网页信息提取技术

邵振凯
(安徽理工大学 计算机科学与工程学院,安徽 淮南 232001)

摘 要:随着互联网的快速发展,Web 页面上的信息量已变得非常巨大,面对网页上海量的信息资源,如何快速有效地检索及发现有价值的信息已成为 Web 研究的一个重要方面。对此提出了一种标签提取方法。利用 JTidy 将网页优化为格式良好的 HTML 文档并解析为 DOM 树,然后用标签提取方法对该 DOM 树中包含有文本信息内容的叶子节点标签进行提取,把用于控制网页交互性和显示的标签删除掉,并运用基于标点符号的信息提取方法去除版权说明等信息。对不同网站的网页进行抽取实验,结果表明标签提取方法不但通用性强,而且能够准确地提取网页的主题信息。

关键词:DOM;标签提取;信息提取;网页净化
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2013)09-0036-03
doi:10.3969/j.issn.1673-629X.2013.09.009

Web Page Information Extraction Technology

SHAO Zhen-kai
(College of Computer Science & Engineering, Anhui University of Technology, Huainan 232001, China)

Abstract: With the rapid development of the Internet, the amount of information in the Web page has become very large, how to quickly and efficiently search and find valuable information has become an important aspect of Web research. In this regard a tag extraction method is proposed. Optimize the Web page into good HTML format documents with JTidy, and resolve to a DOM tree. Then use tag extraction approach to extract the tags contain the text message content from DOM tree, remove the tags used to control the Web interaction and display, and use the method based on the punctuation information extraction method to remove the copyright notice and other information. The results on a number of different sites extraction show that the tags extraction methods not only have a great generality but also can accurately extract site theme.

Key words: DOM; tags extraction; information extraction; Web page purifying

0 引 言

互联网技术的迅速发展,以网页数据为数据源的研究越来越多,但网页中的数据和纯文本不同,网页中包含了各种 HTML 标签,并且大都是半结构化的数据,互联网还具有开放性、动态性与异构性、内容变化迅速等特点,所有这些都导致了网页数据的分散化,没有统一的管理布局风格,网页中还包含大量广告、著作所有权等信息,这些都是与文本本身无关的信息。此外,网页中大都包含模版,模版中又包含了导航条、网站 logo、组织标志和联系信息等,这些信息还会频繁地出现在同一个网站的所有网页中,这些内容都是噪音元素。信息提取的目的就是去除网页中的干扰信息,保留和页面主题相关的文本内容。在以网页数据为对象的相关研究和应用中,信息提取一直是近年研究的热点和重点。

1 研究概述

目前已有很多关于信息提取方面的研究,并且提出了许多关于网页净化的方法,其中主要包括基于统计的方法[1]、基于模版的方法[2]、基于网页内容分块的方法[3]和基于学习的方法[4]等。如宋明秋等人利用已有网页信息提取知识,针对中文网页布局的特点,先将 HTML 文件规范化[5]以构造 HTML 树,并提取结构树中的文字内容及其链路结构;然后根据中文句号的出现频率来确定一部分正文内容;最后根据正文内容链路结构的相似性获取其余正文内容[6]。基于模版[7]的方法主要是利用模版检测进行去噪,即一个网站中重复出现在很多网页中的内容,这些模版内容即是噪

收稿日期:2012-11-23 **修回日期**:2013-02-24 **网络出版时间**:2013-05-09
基金项目:安徽省自然科学基金(11040606M135)
作者简介:邵振凯(1986-),男,硕士,研究方向为计算机监控。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20130509.1057.018.html

音的主要部分,这种方法只适合具有相似模版的网页。还有基于网页分割模型[8]的方法,主要是将网页根据逻辑区域进行分割,逻辑相关的放到一个块里面,作为一个基本的数据块,然后将这些块当作叶子节点构造出一棵 DOM 树结构,进而进行网页的信息提取。

