数据库，构建一个完备的数据库需要专家级别的工作人员花费很多时间完成，得不偿失。

1.3 XWRAR 工具简要介绍

XWRAP [10]系统中的分装器无须全手工编写规则，它通过半自动化的方式获取规则。该分装器首先检查网页标签的规范性，使用网页净化工具如 tidy 净化网页；在对网页标签规范化后，将网页解析成 DOM 树。用户使用该分装器时，它会提供良好的用户界面，用户根据系统的指导完成规则的编写，实现了半自动化的规则编写过程；最后系统自动生成一个针对某一类网页的 java 语言编写的分装器。

1.4 基于模板的网页正文提取算法总结

使用基于模板的网页正文提取算法可以快速的提取出格式较为规则的网页正文内容，一旦模板制定，抽取速率很快。但是模板的制定过程复杂而且耗时很大。同时基于模板的网页正文提取算法都是线性的处理 HTML 文档，通过字符串模式匹配到关键信息，而忽略了 HTML 文档本身的语法。同时字符串模式很难保证匹配的准确性。为了改进网页正文内容提取的效率，研究人员提出了基于语义信息的网页正文提取方法。下面详细介绍。

2 基于语义信息的网页内容提取

所谓的语义信息是指除了网页中的视觉信息（字体大小，颜色，背景色等信息）之外的所有信息。包括 HTML 的标签信息，网页的文字信息，HTML 的 DOM 树信息。

基于语义信息的网页内容提取方法可以分为两类，一类是基于去除 html 标签的网页正文提取算法；另一类是基于统计的建立 DOM 树的网页正文提取算法。

下面详细介绍这两种方法。

2.1 基于去除 html 标签的网页正文提取算法

本节介绍基于去除 html 标签的网页正文提取算法的原理和特点[14][15][16]。

2.1.1 算法原理

基于去除 html 标签的网页正文提取算法是一种不必建立 DOM 树的网页正文提取算法[15][16]。该算法的主要思想是首先去除 html 标签，根据去除 html 标签后的文字密度判断出正文区域。最后将所有的正文区域合并，取得网页正文内容。下面详细介绍具体的算法流程

第一步首先需要去除 html 标签和 javaScipt 代码。采用正则表达式与栈的联合使用方法去除 html 的标签和 javaScipt 代码。例如对于<script>…</scripte>这种类型的标签，使用正则表达式<script>:?</script>去除标签以及标签中的内容，对于 … … <./div>这种类型，使用栈，完成按顺序取出 html 中的非标签文字的需求。

第二步根据去除 html 标签后的文字密度获得正文区域。对于去除 html 标签后的文字，判断每一行文字的长度，判断每一行的文字数，如果没有两个以上连续的文字长度为零，继续执行该操作，直到遇到至少连续两行出现文字长度为空的情况，这时判断该过程中取出的文字的总长度是否超过某阈值，如果超过阈值，该部分为网页正文，否则不是。按照上面的方法继续执行，直到文件的结尾处结束。

第三步将所有的正文区域合并，取得网页正文内容。具体为按照第二步获得的所有正文区域的先后顺序合并所有得到的正文区，得到最终的网页正文内容。

2.1.2 算法特点

基于去除 html 标签的网页正文提取算法[15][16]仅利用 html 文字行之间的间隔进行解析，对于结构相对简单的网页来说，效果很好。然而现在流行的网页结构是不规则的，有些正文块仅有一行，但该行文字数较多，根据该方法如果阈值设置不合适，这个正文块将被忽略不会被获取。同时去除 html 标签的过程也是一个较为费时的过程。目前该算法可以与其他算法（比如下文提到的基于统计的建立 DOM 树的网页正文提取算法，基于视觉的网页正文提取算法）相结合进行网页正文提取。

2.2 基于统计的建立 DOM 树的网页正文提取算法

本节介绍基于统计的建立 DOM 树的网页正文提取算法的原理和特点。

2.2.1 算法原理

基于统计的建立 DOM 树的网页正文提取算法主要是通过建立 DOM 树，然后根据 DOM 树中每个节点标签的文本数量与链接数量，正文长度与链接文字长度的比例为标准，判断该节点是不是正文节点。最后先序遍历 DOM 树，将所有的正文节点取出，完成正文提取功能。

