文章编号: 1003 - 0077(2006) 06 - 0001 - 09

基于双层决策的新闻网页正文精确抽取①

胡国平, 张 巍, 王仁华

(中国科学技术大学 电子工程与信息科学系 讯飞语音实验室, 安徽 合肥 230027)

摘要: 本文提出了基于双层决策的新闻网页正文的精确抽取算法, 双层决策是指对新闻网页正文所在区域的全局范围决策和对正文范围内每段文字是否确是正文的局部内容决策。首先根据实际应用的需要给出了新闻网页正文的严格界定, 然后分析了新闻网页及其正文的特性, 提出了基于双层决策的正文抽取策略, 基于特征向量提取和决策树学习算法对上述双层决策进行了建模, 并在国内 10 个主要新闻网站的 1687 个新闻页面上开展了模型训练和测试实验。实验结果表明, 上述基于双层决策的方法能够精确地抽取出新闻网页的正文, 最终正文抽取与人工标注不完全一致的网页比例仅为 18. 14%, 比单纯局部正文内容决策的方法相对下降了 29. 85%, 同时抽取误差率大于 10%的网页比例更是仅为 7. 11%, 满足了实际应用的需要。

关键词: 计算机应用; 中文信息处理; 信息抽取; 特征向量; 决策树; 正文抽取

中图分类号: TP391 文献标识码: A

Precise Content Extraction from News Web Page Based on Decisions of Two Layers

HU Guo-ping ZHANG Wei WANG Ren-hua

(iFly Speech Lab, University of Science and Technology of China, Hefei, Anhui 230027, China)

Abstract: This paper concerns content extraction from news web pages based on decisions of two layers. The first layer of decision is introduced to predict the scope of content in a webpage, and the second layer is employed to judge whether the paragraph within predicted scope is content or not. We firstly present a strict definition of content for web pages orienting to the practical applications, then analyze the characteristics of news web pages and their contents. Based on the analysis, we propose a content extraction method based on decisions of two layers, and carry out experiments on a corpus of 1867 HTMLs collected from 10 main news web sites in China. The experiment results show that our method can predict the content of news web pages quite well: the percentage of web pages which contain mismatching in extracted content is only 18. 14%, which decreases 29. 85% compared to that just based on the second layer prediction, and only 7. 11% of extracted pages are with more than 10% mismatching, indicating that this method could be applied to practical applications.

Key words: computer application; Chinese information processing; information extraction; feature vector; decision tree; content extraction

1 引言

互联网的飞速发展给自然语言的研究带来了新的机遇和挑战, 互联网上 HTML网页的处

① 收稿日期: 2005-10-09 定稿日期: 2006-06-23
基金项目: 国家自然科学基金资助项目(69975018)
作者简介: 胡国平(1977—), 男, 博士研究生, 主要研究方向为语音合成系统、人机对话系统和信息抽取.

理技术得到了蓬勃的发展，由此产生的网页检索、网页分类、信息抽取、新词预测和新闻摘要等应用也得到了极大的关注。其中不少研究和应用都涉及到了网页页面分析这一基础技术的研究。

互联网上的每个网页大致由以下几类信息组成：导航信息、网页标题、网页正文、广告信息、版权信息和相关链接等。不同的应用关注不同部分的网页信息，如网页收集和检索往往需要用到导航信息、相关链接等信息，而对于新闻网页的自动分类、信息抽取、自动摘要、语音访问互联网（用语音合成技术将网页正文合成为对应的语音播放给用户的技术）而言，网页的标题和正文则是主要关心的信息。进一步，对于不同的应用，网页正文的界定和对正文自动抽取的精度要求也不完全一致。例如对于网页分类、信息抽取来说，正文的界定可以比较模糊而且自动抽取精度对最终效果的影响也较小；而对于自动摘要和语音访问互联网等应用，则对网页正文进行了严格的界定并且对正文自动抽取技术提出了很高的精度要求。本文就是面向自动摘要和语音访问互联网的应用，研究如何精确地从新闻网页中抽取严格界定的正文信息。

这里我们给出新闻网页正文的严格界定：新闻网页的正文指网页中新闻报道具体内容的全部文字，而相关专题、相关报道、版权说明、广告文字、插图说明等都不属于正文。如图 1所示的新闻网页，其正文是“据俄塔社 8月 21日报道，……（编辑：闻问）”共三个段落的文字。

