● 吴思竹[1,2]，张智雄[1]
(1. 中国科学院　国家科学图书馆，北京　100190；2. 中国科学院　研究生院，北京　100049)

基于网页特征识别的噪音网页过滤方法研究*

摘　要：本文通过对网页结构和内容特征的深入分析和识别，对噪音网页的过滤方法进行研究和实验。首先利用阈值过滤具有明显特征的噪音网页，而后建立网页特征向量，利用 SVM 对网页进行分类。采用采集自 Web 的网页数据进行实验分析，最后得出研究结论，并展望下一步工作。

关键词：网络资源；噪音网页；过滤方法

Abstract: With an in-depth analysis and identification of web page structure and content characteristics, this paper makes a study of and an experiment on noise web page filtering method. The paper first uses the threshold to filter the noise web pages with distinctive characteristics, then establishes the characteristic vectors of web pages, and classifies the web pages with Support Vector Machine (SVM). It makes an experimental analysis of the data collected from the web pages, and comes to the final research conclusions. The paper also predicts the future work.

Keywords: network resources; noise web page; filtering method

网络成为快速增长，蕴涵丰富内容的、巨大的信息仓库，逐渐成为研究者进行数据抓取采集、深入知识发现和挖掘的重要来源。但是，并不是所有网页中都包含有用的信息，现有网页类型可以划分为两种：主题网页与噪音网页。主题网页指描述一个或多个实质主题内容的具有较多描述文字的页面；而噪音网页指不包含或包含极少的叙述文字、包含密集链接的页面。真正对于研究具有可用性的是主题网页，而噪音网页在面对以 Web 页面文本数据为基础的应用研究中造成了很大的干扰。因此，采集 Web 页面数据后，在进行深入的分析、挖掘之前，需要对网页的类型进行筛选、过滤，去除噪音页面，保留有效的主题网页，才能确保后期深入分析研究工作的质量和效果。

当前过滤噪音网页的相关研究主要分为两方面，一方面通过提取页面特征，人工设置阈值过滤噪音网页，特征包括网页正文文字长度、标点符号数量、链接数等。这类方法虽然简单，但是能够较为有效地过滤具有明显特征的噪音网页。然而，阈值的设定依靠人工经验，由于网页格式多样化和动态性，阈值的准确性和适合度较难确定。另一方面，利用支持向量机（Support Vector Machine，SVM）、决策树等方法对网页进行分类，辨识主题网页和噪音网页，这些方法需要高质量的训练数据集和对网页特征的有效识别。基于现有研究，本文聚焦于综合两种方法：通过对噪音网页的结构和内容特征进行针对性的深入分析，首先通过网页显著特征对网页进行阈值控制，过滤具有明显噪音特征的一部分网页，缩小数据集；而后选择显著、适合的网页特征生成向量，利用 SVM 对网页进行分类，过滤噪音网页。由于同一站点的网页模板相同或相似，因此，噪音网页也具有相似的特征，通过训练分类模型，可以将其有效去除。相关研究已经应用类似方法对主题网页识别和文本抽取进行了研究[1]，本文进一步归纳和分析噪音网页的特征及其与主题网页的区别，通过生成更丰富的特征向量进行噪音网页的识别，提高分类效果。

本文首先深入分析噪音网页和主题网页的特征，进而描述噪音网页过滤方法的研究框架及具体流程，而后利用采集自 Web 的网页数据进行实验分析，最后给出研究结论和后续研究方向。

1　噪音网页及主题网页的特征分析

本文将噪音网页细分为索引页、列表页、表单页、图片页、视频页、正文文字极少的网页等。索引页指网站中起导航作用、包含大量导航链接的网页；列表页指包含带有链接标题的列表及少量描述文字的网页；图片、视频页指含较少文字、较多图片或视频的网页；表单页指用于填写、提交信息的网页。这些噪音网页和主题网页在结构和内容特征上包含很多明显的区别（见表 1），并且每类噪音网页又存在一些独特的特征，因此噪音网页的过滤基于对这些特征的分析和识别。

