基于规则的网络文本资源标题快速自动识别方法*

刘建华[1] 张智雄[1] 谢 靖[1] 邹益民[1,2]
[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)

【摘要】选取网络文本资源的标题识别作为切入点,除考虑多数研究关注的文本的格式信息(如字体)、位置信息等特征外,加入对标题与网页正文内容的相关度的考虑,利用科技监测项目采集到的大量历史数据作为统计分析的基础,从候选标题的可能来源和特征方面,构建基于规则的网络文本资源标题快速识别方法,并给出该方法的时间效率和识别准确率测评结果。

【关键词】网络文本资源 标题识别 标题来源 标题特征
【分类号】G203

Automatic Identify Title of Web Text Resource Based on Rules

Liu Jianhua[1] Zhang Zhixiong[1] Xie Jing[1] Zou Yimin[1,2]
[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[2](Craduate University of Chinese Acadeny of Sciences, Beijing 100049, China)

【Abstract】As the important role of titles of Web resource for information retrieval, text cluster and so on, this paper proposes a method to identify the titles automatically and quickly based on the style information(such as font) and location information of text which are used by many other researchers. Besides, it considers the relevance between the title candidates and text content. Lastly, this paper implements the title identification component and does some experiments to show the effectiveness of this method.

【Keywords】Web text resources Title identification Title source Title feature

1 引 言

随着科技水平和网络传播方式的迅猛发展,在正式的科技论著产量急剧增长的同时,越来越多的管理机构、科研机构等也选择网络这类非正式交流平台发布、共享相关的科研进展新闻、重大研究成果或研究报告、年度预算等。与正式交流渠道(如期刊、著作等)发布的研究成果相比,网络平台发布的各类文本资源信息具有以下特征:涉及内容范围广,信息资源丰富;动态变化更新及时,时效性强;载体形式丰富(包括HTML、PDF、DOC、PPT等各种文本类型),无法用统一的方法进行处理;各类描述性元数据信息不完整[1],如标题、创建时间、作者等。

为了使庞杂无序的网络文本资源便于机器理解,从而进一步推动信息检索、文本聚类、主题监测等工作,自动有效地识别获取网络文本资源的相关元数据信息非常必要。本文选取文本的标题识别作为切入点,着重探索HTML、

收稿日期:2011-05-05
收修改稿日期:2011-05-26
*本文系中国科学院资助项目"科技机构自动监测服务系统"的研究成果之一。

PDF、DOC 及 PPT 等几类文本类型网络资源的标题识别方法。此外,由于不同语种在自然语言表现上各有特点,本文目标是英文文本,即相关研究针对英文文本展开。

2 标题识别的相关研究

标题标识了文章的主要内容,即文章精要内容的提炼、概括与浓缩。基于此,自动识别文本中的标题吸引了不少研究者的目光。在近年的相关研究中,文本标题的自动识别主要包括两类:对科技论文标题的识别(主要是 PDF 格式的文本)和对网络文本标题的识别。前者由于其格式、载体类型、文档内容结构等相对固定、单一,因此在方法上可以基于简单的模式规则,如 Giuffida 等开发的基于规则从科技论文中识别出其标题的系统,主要采用的规则有"标题出现在首页的最上面"、"标题的字体是文本中最大的"等[2]。为了提高方法的适应性,也有研究人员提出了使用机器学习的方法来获取科技论文中的标题,如 Peng 等利用条件随机场的方法来获取科技论文中的标题[3]。另外,朱海军等还依靠科技论文的结构、数字标识等特征,针对科技论文中的小标题提出了基于特征词的识别方法,较有效地获得了科技文献的小标题[4]。尽管上述的这些方法在科技论文的标题识别中能发挥较好的作用,但针对格式不规范、内容结构多变、来源复杂的网络文本,这些方法还需要进行调整。目前,网络文本标题的识别研究主要分为两类:

(1)从科技论文标题的识别进行延伸,探索普适的文档(主要是 PPT、PDF、DOC 等)标题识别的方法。如 Hu 等利用文档的格式信息,构建机器学习模型[5],从而实现通用文档的标题识别方法。

(2)围绕 HTML 页面的标题提取展开的相关研究。在这方面的研究比较多,常用的方法主要集中于基于规则和机器学习两类,如 Xue 等从 HTML 节点的位置信息、内容信息和视觉信息等方面的特征入手,构建机器学习模型识别网页标题[6]。朱青等也做了相应的研究探索[7]。但上述的两种方法主要考虑的还是 HTML 标签的统计特征信息,却没有考虑到标题与网页正文内容的相关性,李国华等从这一点入手,通过计算句子与句子之间的相似度,从而确定网页的标题[8]。