图 1 新闻网页示例

网页正文信息的自动抽取方面，国内外开展较多的是界定比较模糊的正文的自动抽取的研究，而采用的方法也相对简单，如基于统计的正文信息抽取方法等。严格界定网页正文并开展精确抽取研究工作的，在我们的调研中还未发现。

人们可以很容易判断出新闻网页的正文，即使该网页是由不认识的文字构成的。这是因为新闻网页的正文在布局上有着许多的共性：如占据网页中心位置、少有链接文字、比较大段的统一格式的文字、正文的不同段落往往依次排列等等。本文正是基于新闻网页正文的上述共性来决策某段文字是否为正文的。同时观察实验员手工标注网页正文时我们发现，人们往往是首先定位出正文的范围，然后在这个范围内，细致地标注具体每个段落是否属于正文。基于这种观察，同时考虑到引入机器学习算法时可以更方便的分别提取正文全局范围特征和正文局部内容特征，我们提出了基于双层决策的新闻网页正文抽取策略：首先决策出正文的范围，然后在该范围内决策每个段落的文字是否真正属于正文。

对于一个新闻网页，双层决策的抽取方法首先经过预处理，将新闻网页按 HTML脚本中各标记的层次关系构建成树状结构，然后合并处于同一段落的叶子节点。其次提取树结构中的每个节点的特征向量，用于决策该节点是正文统领节点（包含了所有正文的最低层次的节点）的概率；同时对每个段落提取另一组特征，用于判断其作为正文的可能性。本文引入了决策树这一机器学习算法，对上述两组特征与其结论属性之间的映射关系进行建模。在 10个新闻网站的 1687个页面上的实验表明，上述双层决策的策略能够很好的匹配新闻网页的特性，

最终正文抽取与人工标注不完全一致的网页比例仅为 18.14%,比单纯局部正文内容决策的方法相对下降了 29.85%,同时抽取误差率大于 10%的网页比例更是仅为 7.11%,满足了实际应用的需要。

本文其余各章节内容安排如下:第二节给出了相关的研究情况,第三节是对新闻网页正文信息抽取的问题定义,基于双层决策的新闻网页正文信息抽取的算法将在第四节中详细的说明,第五节是实验方法和实验结果,最后是我们的结论。

2 相关研究

网页信息的自动抽取按所抽取的内容可以分为两大类:一类是对网页中包含的特定知识(领域知识或世界知识)的抽取,如商品信息[1,2]、会议论文信息[3]、新词[4]、术语定义[5]等;另一类是对网页中特定属性的抽取,如网页标题[6]、正文[7]、作者、更新时间等。本文对新闻网页中正文属性开展自动抽取研究,属于后一类。

对网页中所包含的特定知识的抽取,就是将数据从缺乏结构约束的网页中提取出来转换为结构的数据,这一过程称为包装。包装一般是根据一定的信息模式识别规则从网页中抽取相关内容,其关键问题是如何获得包装器所需规则。在这一问题上已经提出了人工书写规则[8]、半自动化规则获取方法[9]和完全自动化的包装器自动生成工具[10]等解决方案。

对于网页中特定属性的抽取,包括网页标题[6]、正文[7]等近年来也被广泛研究。虽然理论上可以采用包装的方法对网页属性进行抽取,但由于网页结构的复杂和不规范性,一个包装器的实现一般只能针对同一网站来源的网页,难以满足面向不同来源的网页信息抽取任务的需求,所以对网页特定属性的抽取一般采用基于统计和机器学习的方法。

文献[6]提出了基于特征向量提取和非对称感知器权值训练算法的面向任意网页的标题自动抽取算法,该论文与本文研究工作的相同之处在于都引入了机器学习的思想来处理网页中信息分类、鉴别的问题,而不同之处在于文献[6]专注于标题的抽取,而本文则关注网页正文的抽取工作。标题一般仅涉及到一个段落,而正文则往往因包含不定个数的段落而增加了抽取的难度。另外,他们针对的是任意网页训练标题抽取算法,而本文针对新闻网页,降低了问题难度的同时却增加了如何针对新闻网页提高识别效果的问题。

