基于机器学习的网页正文提取方法

安增文，王 超，徐杰锋

（中国石油大学(华东) 计算机与通信工程学院 计算机科学与技术系，山东 东营 257000）

摘 要：先将网页转换为规范的 DOM 树，然后计算每行文本的文本密度、与标题相关度等值，并将其作为输入参数利用 BP 神经网络进行训练，进而形成抽取规则，最后通过实验验证该方法的可行性。

关键词：信息提取；神经网络；统计学习

中图分类号：TP391 **文献标识码**：A **文章编号**：1674-7720(2010)12-0004-03

An approach based on machine learning for information extraction method

AN Zeng Wen, WANG Chao, XU Jie Feng

(Computer Science and Technology Department, College of Computer & Communication Engineering, China University of Petroleum, Dongying 257000, China)

Abstract: We firstly translate the HTML to a DOM tree, and then compute the text density, the correlation between the words and the title of each line, and train with them by BP neural network, then we get the extract rules. At last we test the feasibility of this method.

Key words: information extraction; neural network; statistical learning

随着互联网的普及，网络成为人们获取信息的重要途径。而互联网上的信息量也与日俱增，网页上的内容除了主题内容外，通常都会在页面中放置导航条以方便用户访问，还有如广告、版权信息、欢迎信息等与主题无关的内容，我们称之为"噪音"。怎样去除这些噪音，将网页中的正文内容提取出来，从而提高人们的阅读效率，这在垂直搜索和数据挖掘方面具有重要意义。在这个领域已经发表了很多的研究成果，这些研究成果从不同的角度入手，有的只利用网页本身的特征，有的还与其他技术相结合，使网页正文抽取的准确性和完整性得到不断提高，但还没有一种方法能达到人们期望的程度，还需要不断地研究和探索。

1 正文抽取相关研究

到目前为止，已经发表的网页正文内容抽取方法有很多种，其分类方式的依据也不尽相同，下面介绍几种较为常用的抽取方法。

(1) 基于模板的方法

这种技术依赖 HTML 文档的内部结构特征来完成数据抽取，需要使用 wrapper(包装器)来抽取网页中的正文内容。包装器可以通过分析网页源代码来手工编写，也可以通过程序自动或半自动的实现。手工编写的方法一般都针对特定的网页模式，其优点是实现简单、准确率高，缺点是对于不同的网页模式或网页结构发生变化时需要重新编写包装器，如果包装器类型很多，包装器的维护代价会很大，但由于该方法的准确性较高，所以在针对特定网站的抽取中应用很广。自动或半自动地生成包装器的方法在一定程度上减轻了维护包装器的工作量，但是需要样本学习，对用户要求较高。

(2) 基于统计的方法

这种方法从页面的不同角度分析它的统计特征，采用统计学的算法抽取正文。例如根据统计的文字数量、链接数量、标签字符数量等计算出文本密度、链接密度等，并通过这些值来判断哪些为正文文本、哪些为噪音内容。参考文献[1]提出一种通过分析页面文本密度进行正文抽取的方法。这种方法实现简单，并且不需要编写包装器，但提取的准确率有限，有时会将与正文无关的版权声明等当作正文内容提取出来。

(3) 基于神经网络的方法

由于神经网络具有优越的非线性处理能力和泛化能力，因此在很多实际领域中都取得了传统符号学习机制难以获得的效果。文献[2]搜索结点的输入连接权，通

过找出权值之和超过阈值的连接权子集来抽取规则。参考文献[3]利用多层网络度量输入之间的接近程度，并利用单层抑制性网络度量输入、输出相关度，从而获得抽取规则。

参考文献[4]针对新闻类网页及类似布局的页面，在对页面文本密度进行统计之后对文本密度与页面标题、正文之间的对应关系进行分析，以对传网络(CPN)为工具，对文本密度在标题、正文等语义块中的分布模式进行拟合，从而达到抽取目标信息的目的。

参考文献[5]中以行为单位对网页源代码中的每一行计算其相关的六个属性，并以此作为BP神经网络的输入参数进行学习。由于该算法未对文本内容和标题的相关度进行判断，所以导致会将一些网站的版权声明当作正文内容错误地提取出来。所以通过计算文本内容和标题的相关度来区别是否为噪音是合理的。本方法以行为单位对DOM树进行处理，将每行的文本密度、文本内容与标题的相关度作为输入参数利用BP神经网络进行训练，从而提高信息抽取的准确度。