文中采用构造 DOM 树[9]的方法对网页信息进行提取,主要对信息提取[10]和噪音信息裁剪[11]方面的算法进行了进一步的研究。首先对网页的 HTML 文档进行处理,转换为一棵 DOM 树结构,然后提取标签简化为一棵只包含根节点和叶子节点的 DOM 树结构,再从这些叶子节点中提取出网页的主题信息。结果证明运用此方法取得了良好的效果,文章对不同网站的网页进行了抽取实验,实验结果表明文中提出的方法极大地提升了网页信息提取的准确度。

2　信息提取

2.1　HTML 预处理

网页是由各种标签组合而成,这些标签通过嵌套而形成一定层次;其次,一个定义良好的网页应该是层次清晰、标签完整的;再次,网页的主要信息都是放在特定的标签里的,对这些特定的标签进行处理即可得到网页的主要信息。

要对网页进行信息提取,首先需要对网页进行处理。目前互联网中大多数网页仍然使用 HTML 格式。在 HTML 格式的网页中,存在一些编码不规范等情况,如标签不匹配、嵌套混乱及标签格式不完整等。这种不规范有时不会影响网页的正常显示,但会给正文信息的抽取带来麻烦,因此首先应对 HTML 代码进行预处理,将其标准化[12]。文中采用工具集 JTidy 将书写不规范的 HTML 文档进行规范化,转化为 DOM 树的主要算法如下:

算法:预处理。

输入:网页的 URL 地址。

输出:网页的 DOM 树。

1. 根据 URL 地址获取 HTML 源码;

2. 修复 HTML 源码;

3. HTML 转化成 XHTML;

4. 对 XHTML 进行处理转化成 DOM 树。

```
IterateNode(Node,D_tree)//从上向下遍历节点
{
    if(节点是有效节点){
            Node=(Name,Attribute,Mul);
            //获取节点信息
            D_Tree.add(Node); //添加节点
        }
    if(节点包含子节点){
            Node.getFirstChild();
            while(Node is Not Null){
        IterateNode(ChildNode,D_tree);
        NextNode=ChildNode.getNext();
        //获取下一个节点
        ChildNode= NextNode;
        }
    }
}
return DOM 树
```

2.2　信息提取

算法中 Name 表示节点标签名称;Attribute 表示节点显示特性,例如节点内容显示的字体、颜色、背景、注释等信息;Mul 为节点的语义信息,通常可分为文本、超链接、多媒体等。在对文本信息进行提取时一般会将 Attr 中的显示特性的信息去掉,保留 Sem 中的超链接信息。因为网页中的文本信息才是主要信息,部分信息有可能要参考页面中的链接。网页中的主要信息是以文本形式展现的,而这些信息一般都位于 DOM 树的叶子节点内,所以在进行提取时主要是对 DOM 树的叶子节点进行提取,这些节点就是算法中的有效节点,即含有文本信息的节点。直接包含文本信息的标签主要是以下几种:<title>、<meta>、<p>、<tr>、<td>。提取包含文本信息的标签的同时把用于控制文件的交互性和显示的标记删除掉,如<form>、<script>、<style>等这些标签不含有有效信息,删除后得到的 DOM 更简洁,也更易于处理。具体的标签提取算法如下:

```
IterateNode(Node,D_tree){
    if(标签节点为注释、script、style 等不含有文本信息的
标签)
        D_tree.delete(Node);
    else if(节点是叶子节点)
        D_tree.add(Node);
}
return D_tree;
```

