基于节点类型标注的网页主题信息抽取方法

谢方立 周国民 王 健
(中国农业科学院农业信息研究所 北京 100081)

摘 要 提出一种基于 DOM 节点类型标注的网页主题信息抽取的方法。首先依据网页中噪声存在的形式,将 DOM 节点划分为 4 种类型:文本型、图片型、链接型和可忽略型,并给出节点内聚度的计算方法。通过给 DOM 节点添加类型和内聚度两个属性,在正文提取阶段选取内聚度大于阈值的文本型节点,最后整合成网页主题信息。将该方法与另外 3 款网页正文提取工具做对比实验,结果显示该方法在 F1 指标上为 95.1%,比 Evernote 工具高出 0.3%,比 YNote 工具高出 5.01%。

关键词 DOM,节点类型标注,主题信息抽取
中图法分类号 TP391 **文献标识码** A

Approach of Extracting Web Page Informational Content Based on Node Type Annotation

XIE Fang-li ZHOU Guo-min WANG Jian
(Agricultural Information Institute of CAAS, Beijing 100081, China)

Abstract An approach based on DOM node type annotation was proposed to extract web page informational content. According to noise patterns in web page, we firstly classify DOM nodes into four types: text, image, anchor and ignorance, and provide a method to calculate node degree of coherence(DoC). By adding two new attributes, type and DoC, to DOM node, we can select text nodes that have greater DoC than threshold during content extraction phase, and then integrate them as Web page informational content. In comparison to three other content extraction tools, the results show that in F1 index the proposed method is 95.1%, which is 0.3% higher than Evernote tool and 5.01% higher than YNote tool.

Keywords DOM, Node type annotation, Informational content extraction

1 引言

随着互联网的快速发展,网页数量呈指数递增,如何从浩如烟海的网页中获取所需的信息是人们面临的一个亟待解决的问题。网页中包含着丰富的内容,既有用户想要浏览的主题信息,也有对用户形成干扰与主题无关的信息,如页面导航条、推荐链接、广告条、版权声明等,后者通常被称为网页噪声。文献[1]估算出噪声数据在网页中占了 40%~50%的比例,并且预测这个比例还以每年 6%~8%的比例增长。网页噪声比重的持续增大给 Web 信息检索带来很大的难题,也对诸如网页分类和聚类、知识挖掘、话题检测、个性化信息推荐、数据挖掘等任务造成很大的影响。如果不将噪声去除,信息检索系统必然会得出很糟糕的检索结果。因此,去除网页噪声,从网页中抽取主题信息是 Web 信息检索的一个重要的基础性工作。

2 相关工作

在 Web 信息抽取领域,前人已经做了大量的研究工作。文献[2]提出 DSE(Data-rich Section Extraction)算法,针对同一个网站中的页面,自顶向下匹配模板相同的页面 DOM 树,将匹配结果中重叠或相同的结构看作非主题信息,其剩余的部分看作主题内容并提取出来。文献[3]提出风格树 SST(Site Style Tree),算法基本思想跟 DSE 类似:首先针对一个网站构建一个页面级别的风格树,风格树中的每个节点依据其内容特征和视觉特征来计算出节点的复合重要度,最后通过比较重要度来识别噪声节点和主题信息节点。类似于文献[2,3]这种基于模板匹配的方法适用于同一模板的网页集,然而 Web 上的模板不计其数,其导致模板维护成为了一个很大的困难。

文献[4]提出采用机器学习的方式生成网页的模板,通过检测网页中链接之间的关系,识别锚文本的特征来建立页面的模板,以及相应的模板提取规则,最后应用模板对网页进行主题信息提取。文献[5]提出一种基于支持向量机的 FIASCO 系统来识别网页中的噪声数据。在训练阶段,将 DOM 中的节点人工标注为 clean 和 dirty,并使用 SVM 训练出分类器。在清理阶段,首先也需要对网页文件做同样的解析分块

本文受国家高技术研究发展计划(2013AA102405)资助。
谢方立(1989-),男,硕士生,主要研究方向为 Web 信息检索,E-mail:fangli_x@163.com;周国民(1969-),男,研究员,主要研究方向为农业科学数据共享系统与技术、农业网络信息智能搜索技术等;王 健(1971-),男,研究员,主要研究方向为农业专业信息搜索理论与技术、信息检索、开放数据组织与共享等。