2 算法描述

2.1 BP神经网络模型

BP算法属于Delta学习规则，是一种有教师的学习算法，是以网络误差平方和为目标函数，按梯度法(gradient approaches)求其目标函数达到最小值的算法。一个典型的BP神经网络包括：(1)由一个输入层x、一个(或多个)隐藏层y和一个输出层o组成的三层或多层结构；(2)处理单元(图1中的圆圈)是网络的基本组成部分，输入层的处理单元只是将输入值转入相邻的联接权重，隐层和输出层的处理单元将它们的输入值求和并根据传递函数计算输出值；(3)联接权重(如图1中v,w)将神经网络中的处理单元联系起来，其值随各处理单元的连接程度而变化；(4)阈值，其值可为恒值或可变值，它可使网络能更自由地获取所要描述的函数关系；(5)传递函数F，它是将输入的数据转化为输出的处理单元，通常为非线性函数。

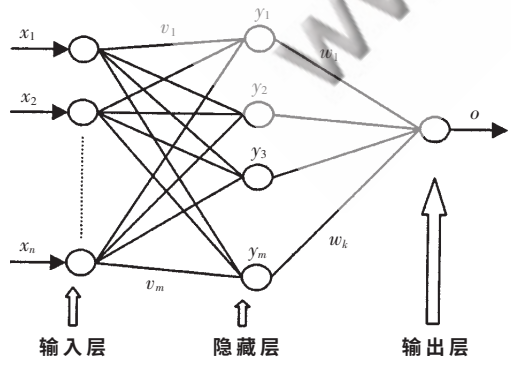

图1 BP神经网络结构图

输入层和输出层的结点个数可以根据训练集来确定，而隐藏层的结点却需要试验判断。如果隐藏层结点数过少，网络就不能具有必要的学习能力和信息处理能力。如果隐藏层结点数过多，不仅会大大增加网络结构的复杂性，网络在学习过程中更易陷入局部极小值，而且会使网络的学习速度变得很慢。

2.2 利用人工神经网络进行正文提取

网页的类型大体上可以分为三类：(1)文字多图片少的内容型网页，如新闻网页；(2)以图片为主文字介绍为辅的图片型网页，如图片新闻；(3)以超链接为主的目录型网页，如新浪首页。试验中我们以内容型网页作为主要研究对象。

2.2.1 网页源文件预处理

随着web2.0的发展，网站为了定制网页的表现形式和提高网页视觉效果，在源文件中加入大量Script脚本和CSS代码。所以在抽取正文之前要对网页源文件进行预处理，去除与正文内容不相关的噪音内容。

首先，由于html语言书写的随意性，导致有些网页源代码的不规范，例如标签对缺失、嵌套不准确等。所以要将缺失的html标签补齐、修改不正确的嵌套关系，并将源代码转换为DOM树的形式。本文采用HTML Tidy工具来处理网页。

其次，要判断网页源文件的编码，否则有可能抽取到乱码。以源文件头中的meta里声明的charset为准，对于编码为GBK、gb2312等格式的网页，都将其转为utf8格式。

最后，Script标签对之间和CSS内容都与正文内容无关，要全部删除。另外，对于<a>等无用的空标签对也一并删除。

2.2.2 神经网络训练过程

(1)页面主题的提取。<title>中的内容一般为文章标题，但现在各大网站一般采用“文章标题+网站名”的形式放在<title>标签中，且用符号“-”或“_”连接。在此将<title>中的文字内容取出，并将“-”或“_”符号后面的文字删除；若有多个这种符号，则将最后一个这种符号后面的文字内容删除，剩下的文字内容作为文章标题。因为标题中的文字内容一般会在正文内容中出现，而非正文内容一般不会包含标题词，所以可以将文本内容与文章标题的相关度作为判断文本是否正文的一个因子。

(2)统计各项值。以行为单位对DOM树进行处理，依次统计每行的文本长度y和字符总长度z，用p表示该段的文本密度，则$p=y/z$，该行的文本内容为c。

(3)计算相关度。分别对文章标题t和每行取出的文本内容c进行分词，得到对应的标题词项$(t_1,t_2\cdots t_m)$和文本词项$(c_1,c_2\cdots c_n)$，然后将每个标题词项t_i和文本词项c_j进行匹配，统计匹配次数并进行加权计算，得出其相关度，记相关度为s。为了提高相关度的准确性，本文借鉴搜索引擎中“倒排索引”的经验，对“的”“是”等停止词放在词库中进行分词，但不对其进行相关度计算。