文献[7]中提出了基于统计的新闻网页正文信息抽取方法,该论文采用了统计的方法并利用新闻类网页的正文大都在一个<TABLE>的 HTML标记所辖范围内这一特性进行正文的抽取。此论文抽取正文是为面向旅游领域的问答系统提供语料加工服务,对正文抽取的准确率要求不是很高,如多取了一个段落或少取一个段落影响不大。而本文抽取正文是为了将新闻网页的正文送入语音合成引擎进行朗读以及进行网络新闻自动摘要实验,对正文有着严格的界定并且对正文自动抽取的精度要求比较高,因此引发了本文的研究工作:基于双层决策的新闻网页正文的精确抽取研究。

在我们的调研范围内,还没有发现有研究机构对新闻网页的正文进行严格定义并开展精确抽取的研究工作。

3 问题描述

本文针对的问题是如何从新闻类型的网页中高精度地抽取严格界定的正文信息。该问题的难点包括:

· 新闻网页正文的严格界定:面向语音访问互联网和自动摘要网页新闻应用的需要,我

们严格界定网页的正文部分。因为界定严格，使得更多的局部段落需要去分析和决策，加大了决策的难度；

· 抽取模型需要能够处理不同新闻网站的各个时期的新闻网页，因此不能依赖分析某个新闻网站某个时刻的新闻编排格式采用包装器的方式来完成抽取工作，必须从新闻网页正文的普遍规律形成抽取算法；

· 人们定位新闻网页往往依据其是否排版在中间重要的位置，但是考虑到效率，难以得到每段文字在屏幕上的显示位置和所占用的面积等重要决策依据信息，而目前的语义分析技术水平更是难以实现从语义上分辨一个段落是否是正文，故只能依赖 HTML标记、段落的长度、字体、颜色、链接文字比例以及在 HTML 节点脚本中的位置等有用但非决定性的信息；

· 正文的界定中把转载说明、版权说明、相关新闻等排除在正文之外，而这些非正文部分与正文没有固定的区分标记，因此很容易与正文产生干扰；

· 正文长短不一：因为无法计算某段文字所占用的显示面积和显示位置，所以计算文字的重要程度时大都依赖非链接文字的长度，而有些页面因为正文太短而难以检出；

· 正文范围内的一些非正文的内容：如画中画、内嵌广告、插图说明文字等等，按严格的正文界定，这些文字也需要决策为非正文，增加了正文抽取的难度。

4 基于决策树的分阶段的新闻网页正文抽取

本文提出的基于双层决策的新闻网页正文的抽取方法，其思想来源于人们定位正文的习惯：人们往往是先大致判断出正文的范围，然后再细致的判断范围内每段文字是否是正文。因此，我们也采取双层决策策略：先决策出全局的正文范围，再局部决策每个段落是否是正文。同时，从特征提取角度出发，这种分层处理也非常有助于从全局和局部上分别提取不同的正文判决属性，使得最终判断一个段落是否为正文时既考虑了该段落本身的特性，又考虑到了该段落所在的上下文环境，有助于正文抽取效果的提高。

