基于启发式规则的网页主题信息精确定位方法 *

胡金柱, 周 星, 舒江波, 熊春秀
(华中师范大学 计算机科学系, 武汉 430079)

摘 要: 目前大部分的信息抽取方法都是针对主题信息块的提取,没有进一步深入到各个单独主题信息的抽取。针对这一问题,提出了一种基于启发式规则的网页主题信息精确定位方法。首先针对各个单独的主题,分析其多方面的特征,制定出对应的启发式规则;然后利用不同的规则对定位主题重要度不同的这一特点,得到启发式规则的权值矩阵;最后利用基于启发式规则的定位算法精确定位各个主题。将该方法用于网页主题信息抽取系统中,抽取系统能够有效地对各个单独的主题进行定位和抽取。实验结果表明,该方法具有很好的有效性和准确性。

关键词: 启发式规则; 信息抽取; 主题信息定位; 模板化网页
中图分类号: TP311 **文献标志码**: A **文章编号**: 1001-3695(2010)02-0494-04
doi:10.3969/j.issn.1001-3695.2010.02.024

Approach of pinpointing subject information in Web pages based on heuristic rules

HU Jin-zhu, ZHOU Xing, SHU Jiang-bo, XIONG Chun-xiu
(*Dept. of Computer Science, Huazhong Normal University, Wuhan* 430079, *China*)

Abstract: At present, most of information extraction methods aim at the extraction of subject information block, not further penetrate into the extraction of each independent subject information. To solve this problem, this article proposed an approach of pinpointing subject information in Web pages based on heuristic rules. Firstly, for each independent subject, it analyzed its various characteristic, and formulated corresponding heuristic rules. Then, it obtained weight matrix of heuristic rules by using the feature that different rules had different importance to locate subject. Finally, according to localization algorithm of heuristic rules, it pinpointed each subject. The method has been applied to an automatic extraction system, and the experimental result shows the effectiveness and accuracy of the method.

Key words: heuristic rules; information extraction; subject information localization; template Web pages

0 引言

信息抽取(information extraction,IE) 是一种直接从自然语言文本中抽取事实信息, 并以结构化的形式描述信息的过程。通常被抽取出的信息以结构化的形式存入数据库中, 可进一步用于信息查询、文本深层挖掘、Web 数据分析、自动问题回答等。Web 页面所表达的主要信息通常隐藏在大量无关的结构和文字中,使得对 Web 文档进行信息抽取十分困难。一般的网页内容包括两部分,一部分是网页的主题信息,如一张新闻网页的新闻标题、新闻正文、发布时间、新闻来源;另一部分是与主题无关的内容,如广告信息、导航条,也称为噪声信息。如何有效地消除网页噪声,提取有价值的主题信息已成为当前信息抽取领域的一个重要课题[1]。

目前在网页信息抽取方面,国外的相关研究有:a) 文献[2]提出了从一种网页中抽取信息块的方法 EIBA,它首先将网页划分为语义块,然后手工标注信息标签和非信息标签,被标注的块用来作为分类模型的训练数据集,最后通过分类模型将信息块抽取出来。b) 文献[3]利用 HTML 文档的文本内容与标记的比率特性从网页中抽取信息,通过计算网页文本与标记的比率将网页聚类成内容和非内容的区域。国内的研究主要有:a) 基于模板的方法,采用机器学习来建立模板库,利用模板来直接提取网页主题信息,如文献[1]。b) 基于 DOM 树的方法,通过将 HTML 文档转换成 DOM 树,并对 DOM 树进行某种扩展,将页面抽取成具有语义特征或视觉特性的离散的信息条,然后通过遍历剪枝过的 DOM 树来实现信息抽取,如文献[4,5]。c) 基于网页布局特征的方法,利用标记在布局方面的作用对页面进行结构分析,区分主题内容和噪声内容,在此基础上抽取主题信息,如文献[6,7]。这些方法有一个共同的问题,都是针对主题信息块的提取,各个主题信息压缩在一起,没有进行进一步的处理。本文提出了一种基于启发式规则的主题信息精确定位方法,该方法对提取的主题信息块进行进一步的定位,分离出单独的主题信息。

