情报学报 ISSN 1000-0135
第27卷第1期 12-17，2008年2月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000-0135
Vol.27 No.1, 12-17 February 2008

一种无词典的从Web新闻页面抽取主题的算法[1)]

蔡 巍 王永成 尹中航
（上海交通大学计算机科学与工程系，上海 200030）

摘要 主题抽取是自然语言处理研究的重要问题之一。目前流行的方法是“词典+匹配”，但该方法用于处理动态变化的网页信息时，词典难于及时更新等弊病就表现出来。本文作者在研究中文新闻网页内容、结构特点的基础上，提出了一种利用Web页面结构无需词典的主题抽取算法。我们使用该方法对新华网财经新闻语料1000篇进行主题抽取实验，并与手工抽取的主题进行比较，结果表明，重合率高达93%以上。

关键词 主题提取 Web页面 超链接

A Practical Algorithm for Extracting Subject from Web Pages without Thesaurus

Cai Wei, Wang Yongcheng and Yin Zhonghang
(*Department of Computer Science & Engineering, Shanghai Jiaotong University, Shanghai* 200030)

Abstract Subject extraction is one of the important problems in natural language processing area. Traditional methods mainly depend on "thesaurus + matching" mode. But problems arise when processing Internet news using this method, one is the limited volume of thesaurus compared with the uninterrupted emergence of new concepts in Internet nearly all the time. According to Web Chinese news page structure, we propose a new practical algorithm for extracting subject from Web pages without thesaurus. We do subject extraction experiment using 1,000 pieces of news corpus, compared with handcraft, coincidence ratio attain 93%.

Keywords subject extraction, Web pages, hyperlinks

1 引 言

主题抽取是自然语言处理领域研究的重要问题之一[1]，它是信息检索、摘要、自动分类的基础，直接影响后者的质量。目前流行的方法是“词典+匹配”，但该方法用于处理动态变化的网页信息时，词典难于及时更新等弊病就表现出来。为克服这些困难，许多研究人员专门探讨从Web页面中抽取主题的方法。清华大学计算机系马少平提出基于统计分词[2]的中文网页分类，使用2-gram语法对中文网页正文进行分词，以抽取主题实现网页分类；长春工业大学大计算机学院许建潮等提出一种用变长度染色体遗传算法提取Web文本特征的方法[3]。本文作者在研究中文新闻网页内容、结构特点的基础上，借鉴了新闻学及网络新闻学、情报语言学研究成果，提出了一种利用新闻标题及超链接从Web新闻页面抽取主题概念的无词典方法。

2 网络新闻页面表现形态及网络新闻标题特点[4]

网络新闻的版面布局及其表现形态就是把新闻

收稿日期：2006年12月11日

作者简介：**蔡巍**，上海交通大学计算机系博士生，研究方向：网络信息处理。E-mail:xsfhcai@sina.com。**王永成**，上海交通大学计算机系教授，博导，研究方向：网络信息处理。**尹中航**，男，1968年生，清华大学智能技术与系统国家重点实验室博士后，研究方向：网络信息智能处理。

1) 本文受国家863项目(No.2002AA119905)及国家自然科学基金项目(No.60082003)资助。

业已成型的产品进行再加工与分类的各种方式的集合，整合新闻的过程实际上并不只是新闻本身因为技术性的处理突出强势吸引网民，也是新闻的表现形态出奇翻新使网络新闻频频亮相的动态情势[4]。下面以新华网为例说明网络新闻页面的版面布局，重点介绍具体某条新闻页面的结构特点。因为正是利用这些结构特征，我们实现了对具体新闻页面主题概念的自动抽取。

具体的新闻页面，上方是新闻标题，接下来是新闻正文，在正文结束后，后面一般都设有相关新闻和相关专题的链接。比如，新华网每一篇新闻都有它的相关稿件和相关专题，超链接连接相关新闻，通过点击这些新闻标题链接即可达到具体新闻页面。下面取新华网 2003 年 5 月 11 日财经新闻“外经贸部公布对己内酰胺反倾销调查初裁决定”作为示例观察其页面特征。

例 1：

新华网（2003-05-11-07 00:10:04）

新闻标题

外经贸部公布对进口己内酰胺反倾销调查初裁决定

新闻正文

× × × × × × × × × × × × × × × × × ×

× × × × × × × × × × × × × × × × × ×

相关新闻

反倾销　中国企业开始大胆叫阵[01-06 09:12]