如下所示代码为未进行处理过的 html 格式的网页文档,文档中的标签有的含有文本信息,有的只是用于控制显示的信息,示例为格式良好的文档,一般网站的 html 文档不是格式良好的,都要经过处理后转换为格式良好的 html 文档。图 1 为对应的 DOM 树结构,html 标签为根节点。Web 中的信息都是存在于叶子节点中的,在对 DOM 树进行优化后保留其叶子节点,得到如图 2 所示的树形结构。该树有一个根节点其余均为叶子节点,这样不仅大大简化了处理的信息,还保留了所有有用的信息,不过这样还是有一些噪音信息包含在里面,还要进行进一步的处理。

```
<html>
    <head>
    <title> web 网页净化 </title>
```

```
    <meta name="Generator" content="EditPlus">
    <meta name="Author" content="">
    <meta name="Keywords" content="">
    <meta name="Description" content="">
</head>
<body>
    <div id="" class="">
      <font size="" color=""> web 网页净化</font>
      <p> web 网页净化</p>
    </div>
    <p> web 网页净化</p>
  </body>
</html>
```

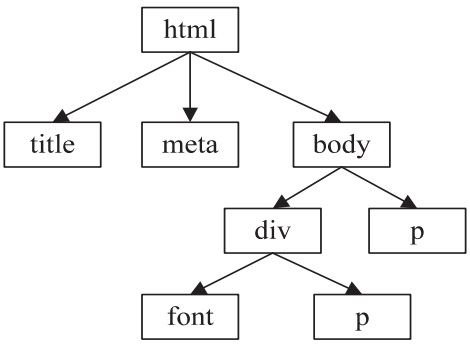

图 1 简化后的 DOM 树结构

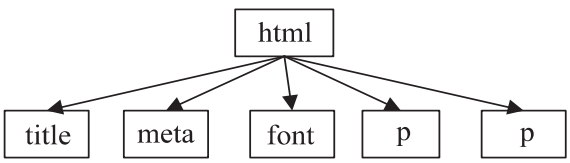

图 2 处理后的树形结构

在通过对节点信息内容的构成进行研究后发现,宋明秋等人提出的基于中文标点符号的信息提取方法不失为一种好的信息提取方法,通过统计发现中文句号出现在网页正文中的概率是所有中文标点符号中最多的,其主要原因是网页正文部分大多由一个个句子组成,所以句号出现较多;而导航信息大多由短语组成;链接部分一般都取自所链接文章的标题,标题中一般不会出现句号;版权部分也基本都没有成行的句子,所以句号较少。根据这样一个特点,用此方法对标签<p>进行处理,可去除其中的版权等噪音信息,实现较好的去噪效果。

2.3 提取结果

如图 3 所示,经过处理后的网页只包含了该新闻的主要内容,诸如广告、导航栏等与主题无关的内容均已去除,达到了预期的效果。运用统计学的基本知识计算文本信息提取率, U_e 为提取的单个网页的主题信息字节数, U_r 为实际的网页主题信息字节数,则单个网页主题信息提取率为 P_1。

$$P_1 = U_e / U_r$$

网页主题信息提取率为 P。

$$P = 1/n \sum_{i=1}^{n} p_i$$

图 3 网页处理前后对比

应用文中提出的算法对新浪、搜狐、腾讯、网易等几种不同的网站中的新闻网页内容进行了信息提取,计算网页的主题信息提取率,结果如表 1 所示。

表 1 网页信息提取率

网站	网页数	提取成功的网页数	准确率/%	主题信息提取率 P/%
www. 163. com	223	219	98.2	94.3
www. sohu. com	197	194	98.5	97.5
www. sina. com. cn	234	223	95.3	93.7
www. qq. com	202	191	94.6	95.1

3 结束语

文中通过对网页结构和现有的关于网页净化算法的分析与研究,提出的对 DOM 树中的叶子节点进行处理的方法实现了较好的网页信息提取效果,并引入了基于标点符号的信息提取方法,这使网页信息提取更加精确,有针对性。文中不足之处是没有对提取出的链接信息进行较好的处理,有一些是和主题信息相关的链接,有一些是无效的广告链接等信息,都需要进行进一步的研究。

参考文献:

[1] 宋明秋,张瑞雪,吴新涛,等. 网页正文信息抽取新方法[J]. 大连理工大学学报,2009,49(4):594-597.