下面详细介绍算法流程。

1.html 网页进行净化，使用 html 规范工具如 tidy 等来修正 html 网页的不规范之处；

2.根据 html 网页建立 Dom 树；

3 遍历一遍 DOM 树，获取每个节点下文本的数目，文本的长度，链接的数目，链接中文本的长度；

4.统计每个节点下文本以及链接出现的次数、文本的长度以及链接下文字的长度，当文本出现的长度以及文本出现的次数比链接出现的长度以及链接出现的次数之比大于阈值时，则将这个节点作为可以选择的内容节点，否则认为该节点所代表内容为广告区等或者版权信息等噪音信息，忽略该节点；

5.前序遍历 DOM 树，将所有被选择为内容节点的节点合并，得到网页正文内容。

2.2.2 算法特点

基于统计的建立 DOM 树网页正文提取方法[19]充分利用了新闻类网页的特性，算法简单便于实现，准确性较高，效率高。对格式相对复杂的页面来说，是比较有效的，它甚至可以抽取出其它方法抽取效果很不好的复杂网页的正文内容，在这种情况下这该算法的效果很好。但该方法也具有很大的局限性，只适合于类似于新闻类的网页，对于论坛类网页的提取效果不是很明显。

2.3 基于语义的网页正文提取方法总结

基于语义的网页正文提取算法相对而言简单便于实现，准确率较高。但他也有无法克服的局限性。对于去除 html 标签的网页正文提取算法，其去除标签所消耗的时间是影响效率的瓶颈之处。而基于统计的建立 DOM 树的网页正文提取算法需要建立 DOM 树，但由于 HTML 的语言版本众多，HTML 语法松散，网页编写没有统一的标准，导致很多网页没有遵循 W3C 标准同时由于各个浏览器对 html 的标签识别不一致，这将增加基于统计的网页正文提取方法的复杂性。在实际编程中，通过使用一些 html 规范工具如 tidy 来修正 DOM 树结构的错误。但不是所有网页使用这些规范工具都可以修正的。同时 DOM 树最初的设计目的是为了显示浏览器中的布局而不是为了显示 html 的语义信息。比如 dom 树两个语义上没有任何联系的结点可能具有共同的父节点，而语义上有联系的两个结点可能在 DOM 上没有任何关联。综合以上分析，基于语义的网页正文提取方法有着一些无法克服的先天性的局限性。

3 基于视觉的网页内容提取

基于语义的网页正文提取算法对于网页结构复杂，正文区分布零散的网页提取效果很差，基于视觉的网页正文提取算法[20]弥补了基于语义的网页正文提取算法的不足。经典的基于视觉的网页正文提取算法称为 VIPS 算法 (Visual Based Page Segment Algorithm)，，中文名为基于视觉的 Web 页面切割算法，或基于视觉的 Web 页面正文抽取算法，该算法是由在微软亚洲研究院实习的北大和清华的学生提出的。

该算法的原理是该算法从用户的观察角度来分析网页的结构，即关心的是网页的视觉信息而不是 web 网页的内部结构，模拟人的眼睛识别语义内容的过程并结合 DOM 树进行分析，可以弥补基于语义的正文提取的不足之处，获得更精确的抽取效果。

3.1 基于视觉的网页正文提取算法思想

互联网用户通常根据 web 页面的布局特征（请参见图 2）感知 web 正文区的内容，基

于视觉的网页正文提取算法的主要思想[20][21]就是模拟互联网用户的判断过程进行正文抽取。该算法的输入为 HTML 源代码，HTML 的视觉属性（需要修改浏览器的内核，使得浏览器提供视觉属性接口，便于开发者获得 HTML 的视觉属性）。