综合分析当前对标题提取的各类研究,可以发现以下几个特点:

(1)文本的格式信息(如字体)、位置信息等是众多研究关注的重点,而对标题与网页正文内容的相关度关注比较少;

(2)当前的研究主要是对离线数据的处理,绝大多数集中于从文本内部的内容中确定标题,在提取的准确性方面都有比较好的评测结果,但在时间效率方面却没有给出有效的评测结果,而如 Google 一类的搜索引擎,在采集过程中即综合采用锚点文本内容(anchorText)和 < title > 标签,以文本的属性信息作为标题的来源,时间效率比较高,但准确性相对较低。

本文基于现有的相关研究成果,通过充分分析,利用网络文本资源的标题来源、标题特征等信息,探索一套快速有效的网络资源标题提取方法。

3 网络文本资源标题的来源及特征分析

当前的网络文本资源主要是两类:纯网页类型的文本资源和 PDF、DOC、PPT 等格式的富文档文件[9]。这两类资源的标题来源和特征有相同之处,也有各自的特点。通过对科技监测项目采集到的数万条资源数据进行分析,笔者得到了相应的分析结果。本文的目标对象是英文资源,因此,所有的分析也仅针对英文资源进行。表 1、表 2 中分别列出了这两类资源的标题来源和特征。

表 1 网络文本资源标题的来源分析

来源 \ 资源类型	纯网页资源	富文档文件
链入页面上的锚点文字	有	有
URL 字段中标出的标题内容	有	有
< title > 标签中的标题内容	有	无
< meta > 标签"title"属性值	有	无
< h1 > 、< h2 > 等标签值	有	无
文本属性中标题属性值	无	有
正文内容中顶部最大字体值	有	有

依据对标题来源和特征的分析,可以进一步利用标题的特征构建相关的规则,从候选的标题(不同来源的标题)中快速确定出最符合文章内容的标题。

4 网络文本资源标题的提取方法

在网络文本资源标题的提取过程中,主要涉及以下三个阶段:标题特征规则撰写、候选标题获取和候选标题规则匹配和修正。

表 2　网络文本资源标题的特征分析

特征		纯网页资源	富文档文件
位置特征	在内容页的顶端	是	是
	不可能出现在边角区域	是	是
形式特征	是内容中最大的字体或者在颜色、字形上与其他内容有所区别	是	是
	不包含较多空行或空格	是	是
内容特征(此类特征中的内容并不是并列出现在实际应用筛选条件中,在实际条件组合时,可能仅需用到其中一个或几个)	通常不包括逗号、句号等标点符号	是	是
	不会少于 3 个分词,也不会超过 30 个分词(包括副标题)	是	是
	不包含“download”等停用词	是	是
	除介词、连接词外,其他单词首字母为大写或全为大写	部分不是	是
	在标题中出现的词或词组会出现在正文中	是	是

4.1　标题特征规则撰写

标题特征规则的撰写过程实际上是决定什么样的内容符合标题的过程,也就是依据资源标题的特征进行相应的匹配、对照。尽管 DOC、PPT、PDF 等富文档文件与 HTML 等网页文档的候选标题来源有所不同,但其标题的特征基本相同,因此其相关的规则除在来源方面有所区别外,主要的特征基本相同。

(1)纯网页资源的标题特征规则

在网页文档方面,通过对科技监测项目采集到的数万条资源数据进行分析,笔者发现:

①来源于 < meta name = " title" > 字段的标题元数据与文档标题匹配的准确率非常高,通过对监测项目中有类似元数据字段的网站进行抽样分析,该字段值为文档标题的准确率达到了 100%,尽管拥有此类元数据的网站在科技监测项目中不到 10%。

②链入页面上相应的链接文字(即锚点信息)与文档标题匹配的准确率也非常高,如图 1 所示:

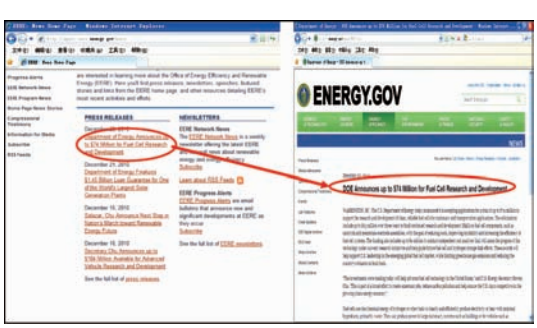


图 1　链接文字与标题的关联

在 http://apps1. eere. energy. gov/news/页面的新闻列表中给出的链接文字即对应链接如 http://energy. gov/news/9923. htm 的新闻标题。但链接文字也可能是“full text”、“download pdf”等一类的共性词,因此,在利用链接文字时需要构建相关的停用词库和停用规则。

③最广为人知的网页文档页面上 < html > 标签内的文本在实际的统计评价中真正能成为文档标题的准确率却不高。因此,该来源的标题重要程度较低。

按照不同来源对网页标题的贡献准确程度,将其进行排序,具体顺序为: < meta name = " title" > 、anchorText、< title > 、URL、< h1 – h6 > 。在实际计算中,同时考虑候选标题的词长(主要指实词的个数)、是否包含停用词、候选标题中的 n 元词组(不含介词、副词等词)在正文中出现的情况几个方面的因素,形成一系列纯网页资源标题判定规则。鉴于分析中发现 < meta > 元数据段和 anchorText 内容对标题的高准确度指示情况,在实际处理过程中,仅对 < title > 标签值、URL、< h1 – h6 > 几个来源的候选标题进行 n 元切分处理以获取其相应的 n 元词组,从而计算 n 元词组在正文中的出现情况。另外,为了避免这种排除法带来候选标题的漏选问题,在经过层层排除还无法确定标题时,可基于候选标题中 n 元词组在正文中出现的频次来确定最终标题。具体判定规则如下:

Declare 网页标题 title;
Declare 停用词表 stopWordSet;
Declare 停用标号集 stopTabSet;
Declare 候选标题的 n 元词组分别为 tagTitleGram,urlGram,tagHGram;
If (存在 < meta name = " title" content = **** > && content 值不为空,且 content 内容长度不小于 10 字符){
　Title = < meta name = " title" content = **** > 中 content 里的内容;
} else If (存在 < meta name = " subject" content = **** > && content 值不为空,且 content 内容长度不小于 10 字符){
　Title = < meta name = " subject" content = **** > 中 content 里的内容;
} else If (anchorText 非空,其分词个数(实词)不少于 3 个且不在 stopWordSet 中){
　Title = < meta name = " subject" content = **** > 中 content 里的内容;
}else If (< title > 标签非空,其分词个数(实词)不少于 3 个且不在 stopWordSet 中,同一个站点下的不同网页标题不同,tagTitleGram 在正文中有出现){
　Title = < title > 标签值;
} else If (< h1 ~ h6 > 标签值循环处理,hn 标签值不为空,其分词个数(实词)不少于 3 个且不在 stopWordSet 中,不包含

stopTabSet,tagHGram 在正文中有出现){

Title = <h1 ~ h6 > 中最符合规则的标签值;

} else If (url 最后一段中取出的不含非正常字符的分词个数(实词)不少于 3 个且不在 stopWordSet 中,urlGram 在正文中有出现){

Title = url 最后一个段落中替换掉非正常字符后的内容;

}else{

title 取 tagTitleGram、tagHGram、urlGram 中在正文中出现频次最多的对应候选标题;

}

(2)富文档文件的标题特征规则

与纯网页资源不同,富文档文件的标题没有相关的标签值可以解析,但是其标题的特征与纯网页资源有共通之处,因此,参考纯网页资源的标题匹配规则,也可以形成类似的富文档文件的标题特征。具体可参见文中的匹配规则,进行相应的修改。

4.2 候选标题的获取

从表 1 中可以看到,候选标题的来源是非常丰富的,不同来源的标题其获取方法也有所不同。具体而言,这些来源可以分为 4 类进行解析获取:

(1)采集过程中进行解析获取

主要包括链入页面上的锚点文字、URL 字段中标出的标题内容两个来源。对于前者,可以通过采集器在采集过程中将链入、链出页面的关系和相应的锚点文字信息进行记录,从而获取到相应的候选标题。对于后者,采集器记录下当前采集页面的 URL,撰写相应的解析器,获取最后一个"/"后的所有内容(若 URL 以"/"结尾,则取倒数第二个"/"后所有内容),并替换其中的非字符编码等内容。