[2] 李文立,王乐超,宋春雷. 基于 HTML 树和模板的文献信息提取方法研究[J]. 计算机应用研究,2010,27(12):4615-4617.

[3] 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):1326-1330.

[4] Ji H, Deng H, Han J. Uncertainty Reduction for Knowledge Discovery and Information Extraction on the World Wide Web

(下转第 42 页)

层上的需求活动集为：

$\{RA_{11},RA_{12},\cdots,RA_{1p}\} = \{f(Ro_{11},I_{11},Ru_{11},G_{11},E_{11}),f(Ro_{12},I_{12},Ru_{12},G_{12},E_{12}),\cdots, f(Ro_{1p},I_{1p},Ru_{1p},G_{1p},E_{1p})\}$

……

系统第 n 层(最低层) 建模为：

$A_n=\{A_{n1},A_{n2},\cdots,A_{nq}\}$，$I_n=\{I_{n1},I_{n2},\cdots,I_{nq}\}$，$Ru_n=\{Ru_{n1},Ru_{n2},\cdots,Ru_{nq}\}$，$G_n=\{G_{n1},G_{n2},\cdots,G_{nq}\}$，$E_n=\{E_{n1},E_{n2},\cdots,E_{nq}\}$

则系统第 n 层的子系统集为$\{S_{n1},S_{n2},\cdots,S_{nq}\}$，该层上的需求活动集为：

$\{RA_{n1},RA_{n2},\cdots,RA_{nq}\} = \{f(A_{n1},I_{n1},Ru_{n1},G_{n1},E_{n1}),f(A_{n2},I_{n2},Ru_{n2},G_{n2},E_{n2}),\cdots, f(A_{nq},I_{nq},Ru_{nq},G_{nq},E_{nq})\}$

系统除第 0 层外，每一层都可以分解为多个子系统，下层子系统关键要素聚集构成上层子系统关键要素，形成跨层连接，同样，下层子系统关键要素所映射的需求活动，聚集构成上层子系统关键要素所映射的需求活动，所有这些需求活动共同组成整个系统的需求活动，为系统需求分析提供了层次化的组织结构分析模型。

3 结束语

基于社会计算的需求工程过程形式化描述框架，将社会科学的理论、方法及研究成果引入需求工程过程中，整个系统被分解为成员、角色、环境、组织，并用层次化的组织结构对系统关键要素：成员(具体的角色)或角色、交互、规则、目标、环境自上而下分解为层次子系统，将子系统关键要素映射为需求活动，用具体的需求活动实现子系统功能，对子系统进行聚集，得到系统全局组织结构，对软件系统建立社会抽象需求框架，为系统需求分析提出新的社会思维方法。

参考文献：

[1] Popova V, Sharpanskykh A. A Formal Framework for Modeling and Analysis of Organizations[J]. IFIP International Federation for Information Processing, 2007, 244: 343-358.

[2] Jonker C M, Sharpanskykh A, Treur J, et al. A Framework for Formal Modeling and Analysis of Organizations[J]. Appl. Intell., 2007, 27(1): 49-66.

[3] Dignum V, Aldewereld H. OperettA: Organization-Oriented Development Environment[J]. Lecture Notes in Computer Science, 2011, 6822: 1-18.

[4] Ebbinghaus M W, Moldt D, Reese C, et al. Towards Organization-oriented Software Engineering[C]//Proc. of Software Engineering Conference. Hamburg: [s. n.], 2007: 205-217.

[5] Ferber J, Gutknecht O. A metamodel for the analysis and design of organizations in multiagent systems[C]//Proc. of the 3rd Int'l Conf. on Multi-Agent Systems. Paris: IEEE Press, 1998: 128-135.

[6] 张 伟，石纯一. Agent 组织的一种递归模型[J]. 软件学报，2002，13(11)：2149-2154.

[7] 金 芝，刘 璘，金 英. 软件需求工程：原理和方法[M]. 北京：科学出版社，2008.

