Computer Science and Application 计算机科学与应用, 2020, 10(1), 51-59
Published Online January 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.101007

Hans汉斯

A Web Page Cleaning Method Based on Template and SVM

Jincheng Yan, Yunfeng Wang*

Department of Computer Science and Technology, Sichuan University, Chengdu Sichuan
Email: *yfwang@scu.edu.cn

Received: Dec. 20th, 2019; accepted: Jan. 2nd, 2020; published: Jan. 9th, 2020

Abstract

This paper presents a method of web page denoising based on template and support vector machine (SVM). This method divides web page noise into common noise and personalized noise. Firstly, a template library from the web page collection is established, and the common noise of web page will be removed by using the template. And then, the features for block-level labels are calculated, with which the SVM model is trained. Finally, the trained SVM model is used to divide block-level labels into noise and main text, achieving the purpose of denoising. This method can effectively remove the copyright, navigation, advertising and other noise information in the web page. Compared with the pure use of SVM for web page denoising, both accuracy and recall rate of this method were improved.

Keywords

Web Page Clean, Template, SVM

基于模板和SVM协同工作的网页去噪方法

严金承，王运锋*

四川大学计算机学院，四川 成都
Email: *yfwang@scu.edu.cn

收稿日期：2019年12月20日；录用日期：2020年1月2日；发布日期：2020年1月9日

摘　要

本文提出一种基于模板和支持向量机(SVM)协同工作的网页去噪方法。该方法将网页噪声分为公共噪声和个性化噪声两类。首先从网页集合中建立模板库，利用模板去除网页公共噪声。对于剩下的个性化噪

*通讯作者。

文章引用: 严金承, 王运锋. 基于模板和 SVM 协同工作的网页去噪方法[J]. 计算机科学与应用, 2020, 10(1): 51-59.
DOI: 10.12677/csa.2020.101007

声，先计算块级标签特征，利用这些特征训练SVM模型，最后用训练好的SVM模型将块级标签分为噪声和正文两类，达到去噪目的。该方法能够有效去除主题型网页中的版权、导航、广告等噪声信息。与单纯使用SVM进行网页去噪相比，查准率和查全率上均有提升。

关键词

网页去噪，模板，SVM

1. 引言

Web 网页蕴含着大量有价值的信息，已成为搜索引擎、舆情分析、问答系统等文本分析领域的天然数据矿场，但 Web 网页中同时掺杂着与正文内容无关的信息，例如：网页版权信息、广告、导航栏等。如何抽取网页正文内容，去除上述“网页噪声”，有着重要的研究意义。目前，网页去噪的方法主要包括：基于规则的方法、基于模板的方法、基于视觉信息的方法和基于机器学习的方法。基于规则的方法需要人工分析网页，手工构建规则，应用受限。基于模板的方法过于呆板，拓展性差。基于视觉信息的方法对现在以 DIV + CSS 为主流布局的网页不太适用且效率低下。基于机器学习的方法依赖于所选取的样本特征，对于短文本噪声识别度低。本文分析了网页噪声规律，将网页噪声分为公共噪声和个性化噪声。由于公共噪声内容固定且文本较短，利用模板法先行对公共噪声去噪，效率高、识别准，能弥补 SVM 对包含短文本噪声的标签块识别不准的问题。同时，利用 SVM 对个性化噪声去噪提高了模板法去噪的拓展性。

2. 研究现状

文献[1]综述了网页去噪方法，从模型数量的角度，将网页去噪分为多模型网页去噪和单模型网页去噪两类。文献[2]主要通过启发式规则去噪，将文档斜率曲线中的“高地”部分确定为正文内容。文献[3]单纯利用模板的方法，首先得到一个文本字符流，选择大小为 W 的窗口在字符流上滑动，每次滑动结果称为一个 Shingle，通过计算 Shingle 集中不同 Shingle 的频率来确定模板，最后通过模板筛选噪声，剩下的内容即为正文内容。文献[4] [5] [6]基于视觉特征对网页去噪，提出了经典算法 VIPS，文献[7]对 VIPS 进行了改进，利用样式特性对样式树进行权重标注，提取正文内容。文献[8]-[15]主要利用了机器学习的方法，将网页去噪问题当作分类问题处理，特别是文献[12]的方法，利用支持向量机对网页去噪，取得了较好的效果。