表 1 利用训练结果进行测试的结果

网页来源	总数	正确个数	错误个数	查准率/P	完整个数	不完整个数	查全率/R	F 值/%
www.xinhuanet.com	40	36	4	95	32	4	88.9	89.4
www.163.com	40	37	3	92.5	33	4	89.2	90.8
www.sina.com.cn	50	43	7	86	40	3	93	89.4
www.sohu.com	50	45	5	90	42	3	93.3	91.6

采用 BP 神经网络作为训练模型，各层的激励函数均为 logsig，目标误差设为 0.05，学习率为 0.2。该模型有 12 个输入结点、5 个隐藏层结点和一个输出结点。其中 12 个输入参数为：每行的文本长度、每行的字符总长度、每行的文本密度、每行文本内容与标题的相关度、上一行的这四个值和下一行的这四个值。具体步骤如下：

(1)获取训练集并做好标记。

(2)对网页源文件进行预处理，生成相应的 DOM 树。

(3)从 DOM 树中读取一行文字，统计相应值，得出输入向量和期望输出。

(4) 输入向量经过隐藏层结点和输出层结点的传递函数得到实际输出。

(5)计算实际输出向量和期望输出向量的误差，并计算各输出误差项和隐藏层结点误差项。如果误差在允许范围内，则回到步骤(3)，从 DOM 树中读取下一行文字继续进行。如果误差不在允许误差范围内，则根据计算出的误差项计算出各权重的调整量，并调整权重。

(6)返回步骤(4)，继续迭代，直到实际输出向量和期望输出向量的误差满足要求。返回步骤(3)读取下一行内容，继续进行学习。

(7)标签树遍历完毕，训练结束。

将 DOM 树的各个元素偶对的相关值作为神经网络的输入，样本标记结果作为输出，通过学习算法自动生成抽取规则，对新的页面应用抽取规则进行测试。

3 测试结果

采用信息抽取技术中常用的查全率(R)、查准率(P)和 F 值三个评价指标对测试结果进行评价。查全率表示被正确抽取的信息的比例、查准率表示提取出来的正确信息的比率、F 值是查全率和查准率的加权几何平均值。用公式表示如下：P=(正确抽取出正文内容的网页数/总网页数)*100%，R=(抽取出完整正文内容的网页数/正确抽取出正文内容的网页数)*100%，在此将查全率和查准率看的同等重要，得出 $F=2PR/(P+R)$。根据 F 值与 1 的靠近程度来判断算法的好坏，越靠近 1 算法越好。

从几大新闻网站随机抽取 20 个网页进行人工分析和标记，按照以上方法进行训练。为了测试抽取方法的可行性，再抽取一定量的网页作为测试集，并利用训练结果进行测试。测试结果如表 1 所示。

在本文中通过统计 DOM 树每一行的文本长度和字符长度，进而计算文本密度以及文字内容与标题的相关度，并将这些数值作为输入参数输入到人工神经网络进行训练。通过计算内容和标题的相关度可以避免将一些标签字符较少、文字内容较多的版权声明等内容提取出来，进而提高正文抽取的准确度。从测试结果看，该方法具有一定的可行性。下一步要寻求更好的相关度计算方法，更准确地计算正文和标题的相关度，进一步提高正文抽取的准确性。

参考文献

[1] ALEXJC. The easy way to extract useful text from arbitrary HTML [EB/OL]. http://ai-depot.com/articles/ the-easy-way-to-extract-useful-text-from-arbitrary-html/. April5, 2007.

[2] FU L M. Rule learning by searching on adapted nets. Proceedings of the 9th National Conference on Artificial Intelligence. Anaheim, CA: AAAI Press, 1991:590-595.

[3] SESTITO S, DILLON T. Knowledge acquisition of conjunctive rules using multilayered neural networks. International Journal of Intelligent Systems,1993, 8(7):779-805.

[4] 陈敬文,彭哲. 基于 CPN 网络的 Web 正文抽取技术研究[J]. 现代图书情报技术,2008(11):65-71.

[5] 游贵荣,陆玉昌. 基于统计和机器学习的中文 Web 网页正文内容抽取[J]. 福建商业高等专科学校学报,2009,4(2):68-72.

[6] 楼顺天,施阳. 基于 MATLAB 的系统分析与设计-神经网络[M]. 西安电子科技大学出版社, 1998.

[7] 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, 2004,18(5):17-22.

(收稿日期：2009-12-01)

作者简介：

安增文，男，1983 年生，硕士研究生，主要研究方向：Web 信息挖掘。

王超，男，1985 年生，硕士研究生，主要研究方向：数据挖掘。

徐杰锋，男，1964 年生，教授，博士，主要研究方向：Web 信息挖掘、数据库应用。