[8] Wooldridge M, Jennings N R, Kinny D. A methodology for agent-oriented analysis and design[C]//Proc. of the 16th National Conf. on Artificial Intelligence (AAAI-99). Orlando, FL: [s. n.], 1999.

[9] Sommerville I. Software Engineering[M]. 8th ed. [s. l.]: Pearson Education Limited, 2007.

[10] 张国生. 基于层次着色 Petri 网的需求工程过程框架[J]. 计算机应用与软件，2011，28(8)：17-19.

[11] 张国生. 基于层次着色 Petri 网的功能需求模型[J]. 计算机技术与发展，2011，21(11)：81-83.

[12] 操龙兵，戴汝为. 开放复杂智能系统：基础、概念、分析、设计与实施[M]. 北京：人民邮电出版社，2009.

(上接第 38 页)

[J]. Proceedings of the IEEE, 2012, 100(9): 2658-2674.

[5] 潘大胜. 计算机半结构化数据源的数据挖掘技术探析[J]. 武汉工业学院学报，2011，30(4)：69-72.

[6] 万 乐，左万利，高 金. 基于主题的网页噪音去除机制[J]. 计算机工程与设计，2008，29(8)：2072-2074.

[7] 黄 荣. 基地模板的网页主题信息抽取模型[J]. 科技信息，2011(4)：250-251.

[8] 汪建伟，杨冬青，高 军，等. 一种基于分类算法的网页信息提取方法[J]. 计算机科学，2008，35(3)：91-93.

[9] Zhang Li, Li Meng, Dong Nannan, et al. An Improved DOM-based Algorithm for Web Information Extraction[J]. Journal of information and computational science, 2011, 8(7): 1113-1121.

[10] 周合明，奚建清. 基于模板的 Web 信息提取系统的设计与实现[J]. 计算机技术与应用，2011，21(11)：105-108.

[11] Lin Shian-Hua, Chu Kuan-Pak, Chiu Chun-Ming, et al. Automatic sitemaps generation: Exploring website structures using block extraction and hyper link analysis[J]. Expert Systems with Application, 2011, 38(4): 3944-3958.

[12] 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)：2646-2649.

网页信息提取技术

作者： 邵振凯， SHAO Zhen-kai
作者单位： 安徽理工大学 计算机科学与工程学院,安徽 淮南,232001
刊名： 计算机技术与发展
ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2013(9)

参考文献(12条)

1. 宋明秋. 张瑞雪. 吴新涛 网页正文信息抽取新方法[期刊论文]-大连理工大学学报 2009(04)

2. 李文立. 王乐超. 宋春雷 基于HTML树和模板的文献信息提取方法研究[期刊论文]-计算机应用研究 2010(12)

3. 邱江涛. 唐常杰. 李川 基于块分布的新闻网页内容提取[期刊论文]-吉林大学学报(工学版) 2009(05)

4. Ji H. Deng H. Han J Uncertainty Reduction for Knowledge Discovery and Information Extraction on the World Wide Web 2012(09)

5. 潘大胜 计算机半结构化数据源的数据挖掘技术探析[期刊论文]-武汉工业学院学报 2011(04)

6. 万乐. 左万利. 高金 基于主题的网页噪音去除机制[期刊论文]-计算机工程与设计 2008(08)

7. 黄荣 基地模板的网页主题信息抽取模型[期刊论文]-科技信息 2011(04)

8. 汪建伟. 杨冬青. 高军 一种基于分类算法的网页信息提取方法[期刊论文]-计算机科学 2008(03)

9. Zhang Li. Li Meng. Dong Nannan An Improved DOM-based Algorithm for Web Information Extraction 2011(07)

10. 周合明. 奚建清 基于模板的Web信息提取系统的设计与实现 2011(11)

11. Lin Shian-Hua. Chu Kuan-Pak. Chiu Chun-Ming Auto-matic sitemaps generation:Exploring website structures using block extraction and hyper link analysis 2011(04)

12. 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[期刊论文]-计算机应用研究 2009(07)