

一种基于可信度分析的 Web 页面新属性发现方法

胡国晴, 李建华

(中南大学 信息科学与工程学院, 湖南 长沙 410075)

摘要:分装器已经越来越多地应用到 Web 信息抽取中,但是当 Web 页面出现新的待抽取属性并且页面结构发生变化时,目前并没有一个完善的分装器能根据这种情况而做出相应调整从而抽取出新属性信息。文中根据待抽取属性自身结构和内容的特点,通过定义一系列规则和证据,提出了一种基于可信度分析发现 Web 页面新属性的方法,并建立了该方法的模型。通过在实际网站中选取网页对本方法进行了实验分析,取得了较好的效果,具有现实可行性。

关键词:可信度;分装器;信息抽取;新属性发现

中图分类号:TP39

文献标识码:A

文章编号:1673-629X(2009)01-0056-04

A Credibility Analysis - Based Method to Discover New Attributes Web Pages

HU Guo-qing, LI Jian-hua

(School of Information Science and Engineering, Central South University, Changsha 410075, China)

Abstract: Although wrapper is applied to Web information extraction much more, when new attributes appear and the structure of Web pages is changed, there is no perfect wrapper to extract the new attributes information corresponding. Based on the attribute's own structure and features of contents, through a set of rules and evidence, a credibility analysis-based method to discover the new attributes of Web pages is proposed in this paper, and established a model of this method. Adopted this method to actual page on the Web site, it plays good and has practical feasibility.

Key words: credibility; wrapper; information extraction; new attributes discover

0 引言

目前,分装器(Wrapper)方法已经广泛应用于 Web 信息抽取中,早期的分装器^[1]是针对某一特定的数据源单独编写的,这样开发的分装器灵活性比较低,一旦数据源的结构发生变化,已建立的分装器将失去抽取数据源数据的功能。

针对这一弱点,研究人员提出了很多的解决办法,如通过训练样本或者已抽取的数据,对样本或者已抽取的数据进行分类^[2~4],当出现新的结构的页面时,将新页面的信息和训练样本等进行比较,判断页面中元素是否需要抽取,进而归纳出新的规则,使一些分装器具有了自主学习的方法,这极大地减少了人工工作量。但是如果在新结构的页面出现的有用信息不能在训练样本等中找到分类,分装器将无法判断信息是否需要抽取,有用信息将被丢弃。

考虑下面情况,已建好的分装器中规则能正确抽取如下表单中公司名、地址、电话三个属于公司信息的属性。

```
<table><tr><td>公司名</td></tr><tr><td>地址</td></tr><tr><td>电话</td></tr></table>
```

当新出现的页面结构如下:

```
<table><tr><td>公司名</td></tr><tr><td>地址</td></tr><tr><td>电话</td></tr><tr><td>email</td></tr></table>
```

在新的页面结构下出现的属性 email 在训练样本等中并没有出现过,将不能准确找到它的分类,这样 email 属性的信息将不会被抽取,从而丢失了有用信息。为解决此类问题,文献[5]通过对待抽取属性分类,并考虑页面特征,提出了利用贝叶斯学习和极大似然估计方法解决此类问题,但是计算过程比较复杂。文中提出基于可信度的 Web 新属性发现方法,在简化了计算量的基础上,以待抽取属性是文本信息为例,获得了比较好的抽取结果。

收稿日期:2008-04-02

作者简介:胡国晴(1982-),男,湖南邵阳人,硕士研究生,研究方向为 Web 信息抽取、垂直搜索引擎;李建华,教授,研究方向为分布式计算、软件工程。

1 相关理论

可信度^[6](CF(H,E))是根据经验(E)对一个事物或者现象(H)为真的相信程度:

$$CF(H,E) = \begin{cases} \frac{p(H|E) - p(H)}{1 - p(H)}, & \text{当 } p(H|E) \neq p(H) \\ 0, & \text{当 } p(H|E) = p(H) \end{cases} \quad (-1 \leq CF(H,E) \leq 1) \quad (1)$$