流程上基于双层决策的新闻网页正文抽取算法分为网页预处理、全局范围决策和局部内容决策三个步骤，如图 2所示。下面将对这三个步骤一一进行详细的介绍。

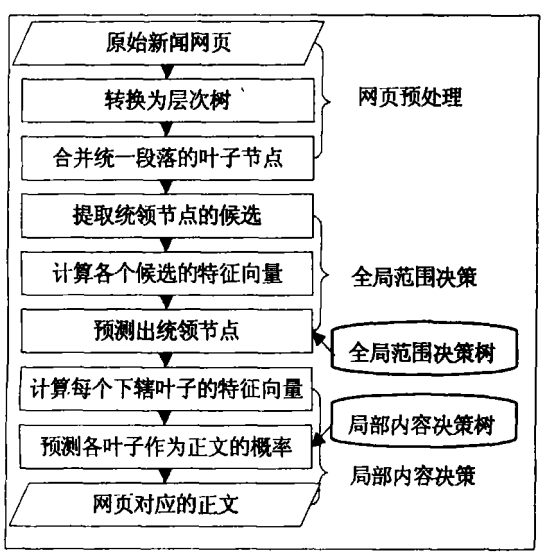

图 2 基于双层决策的新闻网页正文抽取算法流程

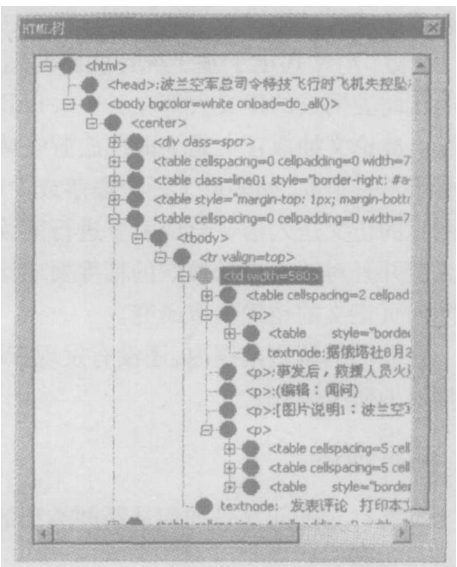

图 3 图 1所示网页对应的 HTML树

4.1 网页预处理

预处理的目标是将新闻网页的 HTML脚本转化为只包含有意义信息的树状结构。具体步骤如下:

首先,删除新闻网页 HTML脚本中的与正文抽取不相关的信息,如注释、Java代码等;

其次,建立树状结构。网页的 HTML脚本本身是嵌套的,因此我们可以比较容易的根据 HTML的脚本建立成对应的树状结构(类似于 DOM 树),并将所有的文字信息挂接在叶子节点,同时把 HTML的转义字符都变换为正常的字符,如"<"变换为"<","&"变换为"&"等,并删除所有不包含任何文字的叶子节点;

最后,合并段落。我们通过观察发现,网页中正文和非正文不会出现在同一个段落中,因此我们的实验以段落为最小决策单元。这里的段落指网页中在同一行内显示,而未被<p>、
、<h1>等 HTML分段指示标记分隔开的文字片段。但是因为一个段落中有时会因为包含一些字体约束标记,链接标记(如图 1中的"相关专题:欧洲新闻")等 HTML标记而在建立树状结构时被拆分成了多个叶子节点,因此我们需要进行段落合并处理,使得每个叶子对应一个段落。

合并段落就是把所有中间没有段落边界标记的相邻叶子节点合并为一个节点。依据 HTML的标记定义[11],通过分析网页在浏览器中的显示与该页面的源代码的对应关系,可以发现下列的标记是分段指示标记:〈P〉、〈BR〉、〈H1〉、〈H2〉、〈H3〉、〈H4〉、〈H5〉、〈H6〉、〈TABLE〉、〈TR〉、〈HR〉、〈DIV〉和〈CENTER〉,而源 HTML脚本中的回车符则是可以忽略的。

定义预处理完成后的树状结构为 HTML树,图 3给出了图 1所示网页对应的 HTML树。

4.2 全局范围决策

新闻网页的正文在网页页面上往往是成块显示的。本文通过定义正文统领节点的概念来描述正文的范围。正文统领节点(简称统领节点)是 HTML树上一个节点,它必须满足两个条件:1)下辖所有正文,2)其子节点都不能下辖所有正文。从图 1中我们可以看出"<body><CENTER><TABLE><TBODY><TR><TD>"HTML脚本解析栈中的最后一个"<TD width=580>"节点(其下辖范围用实线条框出)就是该页面的统领节点,而从其 HTML树我们更可以确认这一点(图 3中选中节点正是"<td width=580>"统领节点),因为该节点包含了所有正文且没有任何一个子节点还包含所有的正文。

文献[7]中提出了基于统计的网页正文信息抽取方法,其基本思想和我们的统领节点的想法一致。文献[7]中定义统领节点必须是 TABLE 节点,并且给出了评判一个节点是不是统领节点的两个最基本的特征:1)本节点所包含的非链接文字的总长度 P;2)该节点所包含的非链接文字占其上层 TABLE 节点的比例 T。在研究中发现,如果希望得到更为精细的正文范围,限制统领节点为 TABLE节点并不十分合适,而且有更多的特征可以参与判决正文的范围。因此,我们引入决策树这一机器学习算法来解决这个问题。具体的训练和处理过程如下:

1. 训练过程:

a)收集一批网页,并手工标注严格界定的正文;

b)计算标注的统领节点:在网页经过预处理后所对应的 HTML树中按统领节点的定义和正文的标注结果计算出各个网页标注的统领节点;

c)计算统领节点的候选:将 HTML树中所有叶子节点(因为某个正文可能仅有一个段落,此时该叶子节点就是统领节点)和包含多个子节点的节点都视为统领节点的候选;

d)形成训练数据:为每个统领节点的候选提取一个特征向量用以表征该候选节点,并将

它是否是标注的统领节点的布尔变量作为结论属性,形成训练数据;

e)引入 C4.5决策树训练算法[12]对上述特征向量和结论属性映射关系进行建模,训练得到全局范围决策树。

2.测试过程:对于新的待决策全局正文范围的网页,首先经过预处理得到对应的 HTML树并计算统领节点的候选,然后对每个候选节点提取特征向量,再依据训练得到的全局范围决策树按 C4.5决策树解释算法[12]计算出每个候选作为统领节点的概率,最后挑选出概率最高的节点作为统领节点决策结果(即全局范围决策的结果)输出。

因此全局范围决策的核心问题是如何定义候选统领节点的特征向量,通过观察分析和实验,我们对每个候选节点提取了如下的特征:

1)本节点包含的非链接文字的长度以及占根节点(整棵 HTML树)的比例;

2)本节点包含的链接文字的长度以及占本节点文字的百分比;

3)本节点包含的非链接文字占上一层候选节点中非链接文字的比例;

4)本节点的子节点个数;

5)本节点到上一层次的候选节点之间的 HTML节点的标记中包含 TABLE、DIV、TR、TD标记中的哪些;

6)本节点下辖范围内出现 P和 BR标记的个数以及占整个页面中 P和 BR出现次数的比例;

7)本节点的上一层次候选节点中除本节点外的范围内链接文字长度及其占该范围内所有文字的比例;

8)本节点是不是在某个 TABLE节点的子孙节点。

4.3 局部内容决策

统领节点决策成功只是找出了正文的范围,但统领节点所辖的所有文字并不都是正文,如图 1的统领节点 TD节点所辖文字中有许多就是相关报道或其他链接。因此在全局正文范围内进一步判定每个段落是不是正文也是我们需要完成的一个决策。

此处沿用全局范围决策的思路,在上述标注了正文的网页集合上,对标注的统领节点下每个叶子节点(即每个段落)提取特征向量,形成训练语料,然后引入 C4.5决策树训练算法,训练得到局部内容决策树。对于新 HTML树,我们将决策的统领节点下辖的所有叶子节点同样抽取特征向量送入正文内容决策树并执行 C4.5决策树解释算法,即可判断出每个叶子节点是否属于正文,并最终形成正文自动抽取结果。

对于每个叶子节点,其正文内容决策树的特征向量定义如下:

1)该叶子节点对应的 HTML标记是 P, BR, H1...H6, LI中的哪个;

2)该叶子节点中链接文字所占比例;

3)该叶子节点非链接文字的长度和链接文字的长度;

4)该叶子节点是不是某个 IMG节点的子孙节点;

5)该叶子节点文字的对齐方式是不是中间对齐;

6)该叶子节点开头的 16个字符中是否包含“讯”、“电”等文字;

7)该叶子节点是否包含“相关新闻”、“下一页”、“版权所有”、“Copyright”、“转载”等特征词,本文简单收集了 22个有助于判决不是正文的特征词;

8)该叶子节点的文字是不是隐藏的(hidden属性)。

5 实验

为了验证基于双层决策的新闻网页正文精确自动抽取算法的性能，我们开展了如下的实验工作。

5.1 语料和标注

我们首先制定了严格、详细的正文标注规范，然后收集了 10个国内主要新闻网站的新闻网页，共计 1687个。在 2名实验员的协助下完成了 1687个页面的正文的精细标注。10个网站具体网址和各自页面的数目如表 1所示。