反倾销产业损害调查与裁决规定

入世第一年我国对外反倾销力度进一步加大

反倾销“不合适宜”？龙永图张玉卿观点对碰对

对外经贸部：入世首年我国反倾销应诉率达 79.3%

入世首年我国反倾销应诉率达 79.3% 比上年大幅提高

应诉反倾销，企业应当做什么？

入世首年 国外对我反倾销明显减少

通过观察，我们发现，仅仅通过浏览网络新闻的标题及超链接连接的相关新闻标题，即可得出这些新闻是报道“反倾销”这一主题概念的。我们的目标就是希望利用这一结构特征设计出能够抽取主题概念的算法，具体算法的实现将在下文中详细介绍。

新闻标题是高度地概括新闻内容、提示新闻内容、评价新闻内容的简短文字，用以引导和吸引读者注意、评介新闻内容或组织新闻内容。

新闻是新近变化的事实的报道，新闻标题则是新闻内容的概括和浓缩。新闻标题必须以新闻中的具体事实为内容来命题，必须标出新闻事实，这是构成新闻标题的必备条件，即用事实说话。新闻标题还必须具备足以把事实表达清楚的必要的新闻要素（在 Who、What、When、Where、Why、How 中，以 Who 和 What 用得最多最频繁，这是人们所公认的新闻标题必须具备的最基本的构成要素），具有确定性，能够给读者一个明确的概念。

网络新闻的标题与传统的新闻标题有很大的区别。海量性的网络新闻信息势必要求新闻标题具有对新闻内容的导引、导航作用，根本区别于报纸那样可以具有整体性浏览的效果，从新闻内容的短暂浏览中即可判别新闻的可读与否。罗列在一起的新闻标题一般仅仅能从标题来判断新闻内容是否需要再读，因此这一点就成了衡量网络新闻标题内容的一个内在的主导性特征。

网络新闻标题要远比传统媒体尤其是报纸的新闻标题更为简洁传神、耐人寻味，而且只能是一行标题：没有引题与副题，只有主题。网络新闻标题必须在有限的一句话里吸引人们阅读的兴趣。

网络新闻标题的特征是：准确的事实提示，标题单行化。这两大特征以及网络新闻页面的结构化特征，为我们自动抽取主题概念提供了坚实的基础。

3　一种利用新闻标题及超链接从 Web 新闻页面抽取主题概念的无词典方法

3.1　从 Web 新闻页面抽取主题概念的前提

信息处理领域的专家学者非常注意文献主题特征的研究[5]。邓顺国先生对文献标题能否较好地反映文献主题进行了研究，结果表明[6]：标题的情报性（即反映文献主题的程度）平均为 86.2%，其中自然科学为 89.2%，社会科学 84.3%。周全明对图书馆学、情报学、档案学专业的1000篇文献的标题的调查研究发现[7]，这些专业的文献标题的情报性高达 93.2%。他们的这些研究表明，绝大部分文献标题可以反映主题。

张琪玉教授对“依据文献题名对文献进行分类和主题标引的可行性”进行了论证[8]，认为文献题名是对文献内容的概括说明，可以认为它是作者对文献主题内容用自然语言表达的“标引句”，文献题名与文献内容的相符率很高，文献题名中的词完全可

表1 新闻要素(2W)在网络新闻标题及链接中的分布情况

序号	项目	作用	出现在标题中(次)	出现在超链中(次)	主题重复的比率
1	谁(Who)	新闻报道的对象(人或机构等)	487	499	97.59%
2	什么(What)	已经发生或正在发生的事情	483	497	97.18%

以作为标引-检索用词。

作者从新华网财经栏目收集2002年11月至2003年5月的新闻语料1000篇,对网络新闻标题反映主题的情况进行人工调查,结果发现,999篇的标题是直接反映新闻主题的,仅有一篇的标题不能反映主题(2002年12月2日的《一个重要的里程碑》)。这个标题内容太空泛,按照网络新闻标题写作的原则[4],属于网络新闻标题不规范,不符合标题写作的基本要求。