其中 $P(H|E)$ 表示在证据 E 出现的情况下现象 H 出现的概率, $P(H)$ 表示现象 H 出现的概率。

2 Web 页面新属性发现方法的实现

在实际应用中, $P(H|E)$ 和 $P(H)$ 值通常难以获得, 因而 $CF(H,E)$ 的值一般由领域专家直接给出。文中提出以下证据和规则, 量化 $P(H|E)$ 和 $P(H)$, 从而可以求出 $CF(H,E)$ 。

2.1 证据定义

2.1.1 定义 1

Web 信息抽取是通过计算机自动地从大量的 Web 数据中抽取感兴趣的信息^[7], 新属性数据是否应当被抽取取决于用户是否对此数据感兴趣, 文中将新属性数据是否属于用户感兴趣信息的程度定义为证据预抽取属性内容与已抽取属性内容是否相关(C)。

2.1.2 定义 2

Web 页面把标记和数据信息按照 HTML 的定义组织在一起, 其中标记反映了页面的结构信息, 同时能通过这些标记来分析页面的布局, 把页面划分为多个区域, 所以新属性是否需要抽取和这些标记之间的关系相关, 文中将这些标记之间的关系对新属性的影响定义为证据预抽取属性信息上下文格式特征(F)。

2.1.3 定义 3

目前大部分 Web 页面都包含大量的无用信息, 如广告等, 所以待抽取属性数据是否需要抽取还和整个页面的内容相关, 文中将它们和新属性数据之间的关系定义为证据预抽取属性内容与整个页面属性内容是否相关(B)。

2.1.4 定义 4

当有多个证据 $E_1, E_2, E_3, \dots, E_n$ 同时出现时, 定义:

$$CF(H,E) = a_1 * CF(H,E_1) + \dots + a_n * CF(H,E_n) \quad (2)$$

其中 $a_1 + \dots + a_i + \dots + a_n = 1$, 且 $0 \leq a_i \leq 1$ 。

2.2 规则介绍

DOM(Document Object Model)树能够很好地反映 HTML 文档结构信息, 目前已有很多的研究人员将

DOM 树应用到 Web 页面信息抽取中, 文中采用开源代码 cobra 对输入的 Web 页面先生成 DOM 树。

2.2.1 规则 1

利用 DOM 树反映的 Web 页面结构信息, 对待抽取属性的结构信息进行分析, 并设定预抽取属性信息上下文格式特征分析规则如表 1 所示。

表 1 预抽取属性信息上下文格式特征分析规则

规 则	影响结构因素	规则值 R		
		相同	不完全相同	不相同
1	字体	1	0.5	0
2	背景	1	0.5	0
3	是否拥有相同的父节点	1	0.5	0
4	是否拥有相同的子节点结构	1	0.5	0
5	本身标记与已抽取属性之间的关系	1	0.5	0

根据上述规则, 设定 $P(H|F) = \frac{\sum R_i}{5}$, 其中 F 为证据预抽取属性信息上下文格式特征, H 为待抽取属性需要抽取现象, 结合公式(1)从而可以得出 $CF(H,F)$ 。

2.2.2 规则 2

根据用户感兴趣的信息, 确定用户感兴趣的信息范围 S, 并规定预抽取属性内容与已抽取属性内容是否相关规则如表 2 所示。

表 2 预抽取属性内容与已抽取属性内容是否相关规则

规 则	影响结构因素	P(H C)		
		属于 S	不完全属于 S	不属于 S
1	用户感兴趣的信息范围 S	1	0.5	0

其中 C 为证据预抽取属性内容与已抽取属性内容是否相关。根据上述规则, 结合公式(1)从而可以得出 $CF(H,C)$ 。

2.2.3 规则 3

对待抽取属性与整个页面主题内容进行分析, 确定待抽取属性数据是否与整个页面主题内容相关, 并规定预抽取属性内容与整个页面属性内容规则如表 3 所示。

表 3 预抽取属性内容与整个页面属性内容规则

规则	影响结构因素	P(H B)		
		相关	不完全相关	不相关
1	待抽取属性内容是否与整个页面主题内容相关	1	0.5	0

其中 B 为证据预抽取属性内容与整个页面属性内容是否相关。根据上述规则, 结合公式(1)从而可以得出 $CF(H,B)$ 。

由规则 1、2、3 和公式(2), 从而可以得出: $CF(H,E) = a_1 * CF(F|H) + a_2 * CF(H|C) + a_3 * CF(H|B)$, $CF(H,E)$ 为待抽取属性是否需要抽取的可信度。

