软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(9):2089−2100 [doi: 10.3724/SP.J.1001.2013.04358]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于网络资源与用户行为信息的领域术语提取*

闫兴龙[1,2,3,4], 刘奕群[1,2,3], 方 奇[1,2,3], 张 敏[1,2,3], 马少平[1,2,3], 茹立云[1,2,3]

[1](智能技术与系统国家重点实验室(清华大学),北京 100084)

[2](清华大学 信息科学与技术国家实验室,北京 100084)

[3](清华大学 计算机科学与技术系,北京 100084)

[4](总参陆航研究所,北京 101121)

通讯作者: 闫兴龙, E-mail: yan-xinglong@163.com

摘 要: 领域术语是反映领域特征的词语.领域术语自动抽取是自然语言处理中的一项重要任务,可以应用在领域本体抽取、专业搜索、文本分类、类语言建模等诸多研究领域,利用互联网上大规模的特定领域语料来构建领域词典成为一项既有挑战性又有实际价值的工作.当前,领域术语提取工作所利用的网络语料主要是网页对应的正文,但是由于网页正文信息抽取所面临的难题会影响领域术语抽取的效果,那么利用网页的锚文本和查询文本替代网页正文进行领域术语抽取,则可以避免网页正文信息抽取所面临的难题.针对锚文本和查询文本所存在的文本长度过短、语义信息不足等缺点,提出一种适用于各种类型网络数据及网络用户行为数据的领域数据提取方法,并使用该方法基于提取到的网页正文数据、网页锚文本数据、用户查询信息数据、用户浏览信息数据等开展了领域术语提取工作,重点考察不同类型网络资源和用户行为信息对领域术语提取工作的效果差异.在海量规模真实网络数据上的实验结果表明,基于用户查询信息和用户浏览过的锚文本信息比基于网页正文提取技术得到的正文取得了更好的领域术语提取效果.

关键词: 领域术语自动抽取;新词发现;Web 数据挖掘;用户行为分析

中图法分类号: TP391 **文献标识码**: A

中文引用格式: 闫兴龙,刘奕群,方奇,张敏,马少平,茹立云.基于网络资源与用户行为信息的领域术语提取.软件学报,2013,24(9):2089−2100. http://www.jos.org.cn/1000-9825/4358.htm

英文引用格式: Yan XL, Liu YQ, Fang Q, Zhang M, Ma SP, Ru LY. Domain-Specific terms extraction based on web resource and user behavior. Ruan Jian Xue Bao/Journal of Software, 2013,24(9):2089−2100 (in Chinese). http://www.jos.org.cn/1000-9825/4358.htm

Domain-Specific Terms Extraction Based on Web Resource and User Behavior

YAN Xing-Long[1,2,3,4], LIU Yi-Qun[1,2,3], FANG Qi[1,2,3], ZHANG Min[1,2,3], MA Shao-Ping[1,2,3], RU Li-Yun[1,2,3]

[1](State Key Laboratory of Intelligent Technology and Systems (Tsinghua University), Beijing 100084, China)

[2](National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China)

[3](Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

[4](Army Aviation Institute of General Staff, Beijing, Beijing 101121, China)

Corresponding author: YAN Xing-Long, E-mail: yan-xinglong@163.com

Abstract: The automatic domain-specific term extraction is an important task in natural language processing, which can be adoptedin domain-specific ontology construction, vertical search, text classification, class-based language model etc. A Web page contains lots of noises and irrelevantcontents, therefore, extracting domain-specific terms from original pages becomes a challenging task. Different from previous works, which rely on the original text of Web pages, this study focuses on anchor text and query log history of pages. This

* 基金项目: 国家自然科学基金(60736044, 60903107, 61073071); 高等学校博士学科点专项科研基金(20090002120005)
收稿时间: 2011-10-28; 修改时间: 2012-05-25; 定稿时间: 2012-12-03

strategy would avoid the trouble of information extraction from the original Web pageand therefore improves the term extraction performance. In this paper, a novel term extraction algorithmis based onanalysis into Web resource and user behaviors. Different Web resources including body of the page data, anchor text of the page and the information of user query data were employed to extract the domain-specific terms and their performances were compared.The result based on scale of the network datademonstratesthe resources of anchor text and the way user query data can obtain a better effect.

Key words: automatic domain-specific term extraction; novel term extraction; Web data mining; user behavior analysis

据 CNNIC 测算,截至 2010 年 12 月,中国网民规模达到 4.57 亿,较 2009 年底增加 7330 万人;互联网普及率达到 34.3%,较 2009 年提高 5.4 个百分点.越来越多的人通过互联网获取各种信息,网络资源正逐渐成为人们获得知识的主要渠道.随着我国科学技术的高速发展,大批新的领域术语出现在学习和生活的各个方面,而这些词语也通过网络传播开来.领域术语集中体现了某个领域的核心知识,对于了解某个领域的发展状况和流行热点都有重要的理论意义和现实意义.领域术语提取主要应用在以下几个方面:

- 领域本体[1,2]:领域本体是对给定领域中存在的概念的一种详尽的特征化描述,它用公认的术语集合和术语之间的关系来反映该领域内的知识和知识结构.本体需要人工进行构建,这是一个繁琐的过程,所以,利用知识获取的技术来降低该过程的繁琐程度成为一种必要.在本体构建过程中,术语抽取是构建本体中最基础的一步,所以,领域术语的抽取在领域本体中有很重要的作用;
- 自然语言处理:领域术语的抽取是自然语言处理中一个重要的基础性问题.分词是自然语言处理的重要方面,如果能够做好分词,就能对于自然语言的理解有很大的提高.术语的抽取对于分词有很大的帮助.因此,得到正确的术语可以更加准确地切分语料;
- 搜索技术:对于自然语言的理解,很大程度上影响搜索引擎的结果.这样,在搜索引擎中引入领域术语集,可以极大地提高检索的准确率,并达到较高的信息覆盖,尤其在垂直搜索方面,如果得到某个领域的术语,对于该领域的搜索可以得到更为精确的信息;
- 文本分类:通过得到领域相关的术语,对于文本的分类有很大的帮助.通过检查文本中出现该领域术语的频率来判别该文本是否属于该领域.文本的分类可以应用在广告投放、垂直搜索等各个方向.

目前,常用的领域术语抽取的方法主要是基于领域语料进行抽取.Web 资源非常丰富,所以,利用 Web 资源抽取领域术语也成为领域术语抽取中一个很重要的方向.有不少文献专门研究如何从 Web 上自动获取领域语料[4].当前所利用的 Web 资源均为网页的正文文本,但是由于种种原因,网页中的正文数据往往被许多噪音数据干扰,如广告、链接、产品推荐、导航条等,如图 1 所示.

从图 1 可以看出,在该网页中,有很大一部分空间被广告、链接、导航条所占用,从该网页中提取网页正文会遇到一些问题.这样的网页在实际网络环境中大量存在,所以网页正文的提取和分析是 Web 数据挖掘的重要问题.由于存在这样的网页正文提取问题,实际应用中往往很难获得基于网页正文的高质量领域语料,基于这样的领域语料的质量问题,便会影响领域术语抽取的质量.Web 资源多种多样,可以利用的资源也有多种形式,不仅仅有网页正文,还有网页的锚文本、查询文本等,而这些文本不存在提取的困难,所以,利用锚文本以及查询文本可以很好地避免信息抽取所遇到的困难,这样会得到所需要的高质量语料.但是,锚文本和查询文本也会有自身的问题存在,如文本长度过短、语义信息不足导致的分词问题等,这些问题也会影响领域术语抽取的质量.本文针对上述问题和挑战,提出适用于上述不同的 Web 资源的领域术语抽取算法并进行领域术语抽取实验,资源的不同对于术语提取质量有很大的影响.本文利用多种 Web 资源进行领域术语的抽取工作,其中主要包括锚文本、从查询日志中得到的相关文本等.实验结果表明,基于锚文本的 Web 资源抽取得到的领域术语有很高准确率和较高的召回率.

本文第 1 节主要介绍提取术语的算法和相关概念.第 2 节主要介绍 Web 资源的类型和特点.第 3 节介绍基于 Web 资源领域术语抽取方法.第 4 节是实验结果分析.第 5 节是对未来工作的展望.

Fig.1 A finance Web

图 1 某财经网页主题部分

1 相关工作

当前,国内外学者对领域术语自动选取工作的研究主要有以下几个方面:

(1) 基于规则的方法[3].规则方法主要通过预先制定规则模板,然后通过模板的匹配和过滤来抽取术语.但是规则的编制主要依赖于语言学知识,也就是由语言学家进行编制.由于规则由人工编制,很大程度上受到人所具有的知识的限制,各个语言学家的观点又各不相同,存在不一致性,手工编制费时费力,制定完备的规则集很难,而且还要考虑多个规则的兼容性.所以,目前单纯使用基于手工编制规则的领域术语获取方法已经较少了.

(2) 基于统计的方法.统计方法很早就被用于术语抽取中,而且取得了不错的效果.Salton[2]使用文档的 TF/IDF 进行术语抽取.Frantzi[5]提出了 C-value/NC-value 评价函数用于领域术语抽取,并取得很好的结果.Pantel[6]采用互信息和对数似然比获取领域术语.Liu[8]采用左右信息熵和对数似然比确定词语边界从而抽取候选术语.而该方法在本文中也有所利用.基于统计的算法在各个语料库中均可使用,并且这类方法效率更高,无需语义分析,所以是一种高效率方法,是当前新词发现的重要的方法之一.

(3) 规则和统计相结合的方法.在实际应用中有很多统计和规则相结合的方法.Thuy[7]先根据规则抽取候选集合,然后使用 C-value/NC-value 和 T 检验的方法计算,最后得到真正术语.这种方法结合上述两种方法的优缺点,得到的结果是相对较好的.