新闻领域里的专家学者都认可新闻的"五要素"构成说,这五个要素是"谁"(Who)、"什么"(What)、"哪里"(Where)、"何时"(When)和"原因"(Why)。后来,人们又加入了"经过"(How)。这样,新闻最多具有六个要素,被称为"6 Ws"或"5 Ws + 1 H"。"6 Ws"是一篇完整新闻的必要条件,这是新闻写作领域里公认的事实。其中"谁"(Who)、"什么"(What)基本就能反映一篇新闻的主题。表1是我们对500篇真实网上财经新闻语料中"谁"(Who)、"什么"(What)在网络新闻标题及链接相关新闻标题中出现情况的统计,其中新闻报道的主体(Who)出现在标题中487次,出现在超链中499次,主题重复的比率达97.59%;作为新闻报道发生的事情(What)出现在标题中483次,出现在链接中497次,主题重复的比率达97.18%。这写事实为通过主题重复来提取存在于标题和超链接中的主题概念提供了坚实的语料基础。

文本主题是一个文本的主要话题和中心思想,它是人们进行相互交流和相互理解的重要基础,同时也是计算机处理自然语言的基本单位。

按照表达形式,文本主题可以大致分成四种类型。

主题词:一个词(也叫关键词)。它能够简单地表达文本的主题。它常常规范成叙词,并构成叙词表。它也是计算机处理一个文本的原始的主题形式。

主题概念:一个或一串词(短语)。就表达主题的能力来说,它具有比单一的主题词更加具体的表达能力。它也可能是一个由几个概念组成的复杂概念。

主题句:一个能够表达文本主题的自然句子。文本标题或子标题就是一种主题句。

主题段:一段能够表达一个较长文本的主题的较短文本。摘要是一种经常使用的主题段。

这四种主题表达方式的关系是前者依次包含于后者之中。换句话说,较长的形式包含较短的形式。例如,主题段通常包含一个或几个主题句,而一个主题句包括一个或几个主题概念,一个主题概念包括一个或几个主题词。

人与计算机在表达文本主题方面的区别在于人经常使用主题句或段落。因为易于组配,计算机早期更多地使用主题词。但是随着计算机处理能力的提高,特别是自然语言处理技术的发展,人们发现主题概念在计算机中具有与主题词相类似的组配能力,同时在表达主题的能力上又比主题词更具体。所以越来越多的自然语言处理任务趋向于用主题概念作为基本的处理单位,例如作为检索、摘要、分类和过滤的语义单位。

通过对多家网站的新闻栏目上的新闻标题以及网络新闻编辑通过超链接推荐的相关新闻标题的观察,我们发现,可以通过网络新闻标题与相关新闻标题进行字符串匹配而不需要词典就可以获取网络新闻的主题概念。这个主题概念可以是一个主题词或者是几个主题词,也可能是一个主题句。下面,我们通过例2,手工实践一下上述过程。

例2:

网络新闻标题:QFII托管人呼之欲出,相关预审工作正在积极推进

相关新闻

链接新闻标题1:关于商业银行申请从事QFII托管业务的通知

链接新闻标题2:商业银行申请成为QFII托管行的有关申请细则出台

链接新闻标题3:中国联通QFII不能投资?

链接新闻标题4:盈利预期比市盈率更重要,谈QFII进入A股市场

链接新闻标题5:QFII最可能相中八只股票

链接新闻标题 6:央行将在近期发布通知,明确 QFII 托管人审批程序

链接新闻标题 7:上证所赴欧推介 QFII 获圆满成功

链接新闻标题 8:受理 QFII 申请首个工作日,银行提出 QFII 托管人申请

首先,网络新闻标题分别与相关链接新闻标题进行字符串匹配,可以匹配出 QFII、QFII 托管、QFII 托管人 3 个主题词。其中,第一个主题词可以包含在后两个主题词中,它们都能正确地反映该篇网络新闻的主题。但是,QFII 出现频次最高,能够作为类主题概念使用。当然这个过程是比较粗糙的,在计算机进行处理中还要涉及歧义处理,从匹配出的主题概念中选择恰当概念等问题。在接下来的小节中,我们就计算机进行网络新闻主题概念提取需要定义的一些概念进行说明,并且描述该算法的流程。通过这个算法,可以在一定程度上解决用于主题抽取词典的不完备问题,抽取出反映主题的名词性短语等未登录词。