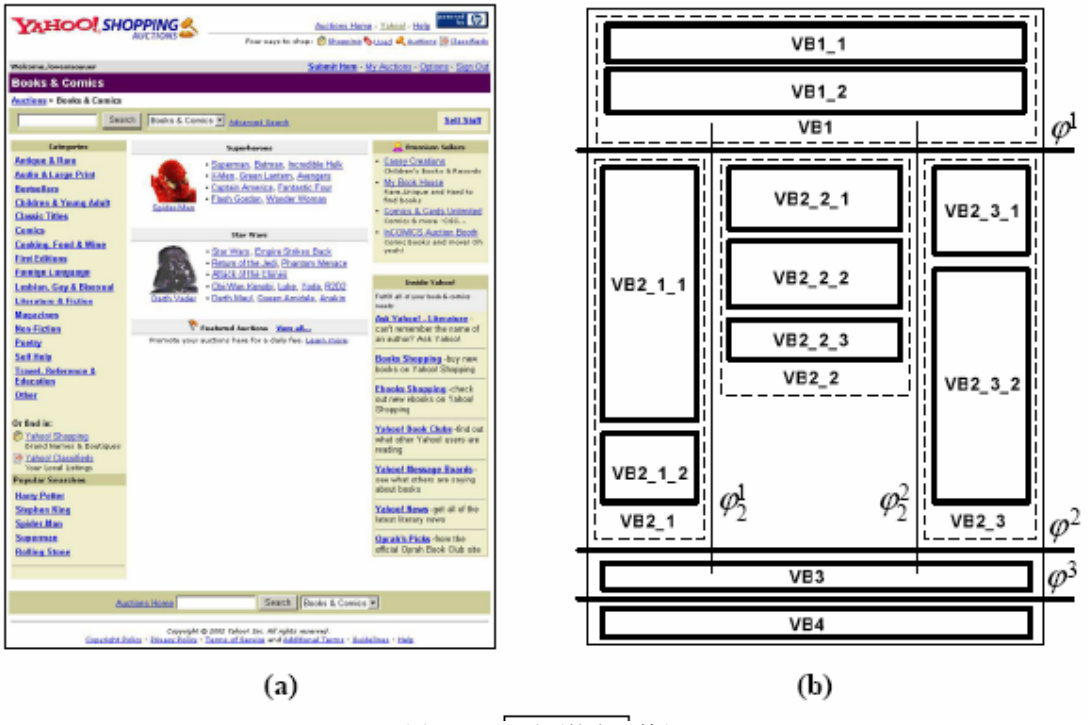

图 2 web 页面的布局特征

Fig. 2 web page layout features

基于视觉的网页正文提取算法流程参见图 3。该算法的流程大体分为三步：首先根据视觉信息和 HTML 源代码将网页分割为多个页面块，分块的过程是一个迭代的过程，初始化页头为一块，中间为一块，页脚为一块。然后将中间块分割为更多的小的页面块。其次当分块完成后，需要对页面块与页面块之间的分隔条（水平和垂直方向分隔条）赋予权值，便于分析页面块之间的关联性。最后，需要根据分隔条的权重进行页面合并，将太小的页面进行合并，使它可以更好地体现页面的主题内容，合并结束后将最后得到的这些数据块提取出来，得到网页正文内容。该算法清晰的展现出 web 页面的主题内容块，完成 HMTL 源码与人类视觉信息的良好沟通。