* 本文为国家“十一五”科技支撑计划子课题“网络科技信息监测与评价”的研究成果，项目编号：2006BAH03B05。

表 1　主题网页和噪音网页的显著特征对比

	主题网页	噪音网页
正文文字（不包含链接文字）	多	少
标点符号	多	少
链接数	少	多
链接文字数量	少	多
网页内容	内容固定、无更新	更新变化频繁
URL 目录层级	深	浅
URL 文字长度	长	短
图片、视频	少	多

噪音网页具体特征分析如下:

1）噪音网页中通常缺乏描述特定主题的文本，没有正文文字（网页解析后去噪，不包含链接的文字）或正文篇幅极短。

2）主题页中包含大段叙述文字，因此包含大量“;”“。”等标点符号；而噪音网页中包含较少成段的文字，不包含或包含较少的标点符号。

3）噪音网页中包含大量导航链接，特别是在索引页、列表页中。索引页，见图 1（a）中的新闻 Index 页，其并不描述某一新闻主题，而是新闻导航页，包括链接到多个新闻标题的锚文本。列表页，如图 1（b）检索结果页，包含多个检索结果的列表和简单描述。这类网页存在密集的导航链接数量，但其本身不具有重要的研究价值。

(a)索引页　(b)列表页　(c)表单页

图 1　三种噪音网页

4）噪音页面的 URL 在网站目录结构的层级中的位置较浅，如 http: //www. number10. gov. uk 是一级目录，通常为网站首页，属于索引页。而主题页包含固定主题，部分网页的主题能在 URL 文本中得到体现，如 http: //www. csiro. au/resources/advanced-processing-technologies. html，其目录层级位置较深，并且 URL 文字长度较长。

5）噪音网页，特别是索引页及列表页的内容更新频繁。如新闻首页中，新闻列表的标题链接会随时间的变化而发生更新，但其网页的 URL 不会发生变化。因此累计采集新页面时，同一网页只要内容发生更新就会被采集多次，其极有可能是索引页或列表页。

6）特定噪音网页的 URL 中包含较明显的特征规律，如表单页的 URL 中包含“sendto_ form”，“pre_ form”；站点地图网页的 URL 中包含“sitemap”等显著特征。

7）表单页（见图 1（c））中包含填写信息的文本框、选项栏、按钮及提交标记。包含“form”，“textarea”，“radio”，“option”，“input”等多种表单标记。

8）多数列表页结构从视觉上看是由多个文字块组成，每个文字块包括列表标题和对该标题的简短描述文字，其标题多包含“h1”，“h2”，“h3”，“h4”标记，并且标题添加了超链接，通过作为锚文本的标题可以链接到描述该标题内容的网页。如 < h2 > < a href = " /Consultations/higher-education-at-work-high-skills-high-value? cat = close-dawaitingresponse" > Higher Education at Work -High Skills: High Value < /a > < /h2 >。因此每个列表页包含多个“< hi > < ahref = ”或“< /a > < /hi >”标记，标题标记 hi 可为“h1”，“h2”，“h3”，“h4”作为标识该类网页的特征之一。

9）噪音网页中包含的图片数量相对主题页较多。

通过对噪音网页的这些显著特征的分析和与主题网页特征的对比，能够有助于对噪音网页和主题网页的有效识别和区分。

2　噪音网页过滤方法的研究思路

通过对噪音网页和主题网页的特征分析，本文采用特征阈值判断和 SVM 分类相结合的方法过滤噪音网页。首先，通过特征阈值控制过滤具有明显噪音特征的网页；而后，通过抽取基于网页结构和内容中的特征属性，生成特征向量，并结合利用训练网页生成的分类模型，通过 SVM 分类器对候选网页进行二级分类，区分主题网页和噪音网页，达到对网页数据进行噪音过滤的目的。过滤方法的研究思路见图 2。