收稿日期: 2009-05-05; **修回日期**: 2009-07-06 **基金项目**: 国家教育部人文社会科学重点研究基地重大项目(07JJD740063);湖北省科技攻关项目(2007AA101C49)

作者简介:胡金柱(1947-),男,教授,湖北宜昌人,博导,主要研究方向为软件工程、中文信息处理;周星(1985-),女,湖北咸宁人,硕士研究生,主要研究方向为软件工程(zhouxing@mail.ccnu.edu.cn);舒江波 (1982-),男,湖北咸宁人,博士研究生,主要研究方向为中文信息处理;熊春秀(1984-),女,湖北黄冈人,硕士研究生,主要研究方向为软件工程.

1　启发式规则及定位算法

本文是在模板化网页主题信息提取的方法[1]基础上进行进一步讨论的,该方法主要是针对同一个站点生成自动的抽取系统。同一个站点的 Web 文档基本上都是由同一个模板生成的。同一个模板生成的 Web 文档布局基本上一样,只是主题信息有所不同,其他信息基本一样。利用这一特点,将同一个站点下的多个 Web 文档组成的文档集转换为对应的 DOM 树集合,并对其进行训练,删除 DOM 树重复内容的子树,得到一个站点的模板树,即主题信息块。此时得到的 DOM 树中,一方面,主题信息(发布时间、来源、正文、标题)各个部分压缩在一起,无法区分各个单独的主题;另一方面仍夹杂着少量噪声,如当前位置、相关链接等。为抽取各个主题信息,就需要利用基于启发式规则的定位算法生成单个主题的抽取规则。

1.1　启发式规则定义

为了更好地描述启发式规则,定义了如下一些概念。

文本节点集合:$LN=\{ln \mid ln$ 是文本节点$\}$。

标题节点集合:$TN=\{tn \mid tn \in LN \wedge tn$ 的文本是标题内容$\}$。

正文节点集合:$CN=\{cn \mid cn \in LN \wedge cn$ 的文本属于正文内容$\}$。

发布时间节点集合:$TMN=\{tmn \mid tmn \in LN \wedge tmn$ 的文本是时间内容$\}$。

来源节点集合:$SN=\{sn \mid sn \in LN \wedge sn$ 的文本是来源内容$\}$。

发布时间特征词汇集合:Tset $=\{$ts $\mid$ ts 具有时间含义的词$\}$,Tset 中元素为“时间”“发布时间”“更新时间”“日期”等。

来源特征词汇集合:Sset $=\{$ss $\mid$ ss 具有来源含义的词$\}$,Sset 中元素为“来源”“转自”“来自”“转贴自”等。

节点偏序关系:对 DOM 树进行先序遍历,得到的节点序列为 $a_1a_2a_3\cdots a_n$,如果 $i \in \{1,2,\cdots,n\}$,$j \in \{1,2,\cdots,n\}$,且 $i<j$,则 a_i 与 a_j 满足偏序关系,记为 $a_i \prec a_j$。

文本节点直接先后关系:对 DOM 树进行先序遍历,取得文本节点序列为 $ln_1ln_2ln_3\cdots ln_m$,对于 $i \in \{1,2,\cdots,m-1\}$,称 ln_i 直接先于 ln_{i+1},记为 $ln_i \prec_w ln_{i+1}$。

经过对大量 Web 页面的分析,针对各个主题(发布时间、来源、正文、标题)制定了如下启发式规则。

Time 规则。a)对 $\exists$ ts $\in T$set,如果 ln. isContain(ts) = true,则 $ln \in TMN$。x. isContain(y)判断 x 是否包含字符串 y,包含则返回 true,否则返回 false。b)如果 ln. isContainFormat("YYYY-MM-DD") = true,或 ln. isContainFormat("YYYY 年 MM 月 DD 日") = true,则 $ln \in TMN$。其中 x. isContainFormat(y)判断 x 是否包含符合 y 格式的字符串,包含则返回 true,否则返回 false。