上述术语抽取的方法是基于传统语料库进行的工作,此类方法的优点是:语法较为严谨、结构规整、噪音较少.但是传统语料库文本中的术语是比较陈旧的,无法得到最新的术语,而且传统语料库中含有领域术语的比例较少.随着网络技术的发展,Web 文本成为人们获取知识的重要资源.相比较传统文本而言,Web 文本资源的优势在于能够体现最新的领域术语变化和更新,所以也有一些文献基于 Web 文本进行领域术语的提取.但是基于 Web 文本的领域术语抽取主要是通过网页的正文进行术语抽取,而网页的正文抽取是一项很复杂的工作.网页正文的抽取需要对网页文本进行解析,网页的制作者并没有一个统一的标准进行网页制作,所以从中提取网页的正文变成一项复杂的工作[10];另外,对于网页正文的定义也存在不确定性,如论坛网站,其包含内容庞杂,无法准确定义其正文,所以,现在仍然没有一个方法可以完全解决抽取正文的问题.而用户的点击和浏览行为信息能够在一定程度上提高 Web 资源的质量[11],所以,利用用户行为信息得到高质量的 Web 资源成为必要.综上所述,本文利用 Web 资源的多样性,对多种类型的网页文本进行领域术语的抽取,采用高质量的 Web 资源可以得到比其他资源更好的结果.

2 Web 资源类型及特征

从网络可以得到的资源非常多,但是网络资源的质量有所不同,本文主要通过 3 类的资源获得相关领域的术语.

2.1 基于搜索引擎查询日志的查询文本

为了改进系统性能,记录运行情况等原因,绝大多数搜索引擎都会记录用户与搜索引擎进行交互的行为日志,这种日志一般被称为搜索引擎查询日志.查询日志通常记录的内容包括:用户使用搜索引擎的时间、用户提交的查询词、用户点击的结果、这些结果在搜索引擎里的排序情况等.本文利用的搜索引擎查询日志主要内容见表 1.

Table 1 Items in search engine logs

表 1 搜索引擎日志包含的信息项

名称	记录内容
query	用户提交的查询
URL	用户点击的结果地址
time	用户点击发生时的日期、时间
rank	该 URL 在返回结果中的排名
order	用户点击的顺序号(这是用户点击的第几个页面)
id	由系统自动分配的用户标识号
submitter information	浏览器信息,计算机信息

由于查询日志是在不影响用户正常使用的情况下完成记录的,因此具有客观、真实保存用户与搜索引擎交互情况的作用.当前,来自研究领域的相当数量的研究者使用查询日志在搜索引擎算法改进、服务质量监控、社会舆情热点分析等方面进行了大量的研究.而微软、美国在线、搜狐等国内外公司也通过各种形式共享了部分查询日志,以便产业界和学术界各方面人员共同对这些资源进行分析与利用.

本文利用搜索引擎查询日志中的用户点击的网页,如果用户最终点击了搜索引擎呈现给用户的相关链接,则认为该查询词和该网页有很高的相关性.查询词也是对于该网页一个很好的描述.如果该网页属于某领域,则认为该查询词和该领域有很高的相关性,定义该查询词为查询文本.由于查询词相当于对网页的一个标签,可以认为是对于该网页的一个说明.所以,收集和该领域相关的查询文本作为语料库,对其进行领域术语抽取,最终得到领域相关的领域术语.

2.2 基于用户浏览日志的锚文本

锚文本是指由网页制作者编写的,用于描述对应的超链接网页内容的文本样式.页面添加的链接一般都与页面有直接相关的联系,搜索引擎可以根据指向某一个网页的链接的锚文本描述来判断该网页的内容属性,所以锚文本在网络搜索中发挥很大的作用.但实际网络环境中的锚文本也包含了大量的噪音(功能性超链接、广告超链接、恶意超链接等),借助用户行为数据可以有效地过滤噪音[11],并保留有意义的超链接信息及锚文本.为了保护用户隐私,数据是在用户体验改进计划的参与者中抽取的,数据收集经过了用户的同意,并删除了用户的 IP、用户名等个人信息.数据使用树结构进行存储.本文利用的日志主要包含的内容见表 2.

因为锚文本由网页设计者编写,锚文本可以作为锚文本所在页面内容的评估.锚文本的信息密度要比网页正文更大,所以其文本的质量要高于网页的正文.如果用户通过点击锚文本来访问某网页,用户实际是被锚文本所吸引而访问该网页的.这种点击的行为也是对锚文本的筛选,如果一个锚文本被用户所点击,则是对该锚文本的一个肯定.通过点击的行为是提高锚文本资源质量的较好方法.经过上述的分析可知,点击的锚文本资源的质量要高于对应网页正文的资源.本文利用用户点击的锚文本作为领域术语抽取的语料库,如果该锚文本对应的网页是某领域的网页,则认为锚文本是对该领域的语料,并对该语料库进行术语的抽取工作.

2.3 网页正文

所谓正文,就是网页呈现给用户的主要内容,其中去除对外的链接的推荐网页,以及相应的一些广告和无用

的信息.基于网页的内容,可以判断其是否为某领域的网页,从而决定是否将该网页正文作为领域文本语料.当前,利用 Web 资源进行的术语抽取都是基于该文本资源的,网页正文的抽取需要对网页代码进行解析.由于网页编写者的习惯不同,则抽取的策略也必须不同,所以正文的抽取是一项复杂的工作,解析的结果也不一定正确,这也在一定程度上影响了对网页正文的领域抽取结果.