处理，然后使用分类器分类，删除那些被标注为 dirty 的块。类似于文献[4,5]这种基于机器学习的方法通常需要大量人工标注的网页训练集，且需要耗费很多资源来训练分类器或模板。

文献[6]提出利用〈TABLE〉标签和信息熵的方法，通过利用〈TABLE〉来划分网页，计算每个页面块的信息熵来将网页分为内容块和噪声块，最后通过比较信息熵的高低来去除噪声。文献[7]根据〈TABLE〉标签将网页分成若干部分，通过计算各个 TABLE 的长宽比，把长宽比很高的部分直接去掉，对其余 TABLE 中的内容根据是否存在和段落文字有关的标签或分隔符等来区分主题内容和噪音内容。文献[6,7]这些方法对〈TABLE〉标签有很大的依赖性，对于不用 TABLE 来布局的网页则不起作用。

文献[8]提出基于视觉特征的页面分块算法 VIPS(Vision based Page Segmentation)。VIPS 以网页的 DOM 为基础，充分利用了页面中的视觉表现，形成了 13 条启发式规则，可将页面划分成满足预设内聚度 PDoC(Permitted Degree of Coherence)的视觉信息块。这种方法能较好地对网页进行分块，然而存在一个弱点：如果两个信息块之间没有太大的、明确的分隔符，可能就识别不出来。另外，VIPS 只是一种网页分块算法，如果对网页作信息抽取则还需要额外的规则集，这进一步增加了算法的复杂度。

上述各类研究都存在一定的局限和不足，本文在前人工作的基础上结合对网页噪声特点以及网页性质的观察和统计，提出了一种基于 DOM 节点类型标注(Node Type Annotation)的主题信息抽取算法——NTA 算法。首先依据网页中噪声存在的形式，将 DOM 节点划分为 4 种类型；其次引用了 VIPS 内聚度的概念并给出本文节点内聚度的计算方法。通过自底向上遍历 DOM 给每个节点添加类型和内聚度两个属性，在正文提取阶段选取内聚度大于阈值的文本型节点，最后整合成网页主题信息。该方法弥补了 VIPS 不能抽取网页主题信息的不足，并且该方法具有较好的算法效率，不依赖特定标签因而也具有更好的通用性。

3 节点类型及内聚度

DOM(Document Object Model)的全称是文档对象模型，它可以独立于平台和语言来访问或修改文档的结构和内容，这里的文档可以是 HTML、XML、XHTML。使用 DOM 表示的文档被描述为一个树结构，DOM 树结构构成的基本要素是节点。在 DOM 标准中，节点的概念很宽泛，它可以是文档、元素、属性、注释等。为便于主题信息的提取，本文在对网页噪声特点观察和统计的基础上，将 DOM 标准中的元素节点进一步划分成 4 种类型：可忽略型、图片型、链接型、文本型，从而通过节点的类型就可以判断哪些是包含主题信息的节点，哪些是噪声节点。

为方便下文做进一步描述，先规定有关节点的一些统计参数及表示符号。LinkChar，LC_i，表示节点 i 中包含的所有链接字符的个数；LinkNumber，LN_i，表示节点 i 中包含的所有链接的个数；TotalChar，TC_i，表示节点 i 中包含的所有字符的个数；NoneLinkChar，NLC_i，表示节点 i 中包含的所有非链接字符的个数；ImageNumber，IN_i，表示节点 i 中包含的所有图片的个数。特别地，LC_b，LN_b，TC_b，NLC_b，IN_b 则都是表示网页〈BODY〉的统计参数。

3.1 节点类型

3.1.1 可忽略型节点

HTML 中的一些修饰性标签对正文提取来说是可以忽略不计的，这些标签不会包含主题信息，因此可以直接忽略掉。文献[9]相对全面地将 HTML 中的修饰性标签与非修饰性标签区别开来。本文在借鉴文献[9]标签分类的基础上，形成了适用于本文目的的可忽略标签集合，即出现在集合中的标签称之为可忽略型节点，如表 1 所列。