Source 规则。a)对 $\exists$ ss $\in S$set,如果 ln. isContain(ss) = true,则 $ln \in SN$。b)对 $\exists tmn \in TMN$,如果 $tmn \prec_w ln$,或者 $ln \prec_w tmn$,则 $ln \in SN$。

Content 规则。a)如果 $\lambda_1 \leqslant$ length(ln),则 $ln \in CN$。其中 length(n)是求节点 n 的文本长度,λ_1 为正文文本长度阈值。b)对 $pn=$ parent(ln),$\exists spn \in \{n \mid n=\langle p\rangle \wedge n \in$ sibling(pn)$\}$,如果 $pn=\langle p\rangle$,且 count(spn) $\geqslant \varepsilon_1$,则 $ln \in CN$。其中 parent(n)为节点 n 的父亲节点,sibling(n)为节点 n 在 DOM 树中的兄弟节点集合,count(n)为统计 n 节点的个数,ε_1 为 spn 节点个数的阈值。c)对 $ps=$ previousSibling(ln),$\exists sps \in \{n \mid n=\langle br\rangle \wedge n \in$ sibling(ps)$\}$,如果 $ps=\langle br\rangle$,且 count(sps) $\geqslant \varepsilon_2$,则 $ln \in CN$。其中 previousSibling(n)为节点 n 前一个兄弟节点,ε_2 为 sps 节点个数的阈值。d)对 $\exists tmn \in TMN$,如果 $tmn \prec ln$,则 $ln \in CN$。

Title 规则。a)如果 $\lambda_2 \leqslant$ length(ln),则 $ln \in TN$。其中 λ_2 为标题文本长度阈值。b)对 $\exists sbn \in$ sibling(ln),$\exists cn \in CN$,如果 $sbn=\langle$strong$\rangle$,或 $sbn=\langle$b$\rangle$,或 fontSize(ln) > fontSize(cn),则 $ln \in TN$。其中 fontSize(n)为节点 n 的字体大小。c)对 $pn=$ parent(ln),$\exists sbn \in$ sibling(ln),$ap=$ attr(pn). class,$as=$ attr(sbn). class。如果 ap. isContain("title") = true,或 as. isContain("title") = true,则 $ln \in TN$。其中 attr(n)为 n 节点的属性。d)对 $\exists tn \in TN$,如果 $ln \prec tn$,则 $ln \in TN$。

1.2　初始化权值矩阵

每个主题都存在各自的启发式规则,每条规则对定位该主题有不同的重要度。例如,正文启发式规则 b)明显比 a)重要度要高很多。所以对于不同的启发式规则,不能等同它们的重要度,必须有所区分。针对该问题,需要初始化一个权值矩阵。初始化权值矩阵算法(IWMA)描述如下:

a)赋予每个主题的每条规则一个经验权值 w。

b)获得所有的主题,针对每个主题生成该主题的权值向量 $\boldsymbol{W}_j=[w_{1j},w_{2j},w_{3j},\cdots,w_{nj}]^{\mathrm{T}}$。其中 $\sum_{i=1}^{n} w_{ij}=1$,j 为主题的序号,i 为启发式规则的序号,n 为主题 j 的启发式规则条数。

c)将各个主题的权值向量进行扩展,构造出一个 $n_1 \times n_2$ 矩阵 $\boldsymbol{A}$,其中 n_2 为主题的个数,n_1 为 $\max(d(\boldsymbol{W}_1),d(\boldsymbol{W}_2),d(\boldsymbol{W}_3),\cdots,d(\boldsymbol{W}_{n_2}))$,$d(\boldsymbol{X})$ 为向量 $\boldsymbol{X}$ 的维数。矩阵的元素满足如下表达式:

$$A_{ij}=\begin{cases} w & \text{存在第 } j \text{ 主题第 } i \text{ 条规则的权值 } w \\ 0 & \text{否则} \end{cases}$$