Table 2 Items in users browsing logs

表 2 用户浏览日志包含的信息项

名称	记录内容
title	用户访问网页的标题
URL	用户浏览的结果地址
time	用户访问网页时的日期、时间
anchor	用户浏览网页时点击的锚文本对应的地址
anchor_text	用户浏览网页时点击的锚文本
id	由系统自动分配的用户标识号
submitter information	浏览器信息,计算机信息

3 基于 Web 资源领域术语抽取方法

3.1 抽取方法框架

本文领域术语抽取方法流程如图 2 所示.

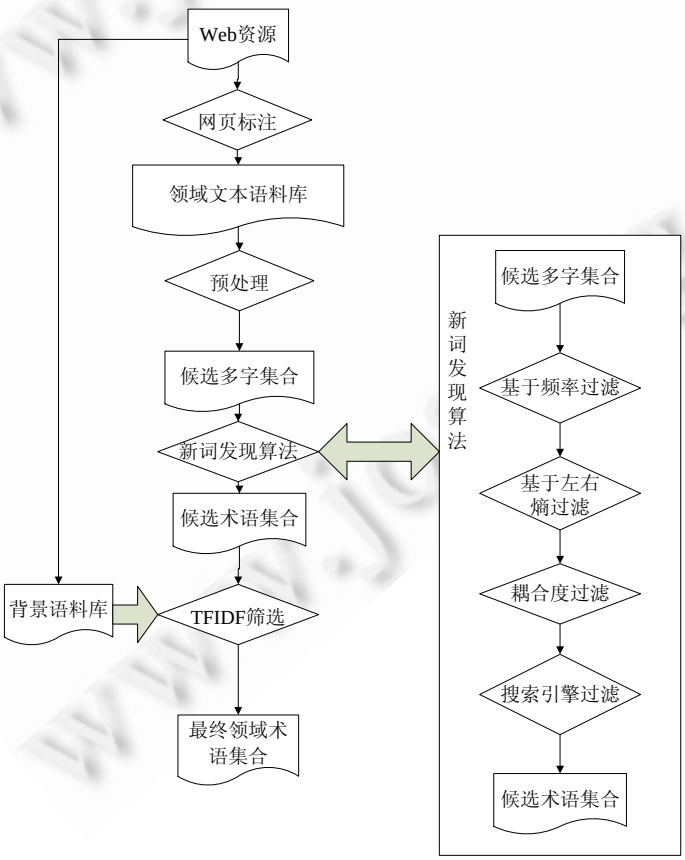


Fig.2 Framework of domain-specific terms extraction based on Web resource

图 2 基于 Web 资源领域术语抽取方法框架图

本文的方法基于多种 Web 语料库,主要通过 4 步运算,可以得到最终领域术语的集合:

第 1 步:网页标注.

网页标注主要通过归纳总结找到某个领域的网页 url 的规律和特点,最终总结出基于 url 的网页筛选方法.通过该方法可以得到某领域相关的 Web 资源.最基本的原理是:大型新闻门户网站某领域网页的 url 均是在某个子域名下,而某领域专业网站下的网页一般为该领域的相关文本.

第 2 步:预处理.

预处理是为新词发现算法处理语料库,对于原有的网络文本进行整理,如对网页正文进行抽取,对原有文本不规则内容进行整理.最后对句子进行切分,得到多字集合,用于新词发现算法处理.

第 3 步:新词发现.

新词发现基于上一步得到的候选多字集合.该方法首先统计候选多字集合中每个候选多字出现的频率,将低于某频率阈值的多字滤除出候选集合;然后,分别计算每个候选多字的左右信息熵,将低于某熵值的多字滤除;基于每个多字左右信息熵以及它出现的频率,本文提出耦合度过滤方法,并将候选多字基于耦合度过滤;进行耦合度过滤之后,将得到的结果放入搜索引擎进行进一步的过滤.实验结果表明,上述两种方法均可以使结果得到质的提升.具体的实现方法将在下一节进行详述.通过上述的过滤,最后得到候选术语集合.

第 4 步:TF/IDF 筛选.

TF/IDF 是一种常用的计算某个词在某篇文档或部分文档集合中重要程度的方法.基于 TF/IDF 筛选,是为了更好地得到与领域相关的术语,通过计算每个候选术语在文本语料库中的 TF/IDF 值,得到每个候选术语在领域文本语料库的重要程度.

3.2 新词发现算法

3.2.1 基于频率过滤

统计该领域多字集合中多字出现的次数,本文将频率大于某阈值的词语作为下一步计算的候选集合.

3.2.2 计算信息熵

信息熵的方法在术语抽取领域用来判别术语的边界,根据词语 w 与邻接的字搭配的不确定性来判断 w 是否是词边界,从而得到候选术语.该方法可以有效抽取出文本中出现的高频词汇.一般而言,领域文本语料中高频词汇往往是该领域的术语.