表 1 各个网站标注的网页数目分布及其涵盖内容

网 址	数目	涵盖内容	网 址	数目	涵盖内容
http://www.sina.com.cn	430	娱乐、军事、体育、科技	http://news.xinhuanet.com	95	各种新闻
http://www.sohu.com	394	娱乐、政治、体育、IT	http://www.zaobao.com	87	经贸、娱乐
http://www.163.com	169	军事、体育、证券	http://world.eastday.com/	75	旅游
http://www.cctv.com	84	体育、军事、法治	http://news.shangdu.com/	100	各种新闻
http://www.ctn.com.cn	81	旅游、海外传真、	http://cn.news.yahoo.com/	83	各种新闻
http://news.china.com	89	各种新闻	合 计	1687	各种新闻

5.2 评测指标

我们分别定义如下指标来评测我们的正文范围和正文的决策性能：

1)采用精度 (*ScopeP*)和召回率 (*ScopeR*)以及 *F*值 (*ScopeF*)评价对一个网页正文范围的决策性能：

$$ScopeP=\frac{|A\cap E|}{|E|}\times 100\% \quad ScopeR=\frac{|A\cap E|}{|A|}\times 100\% \quad ScopeF=\frac{2\times ScopeP\times ScopeR}{ScopeP+ScopeR}$$

其中，*A*为该网页的手工标注的段落集合，*E*为该网页决策的正文范围所包含的段落集合，$A\cap E$表示 *A*和 *E*两个集合的交集，$|E|$、$|A|$等表示对应集合中所包含段落的数目。

2)对整个网页集合正文范围决策性能采用平均精度 *AveScopeP*、平均召回率 *AveScopeR*和平均 *F*-值 *AveScopeF*来评测：

$$AveScopeP=\frac{1}{N}\sum_{i=1}^{N}ScopeP_i \qquad AveScopeR=\frac{1}{N}\sum_{i=1}^{N}ScopeR_i \qquad AveScopeF=\frac{1}{N}\sum_{i=1}^{N}ScopeF_i$$

其中，*N*为网页的数目，$ScopeP_i$、$ScopeR_i$和 $ScopeF_i$分别是第 *i*个网页的正文范围决策的精度、召回率和 *F*值。

3)采用抽取正文与手工标注的误差率 *MCR*(简称抽取误差率)来评价最终对于一个网页的正文抽取性能：

$$MCR=\frac{REL+MEL}{AL}\times 100\%$$

其中，*REL*表示标注为非正文但被抽取为正文的文字长度，*MEL*表示标注为正文但却未被抽取为正文的文字长度，*AL*表示标注为正文的文字长度。

4)对整个网页集合采用抽取误差率大于某一门限 *t*的网页占总网页的比例 MHR_t来做为最终正文自动抽取系统的性能：

$$MHR_t=\frac{1}{N}\sum_{i=1}^{N}\delta(MCR_i>t)\times 100\%$$

其中，*N*为网页的数目，$\delta(MCR_i>t)$当 MCR_i大于 *t*时，取值为 1，否则为 0。*t*取值 [0，1]的正实数，如 $MHR_{0.0}$表示抽取内容与手工标注内容存在误差的网页的比例。MHR_t值越小表

示抽取性能越高。

5.3 全局范围决策实验

为了验证基于决策树的全局范围决策的效果，我们进行了三组实验：

1)认为整个页面都是正文的可能范围，也即不做任何范围决策；

2)采用文献[7]的方法，并根据我们的数据优化其两个参数为 $P=55$，$T=0.60$；

3)基于全局范围决策树的决策方法，采用 4-Fold交叉验证的方法测试其性能。

表 2给出了上述三种方法的性能对比实验结果：

表 2 三种全局范围决策方法的性能

范围决策方法	AveScopeP	AveScopeR	AveScopeF	AveScopeF相对提升
不做决策	25.23%	100.0%	35.55%	-52.62%
文献[7]	65.06%	96.50%	75.03%	基线系统
基于全局范围决策树	76.82%	98.80%	83.30%	+11.03%

从实验结果可以看出基于决策树的正文范围决策比其他两种方法有明显的效果提升。

5.4 局部内容决策实验

我们在上述标注数据集合上采用了 4-Fold交叉校验的方式验证了基于双层决策的新闻网页正文抽取方法，我们共尝试了以下四组实验：

1)不做局部内容决策，即认为所有全局正文范围内的段落都是正文；

2)基于决策树的内容决策：采用 4-Fold交叉验证的方法测试其性能；

3)基于决策树的内容决策，但是范围决策改为文献[7]的方法。

4)不做全局范围决策，直接采用局部内容决策树决策 HTML的正文内容。当然在训练数据生成时所有的叶子节点(而不仅是标注的统领节点之下的叶子节点)都作为训练样本。此组实验也就是单层决策的新闻网页正文抽取算法。