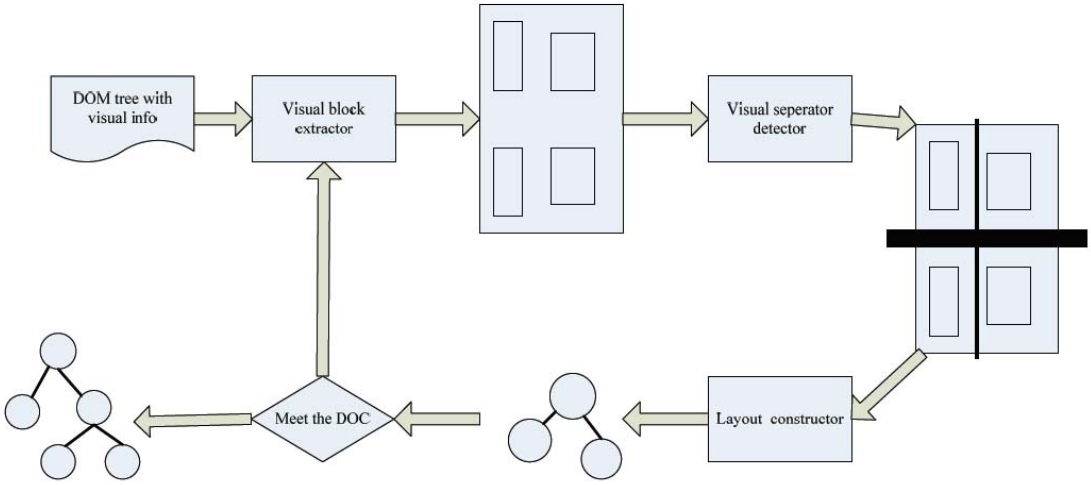

图 3　VIPS 算法流程
Fig. 3　VIPS algorithm process

具体算法流程为：

第一步：将 web 页面进行分块，获得所有的有效页面块。页面的分割方法是根据 Dom 树的各种属性，为每个节点所代表的内容块计算 DOC（Degree of Coherence）值，该值用来描述节点内容块与块之间的耦合程度，DOC 值越大，意味着在当前内容块中的子内容块之间的耦合度很紧密，被分割的概率降低。当 DOC 值降低到阈值以下，认为该节点需要继续分割。该过程迭代进行，直到所有的内容块 DOC 值都在阈值以上，即完成 web 分块功能。

DOC 值的根据 Dom 树的结点属性（包括结点的标签，背景色，该结点所表示的页面的大小形状，Dom 树的结点的子节点标签，子结点所表示区域的背景色，区域的大小）计算，具体的计算公式可以根据经验值或者其他特殊需求制定）。

第二步：需要对页面块与页面块之间的分隔条赋予权值。所谓的分隔条是不同语义的页面块的分隔标志，根据分隔条两边的语义块在视觉上的差异，设置分隔条的权重。分隔条的权值大意味着该分隔条两侧的页面块属于不同语义块的概率大，反之亦然。分隔条的设置与分割条两侧的内容块的视觉属性有关。

分隔条的权值设置过程为：获取分隔条两侧的页面的视觉属性。跟设置权值相关的属性有分隔条两侧的页面块的距离（距离越近，权重越小，合并的概率越大）。分隔条两侧的页面块的背景色（如果背景色是不同的，权重减小，合并的概率减小）；分隔条两侧的页面的字体属性（对于水平分隔符，字体大小，颜色等属性如果不同，分隔条的权重增加;对于垂直分隔条,上侧的字体小于下侧的字体，分隔条的权重增加）；分隔条两侧的页面结构（比如都是图片，或者都是文字，该分隔条的权重相应的减少）。

第三步：进行页面的合并。根据分隔条的权值信息，从权值最小的分隔条开始合并。将分隔条两侧的页面合并成一个新的页面块。该过程进行迭代。直到所有分隔条的权值都小于阈值为止。将这些页面块的内容提取出来作为网页正文内容。

3.2　基于视觉的网页正文提取方法特点

基于视觉的网页正文提取算法充分利用了网页的框架信息和视觉信息，相比于基于语义信息的网页正文提取算法，对于结构较为复杂，正文内容分散的网页可以提高提取的准确性。但基于视觉的网页正文提取算法也有其先天不足之处：

首先，基于视觉的网页正文提取算法需要多次的迭代，最后需要语义块的合并。相比于基于语义信息的网页正文提取算法其迭代次数较多，实现起来更为复杂增大了时间复杂度。

其次，网页视觉信息的提取是费时费力的。因为网页视觉信息的获取与浏览器本身，css 文件，javascript 文件有关，获取视觉信息之前需下载这些文件，之后浏览器的内核将调用这些文件，最后从浏览器的对外接口中获取网页的视觉信息。这个过程依赖于浏览器的内核代码，同时非常耗时。目前某些已经实现的 VIPS 算法是在 Windows 编程环境下实现的，实现中调用 IE Com 接口，而该算法的发现者微软使用的是修改后的 IE 内核，使得 IE 内核可以提供相应的网页视觉信息提取接口。在 linux 编程环境下，目前只有 Firefox（Mozilla）浏览器的源代码开放，如果需要在 linux 环境下实现网页视觉信息提取，需要修改 Firefox 的源代码，这样在通用性和可扩展性方面的优势降低。

4 总结和展望

本文对现在流行的网页正文提取方法进行了分析和总结。经过对基于模板的网页正文提取方法，基于语义信息的网页内容提取方法，基于视觉的网页内容提取方法介绍可知不同复杂性不同结构的网页应采用不同的网页正文提取方法。下面对各种网页提取方法的应用场景进行介绍。

对于结构比较规范的网页，适用于基于模板的网页正文提取方法；对于结构比较简单的网页，适用于基于去除 html 标签的网页正文提取算法；对于网页结构比较复杂且正文内容集中的网页，适用于基于统计的建立 DOM 树的网页正文提取算法；对于页面结构比较复杂且网页内容分布在多个标签中的网页，适用于基于视觉的网页正文提取算法。