信息熵的计算方法:建立词语对应的左右单字统计数据.主要的方法就是遍历所有的文档,然后统计每个词语左边及右边出现的每个单字的频率.计算对应的熵.

定义 1. 假设词语 w 属于候选集,另外,$A=\{a_1,a_2,a_3,\ldots,a_m\}$和 $B=\{b_1,b_2,b_3,\ldots,b_n\}$分别为该词语对应的左右单字集合,则左右熵的定义如下:

$$LE(w)=-\frac{1}{n}\sum_{a_i\in A}C(w,a_i)\log\frac{C(w,a_i)}{n} \tag{1}$$

$$RE(w)=-\frac{1}{n}\sum_{b_i\in B}C(w,b_i)\log\frac{C(w,b_i)}{n} \tag{2}$$

其中,$n=\sum_{a_i\in A}C(w,a_i)=\sum_{b_i\in B}C(w,b_i)$,$C(w,a_i)$和 $C(w,b_i)$分别为对于词 w 而言的左单字 a_i 和右单字 b_i 出现的次数.

对于一个实际存在的词而言,如果其出现频率较高且左右单字集的频率也很高,则可以通过其左信息熵和右信息熵的方法进行过滤.

由于本文采用语料库自身有特点,查询词往往不是一个句子,所以对于某个词而言,其独立成词很有可能不存在左(右)的单字.如在处理的查询语料库中,“京东方”共出现 532 次,而其左右单字一共只有 22 个,所以用信息熵并不能反映其成词的概率,所以我们在这里采用了如下的策略(其中,L,R 为标志位,α为阈值):

- 如果 $\frac{n}{N}<\alpha$,则 L=1;否则,L=0.其中,N 为该词一共出现的频率,n 为该词左单字出现的频率.同理,如果

$\frac{n}{N}<\alpha$,则 R=1;否则,R=0.n 为该词右单字出现的频率;

- 如果 L=R=1,则认为该词放入候选集中,进行下一步过滤;否则,如果 L=0 或 R=0,则通过判断其左信息熵或者右信息熵的方法进行过滤.

依据信息熵进行过滤的策略为:在上文中抽取出候选集之后,对于 L=0 或 R=0 进行判断,如果 $LE(w)>\beta$或 $RE(w)>\beta$,则将该词放入候选集中,进行下一步过滤;否则,将该词去除.另外需要指出的是,如果该侧的熵值不存在,则将其定义为无穷小.只有 w 满足两侧成词的阈值,才能将其放入候选集中.

3.2.3 基于递推的耦合度过滤算法

虽然上一步的方法可以很好地找到过滤后词语集合,但是仍然存在很多噪音,在候选集中可以找到表 3 所示的情况.

Table 3 Exception phrases in the candidate set

表 3 过滤后候选集中的某些异常词组

词语	左信息熵	右信息熵
性突破	1.546	不存在
业之一	1.077	不存在
股频道	1.200	不存在
榜出炉	1.502	不存在
大行业	2.02	2.838
明天大	3.556	1.798
化股份	2.266	2.346

需要说明的是,不存在右信息熵是因为其满足上一节中对于频率的过滤规则.如“化股份”这个词,从语义的角度来讲,该候选词中的“股份”应该和“化”分开,之前没有分开的原因是该词的左信息熵过大,这样依据上一步的规则无法滤除.依据已有的信息熵和候选词的出现频率,提出基于递推的耦合度过滤算法,具体算法如下:

对于字长为 3 的 w,如果存在 $w_1\in T_2$(T_2 为长度为 2 的候选词集合),w 可分解为 pw_1,p 为单字.计算 p 和 w_1 的耦合度公式如下:

$$Co(p,w_1)=\left(\frac{C(w)}{C(w_1)}<\lambda\right)\cap(LE(w)<LE(w_1))\cap(LE(w)<\gamma) \tag{3}$$

如果存在 $w_1\in T_2$(T_2 为长度为 2 的候选词集合),w 可分解为 w_1p,p 为单字.计算 p 和 w_1 的耦合度公式如下:

$$Co(p,w_1)=\left(\frac{C(w)}{C(w_1)}<\lambda\right)\cap(RE(w)<RE(w_1))\cap(RE(w)<\gamma) \tag{4}$$

其中,γ和λ为一个参数阈值.如果耦合度的值等于 1,我们认为 w 不应该为词.

对于字长为 4 的 w,如果存在 $w_1\in T_3$(T_3 为长度为 3 的候选词集合),w 可分解为 pw_1,p 为单字.计算 p 和 w_1 的耦合度公式如下:

$$Co(p,w_1)=\left(\frac{C(w)}{C(w_1)}<\lambda\right)\cap(LE(w)<LE(w_1))\cap(LE(w)<\gamma) \tag{5}$$

如果存在 $w_1\in T_3$(T_3 为长度为 3 的候选词集合),w 可分解为 w_1p,p 为单字.计算 p 和 w_1 的耦合度公式如下:

$$Co(p,w_1)=\left(\frac{C(w)}{C(w_1)}<\lambda\right)\cap(RE(w)<RE(w_1))\cap(RE(w)<\gamma) \tag{6}$$

其中,γ和λ为一个参数阈值.如果耦合度的值等于 1,我们认为 w 不应该为词.