2.3 Web 页面新属性发现模型

根据上面定义的证据和规则确定 Web 页面新属性发现模型如图 1 所示。

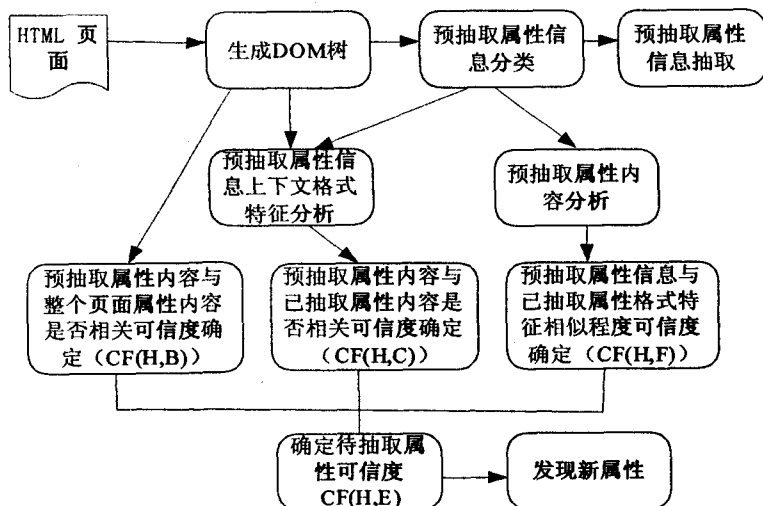


图 1 基于可信度的 Web 页面新属性发现模型

其中预抽取属性信息分类是根据训练样本或者已经抽取的数据信息与待抽取的属性数据进行比较,以确定待抽取属性的数据类别。目前已经有了很多的文本分类方法,主要分为两种方法:一种是基于训练集的文本分类方法;另一种是基于分类词表的文本分类方法。

文中采用基于文本特征向量相关性的方法^[8],它属于从训练集 M 中得出分类模式的方法,设: $M = \{M_1, M_2, \dots, M_i, \dots, M_n\}$, M_i 代表训练集中第 i 个分类的中心向量。 $M_i = \{m_{i1}, m_{i2}, \dots, m_{ik}, \dots, m_{in}\}$, m_{ik} 代表分类 i 中比较强说服力的词汇。

其具体算法如下:

(1) 将待抽取属性信息文本中对表达文本所属类别有比较强说服力的词汇从文本中抽取出来,形成向量 T_i , 其中 $T_i = \{t_1, t_2, \dots, t_k, \dots, t_n\}$ 。

(2) 利用向量之间夹角的余弦值计算新文本特征向量 T_i 和每类中心向量 M_j 间的相似度: $\text{Sim} = (M_j, T_i) = \frac{M_j \cdot T_i}{||M_j|| * ||T_i||} = \frac{\sum m_{jk} * t_{ik}}{\sqrt{\sum m_{jk}^2} * \sqrt{\sum t_{ik}^2}}$

(3) 令 $\text{Sim}(M_k, T_i) = \text{Max}\{\text{Sim}(M_j, T_i)\}$, 当 $\text{Sim}(M_k, T_i) > \text{Threshold0}$ (Threshold0 为用户自定义阈值) 时, 文本 T_i 属于训练集 M 中第 k 个分类, 否则文本 T_i 不属于训练集中任何分类。

如果待抽取属性属于训练集等数据中某个分类, 则抽取待抽取属性数据信息, 否则需要根据上面定义的规则和证据, 判断待抽取属性的可信度 $\text{CF}(H|E)$, 当 $\text{CF}(H|E) > \text{Threshold1}$ (Threshold1 为用户自定义阈值) 时, 抽取待抽取属性, 否则放弃抽取。