表 1 可忽略型节点

类型	标签名
可忽略型节点	SCRIPT，IFRAME，STYLE，NOSCRIPT，BR，BUTTON，INPUT，SELECT，OPTION，LABEL，FORM，COMMENT，MAP，AREA，EMBED

3.1.2 图片型节点

图片是网页中最常见的一类修饰元素，也以各种各样的形式存在，例如背景图片、广告、小图标、链接或正文内容的一部分。特别地，本文把以链接形式存在的图片看作噪声信息。除了在层叠样式表中定义之外，图片被包含在独有的〈IMG〉标签中。一般地，对于 DOM 结构中的非叶子节点 i，如果满足下面的不等式，则认为节点 i 是图片型节点。

$$\frac{LN_i}{LN_b}=0 \text{ 且} \frac{NLC_i}{NLC_b}=0 \text{ 且} \frac{IN_i}{IN_b}>0 \tag{1}$$

具体来说，如果非叶子节点 i 中既不包含链接字符，也不包含非链接字符，并且只有图片，那么认为该节点是图片型节点。对于 DOM 结构中的叶子节点，当其为〈IMG〉时才被认为是图片型节点。本文将图片型节点当作主题信息的一部分。

3.1.3 链接型节点

网页中的噪声大都以链接形式存在。如果一个节点中的链接比重较高，那么该节点更有可能是噪声信息。对于一个节点 i，通过计算节点中的链接字符、链接个数的比重，将其与非链接字符的比重进行比较可以判定节点是链接特征明显，还是文本特征明显。对于 DOM 结构的叶子节点，如果节点是〈A〉标签，那么称之为链接型节点；对于 DOM 结构的非叶子节点，如果满足下列不等式中任何一个，则认为该节点是链接型节点。

$$\frac{LN_c}{NLC_b}>\frac{NLC_i}{NLC_b} \text{ 或} \frac{LN_i}{NLC_b}>\frac{NLC_i}{NLC_b} \tag{2}$$

具体来说，当节点 i 中所含的链接字符的比重大于非链接字符的比重，或者节点 i 中包含的链接个数占据当前网页所有链接个数的比例大于节点包含的非链接字符个数占当前网页所有的非链接字符个数的比例，那么则认为节点的链接特征明显。

3.1.4 文本型节点

文本节点反映节点的文本特征比较明显，即节点中的非链接字符比重大于链接的比重，网页的主题信息更有可能存在于文本型节点中。对于 DOM 结构的叶子节点，如果该元素节点内部不为空并且只含有非链接文本，那么该元素节点是文本型节点。特别地，如果叶子节点中有孤立存在的非链接文本，这些字符仍被看作“文本型节点”，在后续处理中会将这些字符串包装成一个元素节点，然后当作文本型节点来处理。对于 DOM 结构中的非叶子节点，如果满足下列不等式，那么将其当作文本型节点。

$$\frac{NLC_i}{NLC_b}>\frac{LN_c}{NLC_b}\text{且}\frac{NLC_i}{NLC_b}>\frac{LN_i}{LN_b} \tag{3}$$

3.2 节点内聚度

3.2.1 内聚度

DOM 树结构中叶子节点是网页内容的具体体现，非叶子节点决定网页的布局。所有的叶子节点被包含在非叶子节点中，非叶子节点可以是一个叶子节点的祖先或者多个叶子节点的祖先。本文将节点分为 4 种类型，通过计算叶子节点集合中与其祖先节点类型一样的节点个数来定义内聚度。内聚度反映节点内容的一致性。内聚度越高，说明节点某一种类型特征越明显，内聚度越低，说明节点内部包含多种不同的类型节点，可认为该节点包含多种的噪音。

对于节点 i，其类型属于 text、anchor、image、ignore 4 种类型之一。为方便计算，为每种节点类型指定一个数字，如式(4)所示：

$$T=\begin{cases}1, & i\text{ 文本片型节点}\\ 2, & i\text{ 为链接型节点}\\ 3, & i\text{ 为图片型节点}\\ 4, & i\text{ 为可忽略型节点}\end{cases} \tag{4}$$