上述启发式规则定义中涉及到四个主题,分别为发布时间、来源、标题、正文。针对这些启发式规则使用 IWMA,得到上述启发式规则的权值矩阵 $\boldsymbol{A}$ 为

$$\boldsymbol{A}=\begin{bmatrix} w_{11} & w_{12} & w_{13} & w_{14} \\ w_{21} & w_{22} & w_{23} & w_{24} \\ 0 & 0 & w_{33} & w_{34} \\ 0 & 0 & w_{34} & w_{44} \end{bmatrix}$$

1.3　基于启发式规则的定位

基于启发式规则的定位算法(HRPA)描述如下:

a)对于一棵已经去掉重复内容被精简过的 DOM 树,按先序遍历获得 DOM 树中所有文本节点集合 ln_list,|ln_list| = n_3;初始化主题集合 topic_list,|topic_list| = n_2。运用 IWMA 获得权值矩阵 $\boldsymbol{A}$。

b)针对主题,获得该主题的可能度向量。可能度是用来衡量该节点能够成为该主题的可能性。可能度定义如下:

$$P_{kj}'=\begin{cases} P_{kj}+A_{ij} & \text{节点 } k \text{ 满足主题 } j \text{ 的规则 } i \\ P_{kj} & \text{否则} \end{cases}$$

其中:P_{kj} 为节点 k 成为主题 j 的可能度,初始时 $P_{kj}=0$。在启发式规则中,涉及到一些被依赖性主题,该类主题应先定位。例

如,标题启发式规则 d)依赖于发布时间,所以发布时间是被依赖性主题,应先定位。选定一个主题后,ln_list 依次通过该主题的启发式规则学习,得到所有文本节点该主题的可能度,即该主题可能度向量 $P_j=[P_{1j},P_{2j},P_{3j},\cdots,P_{n3j}]^T$。

c)根据可能度向量定位主题节点,不同主题对应的节点个数不同。例如,发布时间、来源、标题只可能对应一个节点,而正文则可以对应多个节点。针对此类情况本文分为两种情况讨论。首先获得最大可能度集合 $\max_j=\{P_{kj}\mid P_{kj}=\max(P_{1j},P_{2j},\cdots,P_{n3j})\}=\{P_{a1j},P_{a2j},\cdots,P_{atj}\}$。

如果主题节点个数 $nc>1$,则判定公式如下:

$$\xi_{kj}=\begin{cases}1 & \text{满足 } p_{kj}\in\max_j\\ 0 & \text{否则}\end{cases}$$

如果 $nc=1$,则判定式如下:

$$\xi_{kj}=\begin{cases}1 & \text{满足 } P_{kj}\in\max_j,\text{且 } k=\min(a_1,a_2,\cdots,a_t)\\ 0 & \text{否则}\end{cases}$$

其中:ξ_{kj}为第 k 个节点第 i 个主题的判定因子。如果 ξ_{kj}为 1,则判定第 k 个节点是第 j 个主题元素节点;如果 ξ_{kj}为 0,则判定第 k 个节点不是第 j 个主题元素节点。

d)判断 topic_list 中所有主题是否遍历完毕,是则转 e);否则转 b)。

e)定位各个主题节点后,获得各个主题节点在 DOM 树中的路径,作为该主题的抽取规则。

2　阈值选取及权值生成

2.1　长度阈值选取

启发式规则中涉及到一些没有确定的阈值,如标题长度阈值、正文长度阈值,长度阈值选取的好坏在一定程度上影响该条启发式规则的效用,以致影响整个主题的抽取效果。本文以标题长度阈值选取为例,给出整个阈值估计的过程。

标题启发式规则 d)以已经确定的发布时间为界限进行区分,标题位于发布时间以上的区域内,经过主题信息块的提取后,在这个区域内对标题抽取造成影响的噪声信息主要是当前位置、点击次数等。把标题区域单独提出讨论,假设一篇文档中只存在标题和影响它的噪声两部分。