本文提出的基于模板和 SVM 协同工作的网页去噪方法，结合模板法和支持向量机对网页去噪的优点，有以下优势：

1) 为每个网站建立模板库，对该网站的公共噪声，诸如：版权信息，部分导航信息去噪，有较精准的去噪能力。

2) SVM 对网站的个性化噪声，诸如：广告等去噪，有较好的适应性，弥补了模板法去噪呆板，不灵活的缺点。

3. 网页数据准备

本文预先爬取了各大主流新闻网站网页共 3000 篇，网页量分布如表 1：

Table 1. Web page volume distribution
表 1. 网页量分布

序号	网站名	网页量
01	光明网	500
02	今日头条	800
03	凤凰网	300
04	搜狐新闻	600
05	腾讯新闻	800
总计		3000

4. 建立模板库

本文将网页噪声分成公共噪声和个性化噪声两类。来自同一网站的不同网页通常有着一些共同信息，比如网站版权信息，此外，由于现代网页通常基于一套样式模板来开发，以保持网站的风格性和美观性，所以这些共同信息还包括统一风格的导航栏，保持网页结构的模板化信息等，这部分噪声称为公共噪声，其他噪声称为个性化噪声，如图 1。其中，红框中内容为公共噪声，绿框中内容为个性化噪声，黑框中内容为正文。由于公共噪声内容固定，利用模板法直接进行比对能精准快速的对其进行识别。

Figure 1. Category of web page noise
图 1. 网页噪声分类

通过 URL 前缀可识别网页的所属网站。下面为网站 X 建立模板库：

1) 初始化已处理网下标 i = 1，每批次处理计数 count = 1。初始化标签集合 S 为空集，S 中的标签记为 S_k(k = 1, 2, 3, …)，该标签的频次记为 F_k。

2) 处理该网站下的第 i 个网页 X_i(i = 1, 2, 3, …)。将网页 X_i 的标签记为 X_iT_j(j = 1,2,3,…,m)。设定 j = 1。对于 X_iT_j，依次与 S 中的所有标签比较，查看是否有标签名相同，内容相近的标签。若存在标签 S_n 与 X_iT_j 标签名相同，内容相近，则 F_n++，否则将 X_iT_j 加入 S。j+ = 1，直到处理完 X_i 的所有标签。

3) 处理完 X_i 后，判断 count 是否达到阈值 N1。若没有达到，则 X_i+ = 1，count+ = 1。返回步骤 2，处理下一个网页。否则进入步骤 4。

4) 查看 S 中所有标签的频次 F_k，若 F_k 达到阈值 N2，则将其对应的标签 S_k 持久化为模板，同时置 count = 1，清空缓存集 S，返回步骤 2，进入下一批次的网页处理，直到处理完在该网站下爬取的所有网页。

例如，网站 X 中可能的两篇网页的部分 HTML 代码如图 2。处理完后，缓存集中的结果如表 2：

网页1：

```
<html><body>
<div><ul>
<li><a>中国</a></li>
<li><a>国际</a></li>
<li><a>军事</a></li>
<li><a>观点</a></li>
<li><a>专题</a></li>
</ul></div>
<div>本文系转载，不代表参考消息网的观点。参考消息网对其文字、图片与其他内容的真实性、及时性、完整性和
准确性以及其权利属性均不作任何保证和承诺，请读者和相关方自行核实。</div>
<div>
<span>来源：新华社</span>
<span>责任编辑：张越</span>
</div>
<div>
<p>正文部分A</p>
<p>正文部分B</p>
<p>正文部分C</p>
</div>
<div><p>国新网备2012001 互联网出版许可证（新出网证(京)字147号）京ICP备11013708 京公网安备110402440030</p></div>
<a>广告A</a>
<a>广告B</a>
</body></html>
```