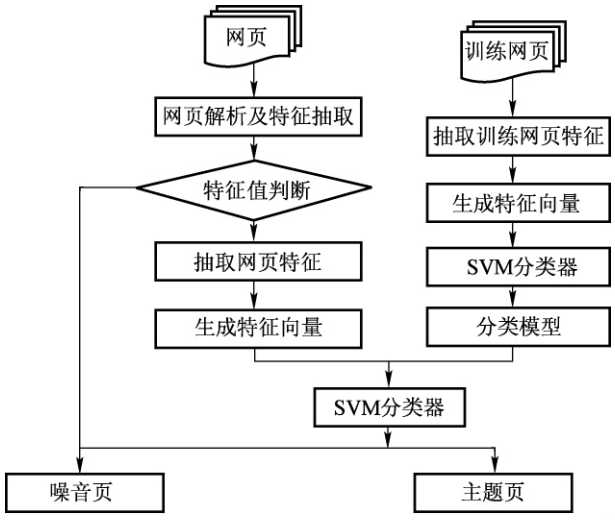

图 2　噪音网页过滤方法的研究思路

3 噪音网页过滤的具体方法

噪音网页过滤，首先需要对网页进行解析和特征抽取，笔者采用轻量级 Html 解析器 Jericho[2] 对网页进行解析和特征抽取。Jericho 与其他解析工具相比，能快速进行页面解析，识别 ASP，JSP，PSP，PHP 的多种标签结构。它没有使用 DOM 树结构和 SAX 对 Html 进行解析，不需要从根节点到叶节点的全部遍历，而是组合使用文本搜索、标签识别和定位的方法，在解析速度上比较快速。笔者利用 Jericho 对采集的网页进行解析和去噪，去除 SCRIPT 和 STYLE 等噪音元素，按“p”，“h1”等标记提取网页正文，并抽取链接文本、计算图片标记、链接数等特征。

噪音网页过滤由两个主要部分组成：①特征阈值判断。②基于 SVM 的网页分类。

3.1 噪音网页特征阈值判断

特征阈值判断考虑绝对特征数量和相对特征数量比例。绝对特征指网页正文长度、网页链接数、网页内容更新频率、URL 目录层次、URL 文本长度等绝对数量的阈值判断；相对特征指网页链接文字与正文文字数量相对比例来作为网页过滤的阈值判断条件。

噪音网页特征阈值判断具体过程如下：

1）判断噪音网页中包含的正文文字长度，设定经验阈值，少于阈值的网页为噪音网页。

2）判断 URL 层次及 URL 文本长度，滤除 URL 层级较浅及 URL 文本长度较短的网页。

3）统计网页内容更新频率，同一 URL 的网页内容如更新频繁，有可能是索引页或列表页。

4）判断网页 URL 中是否包含噪音特征，是否包含“sendto_form”，“sitemap”等明显特征，如包含，网页为噪音网页。

5）计算网页内包含的导航链接数量，如包含链接数量大于设定的经验阈值，则为噪音页面。

6）计算网页内链接文字与正文文字长度的比例。噪音网页中包含过多导航链接，其链接文字长度与正文文字长度比例远大于主题网页中的文字比例，如比值大于设定阈值则判断为噪音页面。

特征阈值判断的方法比较简单，但是较为有效，能够较好地过滤出大部分噪音特征明显的网页，特别是具有突出规律性特征的噪音网页，并且在一定程度上缩小了数据集。但是由于阈值设定是经验性的，选取的阈值要适度，否则将滤出较多主题页。因此，第二步是对初步去噪后的网页通过识别更丰富的特征生成特征向量，利用 SVM 分类器进行分类过滤。

3.2 基于 SVM 的网页分类

SVM 是 Vapnik 等根据统计学习理论的结构风险最小化原理提出的一种机器学习方法[3]。它具有最优分类能力和最广泛化能力，解决了传统算法面对的局部最优以及过拟合等问题。SVM 的主要思想是通过选择适合的核函数，将输入向量映射到高维空间，并在高维空间中构建最优分类超平面。本文选择 Libsvm 分类软件进行分类[4]，具体步骤如下。

1）特征向量选择。SVM 数据输入的是表示网页的特征向量模型，通过对噪音页面和主题页面的特征对比和深入分析，笔者选择 18 个较为突出的特征属性作为网页分类的特征向量（见表 2）。

