

基于CRFs的领域爆发词识别的研究与实现

逯万辉^{1,2}, 马建霞¹

(1.中国科学院国家科学图书馆 兰州分馆/中国科学院资源环境科学信息中心, 甘肃 兰州 730000;
2.中国科学院研究生院, 北京 100080)

摘要:通过对爆发词识别问题的研究和剖析, 本文采用了基于条件随机场模型的方法进行爆发特征提取, 在此基础上设计了频次、频率和词频文档比三个指标进行计算, 选取镍钴产业专利文本为例进行了领域爆发词识别实验, 并实现了爆发词识别系统原型的开发。

关键词:爆发词; 爆发特征; 条件随机场; 原型系统

中图分类号:G254.9 **文献标识码:**A **文章编号:**1007-7634(2014)01-89-05

Research and Implementation on the Domain Burst Word Recognition
Based on CRFsLU Wan-hui^{1,2}, MA Jian-xia¹

(1.Lanzhou Branch of the National Science Library / Scientific Information Center for Resources and Environment, Chinese Academy of Sciences, Lanzhou 730000, China; 2.Graduate University of Chinese Academy of Sciences, Beijing 100080, China)

Abstract: On the base of research and analysis the problem of burst word recognition, this paper extracted the burst feature based on CRF model, then designed three indexes to calculate the weight of burst word, developed a prototype system and experimented on patent text of Ni/Co.

Key words: burst word; burst feature; CRFs; prototype system

1 引言

在信息化高速发展的今天, 人们获取信息的途径和方式越来越多、获得的信息量倍增, 但却容易卷入无序的信息海洋中, 难以获得有用的知识, 因此, 探索基于人工智能的自动知识发现技术一直是目前的研究热点, 并出现了话题检测与追踪、舆情监控等众多新兴研究领域, 将网络信息的处理问题转化为通过程序的方法自动识别话题及其演变的过程。爆发词作为信息意图的最直观表达, 正确识别并处理爆发词对认识事件进展和了解事物变化有重要的借鉴意义。爆发词是指那种在一段时间大量出现的有意义的代表话题走向的词^[1]。从有关

爆发词的描述可知, 关于网络环境下爆发词的识别, 需要进行候选爆发特征识别、标记特征出现的时间、统计并得到爆发词等三个部分。

爆发词识别作为突发监测方法的基础性工作, 正确识别爆发词对突发主题监测和话题追踪具有重要作用, 但也是整个工作的难点和重点所在。除了应用于话题检测与跟踪、舆情监控等领域之外, 在情报分析与应用方面, 基于文本内容分析的知识发现研究也是一种重要的情报研究方法^[2], 其基础工作也是文本词汇的识别和处理。识别科技爆发词可以作为技术预测的前期准备和基础性工作, 对研究热点和研发机会的发现有重要作用, 因此, 正确、有效地捕捉潜在科技爆发词对科学研究趋势预测、研究热点和研发机会发现、科技监测等均有重

收稿日期: 2012-05-12

基金项目: 中国科学院西部之光联合学者项目“基于计算情报方法的甘肃省战略新兴产业竞争发展研究”项目

作者简介: 逯万辉(1987-), 男, 河南人, 报学硕士生, 主要从事信息抽取与数据挖掘研究。

要的研究意义和现实意义。

爆发词识别的基础是词的处理和计算,词单元作为爆发特征的最基本特征,也是携带信息的最小语义单元,处理中只需要在文本切分的基础上剔除无意义的停用词即可获得特征词,但是针对具体领域内容,单个的词语已不能表达完整的语义信息,更多的需要从领域术语上探讨该领域知识的演变和进化,因此就需要进行未登录术语的自动识别,即在词语的基础上探索携带完整信息的特征词或短语。本文将研究重点侧重在爆发特征识别及统计处理工作上,在文本粗切分的基础上进行特征词识别、采用基于条件随机场(Conditional Random Fields, CRFs)的方法进行未登录术语识别研究,从而得到候选爆发特征,以此为基础进行统计分析得到爆发词,并以专利文本为例进行领域爆发词识别实验和实现。