网页2

```
<html><body>
<div><ul>
<li><a>中国</a></li>
<li><a>国际</a></li>
<li><a>军事</a></li>
<li><a>观点</a></li>
<li><a>专题</a></li>
</ul></div>
<div>有消息称...</div>
<div>
<span>来源：新华社</span>
<span>责任编辑：王兵</span>
</div>
<div><p>国新网备2012001 互联网出版许可证（新出网证(京)字147号）京ICP备11013708 京公网安备110402440030</p></div>
</body></html>
```

Figure 2. A part of web page code
图 2. 某网页部分代码

Table 2. Labels in cache S
表 2. 缓存集 S 中的标签

序号	叶子标签名	内容	频次
01	a	中国	2
02	a	国际	2
03	a	军事	2
04	a	观点	2
05	a	专题	2
06	div	本文系转载，不代表参考消息网的观点，参考消息…	1
07	span	来源：新华社	2

Continued

08	p	责任编辑：张越	1
09	p	正文部分 A	1
10	p	正文部分 B	1
11	p	正文部分 C	1
12	p	国新网备 2012001 互联网出版许可证…	2
13	a	广告 A	1
14	a	广告 B	1
15	div	有消息称…	1
16	span	责任编辑：王兵	1

N1、N2、判断文本内容相近算法的选取对模板库质量有重要的影响。N1 过小，统计后，公共噪声标签频次与正文内容标签频次差别将会很近，导致没有合适的 N2 值对二者进行区分，无法识别出公共噪声。N1 过大将大大增加算法的计算量，导致性能降低。

缓存集大小与正文内容和公共噪声的频次关系如图 3，可以看出正文内容和个性化噪声在频次统计中不会高于 3，故在 N1 适中的情况下，比如取到 10，将 N2 设置为 3 可很好的区分公共噪声。

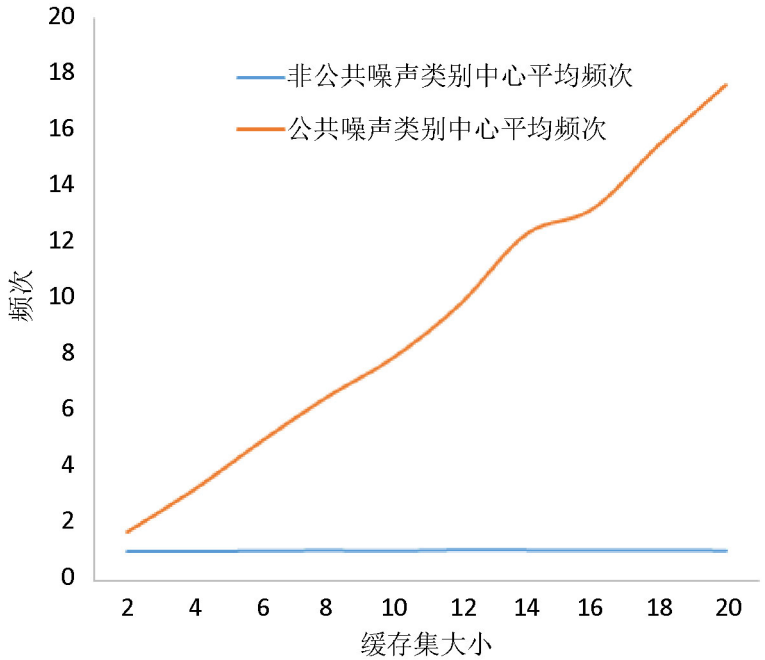

Figure 3. Cache set size-category center diagram
图 3. 缓存集大小-类别中心关系图

对于文本内容相近的判断，可借用编辑距离算法[13]。为保证模板库中模板的准确性，本文采用较为严格的策略，以两文本中文本长度较短的文本作参照，对于文本长度不超过 8 的文本，可将编辑距离设定为 0，即文本必须相同，才判断为相近。对于文本长度超过 8 的文本，每超过 8 个字，可增加一个编辑距离。若用 D 表示文本 TA 和文本 TB 的编辑距离，即：$D \leq (\text{Min}(\text{Len}(TA), \text{Len}(TB)))/8$ 时，可判断为 TA、TB 相近。其中，Len()表示获取文本长度的函数。