通过统计的方法得到标题长度和噪声长度的概率密度函数。噪声长度的密度函数为 $p_1(x)=[1/(\sqrt{2\pi}\sigma_1)]\exp[-(x-\mu_1)^2/(2\sigma_1^2)]$;标题长度的密度函数为 $p_2(x)=[1/(\sqrt{2\pi}\sigma_2)]\exp[-(x-\mu_2)^2/(2\sigma_2^2)]$。其中 μ_1、μ_2 为均值,σ_1、σ_2 为均值的标准偏差。它们的密度函数的曲线图如图 1 所示。

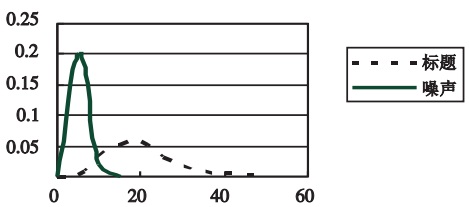

图1　标题和噪声长度概率密度函数图

从图 1 中可以看出,标题长度与噪声长度中间有一个重叠区,而且噪声的长度较短,而标题长度较长。在这种情况下 $\mu_1<\mu_2$,可以定义一个阈值 T,使得所有长度小于 T 的被认为是噪声,而长度大于 T 的为标题。此时,将标题误判为噪声的概率为 $E_1(T)=\int_{-\infty}^{T}p_2(x)\,dx$,将噪声误判为标题的概率为 $E_2(T)=\int_{T}^{+\infty}p_1(x)\,dx$。因此,总的误判概率为 $E(T)=P_2E_1(T)+P_1E_2(T)$。其中 P_1 为噪声信息的概率,P_2 为标题出现的概率,并且 P_1 和 P_2 满足限制条件 $P_1+P_2=1$。本文假设只存在标题信息和影响它的噪声信息两部分,所以 P_1、P_2 通过统计后可以得出。

为了找到一个阈值 T 使得上述的总误判率最小,将 $E(T)$ 对 T 求微分,并令其结果等于零,得到 $P_1p_1(T)=P_2p_2(T)$。将噪声长度和标题长度密度函数代入,取其自然对数,通过化简可以得到方程:$AT^2+BT+C=0$。其中:$A=\sigma_1^2-\sigma_2^2$;$B=2(\mu_1\sigma_2^2-\mu_2\sigma_1^2)$;$C=\mu_2^2\sigma_1^2-\mu_1^2\sigma_2^2+2\sigma_1^2\sigma_2^2\ln(\sigma_2P_1/\sigma_1P_2)$。解出 T,即为所求的标题长度阈值 λ_2。本文中计算得出 $\lambda_2=11$。

2.2　权值动态生成算法

上述 IWMA 中权值的选择根据经验人工制定,但固定的权值很难处理不同类型和不同风格的网页。对于某些网页它们符合标题启发式规则 b),则标题启发式规则 b)权值比较高,标题提取效果会比较好;对于另一些网页它们符合标题启发式规则 c),则标题启发式规则 c)权值比较高,标题提取效果会比较好。权值应根据网页特征自动进行调整,从而得到更好的信息提取效果。权值动态生成算法如下:

```
weight_produce( )
〈输入〉主题集合 topic_list, 其中|topic_list| = n2,由每个主题的启发式规则条数构成的向量 J=[j1,j2,…,jn2]。
〈输出〉权值矩阵 A。
begin
for i=1 to n2;
{c=1;
for w1i=0 to 1 /* k∈{1,2,…,ji},wki ∈{0,0.1,0.2,…,0.9,1} */
    for w2i=0 to 1-w1i
      ú
        for w(ji-1)i=0 to 1-Σ_{k=1}^{ji-2} wki
        {wji-1i=1-Σ_{k=1}^{ji-1} wki;
        投入测试网页抽取主题 i;
        S=EN/TN,其中 EN 为抽取成功页面数,TN 为进行测试的总的页面数;
        Vc=[w1i,w2i,…,wji-1i,S]^T;
        c=c+1;}
```