目前，影响网页内容抽取技术的两个关键因素为：执行性能以及可移植性[22]。今后网页正文抽取方法的研究将围绕如何增强这两个方面展开，重点解决知识获取、篇章分析、高效句法，语义特征标注、共指消解[22]等问题，不断的提高网页正文抽取系统的性能，增强系统的性能和可移植性。

[参考文献] (References)

[1] Huberman,Adamic. Evolutionary Dynamics of the World Wide Web [OL].[1999]. http://www.parc.xerox.com/istl/groups/iea/www/growth.html
[2] Zhifeng Yang, Qichen Tu, Kai Fan, Lei Zhu, Rishan Chen, Bo Peng. Performance Gain with Variable Chunk Size in GFS-like File Systems[J]. Journal of Computational Information Systems,2008,4(3): 1077-1084
[3] Jiying Wang,Fred I-I. Lochovsky. Data-rich section extraction from HTML pages[A]. Pmc 3rd Int Conf on Web Info Syst Eng(wIsE. 02)[C].Singaporre:IEEE Computer Society Press，2002. 1-10.
[4] J. Hammer,H. Garcia Molina and J. Choeta1. Extracting Semi-structured Information from the Web[J].In the proceedings of the Workshop on Management for Semistructured Data,1997,23(2):1 8-25.
[5] B. Adelberg,NoDoSE.A Tool for Semi-automatically Extracting Structured and Semi-structured Data from Text Documents[J].in the proceedings of ACM SIGMOD Conference on Management of Data,1998,24(3)：283-294.
[6] N. Ashish,C. A. Knoblock.Semi-automatic Wrapper Generation for Intemet Information Sources[J].in the proceedings of the Conference on Cooperative Information Systems，1997,24(1):60-69.
[7] N. Ashish,C. A. Knoblock.Wrapper Generation for Semistructured Intemet Sources[J].SIGMOD Record,1997,1(26),1997:8-15.
[8] Garcia. Molina,Hammer,Lreland K. Integrating and Accessing Heterogeneous Information sources in TSIMMIs[J].Proceedings cff the AAAI SyInposium on Information Gathering, 1995,16(2)：61-64.
[9] Gruber T. A Translation Approach to Portable Ontology Specifications[J].Knowledge Acquisition ,1993 ,25(5) :199-222.
[10] LiuL,Puc.XWRAP:An XML enable Wrapper Construction system for the web information source[A],proceeding soft th16th IEEE international conference on Data Engineering[C],2000.

[11] 欧健文,董守斌,蔡斌.模板化网页主题信息的提取方法[N].清华大学学报,2005,45(S1).
[12] Gruber T. A Translation Approach to Portable Ontology Specifications[J] .Knowledge Acquisition,1993,01(5) :199-222.
[13] Crescenzi V,Meeea G.RoudRunner.Towards Automatic Data Extraction from Large Web Site[A].proceeding of the 26th International Conference on very Large Database Systems[C],2001.
[14] Bernard, M.L.,Criteria for optimal web design[OL].[2002] http://psychology.wichita.edu/optimalweb/position.htm.
[15] 张志刚，陈静，李晓明，一种 HTML 网页净化方法[N]，情报学报，2004，23(4).
[16] 宋睿华,马少平,陈刚,李景阳.一种提高中文搜索引擎检索质量的 HTML 解析方法.中文信息学报,2003,20(6).
[17] Daisuke Ikeda,Yasuhiro Yamada. Expressive Power of Tree and String Based Wrapper[A] . Daisuke Ikeda.on line proceedings of IJCAIp03 workshop on Information Integration on the Web[C] . 2003.
[18] 丛艳.自动文本摘要方法的研究及应用[D].北京:华北电力大学,2004 年.
[19] Line Eikv il. Information Extraction from World Wide Web A Survey[Z]. 1999.
[20] D. Cai, S. Yu, J.-R. Wen, W.-Y. Ma. VIPS a vision based page segmentation algorithm[R], Microsoft Technical Report,MSR-TR-2003-79, 2003.
[21] D. Cai, X-F. He, J.-R. Wen,W.-Y. Ma, Block-level LinkAnalysis[R], In Proceedings of the ACM-SIGIR, 2004.
[22] 李保利,陈玉忠,俞士汶. 信息抽取研究综述[J]. 计算机工程与应用, 2003, 39(10) :24 - 27