表 2 网页分类特征选择

序号	特征	序号	特征
1	原始文本长度（网页源文本）	10	带链接的 < h1 > 标记数
2	正文文本长度	11	带链接的 < h2 > 标记数
3	标点数	12	带链接的 < h3 > 标记数
4	链接数	13	带链接的 < h4 > 标记数
5	链接文本长度	14	< Input > 标记数
6	图片数	15	< textarea > 标记数
7	URL 文本长度	16	< select > 标记数
8	URL 目录层级	17	< opition > 标记数
9	正文文本长度/（正文文本长度 + 链接文本长度）	18	< radio > 标记数

按照数据格式生成特征向量矩阵（见图 3），每行表示一个网页，第一列表示人工分类，1 表示主题网页，0 表示噪音网页，其余列为每个网页的特征属性，如该属性存在则以 < 特征序号：数值 > 形式输出，否则无须输出跳到下一个属性，保证输出矩阵除第一列外为非 0 值矩阵。

```
0.0 1:6471.0 2:46579.0 3:4229.0 4:9.0 5:165.0 6:0.5641 7:176.0 11:4.0 12:49.0
0.0 1:5416.0 2:44681.0 3:3689.0 4:9.0 5:162.0 6:0.5601 7:167.0 11:4.0 12:49.0
1.0 1:2422.0 2:27691.0 3:1304.0 4:9.0 5:154.0 6:0.6064 7:90.0 11:10.0 12:41.0
1.0 1:3110.0 2:29132.0 3:1443.0 4:9.0 5:139.0 6:0.6447 7:91.0 11:10.0 12:41.0
1.0 1:3513.0 2:29361.0 3:1362.0 4:9.0 5:155.0 6:0.6844 7:92.0 11:10.0 12:41.0
1.0 1:4097.0 2:29429.0 3:1298.0 4:9.0 5:159.0 6:0.7286 7:90.0 11:10.0 12:41.0
1.0 1:8957.0 2:35437.0 3:1393.0 4:9.0 5:230.0 6:0.8449 7:91.0 11:10.0 12:41.0
1.0 1:3531.0 2:29586.0 3:1417.0 4:9.0 5:169.0 6:0.6773 7:90.0 11:10.0 12:41.0
0.0 1:4156.0 2:41058.0 3:2690.0 4:9.0 5:230.0 6:0.5767 7:146.0 9:1.0 10:2.0 1
1.0 1:2932.0 2:157488.0 5:1032.0 6:1.0 12:59.0 13:2.0
1.0 1:2883.0 2:5464.0 5:36.0 6:1.0 12:59.0 13:2.0
1.0 1:11318.0 2:202822.0 5:1581.0 6:1.0 12:59.0 13:2.0
1.0 1:9274.0 2:16291.0 5:121.0 6:1.0 12:59.0 13:2.0
1.0 1:1815.0 2:4098.0 5:19.0 6:1.0 12:59.0 13:2.0
1.0 1:11382.0 2:203061.0 5:138.0 6:1.0 12:59.0 13:2.0
1.0 1:7298.0 2:68523.0 5:327.0 6:1.0 12:59.0 13:2.0
0.0 1:3386.0 2:40663.0 3:723.0 4:22.0 5:740.0 6:0.7982 7:84.0 10:8.0 12:41.0
```

图 3 网页特征向量片断

2）训练集准备和数据归一化。通过人工分类主题网页和噪音网页构建训练数据集，其中部分噪音网页来自阈值判断阶段反馈的噪音网页。Libsvm 提供简单的数据缩放功能，利用 svmscale 对训练数据进行缩放，目的是：①避免特征值范围过大而另一些特征值范围过小。②避免训练

时为了计算核函数而计算内积的时候引起数值计算困难[5]。将数据缩放到 [-1，1] 或 [0，1] 之间，本文选择缩放到 [-1，1] 之间。

3) 网页分类器的核函数选择及设置。SVM 准确性关键在于核函数的选择和参数的设置，Libsvm 中包含线性核函数、多项式核函数、径向基核函数和 sigmoid 核函数，本文使用多项式核函数 = $(\gamma^* u^{*} v + coef0)^{degree}$，因为在实验中，它与其他核函数相比具有较好的全局性质，其生成的模型在实施分类中具有较高的准确率。