并且定义了一个计数函数 $Count(s,t)$，当类型一样时就进行计数，否则就不计数。如式(5)：

$$Count(s,t)=\begin{cases}1, & s=t\\ 0, & s\neq t\end{cases} \tag{5}$$

节点 i 的内聚度使用式(6)来计算：

$$DOC_i=\begin{cases}\dfrac{\sum_{j=1}^{s_i}Count(T_j,t)}{s_i}, & S_i>0, T_j\in\{1,2,3,4\},\ t\in\{1,2,3,4\}\\ 1, & S_i=0\end{cases} \tag{6}$$

其中，S_i 是节点 i 包含的所有叶子节点的个数。当 S_i 等于 0 时，说明节点 i 本身是叶子节点，规定叶子节点的内聚度为 1。T_j 表示第 j 个叶子节点的类型对应的值，t 表示当前节点 i 的类型对应的值。

具体地说，内聚度是指跟当前节点的类型一样的叶子节点占当前节点所包含的所有叶子节点个数的比重。跟 VIPS 内聚度不同的是，本文方法的内聚取值只在 0 到 1 之间，内聚度越接近 1，说明节点的某一类型的特征越明显，越接近 0，则节点内部混合多种类型节点。

3.2.2 文本密度

文本密度反映节点中包含的非链接文本的比重。节点 i 的文本密度(TextDensity，TD_i)计算方式如下：

$$TD_i=\frac{NLC_i}{NLC_b} \tag{7}$$

其中，NLC_i 表示节点 i 包含的所有非链接文本字符的个数，NLC_b 表示网页中(即〈BODY〉标签)包含的所有非链接文本字符的个数。当节点为图片型叶子节点或者链接型叶子节点时，文本密度为 0。对于非叶子节点，当文本密度越接近于 1，那么该节点包含主题信息内容的比重就越大。在节点类型标注过程中，把文本密度高于阈值的节点单独保存起来，用于后续的主题信息抽取。

3.2.3 阈值

在节点类型标注阶段，阈值被用来识别最可能包含主题信息内容的节点，如果文本密度高于阈值，那么单独保存该节点以备后续主题信息抽取。在主题信息抽取阶段，阈值被用来提取具有较高内容一致性的节点，如果节点的内聚度高于阈值，那么认为该节点包含的所有文本内容是主题信息的一部分，可以不再对该节点进行遍历，而是直接提取和保存节点的内容。

4 基于节点类型标注的算法

4.1 节点类型标注

当网页解析成 DOM 结构之后，利用 DOM 提供的操纵文档的接口将节点类型和内聚度以自定义属性的形式添加到节点属性列表中，实现节点类型标注的效果。采用自底向上遍历 DOM 结构的方法，依次计算节点类型、内聚度以及文本密度。节点类型标注算法流程如图 1 所示。

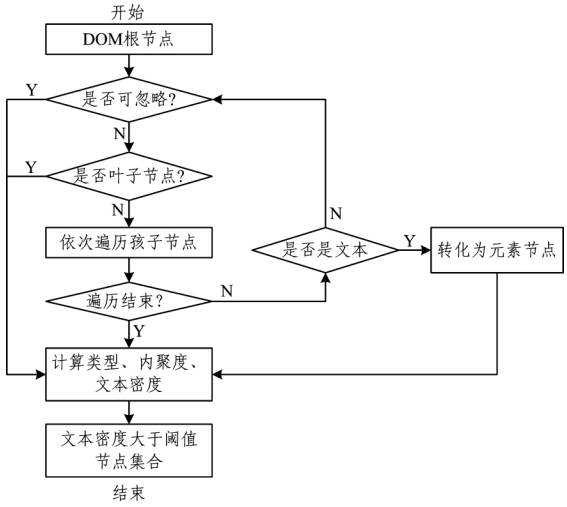


图 1 节点类型标注算法流程

之所以采取后序遍历是因为父节点内聚度的计算需要从子节点获取类型。经过上述流程后 DOM 每个节点都具备类型和内聚度，并且得到一个文本密度超过阈值的文本型节点列表。

4.2 主题信息抽取

主题信息抽取是基于上一步骤所得到的文本密度超过阈值的节点列表的，首先选取最佳剪枝文本节点，然后从该节点出发对 DOM 树进行剪枝，获取内聚度超过阈值的包含正文的文本节点。

