一种基于文本相似度的网页新闻标题自动抽取算法

何春辉
(湘潭大学 数学与计算科学学院，湖南 湘潭 411105)

摘 要：随着互联网技术的发展，网页新闻的标题抽取已经成为了信息抽取和网络爬虫中不可避免的一个环节. 通过分析，发现目前已有的方法存在准确率和通用性无法共存的问题. 因此，提出了一种基于文本相似度的网页新闻标题自动抽取算法，它通过结合目录型新闻网页的外部标题来抽取详情型新闻网页的真实标题. 试验结果表明，相对现有方法来说，新算法具有较好的通用性且平均*F*1值达到了97.58%.

关键词：网络爬虫；新闻标题抽取；调和相似度
中图分类号：TP391 **文献标识码：**A doi:10.3969/j.issn.1672-7304.2019.01.0011
文章编号：1672-7304(2019)01-0058-04

An Automatic Extraction Algorithm for Web News Headlines Based on Text Similarity

HE Chunhui
(School of Mathematics and Computational Sciences, Xiangtan University, Xiangtan, Hunan 411105, China)

Abstract: With the development of the Internet technology the title extraction of web news has become an inevitable part of information extraction and web crawling. By means of analysis it is found that the existing methods have problems that the accuracy and versatility cannot coexist. Therefore, an automatic extraction algorithm for the web news headlines based on text similarity is proposed, which extracts the real title of the detailed news web page by combining the external title of the catalogue news web page. The results show that compared with the existing methods it is universal and the average *F*1 value of algorithm has reached to 97.58%.

Key words: web crawler; news headlines extraction; harmonic similarity

信息抽取[1]的核心目标是从大量的载体中快速准确地抽取出对用户有价值的少量信息. 随着Web技术的发展，人类的日常生活方式已经发生了巨大的变化. 这些变化使得许多传统的纸质载体都被电子载体取而代之，许多信息都通过互联网上的网页来传递和展示，而网页新闻的标题通常会被认为是一张网页的“眼睛”，它能较好地揭示网页主题信息，因此，对网页新闻的真实标题进行自动抽取是一项非常有意义和有挑战性的任务. 考虑到目前学术界和工业界还没有完全成熟的通用解决方法，故本文提出了一种基于文本相似度的网页新闻标题自动抽取算法，它可以准确地抽取网页新闻真实标题，为网页新闻标题的抽取提供了新途径.

1 研究背景

1.1 相关文献

对于网页标题的抽取，国内外已有一些相关的研究成果，大致可分为以下几大类：

1)基于网页标签信息的标题抽取[2-3]. 此类方法依赖于HTML中的标签对，它对早期简单网页标题的抽取效果较好，但由于CSS技术的发展，新网页中加入了大量的新元素，这会影响网页标题抽取的准确性.

2)基于规则的标题抽取[4-5]. 此类方法通过制定大量的规则模板来抽取相关信息. 虽然抽取准确率比较高，但是需要大量的人力来维护规则和模板，因此只适用于少量的网页信息抽取任务，并不具备大规模网页标题抽取的能力.

3)基于网页内容的标题抽取[6-9]. 此类方法先是通过分析网页中的文本内容来挖掘出一些特征，然后结合这些特征来抽取网页的标题. 这类方法一般依赖 NLP 的底层处理技术和相似度计算算法，需要标注大量的语料来训练用于特征抽取的模型，比较依赖标注语料，可移植性差，且存在一定的误判率.

收稿日期：2018-09-25
作者简介：何春辉(1991-)，男，湖南永州人，工程师，硕士，主要从事数据挖掘及信息处理研究. E-mail: xtuhch@163.com

4)基于 DOM 树的标题抽取[10-11]. 此类方法主要是通过获取网页的 HTML 源码来构建 DOM 树，然后深度遍历 DOM 树节点，来提取网页标题. 这类方法的优点是对特定类型的网页标题抽取准确率较高，但其通用性较差.

5)基于机器学习的标题抽取[12-14]. 此类方法一般会根据 HTML 源码的标签特征以及网页内容等多个角度来构造机器学习的特征，然后利用机器学习算法来自带抽取网页标题. 这类方法准确性比较高，但它对特征构造和机器学习算法具有一定的依赖性.