4 实验分析

实验数据利用采集工具 Nutch 从 Web 采集来自 24 个科研机构站点的 3 000 篇包含科技政策、新闻、报告等内容的网页，对其进行手工分类，选择 1 500 篇作为数据集，其中 500 篇作为训练集，1 000 篇作为测试集，训练集和测试集的手工分类情况见表 3。

表 3 数据集手工分类情况

	训练集	测试集
主题页	200	600
噪音页	300	400

实施本研究的噪音网页过滤，通过第一步的阈值判断，滤出 127 个噪音页面，剩余 873 个网页，其中发现对 URL 层次、URL 文本长度、正文长度进行阈值控制，能够较为有效地过滤出噪音特征明显的网页。而后利用训练模型对 873 个网页进行 SVM 分类，根据分类结果过滤噪音网页。所有噪音网页的过滤结果评价指标选择精确率、召回率和 F-measure 来计算，公式为:

精确率 = 正确分入该类的网页数/所有分入该类的网页数 × 100%

召回率 = 正确分入该类的网页数/实际应分入该类的网页数 × 100%

F-measure = 2 × 精确率 × 召回率/（精确率 + 召回率)

表 4 网页过滤结果中主题网页与噪音网页的分类结果

	主题网页	噪音网页
准确率（%）	86. 98	95. 43
召回率（%）	93. 50	90. 67
F-measure（%）	89. 98	92. 99

通过实验结果（见表 4) 可以看出，本方法能够较好地滤除噪音网页，获得较高的准确率和召回率。此外，在实验中为观察本文增加的特征向量对分类结果的影响，在分类时对同样的训练集和测试集数据去掉表 2 中的后 8 个特征，生成特征向量并进行训练和分类。分类结果与之前对比（见表 5)，对噪音网页和主题网页识别的准确率和召回率都有较大下降。本文增加的 8 个特征，主要是对列表类网页和表单型网页的标识特征，实验结果表明它们能够有助于对主题网页和噪音网页的分类。

表 5 选取不同特征分类的对比结果

	18 个特征向量		10 个特征向量	
	主题页	噪音页	噪音页	主题页
准确率（%）	86. 98	95. 43	73. 36	77. 75
召回率（%）	93. 50	90. 67	84. 05	78. 17
F-measure（%）	89. 98	92. 99	78. 34	77. 56

5 结束语

在面对以 Web 页面数据为基础的应用研究中，噪音网页形成了较大的干扰。笔者深入分析了噪音网页和主题网页的特征，通过页面特征识别，对各特征设置阈值判断能够有效去除具有明显特征的噪音网页，缩减数据集。而后，利用显著特征生成更丰富的特征向量，利用 SVM 分类器对网页进行分类，过滤噪音网页，研究取得了一定的效果。研究中采用人工设置特征阈值，阈值设定尚需人工经验，下一步工作将继续研究更科学、更适合的方法来改进阈值设定方式。实验中，当训练集过大时，训练模型比较耗时，因此，未来将进一步考虑训练模型过程的优化并采用更大数据集进行测试。□

参考文献

[1] 蔡捷飞，陈啟泓，梁志宏，等. 主题型网页发现以及网页内信息块发现 [EB/OL]. [2010-07-17]. http://net.pku.edu.cn/~webg/cwt/2008WebTrack/result/task1_quark/SE-WM2008_SCUT1.ppt.

[2] Jericho HTML Parser [EB/OL]. [2010-07-13]. http://jericho.htmlparser.net/docs/index.html.

[3] VAPNIK V N. The nature of statistical learning theory [M]. New York: Springer，1995: 188.

[4] Libsvm [EB/OL]. [2010-07-15] http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[5] CHIH CHUNG C，CHIH JEN L. LIBSVM: a library for support vector machines [EB/OL]. [2010-07-18]. http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf.

作者简介：吴思竹，女，1981 年生，博士生。

张智雄，男，1971 年生，研究馆员。

收稿日期：2010 - 11 - 30