4.2.1 DOM 剪枝

最佳剪枝文本节点指从节点密度超过阈值的列表(以下简称节点列表)中找到最适合作为剪枝起点的文本节点，从该节点出发，既能最大程度地获取包含主题信息的文本节点，也能最大程度地提高剪枝的效率。由于网页主题信息除了正文，还包含标题、发表日期、作者、来源等附加信息，这些附加信息跟包含正文内容的节点往往不在同一 DOM 层次。根据对多家门户网站主题型网页的调查，这一现象发生的概率是 99.7%。节点列表中的节点按加入的先后顺序排列，具有明显的父子关系。本文选取节点列表中处于第二顺位的节点(当列表节点个数小于 2 时则选择仅有的节点)作为最佳剪枝文本节点。

DOM 剪枝是从 DOM 树中寻找文本型节点的过程。与节点类型标注过程相反，DOM 是一个剪枝自顶向下的过程。如果节点是文本型节点并且内聚度大于阈值，则认为该节点中的内容是主题信息内容，将其保存起来；如果节点是图片型节点，那么将图片提取出来作为正文；如果节点是链接型或可

忽略型，则直接跳过。如果文本型节点的内聚度低于阈值，则继续遍历该节点的子节点，直至遍历完整个 DOM 树。

4.2.2 图片与链接问题

本文将图片型节点作为主题信息的一部分。然而网页中通常存在一些交互友好的小图标，例如分享、改变页面字体大小等图标，这些小图标可能会出现在最后的提取结果中，但并不是我们想要的。根据对几大门户网站主题型网页中图片和图标的大小进行调查，结果如表 2 所列。基于调查结果，设定噪声图片宽度阈值为 100 像素。在 DOM 剪枝过程中，如果图片型节点宽度小于 100 像素，那么认为该图片节点为图标类型节点，忽略不作处理；反之则保留该节点，将其作为主题内容的一部分。

表 2 网页中图标与图片宽度比较(单位 px)

网站	图标平均大小	图片平均大小
新华网	30～100	500～550
新浪网	45～90	478～650
搜狐网	40～80	300～520

主题内容中包含少量的链接文本是很常见的情况，这些链接文本往往具有较重要的提示，因此在主题信息提取时不能把这些链接文本遗漏掉。根据对几大门户网站主题型网页中链接文本占网页正文比重的调查，结果如表 3 所列。正文中的链接文本的邻近节点都是文本型节点，并且链接文本只占网页内容的很小的一个比例。在 DOM 剪枝过程中，当遇到链接型节点时，判断该节点邻近的兄弟节点是否为文本型节点，如果是文本型节点，则将该链接型节点保留视作主题内容的一部分；如果不是，则可跳过不作处理。

表 3 主题内容中链接文本的比重

网站	链接文本平均比重
新华网	0.053
新浪网	0.008
搜狐网	0.066

5 主题信息抽取实验

5.1 评价方法

使用查全率(Recall)、查准率(Percision)、F-1 来评价提取效果。网页中包含的主题内容字数用 T 表示，正确提取的字数用 TP 表示，提取出来的字数用 TR 表示。按照下列公式分别计算 $Recall$，$Percision$，以及 $F1$ 值。

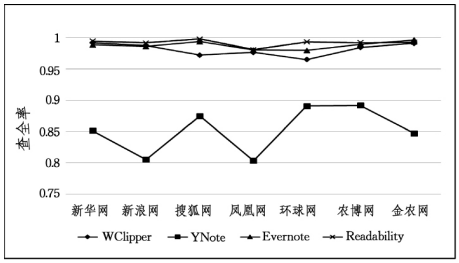
$$Recall=\frac{TP}{T} \tag{8}$$

$$Percision=\frac{TP}{T} \tag{9}$$

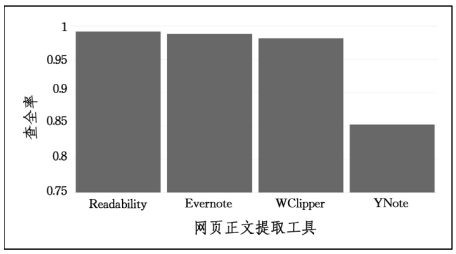
$$F1=\frac{2*Recall*Percision}{(Recall+Percision)} \tag{10}$$

