维普资讯 http://www.cqvip.com

一种基于分类算法的网页信息提取方法*)

汪建伟[1,2] 杨冬青[1] 高 军[1] 王腾蛟[1]
(北京大学信息科学技术学院 北京 100871)[1] (军事交通学院 天津 300161)[2]

摘 要 在目前的Web信息提取技术中,很多都是基于HTML结构的,由于HTML结构的经常变化,使提取模板需要经常更新,而提取模板的更新需要很多领域知识。本文提出一种基于分类算法的Web信息提取方法,通过将网页文本按照其显示属性的不同进行分组,以显示属性值为基础对Web页面文本进行分类,获取所关注文本,从而完成对Web页面的信息提取。这种提取方法操作简单,易于实现,对网页结构的依赖性小。

关键词 信息提取,属性向量,Wrapper,显示属性

A Method of Web Information Extraction Based on Classification Algorithm

WANG Jian-Wei[1,2] YANG Dong-Qing[1] GAO Jun[1] WANG Teng-Jiao[1]
(School of Electronics Engineering and Computer Science, Peking University, Beijing 100871)[1]
(Military Traffic College, CPLA, Tianjin 300161)[2]

Abstract In the research of Web information extraction, most of the existing algorithms are based on HTML structure. As the structure of HTML files changes frequently, wrapper must be updated accordingly. But the update of wrapper needs a lot of domain knowledge. In this paper, a new Web information extraction method based on classification algorithm is provided, which can group the Web text by HTML text display attributes. The information extraction of Web pages is finished by classifying the Web text with different values of the display attributes and acquiring desired text. This algorithm is easy to implementation and small-dependent of the HTML structure. Experiments prove its good performance.

Keywords Web information extraction, Attribute vector, Wrapper, Display attributes

1 引言

随着Internet的快速发展和网络信息量的增长,Web信息已经成为当前人们生活必不可少的信息源。为了在Web这个大的信息库中查找信息,我们通常使用搜索引擎来搜索信息,如Google[1],百度[2]等。搜索引擎返回用户指定关键字的网页信息。这种基于关键字的查询方式不足以反映用户的查询需求。

产生这种局限性的根源是Web环境中网页绝大多数是采用HTML这种半结构化语言编写的,不同网站的网页结构是异构的,应用程序难以采用传统的SQL查询语言对这种数据源进行精确查询和进一步应用。在这种情况下,为了更好地利用Web中的数据,我们需要将信息从无结构化、半结构化数据中提取出来,形成一个统一的模式,建立一个集成平台,通过这个平台,满足用户更精确的查询要求,例如查找"新华社发布的所有标题含物权法的新闻","找出价格低于10000元的戴尔笔记本"等。这就是Web信息提取和数据集成系统。其中,信息提取技术(Information Extraction)是Web信息提取和数据集成系统的基础。其工作就是从Web网页中提取出系统需要的数据,并将其赋予一定的数据模式,提供给数据集成系统。

经过众多研究者的大量研究,出现了很多信息提取的技术。按照采取的技术来区分,Web信息提取可分为以下几种[4]:基于HTML分析的提取方法、基于提取语言的方法、基于NLP的提取方法、Wrapper归纳提取方法、基于模型的提取方法、基于Ontology的方法等。

在实际的应用中,由于数据源经常发生变化,从而使提取方法失效,导致提取错误。此外为了更新Wrapper,需要大量的人工参与和相关的领域知识。这些都为提取技术的推广和应用带来了困难。

在当前的研究中,研究者已经开始关注根据Web页面的显示特点来对Web页面进行分析,但仅限于利用Web页面的显示特点对Web页面进行页面语义分块[5]和使用文本的显示属性来描述文本块的差异[6]。

本文从近年来研究者使用Web页面的显示属性来对Web页面进行分析的基础上出发,提出了一种基于分类算法的Web信息提取方法。这种提取方法按照Web页面中文本的显示属性的不同来对Web页面文本进行分类,通过机器学习的方法来提取网页中的数据。这种方法对网页的结构的结果依赖较小,操作简单,适应性强,易用性和健壮性得到了提高。