以此类推,得到长度更长的词.

3.2.4 搜索引擎过滤

依据搜索引擎的构建机制,将多字放入搜索引擎,如果得到的结果很少,说明该多字无法独立成词.根据这个原理,可以对结果进行进一步的过滤.

本文利用某商业搜索引擎得到的网页数对最后结果进行过滤,去除不能独立成词的多字.实验结果表明,这

种方法是可以滤除非词语.

3.3 TF/IDF筛选

TF/IDF 方法是一种非常有效的方法,广泛地应用于信息检索及文本分类等自然语言处理研究中.术语的一个明显特征就是在本领域的文本中多次出现,而在其他领域中出现的次数较少,而 TF/IDF 特征则能在一定程度上反映术语的这个特征.对于文档 d,候选术语 w 对应的词频及文档频度倒数特征计算公式如下:

$$Tfidf_d(w)=tf\times idf \tag{7}$$

$$tf=\frac{f(w)}{\sum f(w)} \tag{8}$$

$$idf=\log\left(\frac{|D|}{|\sum d|}+0.01\right) \tag{9}$$

其中,$f(w)$为词 w 在文档 d 中出现的次数,$\sum f(w)$ 为一篇文档的总词数,$|D|$为语料库中的文件总数,$|\sum d|$ 为包含词语 w 的文件数目.而使用类别的 TF/IDF 作为候选词的评价函数,其公式如下:

$$TFIDF(w)=\sum_{d\in D} tfidf_d(w) \tag{10}$$

4 实 验

4.1 实验数据

该实验一共有 3 类语料库作为实验数据,这 3 类实验数据分别为:

(1) 查询日志.具体是指用户将某查询提交搜索引擎,用户在检查搜索引擎的结果后点击某个链接,而该网页是财经领域的网页,则将该查询放入语料库;

(2) 点击的锚文本.具体是指用户在查看网页时,通过点击锚文本跳转到别的网页,而跳转后的网页为财经类网页.将该锚文本放入语料库;

(3) 用户通过锚文本浏览的网页全文.具体是指用户通过点击锚文本后看到的网页正文放入语料库.

本文采用某商业搜索引擎一周(2010 年 10 月 11 日~2010 年 10 月 17 日)的查询日志,以及该周用户浏览行为日志.本次实验 3 个语料库的大小见表 4(本文将全网络的网页作为全文的背景语料库).

Table 4 The size of each corpus

表 4 各语料库的条目和规模

语料库领域	查询	锚文本	全文
背景语料库条目	412 545 654	114 579 967	∞
背景语料库规模	5.5G	2.9G	∞

4.2 实验标注

本文采用对 url 域名的标注方法判断某个网页是否属于某个领域.采用财经类的网页作为抓取领域术语的语料库,如“东方财富网”是专业的财经类网站.本文认为,包含“eastmoney.com”的网页 url 属于财经类的网页;对于一些大的门户网站,如“sina”,“sohu”及“qq”等,本文采用归纳总结的方法得到该门户网站的财经领域网页,如“sohu”门户中的子域名,见表 5.

将该时间段的所有的网页作为背景语料库.为了检查标注后的结果准确率,即标注后的网页属于财经类网页的比率,本文通过多次重复随机抽样,在 66 万的网页中,每次抽样 100 个网页,其准确率达到 96%.通过标注后的财经领域语料库数量和大小见表 6.

4.3 实验结果和分析

通过新词发现算法后,分别得到了 3 组结果.基于术语频率排序后的结果见表 7~表 9.

Table 5 Finance Web of sohu

表 5 sohu 网站下某些财经类网页

url 包含的串	该网页的主要内容
q.stock.sohu.com	查看某支股票的网页
business.sohu.com	财经类资讯的网页
stock.sohu.com	股票类资讯的网页

Table 6 The size of finance corputs

表 6 财经领域语料库条目和规模

语料库领域	查询	锚文本	全文
财经领域条目	2 047 553	660 244	660 244
财经领域规模(M)	2	17	210

Table 7 The frequence and information entropy of the words in query corpus

表 7 基于查询语料库的词语的信息熵和频率

词语	频率	左信息熵	右信息熵	网页数
茅台	7 085	0.61106	1.33166	3 410 694
中国	7 044	3.44439	4.62759	764 219 624
基金	3 735	3.13219	2.81496	54 365 102
公司	3 580	2.88908	4.40276	578 556 549

Table 8 The frequency and information entropy of the words in anchor corpus

表 8 基于锚文本语料库词语的信息熵和频率

词语	频率	左信息熵	右信息熵	网页数
下一页	21 987	不存在	不存在	26 144 336
股份	21 062	4.94021	3.52753	26 144 336
中国	21 033	5.29991	5.02516	764 219 624
公司	20 081	3.48893	4.75456	578 556 549

Table 9 The frequency and information entropy of the words in content corpus

表 9 基于网页正文语料库词语的信息熵和频率

词语	频率	左信息熵	右信息熵	网页数
公司	489 850	4.13603	5.34761	578 556 549
时间	195 472	1.48874	4.64484	387 914 356
投资	183 939	4.77629	4.08081	188 029 863
作者	155 169	5.58619	0.77377	96 825 229

基于上述算法在 3 种语料库上所做的实验得到的统计结果见表 10.