(2)对 HTML 页面编码进行解析,获取相应候选标题

这种方法主要针对纯网页资源。在解析过程中,可以借助 HTMLParser[10] 或 Jericho[11] 等分别解析获得 < title > 、< meta > 中的"title"属性、< h1 - h6 > 标签的值,也可以综合加入 < font > 、< strong > 等描述字体、字形等的特征标签对结果进行过滤。

(3)对文本属性进行解析,获取相应候选标题

这种方法主要针对 DOC、PPT、PDF 等富文档文件,目标是要从其相关的文本属性信息中解析获得相应的标题信息。PDF 和 PPT 两个类型的富文档文件的标题文本属性源,如图 2 所示。

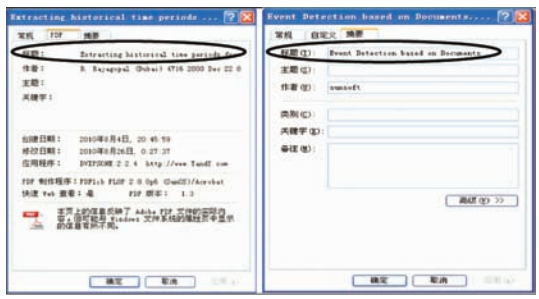


图 2 富文档文件的标题文本属性源

针对这类信息,可以分别利用 Apache PDFBox[12] 和 Apache POI[13] 两个开源工具进行相应的处理和获取。但是这一来源的信息往往较少,多数情况下该来源值为空。

(4)对正文内容进行解析,获取正文顶部最大字体值

无论是 DOC、PPT、PDF 等富文档还是 HTML 页面,正文顶部最大字体且无结束标点符号(如句号等)的单独成段文字通常可能是文档标题。对于 DOC、PPT、PDF 等富文档文件而言,可以分别利用 Apache PDFBox 和 Apache POI 解析获取其首页的内容并从首页内容中按行取出字体中的最大字体内容。对 HTML 而言,则可以内容块判定的方法确定出主体内容最大的区域,在该区域中依据 < font > 等属性特征筛选出最大内容。

4.3 候选标题规则匹配与修正

获选标题规则匹配的过程即按照拟定的标题筛选规则,从候选标题中筛选出最符合规则描述的标题资源作为对应网络文本资源的标题。在规则匹配阶段,需要先对候选标题进行相应的预处理,包括 N - gram 分词、特殊字符的转换、多空格的替换、换行的替换等处理。

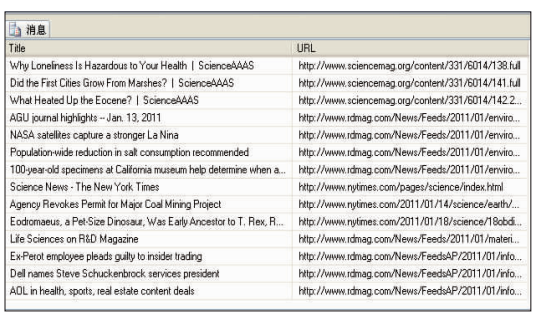
5 网络文本资源标题识别的实现与实验

基于网络文本资源标题的识别方法,笔者利用 HTMLParser、Apache PDFBox 和 Apache POI 等开源工具构建了相应的标题识别组件。通过该组件在实际的网络监测系统中的应用,分析测试统计了其时间效率、识别准确率。在计算准确率的过程中,主要通过随机抽样分析,其中富文档文件随机抽取了 1 000 篇,纯网页资源随机抽取了 1 500 篇,通过人工比对的方式筛选出识别标题与人工所判定的标题相吻合的记录。在实际比对中,如果在自动识别出的标题中包含了一些小的错误,如没

有完整识别出文档标题中的副标题、包含了一些特殊符号“*”等,这样的抽取结果也认为是正确的。具体的测试结果和抽取的样例分别如表 3 和图 3 所示:

表 3　网络文本资源标题识别效率测试结果

资源类型	样本总数	时间效率	识别准确数	识别准确率
富文档文件	1 000 篇	0.15s/篇	852 篇	85.2%
纯网页资源	1 500 篇	0.05s/篇	1 271 篇	84.7%