5.2 实验结果

实验运行环境为 AMD FX(tm)-4300 CPU 处理器，CPU 为 3.80GHz，内存为 4.00GB，操作系统是 windows。基于 NTA 算法开发了一款网页正文提取工具 WClipper。为评估本文所提算法的有效性，从新华网、新浪网等 7 大门户网站上选取了 100 多个网页作为测试数据，同时将 WClipper 与其他 3 款网页正文提取工具——有道剪报工具(YNote)、印象笔记工具(Evernote)、Readability 进行对比实验。实验先后从 0.6 到 1 之间选取多个值作为阈值，通过对比发现当阈值设为0.9 WClipper 能取得最佳的提取效果。实验结果如图 2－图 4 所示。

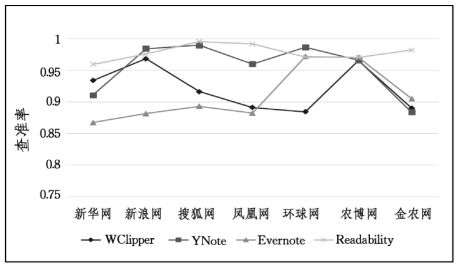


(a)

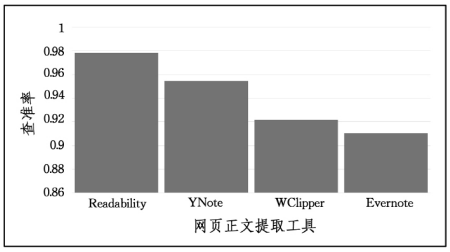


(b)

图 2 查全率对比图

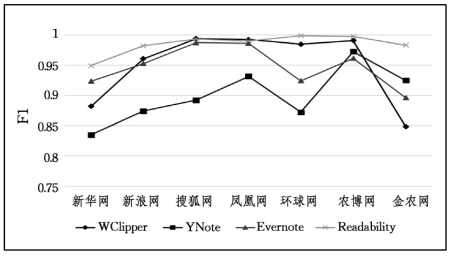


(a)

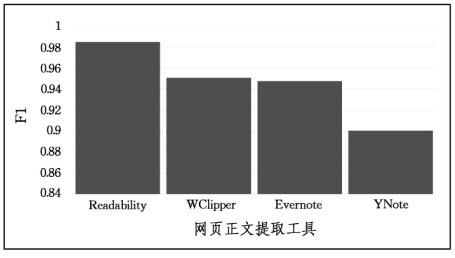


(b)

图 3 查准率对比图



(a)



(b)

图 4 F1 对比图

(下转第 49 页)

随机梯度下降法能够对模型进行优化。在小规模连续语音识别实验中，均值归一化的随机梯度下降法能够在一定程度上降低语音识别系统的识别错误率，在与多状态激活函数相结合后，系统的识别错误率再一次降低。

(3)使用奇异值分解与重构法可以对网络进行降维。在小规模连续语音识别实验中，使用这种方法能够使得网络的参数数量减少为原有的 0.49 倍，并且性能仅受到轻微的影响。

本文主要研究了语音识别系统中的声学模型，而语音识别系统的性能不仅受到声学模型的影响还受到语言模型的影响。为了进一步提升系统性能并且降低系统的硬件资源消耗，未来将考虑对语言模型的解码算法进行优化。

参 考 文 献

[1] Vincent P, Larocheiie H, Lajoie I, et al. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion[J]. Journal of Machine Learning Research, 2010, 11: 3371-3408

[2] Martens J. Deep learning via hessian-free optimization[C]// Proceedings of the 27th International Conference on Machine Learning (ICML-10). Israel: Haifa, 2010: 735-742

[3] Dean J, Corrado G, Monga R, et al. Large scale distributed deep networks[C]// Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Lake Tahoe, Nevada, United States: Proceedings of a meeting held. 2012: 1232-1240

[4] Deng W, Qian Y, Fan Y, et al. Stochastic data sweeping for fast DNN training[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Florence, Italy: ICASSP 2014, 2014: 240-244

[5] You Zhan, Wang Xiao-rui, Xu Bo. Exploring one pass learning for deep neural network training with averaged stochastic gradient descent[C]// ICASSP 2014. 2014: 6854-6858