Table 10 The new-word number and probability in different corpus

表 10 不同语料库新词数和成词概率

语料库	词语数	成词概率
查询	2 915	**0.952**
锚文本	12 228	0.836
全文	19 031	0.829

TF/IDF 是词语在文档中重要性的一种指标.术语的明显特征是在本领域出现的次数远远高于非本领域出现的次数.TF/IDF 在一定程度上能够反映这样的特征.

依据词语出现的次数进行排序,在不同的语料库得到有序的结果.对结果进行标注,检查其是否为财经类词语.取前 10 位、前 100 位、前 1 000 位(如果存在)及全部分别标注,并计算其精确度见表 11(其中,相对频率阈值表示滤除的比例).

Table 11 The accuracy of finace words in different corpus

表 11 不同语料库在财经领域词语准确率

相对频率(%)	语料库	成词概率	*P*@10	*P*@100	*P*@1000	*P*
0	查询	**0.952**	**0.8**	**0.66**	0.471	0.344
	锚文本	0.836	**0.8**	0.63	**0.524**	**0.595**
	全文	0.829	0.7	0.53	0.416	0.423
10	查询	**0.963**	0.8	**0.65**	0.473	0.352
	锚文本	0.884	**0.9**	0.62	**0.602**	**0.625**
	全文	0.851	0.6	0.59	0.424	0.436
30	查询	0.98	0.8	**0.72**	0.537	0.385
	锚文本	**0.917**	**1**	**0.72**	**0.69**	**0.823**
	全文	0.852	0.6	0.6	0.506	0.445
60	查询	**0.982**	0.8	0.82	0.652	0.638
	锚文本	0.892	**1**	**0.96**	**0.836**	**0.865**
	全文	0.864	0.8	0.83	0.644	0.465

由于没有完整的标注好的财经领域术语集合,计算抽取出领域术语的召回率需要用特殊的方法实现,我们采用 TREC 中使用的标注池的方法.将相对频率过滤 60%后的结果的前 1 000 个放入标注池中进行标注,将该集合作为计算召回率时的术语集合,计算得到各个语料库的召回率见表 12.

Table 12 The recall of finace words in different corpus

表 12 不同语料库在财经领域词语召回率

语料库	召回率(%)
查询	39.9
锚文本	51.2
正文	39.6

就新词发现算法而言,其在各个语料库表现均不错,都得到了较好的结果.基于相对频率过滤后,成词概率均在 0.8 以上,其在查询语料库中的表现最好.分析其主要原因在于,查询往往是以词的形式出现,而不是以句子的形式出现,所以其以更大概率成词.

基于上述的新词发现算法,对 3 种语料库进行实验,最终得到的结果见表 11 和表 12,从表中可以发现以下几种情况:

(1) 3 种语料库都有很大一部分词语是财经类词语;

(2) 对比每种语料库在不同相对频率过滤下的准确度,在各个区间的准确率总体的趋势为上升的,尤其是在过滤 60%之后,各个语料库的领域术语概率都有很大的提高.所以,相对频率过滤的方法是适用于领域术语的抽取;

(3) 对于每种语料库而言,其前 1 000 位的准确率不一定比最终的准确率高,主要原因是标注时的排序是基于词语在语料库中出现的次数,而两字词在语料库出现的次数会明显高于多字词语.所以在标注时,排名前 1 000 位的词语往往是两字词居多;而通过抽样标注,两字词的领域术语概率低于 4 字词领域术语概率.所以,排名较后的多字词能够提升总体的领域术语概率;

(4) 对于每种语料库而言,其前 1 000 位的召回率结果都不是特别高,主要原因有:各个语料库得到的结果交集较少.基于标注池的方法,计算召回率的候选全集较大,所以每种语料库得到的召回率均不高.

对比 3 种语料库的领域术语概率,在不同的相对频率过滤下,锚文本的领域术语概率均高于查询文本和全文本,而且从表 11 中可以看到,锚文本的成词概率并不是明显高于其他语料库,而且查询文本的语料库的成词概率远远高于其他语料库,所以,并不是成词概率导致锚文本语料库的领域术语准确率提高.对于全文文本,其内容庞杂,往往很多常用词出现的概率更高,相对频率过滤中的全文频率只能通过搜索引擎得到,并不能得到一个精确的数字,影响了该过程对常用词的过滤.