3 实验分析

为验证实验结果, 文中采用与文献[4]相似的实验方法, 并利用准确度和召回率作为衡量实验结果的指标, 抽取部分网站的诺基亚手机参数信息, 召回率和准确度定义如下:

$$\text{召回率} = \frac{\text{抽取的有效记录数}}{\text{抽取的全部记录数}}$$

$$\text{准确度} = \frac{\text{抽取的正确记录数}}{\text{抽取的有效记录数}}$$

表 4 描述了选取的网站, 第一列表示选取网站的标识符, 第二列表示选取的网站网址, 第三列表示对应网站选取的页面数, 第四列表示对应网站选取的记录数, 图 2 表示了部分抽取页面。

由于 Web 页面内容的主体部分在标记 $\langle \text{body} \rangle \langle / \text{body} \rangle$ 之间, 文中只选择 Web 页面 $\langle \text{body} \rangle \langle / \text{body} \rangle$ 标记之间内容进行分析; 选取的网站属于中文网站, 文中采用 CSW 中文分词组件以便于预抽取属性信息分类中中文文本的处理。

表 4 选取网站描述

标识符	网站网址	网页数	记录数
W1	http://www.pcpop.com/	4	9035
W2	http://www.21cn.com/	5	5500
W3	Phhttp://www.pchome.com/	5	5062
W4	http://www.pconline.com.cn/	5	1945

四个网站的手机基本参数信息都包含了手机名称、参考价格, 开始建立的分装器中规则只是抽取这两个属性, 采用文中的方法, 当 $a_1 = a_2 = a_3 = 1/3$, $\text{Threshold0} = P(H) = 0.5$, $\text{Threshold1} = 0$ 时, 实验结果如表 5 所示。

表 5 Web 页面新属性发现实验结果

标识符	新发现的属性	未发现的属性	发现错误的属性	准确度(P)	召回率(R)
W1	商家报价、手机制式、主屏尺寸、主屏分辨率、主屏颜色、手机通话时间、摄像头像素、产品尺寸、音乐播放	-	价格纠错、商家促销	93.75%	1.42%
W2	主屏尺寸、手机类型、和弦铃声、手机制式、操作系统、屏幕色彩	-	-	100%	2.82%
W3	参考价格、商家报价、产品天线、机身颜色、产品尺寸、手机重量、外观设计、理论通话时间	-	-	100%	3.06%
W4	商家报价、上市时间、网络制式、手机外形、主屏参数、数据业务、摄像头	-	报价	90.91%	5.66%

从表中可以看出, 对应的各个网站的召回率都比较低, 这是由于页面中真正需要抽取的属性比较少, 而每个页面选取的记录都太多; 在 pcpop 和 pconline 两个