3.2 主题概念单元

主题重复是我们研究的基础。为了得到重复的字符串,我们利用了主题概念单元(在文献[9]中称之为主题基因,本章中我们推广了这一概念形式,即相邻的主题基因构成一个有意义的主题概念单元)这一形式,现将其定义如下:

定义 1:主题基因(SG)是两个或多个不间断的、同时出现在标题和正文中的字符。两个相邻的主题基因可以连接成一个更长的基因,形成一个主题概念单元。作为主题概念的基本单位,它可以是一个完整的概念,也可以是一个概念的一部分。

我们提出主题基因的目的是把所有可能的基因尽可能长地连接在一起,以便形成能够表达新闻要素的主题概念串。这个主题概念串就是一个主题概念单元,它表达了一个主题概念。下面,我们通过一个例子进行说明。本示例取自新华网财经频道的真实网页新闻。

例 3:

国务院发展中心:完善公司治理是国企改革核心

相关新闻

链接新闻标题 1:深圳国企改革引发外商投资潮 开辟赢利新来源

链接新闻标题 2:深圳国企改革大突破 部分股权转让国际投资者

链接新闻标题 3:一万个“主人”为了一个企业 吉林化纤采访手记

链接新闻标题 4:吴邦国强调:努力做好国企改革发展各项工作

链接新闻标题 5:吴邦国在吉林省考察强调:努力做好国企改革工作

链接新闻标题 6:中国华录集团总经理袁义祥谈如何“拧”出利润来

链接新闻标题 7:吴邦国强调,以“三个代表”重要思想为指导 努力做好国企改革各项工作

链接新闻标题 8:吴邦国强调:努力做好国企改革发展各项工作

链接新闻标题 9:天津动员上市公司治理大检查

如果使用例 3 标题对“相关新闻”(可以使用 HTML 超链接标识识别)中的链接新闻标题进行字符串的匹配操作,就会发现中文字符“国”与“企”是两个连续的、且在标题和正文中同时出现的字符,所以它们构成了一个主题基因“国企”。同样,“企”与“改”也构成了一个基因“企改”。由于这两个基因“国企”与“企改”是相邻的,把重复的“企”删去,它们就可以连接起来形成一个较长的基因“国企改”。当把所有可能的相邻基因都连接起来后,就可以得到主题概念单元“国企改革”(国有企业改革的简称)。

3.3 网上新闻主题概念单元自动抽取算法

网上新闻主题的抽取方法与单篇新闻主题的抽取方法在实现上既有相似之处,又有不同点。我们课题组在论文[9]中曾详细地介绍了单篇新闻主题提取的算法,在这里主要介绍两者的相异之处,对相似之处即实现的数据结构部分仅做简要介绍。

3.3.1 数据结构

为了实现快速主题提取,我们利用了两个线性表作为基本数据结构,如图 1 和图 2 所示。在这两个图中,m 是标题中字符的数量,n 是超链接标题中字符的数量,s 是原始主题基因的个数。

图 1 是一个标题字符的记录,总的记录数小于或等于 m,每个记录的项数不超过 $n+2$。它记录了每一个标题字符在超链接标题中的位置。通过检查每个记录表,我们就能找到所有的原始主题基因串。图 2 是得到的一个主题基因,其中 Hpt、Hph、Tpt、Tph 分别代表每个基因串的头和尾在标题和超链接标题

中的位置。通过检查这个基因表中的项目，我们可以判断两个基因串是否相邻(相邻意味着一个基因串的头是另一个基因串的尾)。至于基因串的长度Len和基因串出现的频率Freq，主要用于在下一步计算基因串的权值。

权值计算的相关问题以及形成主题概念的后处理可参考论文[9]。

Char(*m*)	Times	Pos1	Pos2	…	Pos*n*

Char (*i*)：标题中的字符；
Times： Char (*i*) 在超链接标题中出现的次数；
Pos：Char (*i*) 在超链接标题中的不同位置。

图1 标题字符线性表

SG(*s*)	Len	Freq	Hpt	Hph	Tpt	Tph

SG：主题基因；Len：基因串长度；Freq：基因串出现频次；
Hpt：基因串头在标题中的位置；Hph：基因串头在超链接标题中的位置；
Tpt：基因串尾在标题中的位置；Tph：基因串尾在超链接标题中的位置。