比较向量 V_c 中 S 的大小,取得最大的 S 值的向量,将该向量中的权值赋予权值矩阵 A 中对应的元素;}

```
end
```

3　实验结果及分析

实验包括两个部分:检验长度阈值估计方法的准确性;测试 HRPA 的效果。前者主要是使用标题启发式规则 a)抽取标题信息,验证该阈值估计方法的准确性;后者主要是将 HRPA 应用到抽取系统中,使用多条规则抽取多个主题,评测该算法的有效性。

为了考察算法的效果,避免单种网站风格带来的影响,本文选取了四种代表性网站的 6 293 个页面进行测试(基本类型及其数据如表 2 所示)。使用标题启发式规则 a)抽取标题信息,根据长度阈值 λ_2 取值变化,实验结果如表 1 所示。

表 1　λ_2对误判率的影响

λ_2	E_1	E_2	E
7	0.076	0.250 9	0.426 9
8	0.095 7	0.158 2	0.353 9
9	0.118 3	0.098 9	0.317 2
10	0.141 3	0.047 0	0.288 3
11	0.169 3	0.022 3	0.066 4
12	0.204 7	0.008 1	0.067 1
13	0.245 6	0.004 2	0.076 7
14	0.296 7	0.002 4	0.090 7

其中:E_1为标题误判为噪声信息的概率;E_2为噪声信息误判为标题的概率;E 为总的误判概率,$E = P_2E_1 + P_1E_2$。其中 $P_1 = 0.7$ 为噪声信息出现的概率, $P_2 = 0.3$ 为标题出现的概率。

通过表 1 所示的标题实验结果,可以得出 E_1与 E_2成反比,E_1随着 λ_2的增大而逐渐增大,而 E_2随着 λ_2的增大而逐渐减小。但由于噪声信息在页面中占的比例 P_1比标题占的比例 P_2大很多,导致 E 主要受 E_2的影响。虽然在 $T = 11$ 时,$E_1 = 0.169\ 3$,标题误判为噪声信息占到了一个比较大的比例,但是总的误判率是最小的,所以 $T = 11$ 是最佳的分割点。这与长度阈值估计方法计算出的 T 值是一致的,证明了该方法的准确性。

将 HRPA 应用到抽取系统中,抽取结果如表 2 所示。

表 2　抽取结果

网站类型	网页数目	tpc/%	tipc/%	cpc/%	spc/%	dc/%
大型新闻类门户网站	1829	91	95	90	95	89
政府类门户网站	1795	94	99	93	98	90
事业机关类门户网站	1457	92	97	89	96	86
学校部门网站	1212	95	100	94	97	92

其中:tpc 为标题抽取正确率;Tipc 为发布时间抽取正确率;cpc 为正文抽取正确率;spc 为来源抽取正确率;dc 为总体抽取正确率。

对各个主题的抽取结果的分析如下:

a)发布时间抽取正确率。tipc 相对较高,平均达到 97%。主要是发布时间特征比较明显,定位比较精确,但是并不是所有网站都达到了 100%。影响发布时间定位的因素主要有:当前时间、正文中的时间、相关链接中的时间三类。通过对出错页面分析,发现影响发布时间抽取的主要是当前时间、正文中的时间,相关链接中的时间由于其位置特性,并没有对抽取效果造成影响。

b)来源抽取正确率。Spc 平均达到了 96%。导致来源抽取错误的因素有两类:来源没有明显的标记,且位置未与发布时间相邻;发布时间的抽取错误,导致来源的抽取错误。

c)标题抽取正确率。Tpc 平均达到了 93%。导致标题抽取错误因素也有两类:当前位置等噪声信息对其造成的影响,这种情况在上面已经讨论了;发布时间抽取错误,导致标题的抽取错误。

d)正文抽取正确率。Cpc 最低,平均达到91.5%。导致内容抽取错误因素也有两类:部分噪声信息夹杂在段落文字中,也就是说存在〈p〉或〈br〉等认为是段落文字的标签中,这样导致误把噪声信息当做正文抽取出来,造成抽取过度;部分正文内容以超链接的形式出现,这部分信息没有被抽取出来,造成抽取不足。

e)总体抽取正确率。理论上总体抽取正确率的计算如下:设 WS 为某一网站的测试网页集,$|WS| = N$;W_i为第 i 个主题信息正确抽取的网页集,$|W_i| = X_i$。那么,第 i 个主题信息抽取的正确率 $P_i = |W_i|/|WS| = X_i/N$;总体抽取正确率 $P = |\bigcap_{i=1}^{4} W_i|/|WS|$。