2 爆发词识别研究进展

关于爆发词识别的研究,其基础是对词正确的切分和获取,重点是计算时间段内爆发特征的爆发强度,进而获取爆发词。爆发词的识别是主题探测技术的一部分,来源于话题检测与追踪领域,因此,了解该领域的研究进展对我们进行爆发词识别有一定的借鉴意义。目前已有较多学者和科技公司进行了话题检测与追踪的研究方法和技术的研究,并已有一些研究成果和系统出现,其中在科学研究领域较为著名的有 ThemeRiver 和 Citespace。

ThemeRiver^[3]是一个可视化的系统,它用河流做隐喻,来描述文章主题随时间的变化,其主题的变化随着外部事件的时间线索而显示出来。主题河是由术语的频次支流组成,支流的宽度依据术语在不同时间段上出现频次的不同而发生变化,某个外部事件的出现可能会伴随着主题强度的突然变化,即在术语的基础上进行频次计算作为主题强度,并研究主题随时间的变化问题。Citespace^[4]中提出运用尽可能广泛的专业术语来确定一个知识领域的思路,通过提取研究前沿术语,(该前沿术语由在题目、摘要、系索词(指标引文献主题的单元词或词组,即关键词)和文献记录的标识符中提取出的突变专业术语而确定),进而将这些术语用做专业术语和文章异质网络中的聚类标注,每一年主题变化趋势是由使用比例最高的前五个标题词来识别,从而获得主题演变轨迹。

从以上关于主题探测的系统中可以看出:主题识别目前都通过主题词的识别来表现,主题探测要以主题词的识别为基础,主题词的获取又需要正确标识领域术语等特征词汇,因此,爆发词的识别也可以说是主题识别的一部分、和主题识别是一个相互交织的问题,其共同的基础点就是都需要以正确标识特征词汇为基础。但爆发词识别是以词语为研究和处理对象,更强调对正确词汇识别的依赖性,正确获取特征词汇对爆发词识别结果具有重要的影响,因此除剔除停用词典的停用词外,正确识别未登录词语也是关键的一个问题。

综上,本文在爆发词识别研究上将从特征词汇的获取入手,以中文文本为对象,在特征词的基础上进行计算从而获得主题爆发词。有关特征词汇的处理,中文信息处理较之英文文本处理有更为复杂的地方,即需要进行文本语句切分,关于中文切分词,中科院计算所的 ICTCLAS^[5]等在内的众多研究机构和公司都进行了较为深入的探索并形成了相应的系统,本文在已有切分词的基础上进行二次处理,采用条件随机场模型(Conditional Random Fields, CRFs)进行识别研究,并将其作为爆发特征为后续爆发词识别提供计算基础。

3 爆发词识别模型及算法设计

通过以上关于主题探测和爆发词识别研究的对比分析和对其研究基础的深入剖析,本文从爆发特征的获取入手,构建了基于条件随机场模型的爆发特征识别模型,并在此基础上设计指标算法来进行爆发词识别实验及系统实现。

3.1 基于条件随机场的爆发词识别模型

本文构建了如下爆发词识别模型:

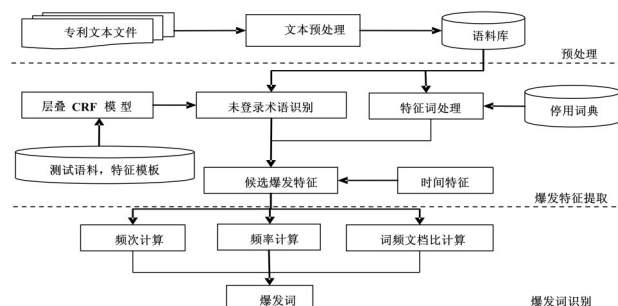


图1 专利文本爆发词识别技术路线

该模型主要包含以下几个步骤:

(1)文本预处理:根据选择的处理对象,获得待

处理文本并进行语句粗切分,即切分词、词性标注等工作(本文采用ICTLAS进行处理),形成语料库。

(2)候选特征提取:即进行停用词标注、未登录识别工作。这一步主要包含以下两个方面的工作:

①构建停用词词典:即对一些常用词,如“的、了、方法”等虚词、无意义的常用词等进行过滤,减小爆发特征集合,简化运算;②进行未登录词(领域未登录术语)识别:这是候选特征提取的难点,也是本研究的一项重点工作,本文采用基于条件随机场的方法进行未登录术语识别研究,通过构建训练语料、构建CRF特征模版,并在此基础上运用CRF++工具^[5]进行未登录术语的训练和测试实验。

(3)时间特征提取:时间信息是爆发词识别中的一项重要内容,特别是在网络文本等非结构化和半结构化文档中,本实验采用半结构化文本进行实验,即采用专利数据库中的专利文本为实验对象,这就为时间信息的获取提供了便利,因为数据库中文本的出现时间有明显的字段标识,只需在抓取文本的同时抽取出相应的字段即可。

(4)构建(特征词,时间)二元组,进行爆发特征的计算、识别爆发词。本研究采用三个指标进行计算,分别是词语的频次计算、频率计算和词频文档比计算,前两个指标主要是基于自然维度和时间维度来获取爆发词,第三个指标的构建目的是获得文档维度上的相对重要词,关于指标算法的说明,将在后续小节进行重点介绍。

3.2 爆发词识别算法设计

关于爆发词的识别,设定时间段及在获取爆发特征集合之后,怎么正确获得爆发词及计算爆发权重就是一个比较重要的问题。本文采用了词语的频次、频率和词频文档比三个指标,下面将解释这三个指标的不同含义以及在计算中的差异:

(1)频次,即某一时间段内词的出现次数,公式表示为:

$weight = n_w(t)$,其中 $n_w(t)$ 表示词 w 在时间 t 内出现的次数。

该指标本文也将其称为爆发词识别的自然维度,即通过计算获得某段时间内大量出现的词,即爆发词。

(2)频率,即词频除以该时间段出现的所有词,该指标称为爆发词识别的内在维度,即通过计算某段时间内词频与总词频之比,得到词的出现频率。该指标是在频次的基础上进行的加工和改造,是对

频次的归一化处理,公式表示为:

$weight = \frac{n_w(t)}{N(t)}$,其中 $n_w(t)$ 表示词 w 在时间 t 内出现的次数, $N(t)$ 表示 t 时间的总词数。

例如,某词 A 在 X 年出现了 20 次,该年共出现了 50 个词;在 Y 年出现了 30 次,该年共出现了 100 个词,那么根据频次和频率的计算结果就是不一样的,该方法能获得该时间段内相对重要程度高的词。

(3)词频文档比,即词频与出现该词的文档数之间运算后得出的值,是在 TF-IDF 基础上的改造,公式表示为:

$weight(w, t) = \frac{n_w(t)}{N(t)} \log \frac{N_D(t)}{DF_w(t)}$,其中 $n_w(t)$ 表

示词 w 在时间 t 内出现的次数, $N(t)$ 表示 t 时间的总词数, $N_D(t)$ 表示 t 时间的总文档数, $DF_w(t)$ 表示 t 时间出现词 w 的文档数。

指标计算词的词频与出现该词的文档数之间的关系,该方法能剔除文档累积效应对词的干扰,通过计算词的逆文档频率,即出现某词 w 的文档数越多,说明该词的文档普遍性越强,该词的相对重要程度就越低,该指标主要用来获得文档间的相对重要程度高的词汇。

4 爆发词识别的系统实现

4.1 爆发特征获取

爆发特征词包含去除停用词之后的词语集合和正确识别出来的未登录词。有关停用词的识别和处理较为简单,本文针对爆发特征的获取主要解决未登录词的识别问题。

目前针对未登录词和新词的研究仍然是存在较大问题,有关该问题的研究,在自然语言处理领域主要存在两种思路:基于规则的方法和基于统计的方法。基于规则的方法通过研究未登录词的构成结构和存在形式、短语型术语的构造等内容设计识别规则,进而实现未登录词和术语的识别,该方法在有限的语料规模下表现不错,但是需要消耗大量的人力资源,同时总结的规则也不够全面、扩展性较差,在网络环境下大规模语料的实验中表现却不尽人意。因此,基于统计的方法由于其严谨的理论基础和良好的表现效果更被多数研究者采用,同时结合基于规则的方法,研究者基于该方法只需构