消息

Title	URL
Why Loneliness Is Hazardous to Your Health | ScienceAAAS	http://www.sciencemag.org/content/331/6014/138.full
Did the First Cities Grow From Marshes? | ScienceAAAS	http://www.sciencemag.org/content/331/6014/141.full
What Heated Up the Eocene? | ScienceAAAS	http://www.sciencemag.org/content/331/6014/142.2...
AGU journal highlights -- Jan. 13, 2011	http://www.rdmag.com/News/Feeds/2011/01/enviro...
NASA satellites capture a stronger La Nina	http://www.rdmag.com/News/Feeds/2011/01/enviro...
Population-wide reduction in salt consumption recommended	http://www.rdmag.com/News/Feeds/2011/01/enviro...
100-year-old specimens at California museum help determine when a...	http://www.rdmag.com/News/Feeds/2011/01/enviro...
Science News - The New York Times	http://www.nytimes.com/pages/science/index.html
Agency Revokes Permit for Major Coal Mining Project	http://www.nytimes.com/2011/01/14/science/earth/...
Eodromaeus, a Pet-Size Dinosaur, Was Early Ancestor to T. Rex, R...	http://www.nytimes.com/2011/01/18/science/18obdi...
Life Sciences on R&D Magazine	http://www.rdmag.com/News/Feeds/2011/01/materi...
Ex-Perot employee pleads guilty to insider trading	http://www.rdmag.com/News/FeedsAP/2011/01/info...
Dell names Steve Schuckenbrock services president	http://www.rdmag.com/News/FeedsAP/2011/01/info...
AOL in health, sports, real estate content deals	http://www.rdmag.com/News/FeedsAP/2011/01/info...

图 3　网络资源标题提取样例

6　结　语

本文主要从资源候选标题来源和资源特征两个方面入手,充分考虑其内容特征(包括与正文的相关度),构建相应的规则来实现快速而准确的网络文本资源的标题识别。通过实验结果发现,该方法在时间效率和识别的准确率上都有不错的表现,为其他后续深度的文本分析提供了较好的支持。但是,通过实验也发现,在实际的应用中,其准确率还有待改进的方面,而基于规则的方法是基于对样例数据进行了大量的统计而形成的,在遇到目前没有考虑到的特征时,识别效率就会受到影响。而且目前的规则中采用的排他性的优先级选择方法对识别准确率的影响程度也未经过实际数据的证明,这些问题都需要在实际应用中总结发现识别错误的原因和规律,从而进一步完善网络文本资源的自动识别方法。

参考文献:

[1] Changuel S, Labroche N, Bouchon - Meunier B. A General Learning Method for Automatic Title Extraction from HTML Pages[C]. In: *Proceedings of the 6th International Conference on Machine Learning and Data Mining in Pattern Recognition*. 2009: 704 - 718.

[2] Giuffrida G, Shek E C, Yang J. Knowledge - based Metadata Extraction from PostScript Files[C]. In: *Proceedings of the 5th ACM Conference on Digital Libraries*. 2000:77 - 84.

[3] Peng F, McCallum A. Accurate Information Extraction from Research Papers Using Conditional Random Fields[C]. In: *Proceedings of the Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics Annual Meeting*. 2004:329 - 336.

[4] 朱海军,张桂平,蔡东风,等. 科技论文的标题识别[C]. 见:*第九届全国计算语言学学术会议论文集*,2007.

[5] Hu Y, Li H, Cao Y, et al. Automatic Extraction of Titles from General Documents Using Machine Learning[J]. *Information Processing and Management*, 2006, 42(5):1276 - 1293.

[6] Xue Y, Hu Y, Xin G, et al. Web Page Title Extraction and Its Application[J]. *Information Processing and Management*, 2007, 43(5):1332 - 1347.

[7] 朱青,吕晓旭. 基于机器学习的 HTML 标题抽取[J]. *微计算机信息*,2010,26(3):15 - 16,11.

[8] 李国华,眷红英. 基于语句相似度的网页标题抽取方法[J]. *中文信息学报*,2011,25 (2): 32 - 37.

[9] Open Document Format for Office Applications (OpenDoeument) v1.0 [EB/OL]. [2011 - 02 - 10]. http://docs.oasis - open.or ~ oficdv1.0.

[10] HTML Parser[EB/OL]. [2011 - 03 - 10]. http://htmlparser.sourceforge.net/.

[11] Jericho HTML Parser[EB/OL]. [2011 - 03 - 10]. http://jericho.htmlparser.net/docs/index.html.

[12] Apache PDFBox - Java PDF Library[EB/OL]. [2011 - 03 - 20]. http://pdfbox.apache.org/.

[13] Apache POI - Text Extraction [EB/OL]. [2011 - 03 - 20]. http ://poi.apache.org/.

(作者 E - mail:liujh@mail.las.ac.cn)