表 3给出了几组全局范围决策与局部内容决策方法组合后的实验结果。从实验结果中我们可以看出，基于局部内容决策树的方法可以明显提高正文抽取的正确率，$MHR_{0.0}$从 57.20%下降到了 18.14%。而对比最后三组实验结果，可以看出在相同的局部内容决策技术基础上，最终正文抽取性能与全局范围决策的性能有着密切的联系。以不做范围决策的实验为基线系统，全局范围决策树方法可以使得 $MHR_{0.0}$相对下降 29.85%，文献[7]的方法对 $MHR_{0.0}$也有少量的贡献，这些实验结果充分证明了本文所提出的双层决策的有效性。在 $MHR_{0.1}$标准上，全局范围决策的作用相对较少，分析原因是因为局部内容决策本身比较倾向于保留大段文本而丢弃小段文本，虽会使抽取的正文不顺畅，但最终表现为抽取误差率很可能还是小于 10%。

表 3 各种方法的性能测试结果

全局范围决策	局部内容决策	$MHR_{0.0}$	$MHR_{0.05}$	$MHR_{0.1}$	$MHR_{0.0}$相对下降
全局范围决策树	不做内容决策	57.20%	44.22%	36.16%	—
全局范围决策树	局部内容决策树	18.14%	9.43%	7.11%	29.85%
文献[7]的方法	局部内容决策树	25.33%	12.62%	9.30%	2.05%
不做范围决策	局部内容决策树	25.86%	12.50%	8.38%	基线系统

5.5 通用性验证实验

为了验证基于双层决策的新闻网页正文精确抽取算法对于未参与训练的网站的通用性，本文采用表 1第 1列的 6个网站的共 1247个新闻网页作为训练集，表 1第 4列的 5个网站的共 440个新闻网页做为测试集，完成集外测试的效果，实验结果如表 4所示。为了便于对比，

表 4中同时给出了 440个新闻网页 4 – Fold的交叉校验的实验结果。

表 4 网站间通用性验证实验结果

方 法	AveScopeP	AveScopeR	AveScopeF	$MHR_{0.0}$	$MHR_{0.1}$
4 Fold交叉验证	74.99%	98.18%	82.22%	22.27%	10.00%
集外测试	68.87%	89.02%	77.39%	30.68%	14.86%

从实验结果中可以看出,本方法在有限几个网站上训练得到的正文自动抽取系统,在其它网站的新闻网页上也有较好的正文抽取效果,说明本文提出的方法具有一定的通用性。

5.6 错误分析

我们统计分析了正文抽取误差率较高 (> 10%)的出错原因,主要可以分为三大类:

1)某些新闻网页正文下有较多的网友评论,而评论在严格界定的正文中被规定为非正文。但因网友评论往往是大段文字,符合正文的特征,所以容易被误抽取为正文的一部分;

2)某些新闻网页正文字数较少,对于这种网页本文提出的方法有时抽取结论是无正文,此时正文抽取误差率为 100%,对总体性能评分影响较大;而有时则因为抽取内容稍多,比如带上了版权信息,也因为标注正文长度很小而使得该网页正文抽取误差率偏高。

3)某些新闻网页包含多块新闻正文,而本文提出的方法在决策正文范围时容易只决策出一个正文块而忽略了其他正文块,使得最终抽取的正文不完整。

以上情况如何处理,还需要进一步的研究和探索。

6 结论

本文首次面向实际应用定义了严格的新闻网页正文,并提出和实现了基于双层决策的正文的自动抽取方法。该方法将对网页正文范围的全局决策和对决策范围内具体段落是否确是正文的局部决策这两个层面的决策有机结合,同时引入特征向量提取和决策树等方法来实现上述两个层面的自动决策。这种分层的思想既符合人们判断正文的步骤,又很好的满足了机器学习从不同层面提取各自特征的需要。实验表明,上述基于双层决策的方法能够精确地抽取出新闻网页的正文,最终正文抽取与人工标注不完全一致的网页比例仅为 18.14%,比单纯采用局部正文内容决策的方法相对下降了 29.85%,同时抽取误差率大于 10%的网页比例更是仅为 7.11%,满足了实际应用的需要。