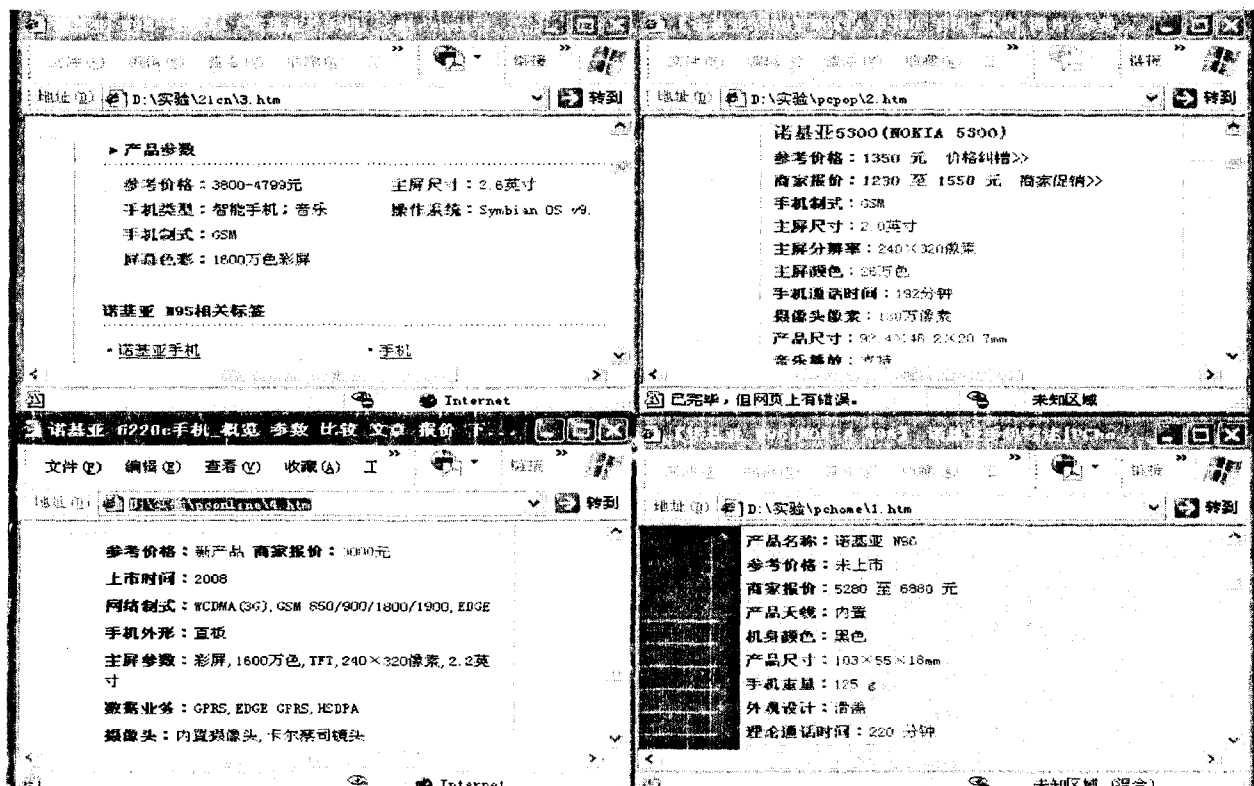


图2 各网站部分抽取页面

网站的页面中发现了错误的新属性,这是由于这些错误的属性结构与已发现属性的结构和内容都比较相似,但是它们的准确率都比较高,具有在实际应用中的可行性。

4 结束语

提出了在 Web 信息抽取中发现 Web 页面新属性的方法,并且在实际网站中选取网页对本方法进行了实验分析,取得了较好的效果。但是文中只是考虑了文本属性发现的情况,对其他如图片、视频等情况并没有分析,具有一定的局限性;为了方便新属性可信度的量化,规则的规定中具有较强的主观性,而这种主观性对实际情况的影响,文中并没有做过多分析,未来需要加强这两方面工作。

参考文献:

- [1] Laender A H F, Ribeiro Neto B A, da Silva A S, et al. A Brief Survey of Web Data Extraction Tools[J]. SIGMOD Record, 2002, 31(2): 84-93.
- [2] Zhai YanHong, Liu Bing. Web Data Extraction Based on Partial Tree Alignment[C]//International World Web Conference

Committee, International World Wide Web Conference 2005. Chiba, Japan: [s. n.], 2005.

- [3] Sahuguet A, Azavant F. Building intelligent web applications using lightweight wrappers[J]. International Journal of Data and Knowledge Engineering, 2001, 36(3): 283-316.
- [4] 陈琼, 苏文键. 基于网页结构树的 Web 信息抽取方法[J]. 计算机工程, 2005, 31(20): 54-55.
- [5] Wong Tak-Lam, Lam Wai. A Probabilistic Approach for Adapting Information Extraction Wrappers and Discovering New Attributes[C]//In: Proceedings of the Fourth IEEE International Conference on Data Mining, ICDM 2004. Washington, DC, USA: [s. n.], 2004.
- [6] 蔡自兴, 徐光佑. 人工智能及其应用[M]. 北京: 清华大学出版社, 2005.
- [7] Baumgartner R, Flesca S, Gottlob G. Visual Web Information Extraction with Lixto[C]//In: Proceedings of the 27th International Conference on Very Large Data Bases. San Francisco, CA, USA: [s. n.], 2001.
- [8] Joachims T. Text Categorization with Support Vector Machines: Learning with Many Relevant Features[C]//In: Proceedings of ECML-98, 10th European Conference on Machine Learning. Chemnitz, Germany: [s. n.], 1998.

(上接第 55 页)

- [6] Asleson R, Schutta N T. Foundations of Ajax[M]. 北京: 人民邮电出版社, 2006.

- [7] 祝青, 向南平. AJAX 技术在 WebGIS 中的应用与研究[J]. 测绘工程, 2007, 16(5): 39-41.