5. 标签特征化

将网页去噪问题当作分类问题，即要利用 SVM 分类器将每个标签及其内容分为噪声和非噪声两类。经统计，噪声内容和非噪声内容一般聚集在某块级标签内，故本文以块级标签为单位计算标签特征。网页中块级标签主要包括：body 标签、section 标签、div 标签。在计算特征时，要去掉嵌套在块级标签的其他块级标签内容。例如，对于如 2 中所述网页，在计算 body 标签的特征时，要抛开其中嵌套的 div 标签。下面介绍各特征属性及计算方法。

5.1. 特征计算方法

1) 外部文本长度比率。外部文本长度比率 R1 是块级标签内所有文本长度 innerLen 和整个 HTML 页所有文本长度 outerLen 的比值，统计发现，包含正文内容的块级标签内所有文本长度占全文所有文本长度的比重一般较大，故标签 R1 特征往往较大。

2) 链接文本长度比率。链接文本长度比率 R2 是块级标签内所有链接文本长度 innerLinkLen 和整个 HTML 页所有链接文本长度 outerLinkLen 的比值，统计发现，包含正文内容的块级标签所有链接文本长度占全文所有链接文本长度的比重一般较小，故标签 R2 特征往往较小。

3) 链接标签数量比率。链接标签数量比率 R3 是块级标签内 a 标签数量 innerLinkNum 和整个 HTML 页 a 标签数量 outerLinkNum 的比值，统计发现，包含正文内容的块级标签内 a 标签数量占全文 a 标签数量比重一般较小，故标签 R3 特征往往较小。

4) 图片标签数量比率。图片标签数量比率 R4 是块级标签内 img 标签数量 innerPicNum 和整个 HTML 页 img 标签数量 outerPicNum 的比值，统计发现，包含正文内容的块级标签内 img 标签数量占全文 img 标签数量比重一般较小，故标签 R4 特征往往较小。

5) 内部文本长度比率。内部文本长度比率 R5 是块级标签内链接文本长度 innerLinkLen 和块级标签内所有文本长度 innerLen 的比值，统计发现，包含正文内容的块级标签内链接文本长度占块级标签内所有文本长度比重一般较小，故标签 R5 特性往往较小

5.2. 数据平滑化

计算上述标签特征时，可能遇到分母为 0 的情况。为简单起见，本文引用自然语言处理领域的加法平滑方法[16]，在计算上述标签特征时，分母统一加上 1。该方法对最终的计算结构影响不大，又避免了除 0 情况的发生。平滑后各特征计算公式如下：

$$\begin{cases} R1 = \dfrac{\text{innerLen}}{\text{outerlen}+1} \\ R2 = \dfrac{\text{innerLinkLen}}{\text{outerLinkLen}+1} \\ R3 = \dfrac{\text{innerLinkNum}}{\text{outerLinkNum}+1} \\ R4 = \dfrac{\text{innerPicNum}}{\text{outerPicNum}+1} \\ R5 = \dfrac{\text{innerLinkLen}}{\text{innerLen}+1} \end{cases}$$

5.3. 特征计算示例

以图 2 中网页 1 所示代码为例，计算各块级标签统计量和特征，其结果如表 3 和表 4。

Table 3. Statistics of block label
表 3. 各块级标签统计量

序号	标签名	文本长	链接文本长	a 标签数	img 标签数
01	body	6	6	2	0
02	div	10	10	5	0
03	div	83	0	0	0
04	div	13	0	0	0
05	div	15	0	0	1
06	div	65	0	0	0
总计	全文统计	192	16	7	1

Table 4. Features of block label
表 4. 各块级标签特征

序号	标签名	R1	R2	R3	R4	R5
01	body	0.031	0.353	0.25	0	0.857
02	div	0.052	0.588	0.625	0	0.909
03	div	0.430	0	0	0	0
04	div	0.067	0	0	0	0
05	div	0.078	0	0	0.5	0
06	div	0.336	0	0	0	0