下一步的研究工作主要包括:1)对于错误分析中指出的问题进行改进,进一步提高正文抽取的效果;2)本文的所有实验都是在有正文的网页上开展的,而实际某个新闻网站下载的页面可能并非都有正文,不少页面是导航型网页,如何自动的区分包含正文的真正新闻网页和导航型页面,也有待更深入的研究。

参 考 文 献:

[1] David Buttler Ling Liu et al. A Fully Automated Object Extraction System for the World Wide Web[A]. In: Proceedings of the 2001 International Conference on Distributed Computing Systems[C]. 2001: 361 – 370.

[2] 高军, 王腾蛟, 等. 基于 Ontology的 Web内容二阶段半自动提取方法[J]. 计算机学报, 2004 27(3): 310 – 317.

[3] 张绍华, 徐林昊, 等. 基于样本实例的 WEB信息抽取[J]. 河北大学学报(自然科学版), 2001, (12): 431 – 437.

[4] 邹纲, 刘洋, 等. 面向 Internet的中文新词语检测[J]. 中文信息学报, 2004 18(6): 1 – 9.

(下转第 103页)

5 结束语

在编码字符集标准相关研究领域,形式化方法一直较少应用,对标准的误读广泛存在,影响了系统实现和相关标准的开发。本文提出的 ISO 2022的有限状态机描述方法,可以准确描述其特征,便于评判实现复杂度,采用该方法对派生标准进行的分析揭示了不同标准间的内在联系,为相关研究工作带来新的思路和方法。

参 考 文 献:

[1] Atkin Steven A Framework for Multilingual Information Processing[D]. Doctor's dissertation Florida Institute of Technology. December 2001.

[2] ECMA-356 th Edition Character Code Structure and Extension Techniques[S].

[3] GB/T 2311-90 信息处理 七位和八位编码字符集 代码扩充技术[S].

[4] Kolman B Busby RC Ross SC. Discrete Mathematical Structures 4th Edition[M]. Beijing Higher Education Press & Prentice Hall Inc., 2001.

[5] Lunde Ken CJKV Information Processing[M]. Sebastopol O'Reilly & Associates 1999.

[6] 陈季雷, 杨裕衡, 林守铿. 洞悉 UNIX 中文系统篇[M]. 台北: 和硕科技文化有限公司, 1994.

[7] Zhu HF. et al Chinese Character Encoding for Internet Messages[S]. RFC 1922 March 1996.

[8] Scheifler Robert Compound Text Encoding Version 1.1[S]. X Consortium Standard X Version 11 Release 6.4. 1989

(上接第 9页)

[5] 许勇, 荀恩东, 等. 基于互连网的术语定义获取系统[J]. 中文信息学报, 2004 18(4): 37-43

[6] Yunhua Hu Guomao Xin Ruihua Song Guoping Hu Shuming Shi Yunbo Cao and Hang Li Title Extraction from Bodies of HTML Documents and Its Application to Web Page Retrieval[A] Proc of ACM-SIGIR'05[C]. 2005.

[7] 孙承杰, 关毅, 等. 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

[8] Valter Crescenzi Giansalvatore Mecca RoadRunner: Towards Automatic Data Extraction from Large Web Site[A]. In: proceeding of the 26th International Conference on very Large Database Systems[C], 2001: 109-118.

[9] Alberto H. F. Laender, Berthier A. Ribeiro2Neto A Brief Survey of Web Data Extraction Tools[J]. SIGMOD Record 2002 31(2): 84-93

[10] Daisuke Ikeda Yasuhiro Yamada Expressive Power of Tree and String Based Wrapper[A]. In: on2line proceedings of IJCAI'03 workshop on Information Integration on the Web[C]. 2003

[11] T. Berners Lee, D. Connolly, Hypertext Markup Language 2.0 MIT/W3C 1995 http://www.w3.org/MarkUp/html-spec/html-spec_toc.html

[12] J. R. Quinlan. C4.5 Programs for Machine Learning[J]. Morgan Kaufmannn Publishers San Meteo California 1992