针对现有方法在网页新闻标题抽取上存在准确性和通用性无法共存的问题，本文提出了一种通用网页新闻标题自动抽取算法，该算法既具有通用性，又能保证抽取准确性.

1.2　新闻网页结构分析

通过分析现有主流新闻网页的构成情况，发现它们一般是由目录型[15]和详情型[16]新闻网页构成. 目录型新闻网页样例中主要包含了网站的栏目信息、一些新闻列表的外部标题和其摘要信息，以及广告信息，并不会包含网页新闻详细内容；而详情型新闻网页样例中主要包含了某一篇新闻的详细信息，即包括新闻真实标题、发布时间、来源、作者、新闻正文及广告和推荐信息.

这 2 种类型的网页存在着映射关系，目录型新闻网页的列表中某一条新闻的资源定位符(URL)和对应的标题，会唯一地映射到一个详情型新闻网页. 考虑到目录型新闻网页的标题带有噪声，例如有时会出现一个 URL 对应多个外部标题的情况，因此在抽取新闻标题时，首先会考虑从详情型新闻网页中抽取真实标题，而不会直接从目录型新闻网页中抽取外部标题作为新闻的真实标题，这样处理可以减少噪声，提高标题抽取的准确性.

1.3　字符串最长公共子序列长度及相似度求解

2 个字符串之间的最长公共子序列可以说明 2 个字符串之间是否有公共交集，这样可从字符层面说明字符串之间是否具有相关性，它可以有效地过滤噪声并找出相关联的信息. 因此，引入字符串的最长公共子序列作为起始行块和结束行块的定位判别准则.

对于 2 个字符串 $I=(s_1, s_2, \dots, s_i)$和$J=(t_1, t_2, \dots, t_j)$之间的最长公共子序列长度[17](也简称 LCS)的计算如式(1)所示.

$$\mathrm{LCS}[I,J]=\begin{cases}0, \ \text{if}(i=0)\text{or}(j=0);\\ \mathrm{LCS}[i-1,j-1]+1, \ \text{if}(i,j>0,\mathrm{s}_i=t_j);\\ \max\{\mathrm{LCS}[i,j-1],\mathrm{LCS}[i-1,j]\}, \ \text{if}(i,j>0,s_i\neq t_j).\end{cases} \tag{1}$$

其中 LCS[*I*, *J*]表示 2 个字符串之间的最长公共子序列的长度，是一个非负整数.

利用式(1)得到 LCS 值之后，就可以结合字符串自身的长度信息来计算 2 个字符串之间的相似度，其具体计算方法如式(2)所示.

$$\mathrm{Sim}(I,J)=\frac{2*\mathrm{LCS}[I,J]}{(\mathrm{len}(I)+\mathrm{len}(J))}, \tag{2}$$

其中，Sim(*I*, *J*)表示 2 个字符串 *I* 和 *J* 之间的调和相似度；LCS[*I*, *J*]表示字符串 *I* 和 *J* 之间最长公共子序列的长度；len(*I*)和 len(*J*)分别表示字符串 *I* 和字符串 *J* 自身的长度.

2　网页新闻标题自动抽取算法

根据新闻网页结构的分析结果，考虑到目录型新闻网页和详情型新闻网页的连接桥梁是同一条新闻的 URL 相同. 因此，可以利用 URL 来结合这 2 种类型的新闻网页. 基于上述分析，提出了通用网页新闻标题自动抽取算法.

该算法具体实施过程分为 2 个阶段. 第 1 阶段需要获取某个指定的新闻网站目录页 URL 的 HTML 源码，利用正则表达式和 DOM 树的 CSS 元素选择器来抽取该页面上所有新闻列表的外部标题 external_Title$_i$ 和对应的 url$_i$. 算法第 1 阶段具体实施步骤如下：

1)首先输入一个目录型新闻网页的 URL，并初始化一个容器 News_List；

2)通过 URL 获取 HTML 源码，结合 DOM 树节点和正则表达式抽取出新闻列表中所有新闻的外部标题 external_Title$_i$ 和对应的 url$_i$；

3)将抽取的结果依次存到容器 News_List 里面，保存备用.