设 $\min P = \min\{P_1, P_2, P_3, P_4\}$,因为 $|\bigcap_{i=1}^{4} W_i| \leqslant \min|W_i|$,$i = 1,2,3,4$,故 $P \leqslant \min P$。

又因为 $|\bigcap_{i=1}^{4} W_i| \geqslant \sum_{i=1}^{4}|WS - W_i|$,所以 $P \geqslant 1 - \sum_{i=1}^{4}(1 - P_i) \Rightarrow P \geqslant \sum_{i=1}^{4} P_i - 3$,所以 $P \in [\sum_{i=1}^{4} P_i - 3, \min P]$。当 P 越接近 $\min P$ 时,说明总体抽取效果越好。

从实验结果可以看出,大型新闻类门户网站、事业机关类门户网站的 dc 相对较低,政府类门户网站和学校部门网站的 dc 相对较高。通过对页面进行分析发现,大型新闻类门户网站夹杂的噪声信息比较多,而事业机关类门户网站不太规范,政府和学校的网站比较规范。同时,总体抽取效果也比较接近理论上的最好水平。

4　结束语

从目前的研究来看,学者们对网页信息抽取的研究大多集中在主题信息块的抽取。因为未见针对各个主题精确定位并提取的研究文献,所以无法进行针对性比较。但是,从各个主题的抽取效果以及总体抽取正确率来看,本文提出的基于启发式规则的网页主题信息精确定位方法能够有效地、准确地分离主题信息块,为网页信息抽取提供了一种有效的处理算法,并为 Web 挖掘中半结构化数据向结构化数据转换提供了方法。

有待进一步研究的内容是:a)本文算法中的阈值选取方法仍需进一步改进,以提高自适应程度和准确性;b)本文算法中制定的启发式规则并不适用于所有的主题信息的提取,还需进一步研究提高启发式规则的通用性。

参考文献:

[1] 欧健文,董守斌,蔡斌. 模板化网页主题信息的提取方法[J]. 清华大学学报:自然科学版,2005,45(S1):1743-1747.

[2] CAO Yu-juan, NIU Zhen-dong, DAI Liu-ling, *et al*. Extraction of informative blocks from Web pages[C]//Proc of International Conference on Advanced Language Processing and Web Information Technology. Washington DC:IEEE Computer Society,2008:544-549.

[3] WEMINGER T, WILLIAM H. Text extraction from the Web via text-to-tag ratio[C]//Proc of the 19th International Conference on Database and Expert Systems Application. Washington DC:IEEE Computer Society,2008:23-28.

[6] 常育红,姜哲,朱小燕. 基于标记树表示方法的页面结构分析[J]. 计算机工程与应用,2004,40(16):129-133.

[4] 王琦,唐世渭,杨冬青,等. 基于 DOM 的网页主题信息自动提取[J]. 计算机研究与发展,2004,41(10):1787-1792.

[5] 王磊,蒋建中,郭军利. 基于扩展 DOM 树的 Web 页面信息抽取[J]. 计算机应用与软件,2007,25(6):137-139.

[7] 时达明,林鸿飞,杨志豪. 基于网页框架和规则的网页噪音去除方法[J]. 计算机工程,2007,33(19):276-278.

[8] 石倩,陈荣,鲁明羽. 基于规则归纳的信息抽取系统实现[J]. 计算机工程与应用,2008,44(21):166-170.

[9] 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,2008,22(1):22-28.

[10] WANG Ji-ying, LOCHOVSKY F H. Data-rich section extraction from HTML pages[C]//Proc of the 3rd International Conference on Web Information Systems Engineering. Washington DC: IEEE Computer Society, 2002:313-322.