2 基于分类算法的网页信息提取方法

我们在浏览新闻或BBS网页的时候,发现同一网站的网页的布局可能发生变化,但显示风格却很少变化。在新闻和BBS网页的信息提取中,我们提取目标的数据模式是不变的。例如,我们希望从网页上提取的信息是新闻或BBS的标题、发布时间、作者、来源、正文、阅读数、评论数等信息。由于这

*)基金资助:国家242基金(课题编号:2005B22,2006B20)。**汪建伟** 硕士研究生,主要研究方向:信息系统与数据库。

维普资讯 http://www.cqvip.com

些条件，我们可以使用分类算法来提取网页中的信息。

2.1 算法综述

基于分类算法的 Web 信息提取方法分为两大步骤：

(1)根据样本网页文本的显示属性，对显示属性进行训练，生成分类器。

①从样本网页的 HTML 源文件中获得所有文本结点的显示属性；

②对显示属性相同的结点进行分组，并进行标注；

③利用分类算法对所分的组进行训练，生产分类器；

(2)对目标网页的文本使用分类器型进行分类，获得关注类别。

①从目标网页的 HTML 源文件中获得所有文本结点的显示属性；

②使用分类器，依据 DOM 文本结点的显示属性对文本进行分类；

③获取标注的类别。

结构框架如图 1 所示。

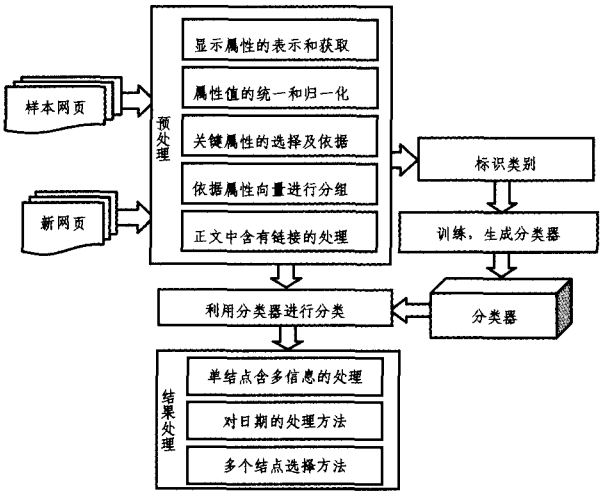

图 1 基于分类算法的 Web 信息提取方法的结构框架

按照结构框架，我们将整个流程分为三部分，第一部分是预处理，包括显示属性的表示和获取，属性值的统一、量化和归一化，以及对正文中包含链接文本的处理；第二部分是分类算法的选择；第三部分是结果处理，包括对单结点含多项信息的处理，以及多个类别的选择问题。

2.2 预处理

2.2.1 显示属性的表示

文本结点的显示属性可包括：字体大小、颜色、对齐方式、是否是链接等属性。将这些显示属性用向量来表示，记为属性向量。

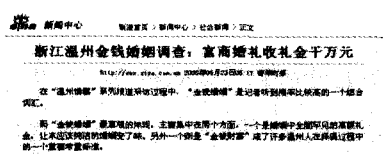
浙江温州金钱婚姻调查：富商婚礼收礼金千万元

图 2 显示属性的表示

如图 2 所示，使用属性向量将其网页文档中的内容表示如表 1 所示。

表 1 使用属性向量表示网页中的文本

结点	属性向量
浙江温州金钱婚姻调查…	(24px，＃05006C,center,0)
… 2005 年 06 月 23 日 05:17	(7px，＃FFFFFF,center,0)
青年时报	(7px，＃A20010,center,1)
在“温州情事”系列报道采访…	(14px,＃FFFFFF,center,0)
而“金钱婚姻”最直观的体现…	(14px,＃FFFFFF,center,0)
……	……

2.2.2 显示属性的获取

Web 文本结点的显示属性获得的准确与否直接决定了提取的质量。文本结点的显示属性在 HTML 中的表示方法通常有如下几种方法：通过属性名获取；通过 text/css 中定义的格式获取；通过结点内嵌的 style 获取；Tag 标签本身就是显示属性，如〈B〉、〈U〉、〈I〉、〈H1〉等。