5 结论和进一步的研究

3 种语料库在该算法下都得到了不错的结果,它们的特点为:

(1) 查询文本.该文本在算法下的成词概率很高,但是领域术语的概率较低,主要原因在于对于是否为该领域的查询标注有可能带来部分噪音;

(2) 锚文本.该语料库在算法下表现最好,能够得到最好的结果,这是因为锚文本有相对专业的网站制作者编写,有较高的质量,所以可以得到较好的结果.值得一提的是,锚文本在网页提取操作中也较为方便,提取该文本作为语料库在工程实践中也是较为可行的;

(3) 全文文本.该语料库在该算法下表现最差,因为全文中所含信息量很大,其中,领域术语在全文文本中所占的比例也是最低的.而且对于全文文本的提取是一项庞杂的工程,很多噪音都无法滤除,如广告、垃圾等.在整个语料库中有太多的无用信息,在提取该语料库中的词语和专业术语时,需要大量的时间,并且无法得到满意的结果.

通过上述实验,基于多重过滤的新词发现算法是可行的,其在各种语料库中均有不错的表现.对比 3 种语料库可知,语料库的质量很大程度上会影响最后的结果,基于锚文本的语料库可以得到更好的结果.随着过滤阈值的提高,其精确度会越来越高,但召回率会相对减少.本文算法在领域词典构建,领域热词新词发现中均有很好的前景.以后的工作将对该算法进一步加以改进,并将该结果用于领域词典的构建和新词热词发现中.

References:

[1] Velardi P, Missikoff M, Basili R. Identification of relevant terms to support the construction of Domain Ontologies. In: Proc. of the Workshop on Human Language Technologies and Knowledge Management. ACM Press, 2001. 1−8.

[2] Maedche A, Staab S. Ontology learning handbook on ontologies in information system. Heidelberg: Springer-Verlag, 2004. 173−190.

[3] Liu J, Liu YC, Jiang W, Wang XL. Research on automatic acquisition of domain terms. In: Proc. of the Int'l Conf. of Machine Learning and Cybernetics (ICMLC 2008). 2008.

[4] Gao R. The study of domain dictionary [MS. Thesis]. Harbin: Harbin Institute of Technology, 2008 (in Chinese).

[5] Frantzi KT, Ananiadou S, Tsujii J. The *C-value/NC-value* method of automatic recognition for multi-word terms. Journal of Natural Language Processing, 1999,6(3):115−130.

[6] Pantel P, Lin DK. A statistical corpus-based term extractor. In: Proc. of the Conf. on AI. 2001. 36−46.

[7] Bonin F, Dell'Orletta F, Venturi G, Montemagni S. A contrastive approach to multi−word term extraction from domain corpora. In: Proc. of the 7th Int'l Conf. on Language Resources and Evaluation (LREC 2010). 2010. 3222−3229.

[8] Liu T, Liu BQ, Xu ZM, Wang XL. Automatic domain-specific term extraction and its application in text classification. Acta Electronica Sinica, 2007,35(2):328−332 (in Chinese with English abstract).

[9] Luo ZY, Song R. An integrated method for chinese unknown word extraction. http://acl.ldc.upenn.edu/W/W04/W04-1122.pdf

[10] Wan J. The research on text extractionfrom Web pages [MS. Thesis]. Nanchang: Nanchang University, 2010 (in Chinese).

[11] Liu YT, Gao B, Liu TY, Zhang Y, Ma ZM, He SY, Li H. BrowseRank: Letting Web users vote for page importance. In: Proc. of the SIGIR 2008. 2008. 451−458.

[12] Hu JD. Research on Web news extraction and duplicates elimination [MS. Thesis]. Hangzhou: Zhejiang University, 2011 (in Chinese with English abstract).

附中文参考文献:

[4] 高锐.基于 Web 的领域词典构建技术研究[硕士学位论文].哈尔滨:哈尔滨工业大学,2008.

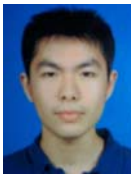
[8] 刘桃,刘秉权,徐志明等.领域术语自动抽取及其在文本分类中的应用.电子学报,2007,2(35):328−332.

[10] 万晶.Web 网页正文抽取方法研究[硕士学位论文].南昌:南昌大学,2010.

[12] 胡金栋.网页正文提取及去重技术研究[硕士学位论文].杭州:浙江大学,2011.

闫兴龙(1986－),男,甘肃民勤人,助理工程师,主要研究领域为信息检索,推荐系统.
E-mail: yan-xinglong@163.com

刘奕群(1981－),男,博士,副教授,CCF 高级会员,主要研究领域为网络搜索性能评价.
E-mail: yiqunliu@mail.tsinghua.edu.cn



方奇(1984－),男,博士,主要研究领域为信息检索,推荐系统.
E-mail: qqlevin@126.com

张敏(1977－),女,博士,副教授,CCF 高级会员,主要研究领域为信息检索,机器学习,网络用户行为分析.
E-mail: z-m@tsinghua.edu.cn

马少平(1961－),男,博士,教授,博士生导师,主要研究领域为智能信息处理,包括模式识别,文本信息检索,图像信息检索,中文古籍的数字化与检索.
E-mail: msp@mail.tsinghua.edu.cn

茹立云(1981－),男,博士,CCF 会员,主要研究领域为信息检索.
E-mail: lyru@vip.sohu.com