算法第 2 阶段具体实施步骤如下：

1)从 News_List 里依次取出第 1 阶段得到的某一条具体新闻的 url$_i$；

2)将该 url$_i$ 对应的外部标题 E_Title$_i$ 加入到该新闻对应的候选标题列表中；

3)由 url$_i$ 获取详情型新闻网页 HTML 源码；

4)结合 DOM 树和 CSS 元素选择器对获取到的 HTML 源码进行元素定位，并依次取出该网页中所有<h*></h*>标签所对应的内容，再加入到候

选标题列表中；

(5)取出详情页中<title></title>对应的标题内容 M_Title，如果该内容为空，就将 url_i 对应的外部标题 E_Title_i 赋值给它，否则就将它本身的内容作为 M_Title；

(6)利用式(1)和式(2)依次计算 M_Title 与候选标题列表中所有候选标题的相似度；

(7)对相似度做降序排序，取出相似度最大的候选标题作为这个 url_i 的真实新闻标题.

3 验证实验

3.1 数据集的选取

为验证算法性能，共选取了 5 000 篇分别来自不同大型新闻网站的网页新闻作为新算法的性能评测样本数据集. 实验最后将新闻样本的人工标注之真实标题与算法自动抽取的标题作准确性对比，以此来衡量算法的抽取准确率. 评测样本数据集如表 1 所示.

表 1 新闻标题抽取性能评测样本数据集 篇

网站名称	样本数量	合计
新浪新闻	500	
腾讯新闻	500	
凤凰资讯	500	5 000
新华网	500	
其它新闻网	3 000	

3.2 评测指标及实验结果

根据 CETR[18]算法使用的相关指标，实验采用 Precision、Recall、F-measure($F1$)这 3 个通用的评测指标来衡量通用网页新闻标题自动抽取算法的性能. 3 个指标的计算如式(3)~式(5)所示.

$$\text{Precision}=\frac{|S_e\cap S_l|}{|S_e|}, \tag{3}$$

$$\text{Recall}=\frac{|S_e\cap S_l|}{|S_l|}, \tag{4}$$

$$F1=\frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}. \tag{5}$$

其中，S_e 表示自动抽取算法抽取的新闻标题所构成的字符集合；S_l 表示人工标注新闻标题所构成的字符集合；Precision 是指正确抽取出的新闻文档数与抽取出的总文档数的比率，主要用来衡量抽取准确率；Recall 是指抽取出的总文档数和所有测试新闻文档总数的比率，主要用来衡量抽取的召回率. 准确率(Precision, 简称 P)和召回率(Recall, 简称 R)被广泛用于信息检索领域的质量度量指标，P 和 R 指标有时会出现矛盾情况. 因此，需要综合考虑它们，最常见的方法就是 F-measure，它是 Precision 和 Recall 加权调和平均的结果. 当参数取 1 时，就是常见的 $F1$，它综合了 P 和 R 的结果，当 $F1$ 值较高时，则能说明算法性能较好. 若实验数据量越大，数据集分布越平衡，则评测指标的可信度也会越高. 最后通过计算，可以得出自动抽取算法在整个样本数据集上的平均准确率 P、平均召回率 R 以及平均 $F1$ 值，实验结果分别如表 2~表 3 所示.

表 2 新算法在不同网站样本集上实验结果 %

评测指标	新浪新闻	腾讯新闻	凤凰资讯	新华网	其它新闻网
Average P	99.5	98.9	99.4	99.8	98.3
Average R	96.6	95.3	96.3	96.9	95.1
Average $F1$	98.1	97.1	97.8	98.3	96.6

表 3 不同算法在 5 000 样本集上实验结果对比 %

评测指标	News_E[2]	CETR[18]	IRNLP[19]
Average P	96.6	97.4	95.7
Average R	93.7	94.6	91.2
Average $F1$	95.1	96.0	93.4

根据表 2 和表 3 的实验结果可知：新算法在新华网上面的抽取结果最好，平均 $F1$ 值达到了 98.3%；新浪、腾讯、凤凰资讯新闻网的抽取结果居中，平均 $F1$ 值分别为 98.1%、97.1%和 97.8%；而在其它新闻网站上面，可能由于各网站类型和风格存在一定差异，导致抽取结果稍差一些，平均 $F1$ 值为 96.6%. 整体上来看，新算法在所有网站上各项综合指标的平均值比现有算法要高，较好地说明了新算法的性能优势.