2.2.3 属性值的统一、量化和归一化

由于 Web 页面中结点的属性值的表示具有多样性，因此我们需要对属性值的表示进行统一，使用统一的度量单位来对属性值进行处理。

数值的不同表示方式的统一，如：字体大小的像素值、逻辑值以等表示方式的统一。

枚举类型的转换和量化，如：对齐方式 left、center、right 等。

颜色值的处理：我们需要将颜色的表示方法转换成数值类型的值。

由于不同属性值的取值范围不同，有的属性值可能的取值范围较大，有的较小，为了平衡属性值取值范围对分类结果的影响，需要对属性值进行归一化处理，这样使不同属性值对分类结果的影响作用一致。

2.2.4 显示属性的选取

显示属性选取的是否合适是提取准确与否的关键。字体大小、字体模式、字体颜色、是否包含某个关键字等属性都是可供选取的重要属性。选择属性的原则就是使用最少的属性将所关注的文本结点同其他结点区分开来。使用的属性越少，提取的效率就越高。

2.2.5 对正文中含有链接的处理

在商业网页中，在文章的正文中经常包含有链接，正文中包含链接的结点由于显示属性的不同，在分组时会将含链接部分的文字从正文截断，使正文不完整。

为了解决这样的问题，我们扩展 STU-DOM 中主题相关度的计算方法[7]，通过判断当前结点所有子结点中含有链接的文本长度和其所有文本的长度的比值的大小来对含有链接的结点进行处理。

$$\text{Separate(STU)}=\frac{\text{LinkLength(STU)}}{\text{ContentLength(STU)}}$$

$$\text{LinkLength(STU)}=\sum_{i=1}^{N}\text{LinkLength(STU}_i\text{)}$$

$$\text{ContentLength(STU)}=\sum_{i=1}^{N}\text{ContentLength(STU}_i\text{)}$$

其中，STU_i 表示是 STU 的第 i 个子树。LinkLength(STU)是 STU 的链接文本的长度值，用其所有子树中的链接文本的字符数长度之和计算。ContentLength(STU)是 STU 的文本长度之和，用其所有子树中的文本的字符数之和计算。

设独立性阈值为 Ls，如果 Separate(STU)$<Ls$，则说明需要将当前结点 STU 的内容作为一个结点来处理。为了避免将非正文的结点当成一个结点来进行处理，我们还需要对

维普资讯 http://www.cqvip.com

ContentLength(STU)进行检查，如果 ContentLength(STU)的值小于某个数值 Cm，系统将不会将该结点看成一个结点来处理。

2.3 分类算法的选择

通过预处理，获得网页文本的特征值，然后通过分类算法对特征值进行训练和分类。分类算法比较多，如何选择分类算法也是需要考虑的问题，我们采用支持向量机、Rocchio、KNN 三种分类算法来进行测试，分析对提取结果的影响。

2.4 对分类结果的处理

对于目标网页，采用分类算法对 Web 页面中的 DOM 文本结点分类后，还需要对分类结果进行处理。

在实际的 HTML 文件中，有时一个文本结点可能会包含多个我们关注的元信息，我们可以根据不同目标的不同的特点，来提取相应的信息。

另外，对于特征值相似或接近的结点，总是会出现分类错误的情况。这样在同一目标模式的分类结果中可能包含两个甚至多个类别，为准确提取数据设置了困难。为了从多个类别中选择和目标模式最接近的类别，我们使用目标模式和待选类别的相似度来进行类别选择。