[6] Xue J, Li J, Gong Y. Restructuring of deep neural network acoustic models with singular value decomposition[C]// INTERSPEECH 2013. Lyon, France: 14th Annual Conference of the International Speech Communication Association, 2013: 2365-2369

[7] He Y, Qian F Y, et al. Reshaping deep neural network for fast decoding by node-pruning[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Florence, Italy: ICASSP 2014, 2014: 245-249

[8] Graves A, Mohamed A R, Geoffrey E. HINTON. Speech recognition with deep recurrent neural networks[C]// ICASSP 2013. 2013: 6645-6649

[9] Sarikaya R, Hinton G E, Deoras A. Application of Deep Belief Networks for Natural Language Understanding[J]. IEEE/ACM Transactions on Audio, Speech & Language Processing, 2014, 22(4): 778-784

[10] Mohamed A, Dahl G E, Hinton G E. Acoustic modeling using deep belief networks[J]. IEEE Transactions on Audio, Speech & Language Processing, 2012, 20(1): 14-22

(上接第 34 页)

5.3 分析与讨论

从图 2 可以看出，WClipper 在新华网、新浪网、凤凰网、农博网、金农网上的查全率表现跟 Evernote 工具和 Readability 很接近，且查全率接近 1。在搜狐网、环球网上的查全率表现不如 Evernote 工具和 Readability，是因为 WClipper 遗漏掉了主题内容的一些附加信息，比如文章来源。这是受限于本文的算法特性，文章来源一般以链接的形式存在，会被识别为链接型节点，而直接忽略掉。从图 3 看出，WClipper 在搜狐网、凤凰网、环球网上的查准率低于其他工具，是因为网站的部分页面含有评论区域并且评论区域含有比较多的文本内容。评论区的内容不属于主题信息的一部分，但由于 NTA 算法并未对文本节点做出进一步的区分，因此最后都整合到一起，影响了查准率。从图 4 中 F1 指标来看，WClipper 工具的综合提取效果比 Evernote 工具高出 0.3%，比 Ynote 工具高出 5.01%，这在一定程度上验证了本文方法的有效性。

结束语 在前人工作的基础上结合对网页噪声特点以及网页性质的观察和统计，提出了一种基于 DOM 节点类型标注的主题信息抽取方法。将 DOM 节点划分为 4 种类型并依据节点内聚度、节点文本密度、阈值等统计信息实现网页主题内容的抽取。初步试验验证了本文方法的有效性，显示了本文方法相比其他同类工具表现出较好的主题信息抽取效果。由于该方法不依赖特定标签且只规定了较少的启发式规则，因此具有较好的通用性和算法效率。但是也存在一些不足，进一步的研究和改进包括：识别和去除网页中不是超链接形式的噪声，如评论文字、版权声明等，可以借助网页中的视觉特征来加以甄别。

参 考 文 献

[1] Gibson, David, Punera K, et al. The volume and evolution of Web page templates[C]// Special Interest Tracks and Posters of the 14th International Conference on World Wide Web. ACM, 2005

[2] Wang Ji-ying, Lochovsky F H. Data-rich section extraction from html pages[C]// Proceedings of the Third International Conference on Web Information Systems Engineering, 2002 (WISE 2002). IEEE, 2002: 313-322

[3] Yi L, Liu B, Li X. Eliminating noisy information in web pages for data mining[C]// Proceedings of the 9th ACM SIGKDD Int Conference on Knowledge Discovery and Data Mining. New York: ACM, 2003: 296-305

[4] 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)，2008(S1)：1743-1747

[5] Bauer, Daniel, et al. FIASCO: Filtering the Internet by Automatic Subtree Classification, Osnabruck. Building and Exploring Web Corpora[C]// Proceedings of the 3rd Web as Corpus Workshop, Incorporating Cleaneval. Vol. 4. 2007

[6] Lin S H, Ho J M. Discovering informative content blocks from Web documents[C]// Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2002

[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] Cai Deng, et al. VIPS: a vision based page segmentation algorithm. Microsoft technical report[R]. MSR-TR-2003-79, 2003

[9] 邹永强，钟志农. 一种高效的新闻网页噪声过滤方法[J]. 微型机与应用，2011，30(16)：64-67