图2 主题基因线性表

3.3.2 网上新闻主题概念单元自动抽取算法实现过程

具体的实现过程如下：

(1)扫描Web网页中文新闻页面。

(2)利用超文本结构，抽取网页中〈title〉〈/title〉之间的网络新闻标题以及〈a〉〈/a〉之间的超链接的个数a。

(3)把所有的单字节字符转换为双字节字符，建立起字符位置表。利用标题中的每一个字符，从第一个超链接标题的开始到结尾进行扫描，把每一个标题字符在超链接标题中的位置放入到图1所示的标题线性表数据结构中，并记录该字符在超链接标题中出现的次数。

(4)形成基本的主题基因。在线性表中从上到下把所有在超链接标题中相连的字符取出作为可能的SG，记录其长度(至少为2)以及SG的头尾字符分别在标题及超链接标题中的位置，把这些信息按顺序放入图2所示主题基因线性表数据结构中。

(5)连接SG。在主题基因线性表中自底向上，比较下面的SG的头字符与上一个SG的尾字符，当两个字符相同且在标题和超链接标题中的位置均相同时，删除重复的字符，并把两个SG连在一起形成新的较长的SG。

(6)不断重复上一步，直到把符合定义1的SG都尽可能长地连接起来，形成主题概念。在连接的过程中，计算各个SG出现的频次。

(7)进行误连接后处理。

(8)输出正式的SG。

(9)然后使用标题对后a－1个超链接标题重复执行步骤(3)～(8)。

(10)对输出的SG进行两两匹配，选取出出现频次最高的SG，即为主题概念。

4 实验结果及结论

我们使用该方法对新华网财经新闻语料1000篇进行主题抽取实验，并与手工抽取的主题进行比较，结果表明，重合率高达93%以上。

结合我们所做研究，获得了如下经验，希望能对今后的Web页面主题特征的提取有所启发。

根据新闻学及网络新闻学的基本原理，新闻标题的作用是高度概括新闻内容、提示新闻内容、评价新闻内容的简短文字，用以引导和吸引读者注意、评介新闻内容或组织新闻内容。网络新闻标题要远比传统媒体尤其是报纸的新闻标题更为简洁传神、耐人寻味，而且只能是一行标题：没有引题与副题，只有主题。另外，Web新闻网页不仅含有网络新闻标题及正文，而且还含有网站新闻编辑通过超链接推荐的相关新闻标题。通过这些链接，我们可以浏览具体的相关新闻。网络新闻标题的特征是：准确的事实提示，标题单行化。这两大特征以及网络新闻页面的结构化特征为我们自动抽取主题概念提供了坚实的基础。经过观察，我们发现通过网络新闻标题与相关链接新闻标题进行匹配以及经过统计排序，可以获取网页新闻的主题概念。为此，作者设计了一种利用新闻标题及超链接从Web新闻页面抽取主题概念的无词典方法，这种方法是我们提出的无词典提取主题概念法[9]在Web新闻页面环境下的推广及应用。另外，情报语言学专家、学者关于标题主题性特征的研究成果为我们提供了坚实的理论保证，使我们更深刻的感觉到“情报语言学是为情报检索提供语言学保证的一门学科”(张琪玉教授语)，在研究相关算法的时候，情报语言学的研究成果值得我们热切关注。

参 考 文 献

[1] 王永成，等. 中文信息处理技术及其基础[M]. 上海：

上海交通大学出版社,1991.

[2] 马少平. 基于统计分词的中文网页分类[J]. 中文信息学报, 2002, 16(6): 25-31.

[3] 许建潮. 中文 Web 文本特征提取与分类[J]. 计算机工程, 2005, 31(8):24-25, 39.

[4] 仲志远. 网络新闻学[M]. 北京:北京大学出版社, 2002.

[5] 张琪玉. 张琪玉情报语言学文集[M]. 北京:北京图书馆出版社,1999.

[6] 邓顺国. 中文期刊论文标题情报性的调查分析[J]. 图书情报知识,1985(1).

[7] 周全明. 论机辅抽词标引及其规则[J]. 图书情报工作,1995(3):44-49.

[8] 张琪玉. 分类法主题法一体化自动标引系统的基本原理和方法[J]. 图书馆论坛,1995(6): 3-4.

[9] 尹中航,王永成,蔡巍. Extract subject from Internet news by string match[J]. 软件学报,2002,13(2): 159-167.

(责任编辑 许增棋)