设 $L1, L2, \cdots, Ln$ 为 n 个待选类别，目标模式类别为 D。使用 $Sim(D,L)$来表示 D 类别和 L 类别的相似度。$Sim(D,L)$的计算使用下面的公式(1)：

$$Sim(D,L)=e^{-D(D,L)} \quad (1)$$

其中，$D(D,L)$表示 D 与 L 的欧式距离。$D(D,L)$值越大，说明 D 与 L 距离越远，$Sim(D,L)$值就越小，越不相似。

计算每一个待选类别 Li 与模板模式 D 的相似度，从 Li 中选择最相似的类别作为最终结果。

3 实验结果

本文分别采用了 SVM、Riccono、KNN 三种分类算法对提取效果进行了实验，其中 Xpath 表示的是采用 Xpath 方法提取网页数据的结果。

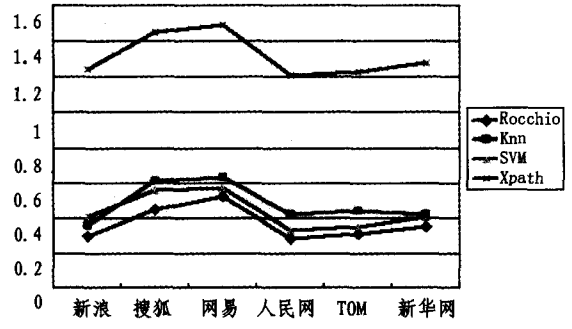

图 3 提取效率

实验测试对象包括了当前主流的门户网站，经过大量的实验发现本方法对于新闻和 BBS 领域的信息提取有着很好的效果，提取正确率能够达到 97%以上，提取一个网页的时间大概在 0.3 到 0.6 秒之间，较 XPath 的方法有更好的效率。从实验结果可以看出，对于网站显示风格比较确定的网站，提取效率较高，提取错误的原因主要是因为某些网站的个别网页显示风格与整体风格不同。

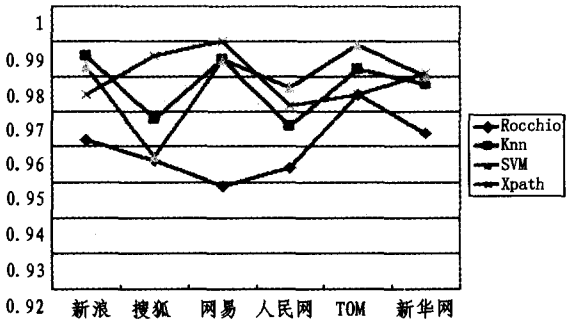

图 4 提取准确率

经大量实验分析发现，独立性阈值对于含链接的正文结点的处理有较大影响。在实验的过程中，独立性阈值的大小设为 0.3，STU 结点 ContentLength 的长度 Cm 设为 70 比较合适。经过多次实验评估，效果较好。

结论和进一步研究 本文提出的基于分类算法的网页信息提取方法能够根据网页文本的显示风格的不同，通过机器学习的方法，利用训练模型能准确地提取出 HTML 文档的信息，并且对网页结构的依赖性较小，该方法易于实现、操作简单，同时本方法的效率较高。

进一步的研究希望针对不同的目标模式选择不同的显示属性集，针对不同的目标模式生成不同的训练集，从而获得不同目标模式的特点，建立统一的模型，来解决异构网站的统一提取问题，从而达到不需要针对某个网站进行定制相应的 Wrapper 就可以完成信息的提取，提高提取的简单化和自动化。

参 考 文 献

1 www. google. com

2 www. baidu. com

3 Chang C H, Kayed M, Girgis M R, Shaalan K. A Survey of Web Information Extraction Systems. IEEE Transactions on Knowledge and Data Engineering, 2006. 1411～1428

4 Laender A H F, et al. A Brief Survey of Web Data Extraction Tools. [J]. ACM SIGMOD Record, 2002, 31(2)

5 DENG Cai, YU Shipeng, WEN Jirong, et al. VIPS: A Vision-Based Page Segmentation Algorithm [R]: [Microsoft Technical Report, MSR-TR-2003-79]. 2003

6 Zhao Hongkun, Meng Weiyi, Yu C. Automatic Extraction of Dynamic Record Sections From Search Engine Result Pages VLDB 2006 Seoul, Korea

7 王琦，唐世渭，杨冬青，王腾蛟. 基于 DOM 的网页主题信息自动提取. 计算机研究与发展，2004，41(10)：1786～1792