注：总计结果略小于 1 的原因由数据平滑造成。

6. 数据训练及模型生成

1) 数据标注。对第 1 章爬取的所有网页进行特征计算后，采用人工标注的方式，标注其块级标签中的内容是否为正文内容。

2) 保证正负样本均衡。一般来说，某网页正文内容块级标签较少，噪声块级标签较多。故爬取得到的噪声样本将远远多于正文样本。在训练前对噪声样本进行抽样，保证正负样本数量均衡。

3) 确定 SVM 核函数[17]，由于训练样本中特征数量较少，故采用径向基核函数将样本映射到更高纬空间，可以取得更好效果。

4) 确定模型参数。采用 10 折交叉验证法进行训练，确定最终模型参数，保证参数较优。

7. 协同去噪方法

该方法分为两阶段对网页进行去噪，具体步骤为：

1) 通过网页的 URL 前缀找到对应网站模板库。

2) 遍历待去噪网页的所有标签，若该标签与模板库中某标签的标签名相同，且文本内容相近。则认为该标签所含内容为公共噪声，将这些标签删除，即除去了公共噪声。内容相近的判断方法与第 3 章所描述的方法一样。

3) 经过步骤 1 后，按第 4 章的方法计算网页剩余的块级标签特征，送入训练好的 SVM 模型进行识别。包含噪声的块级标签标记为 0，包含正文内容的块级标签标记为 1。

4) 保留标记为 1 的块级标签，将其所包含的内容提取出来，去噪完成。

8. 实验结果与分析

8.1. 评价指标

以叶子标签为单位，对协同方法去噪结果采用查准率和查全率[18]进行评价。由于 SVM 模型对块级标签进行分类，故将块级标签下的叶子标签分类结果设为该块级标签分类结果。下面通过混合矩阵介绍这两种评价指标，如表 5：

Table 5. The mixed matrix
表 5. 混合矩阵示意

	判定位正文	判定为噪声
实际为正文	TP	FN
实际为噪声	FP	TN

根据上述混合矩阵，查准率 P 和查全率 R 的计算方式如下：

$$P = \frac{TP}{TP + FP} \tag{1}$$

$$R = \frac{TP}{TP + FN} \tag{2}$$

8.2. 实验结果

根据上述评价指标，本文采用第 6 章介绍的方法对爬取的 3000 篇网页进行去噪，其结果如表 6：

Table 6. The result of de-noising
表 6. 协同方法去噪结果

网页来源	网页量	查准率/%	查全率/%
光明网	500	96.2	95
今日头条	800	97.6	98.2
凤凰网	300	95.4	94.4
搜狐新闻	600	94.9	96.3
腾讯新闻	800	95.5	96.7

为说明协同去噪方法对去噪效果的提升，本文与文献[15]的方法进行了对比实验，实验结果如表 7：

Table 7. The result of comparing
表 7. 对比实验结果

方法	网页量	查准率/%	查全率/%
本文	3000	96	96.5
文献[15]	3000	94.7	90.2

8.3. 实验分析

从实验结果来看，本文方法在查准率和查全率都有较好的表现。对于搜狐新闻，查准率略低，是因为爬取的网页中有较多的链接型网页和图片型网页，所以本文方法对于主题型网页有较好的效果，而对

图片型和链接型网页表现略差。与文献[12]方法的对比实验中，本文方法在查全率上有明显提升，是因为文献[12]需要对正文信息定位，一是可能发生定位错误，二是有少许正文信息存在于主要 div 标签之外。其次，查准率的略微提升得益于模板法对固定短小的公共噪声识别有较高的准确性，弥补了单纯使用 SVM 不能较好的识别短文本噪声的不足。

9. 结束语

本文提出了模板与 SVM 协同工作的网页去噪方法，利用事先建立好的模板库识别网页中的公共噪声信息，再通过 SVM 模型对网页中个性化噪声进行识别。实验表明，该方法整体效果较好。但是本文在训练 SVM 模型时，计算的标签特征量较少，没有结合文本内容的语义信息进行考虑。其次，SVM 模型识别阶段是以块级标签为单位，对于块级标签中正文内容和噪声信息混合的情况，无法将二者分开，后续将对这些内容继续研究。

基金项目

成都市科技计划项目资助(2019-RK00-00015-ZF)。

参考文献

[1] 毛先领, 何靖, 闫宏飞. 网页去噪: 研究综述[J]. 计算机研究与发展, 2010, 47(12): 2025-2036.