4 结论

通过分析新闻网页的结构，发现新闻网页存在目录型和详情型 2 种结构类型，且它们之间存在唯一的 URL 映射关系. 文中提出了一种基于文本相似度的网页新闻标题自动抽取算法，算法在测试样本数据集上的平均 $F1$ 值达到 97.58%.

本文算法还可应用于网络爬虫领域，理论上支持所有语种的新闻标题抽取，但本实验中只对中文新闻网站做了相关测试，未来会考虑增加其它语种的实验，以便进一步印证算法的性能.

参考文献：

[1]JI H. Information extraction[M]. Boston: Springer, 2002.

[2]向菁菁, 耿光刚, 李晓东. 一种新闻网页关键信息的提取算法[J]. 计算机应用, 2016, 36(8): 2082-2086.

[3]SHUKLA S, NITIN N, TAMRAKAR S. Web information extraction: tag density and keyword approach[J]. International Journal of Computer Applications, 2013, 61(12): 28-30.

[4]GUO T F, HE J Y. Inductively learn XPATH web information extraction rules[J]. Computer Technology & Development, 2007, 17(3): 98-101.

[5]JIMÉNEZ P, CORCHUELO R. On learning web information extraction rules with TANGO[J]. Information Systems, 2016, 62: 74-103.

[6]REZAEI M, GALI N, FRÄNTI P. ClRank: a method for keyword extraction from web pages using clustering and distribution of nouns[C]. International Conference on Web Intelligence and Intelligent Agent Technology, 2015, 1: 79-84.

[7]GALI N, MARIESCU-ISTODOR R, FRÄNTI P. Using linguistic features to automatically extract web page title[J]. Expert Systems with Applications, 2017, 79: 296-312.

[8]李国华, 昝红英. 基于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]MOHAMMADZADEH H, GOTTRON T, SCHWEIGGERT F, et al. TitleFinder: extracting the headline of news web pages based on cosine similarity and overlap scoring similarity[C]. The Twelfth International Workshop on Web Information and Data Management, 2012: 65-72.

[10]张兵, 汤进, 罗斌. 基于超链接和DOM结构树的网页标题实时抽取方法[J]. 计算机与现代化, 2015(8): 84-88.

[11]WU G Q, LI L, HU X G, et al. Web news extraction via path ratios[C]. International Conference on Information & Knowledge Management, 2013: 2059-2068.

[12]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.

[13]罗永莲, 赵昌垣. 突发事件新闻标题与正文提取方法[J]. 计算机应用, 2014, 34(10): 2865-2868.

[14]MA W Q, WU B, GAO W X, et al. Title extraction using natural language processing: US9946703[P]. 2018-04-17[2018-09-25].

[15]凤凰网. 大陆资讯频道[EB/OL]. (2018-07-17)[2018-09-25]. http://news.ifeng.com/mainland/.

[16]凤凰网资讯. 中国气象局启动Ⅳ级应急响应应对台风“山神”[EB/OL]. (2018-07-17)[2018-09-25]. http://news.ifeng.com/a/20180717/59232582_0.shtml.

[17]APOSTOLICO A, GUERRA C. The longest common subsequence problem revisited[J]. Algorithmica, 1987, 2(1-4): 315-336.

[18]WENINGER T, HSU W H, HAN J W. CETR: content extraction via tag ratios[C]. The 19th International Conference on World Wide Web, 2010: 971-980.

[19]大数据搜索与挖掘平台. IRNLP语义分析系统[DB/OL]. [2018-09-25]. http://ictclas.nlpir.org/nlpir/.

(责任编校：龚伦峰)