建一个合适的测试语料库,通过调整机器自动学习的参数就可快速获取结果。目前在基于统计的未登录术语识别的研究中,基本的研究思路是将未登录术语看成一个词语组合,将未登录术语的识别看成一个词语边界标识的问题,通过特定的标识符将未登录术语的起始位置、中间位置、结束位置进行标记,然后进行未登录术语的提取,这就将该问题转化成了一个文本序列分类的问题。

```

for (String s : list) {
    if ("".endsWith(s.trim())) {
        isNewDoc = true;
        resultList.add(tmpList);
        tmpList = new ArrayList<String>();
    } else {
        String[] stringArray = s.split("\\t");
        String tmp = stringArray2 == null ? "" :
stringArray2
                .trim();
        // 如果一行的第三个字母是B或者BI,那么将该
// 行的第一个字符串记录下来
        if (("B".equals(tmp) || "BI".equals(tmp))) {
            if (isLastTerm) {
                isNewTerm = true;
                tmpList.add(toRecord);
            }
            isLastTerm = true;
            toRecord = stringArray[0];
        }
        // 如果一行的第三个字母是I,那么将该行的第
// 一个字符串加入到已经有 toRecord 的末尾
        if ("I".equals(tmp) && isLastTerm) {
            toRecord=toRecord.concat(stringArray[0]);

```

图2 CRF标注结果处理关键代码

当前在解决基于文本序列标注的未登录词识别时,主要采用隐马尔科夫模型(Hidden Markov models, HMM)、最大熵马尔科夫模型(Maximum Entropy Markov Models, MEMM)和条件随机场模型(CRF)这三种方法,其中HMM属于产生式模型, MEMM和CRF属于判别式模型, MEMM模型是对转移概率和表现概率建立联合概率,计算时统计的是条件概率,但MEMM容易陷入局部最优,是因为MEMM只在局部做归一化,相反,条件随机场模型在观测序列的基础上对目标序列进行建模,重点解决序列化标注的问题。该模型统计了全局概率,在做归一化时,考虑了数据在全局的分布,而不是仅仅在局部归一化,它既具有判别式模型的优点,又具有产生式模型考虑到上下文标记间的转移概率,以序列化形式进行全局参数优化和解码的特点,解

决了其他判别式模型(如最大熵马尔科夫模型)难以避免的标记偏置问题^[6],因此,本文将采用基于条件随机场模型的方法进行爆发特征词识别实验的研究。

爆发特征的获取主要包含以下步骤:

(1)获取实验语料:本文采用专利文本为例进行了实验,选取镍钴产业领域专利文本摘要(专利号:C22B23/00)下所有专利文本,进行切分词、词性标注,及转化为CRF语料格式,形成语料库S:

(2)从语料库S中选取部分语料进行未登录术语标注,实验中采用BIO标注法,B表示术语块的开头、I表示术语块的后续、O表示其他非术语类型;

(3)构建CRF特征模板,采用CRF++工具对已标注语料进行训练,并生成模板文件;

(4)再次利用CRF++工具对未标注语料进行测试,生成CRF标注结果,对此进行整理,得到未登录术语,并将此作为候选爆发特征进行存储以供后面计算。有关标注结果处理的的关键代码如图2所示:

4.2 爆发词识别系统实现

在获取爆发特征的基础上,本系统采用3.2节中设计的爆发词识别模型和介绍的几个统计指标,对进行未登录词识别和停用词剔除后的爆发特征词进行了量化计算处理,以“年”为时间单位,分别计算词的频次、频率以及词频文档比,通过这些指标的计算和对文档累积效应和时间累积效应的处理,在此基础上获得爆发词,计算结果如图3所示:



图3 爆发特征各指标计算结果

在爆发词的识别上,本文通过对以上三个指标的计算,选取在每个指标下排名靠前的词汇再进行时间轴上的对比分析,进而获得爆发词,具体的实现方法就是:

(1)选取时间范围,获取高频词。本文的时间单位是“年”,根据指标得出该年词汇在该指标下的计算结果,即可以获取该年的高频词:

(2)该高频词并不是爆发词,只是在时间累积效应下文档累积效应较高的词,也可以看作是该词在

该年的权重,要识别爆发词,则需要将该词的权重放在时间轴上进行演绎和回溯,通过时间轴上的纵向比较获得爆发词。

4.3 结果分析与展示

通过以上指标权重的计算,仍不能得到爆发词,还需要通过对该年排名靠前的词的时间序列演化进行分析,即通过比较该词之前在时间轴上的表现,判断该词是不是爆发词,显然,根据不同的指标选取的爆发词是不同的,可以根据不同情况选取不同指标进行对比研究。图4是在不同指标下得到的爆发词识别结果:



图4 爆发词识别结果

通过选取爆发词识别结果中的某几个词进行时间轴上演变的可视化展示,可以更加清晰的看出该词的爆发效果,如图5所示(该图通过EXCEL生成)。

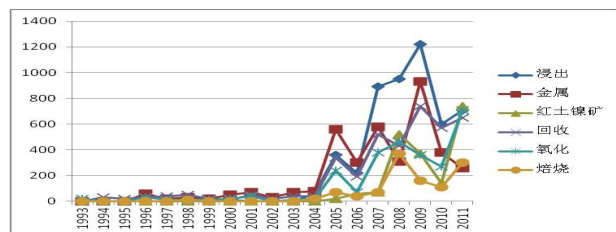


图5 部分爆发词可视化展示

可以看出,从前几年的变化来看,“浸出”一词在2009年呈现爆发态势,随后又出现减少,分析该词背后的含义,即运用化学还原法进行镍钴的提取研究和专利申请量在2009年呈现了一个较大的量;同样,“焙烧”一词在1993-2007年都处于较低的水平,但是在2008年也突然达到了一个小峰值,说明经过前几年的研究,基于“焙烧”的技术和方法正逐渐成熟,采用该技术的研究也在逐渐增加。

5 结 语

本文从爆发词识别的研究现状入手,经过调研分析和对该研究问题的深入剖析,从爆发词的基础性工作——特征词的识别开始,采用条件随机场模型的方法进行了候选特征词提取,并在此基础上设计了指标进行计算,并在Java环境下进行了开发试验,设计了一套爆发词识别的原型系统,从实验结果来看,该系统在解决爆发词识别的问题上是有效果,但是系统的结果和展示也有需要改进的地方:

(1)关于爆发特征词的语义化合并问题:从识别结果来看,一些意义相同的词的不同表现形式需要进行合并或区分,如“镍红土矿”和“红土镍矿”就需要合并、“炉子”“冶炼炉”和“熔炼炉”等在词义上的则需要区别处理,该步骤需要进行基于文档的篇章语义分析,较难处理,需要后续进一步的研究;

(2)关于结果的可视化展示问题:本文是将爆发词识别结果导入Excel进行了可视化展示,从效果来看更加直观方便,因此,后续本系统仍将研究增加爆发词识别结果显示功能。

参考文献

- 1 逯万辉,马建霞,赵迎光.爆发词识别与主题探测技术研究综述[J].情报理论与实践,2012,(6):125-128.
- 2 冷伏海,冯璐.情报研究方法发展现状与趋势[J].图书情报工作,2009,53(2):29-33.
- 3 Susan Havre, Elizabeth Hetzler, Paul Whitney, Lucy Nowell, ThemeRiver: Visualizing Thematic Changes in Large Document Collections[J].IEEE Transactions on Visualization and Computer Graphics, 2002,8(1): 9-20.
- 4 陈朝美.CiteSpace II:科学文献中新趋势与新动态的识别与可视化[J].陈悦,等,译.情报学报,2009,28(3): 401-421.
- 5 ICTCLAS 汉语分词系统.[EB/OL].<http://ictclas.org/>, 2012-09-12.
- 6 基于CRF的中文分词[EB/OL].<http://blog.csdn.net/wen718/article/details/5960820>, 2012-09-05.

(实习编辑:毛秀梅)