[2] Finn, A., Kushmeric, N. and Smyth, B. (2001) Fact or Fiction: Content Classification for Digital Libraries. *Proceedings of the 2nd DELOS Network of Excellence Workshop on Personalization and Recommender Systems in Digital Libraries*, Dublin, Ireland, 1-6.

[3] Gibson, D., Punera, K. and Tomkins, A. (2005) The Volume and Evolution of Web Page Templates. In: *Proceedings of the 14th International Conference on Word Wide Web*, ACM, New York, 830-839. https://doi.org/10.1145/1062745.1062763

[4] Cai, D., Yu, S., Wen, J.R. and Ma, W.-Y. (2003) Extracting Content Structure for Web Pages Based on Visual Representation. In: Zhou, X., Orlowska, M.E. and Zhang, Y., Eds., *Web Technologies and Applications. APWeb* 2003. *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, 406-417. https://doi.org/10.1007/3-540-36901-5_42

[5] Cai, D., Yu, S., Wen, J.R. and Ma, W.-Y. (2003) VIPS: A Vision-Based Page Segmentation Algorithm. Microsoft Research.

[6] Debnath, S., Mitra, P., Pal, N. and Giles, C.L. (2005) Automatic Identification of Informative Sections of Web Pages. *IEEE Transactions on Knowledge and Data Engineering*, **17**, 1233-1246. https://doi.org/10.1109/TKDE.2005.138

[7] 王健, 张金. 基于节点权重的网页去噪方法的研究[J]. 计算机技术与发展, 2017, 27(10): 83-86.

[8] 伊政, 徐武平, 徐爱萍. 一种基于结构分析的网页主题区域发现方法[J]. 计算机工程与应用, 2015, 51(6): 227-230+259.

[9] 郗家贞, 郭岩, 黎强, 等. 一种短正文网页的正文自动化抽取方法[J]. 中文信息学报, 2016, 30(1): 8-15.

[10] 周艳平, 李金鹏, 宋群豹. 一种基于 SVM 及文本密度特征的网页信息提取方法[J]. 计算机应用与软件, 2019, 36(10): 251-255+261.

[11] 李桐宇, 任锐, 蔡鸿明, 等. 基于文本对象模型的自动化网页内容提取方法[J]. 上海交通大学学报, 2018, 52(10): 1363-1369.

[12] 杨贤, 唐超兰, 李航. 基于文本块密度与标签路径等特征的正文提取[J]. 广东工业大学学报, 2018, 35(2): 51-56.

[13] 陈雪, 徐慧, 沈家峻. 基于网页结构的网页去噪算法设计[J]. 软件, 2013, 34(8): 95-97.

[14] 宋鳌, 支琤, 周军, 等. 基于 LCS 的特征树最大相似性匹配网页去噪算法[J]. 电视技术, 2011, 35(13): 44-48+63.

[15] 梁东, 杨永全, 魏志强. 基于支持向量机的网页正文内容提取方法[J]. 计算机与现代化, 2018(9): 21-26+31.

[16] W. Bruce Croft, Donald Metzler, 等. 搜索引擎信息检索实践[M]. 北京: 机械工业出版社, 2010.

[17] 刘春卫, 罗健旭. 基于混合核函数的 PSO-SVM 分类算法[J]. 华东理工大学学报(自然科学版), 2014, 40(1): 96-101.

[18] Raghavan, V. and Wang, G.S. (1989) A Critical Investigation of Recall and Precision as Measures of Retrieval System Performance. *ACM Trans on Information Systems*, **7**, 205-229. https://doi.org/10.1145/65943.65945