基于层叠隐马模型的屏蔽关键词研究

陶非凡
(上海海事大学 信息工程学院,上海 201300)

摘 要:信息时代给人们的生活带来巨大改善,但同时也伴随一系列问题的产生,其中如何对网络中产生的大数据量的言论信息进行过滤的问题是研究的一大难点。传统的屏蔽法效率较低而且不够准确,因此文中提出了一种新的关键词屏蔽技术。主要采用二元语法模型结合层叠隐马可夫分词技术,首先运用二元语法模型在大量语料中得到普通词和关键词的构成概率,建立一个有普通词和关键词分类的词典,再结合层叠隐马可夫模型对具体句子进行分词处理,对分词后的结果计算其关键词屏蔽概率,最终得到一个科学的屏蔽概率,可以大大提高关键词屏蔽的准确性。

关键词:关键词屏蔽;概率;层叠隐马尔可夫模型;分词
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2014)11-0167-03
doi:10.3969/j.issn.1673-629X.2014.11.042

Research on Shielded Keywords Based on Cascaded Hidden Markov Model

TAO Fei-fan
(College of Information and Engineering, Shanghai Maritime University, Shanghai 201300, China)

Absrtact: The information age brings a huge improvement in people's lives, but also accompanied by a series of problems arising, in which how to filter a large amount of information the network's remarks generated is a major difficulty. The traditional method of shielding has low efficiency and is not accurate enough, so propose a new keyword shielding technology. Mainly use binary syntax model combined with layered hidden Markov model segmentation techniques, first utilize binary syntax model to get the constitute probability of the common words and keywords in a large corpus, creating a dictionary of common words and keywords classified, then combined cascading hidden Markov model for the specific sentence word processing, calculate the probability of its keywords shield for segmented result, finally get a scientific shielding probability, which can greatly improve the accuracy of keyword shield.

Key words: keywords shield; probability; cascading hidden Markov model; word segmenting

0 引 言

随着信息时代的到来,网络上的信息量也在剧增,每天都有数以千万的人们在网络上发表言论,这些言论有些文明得体,积极向上,有些却对社会造成一定的危害,因此对于网络上大信息量的过滤屏蔽问题显得尤为重要。在国内,这些信息主要以汉字的形式体现,如何实现对汉语句子的过滤和屏蔽成为研究的重点。

传统的关键词屏蔽是通过对句子中的词进行搜索匹配的方法,通常都被放在汉语分词前或后,不能和分词有机地结合起来。文中提出的这种方法是关键词屏蔽和汉语分词相结合,运用成熟的汉语分词技术来解决大信息量的关键词屏蔽的一些难题,大大提高关键词屏蔽的速率,并且结合分词技术提出一种科学的关键词屏蔽机制,抛弃传统的蛮力搜索匹配方法,使关键词屏蔽更具准确性。

中文分词[1-2]主要有基于词典和基于统计两种方法,但通常都是两者相结合。比较成熟的有层叠隐马尔可夫模型分词[3-4]、句法树分词[5]、N-最大概率法分词[6]、基于字位标注的分词[7]等。文中采用的是基于层叠隐马尔可夫模型分词,因为这种方法提出分类词法,对词语进行分类后可以更容易处理其中的关键词屏蔽问题。

收稿日期:2013-11-26 **修回日期**:2014-03-03 **网络出版时间**:2014-09-11
基金项目:国家自然科学基金资助项目(61070154,61373028)
作者简介:陶非凡(1991-),男,安徽枞阳人,硕士研究生,研究方向为云计算与云存储。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20140911.0942.005.html

1 二元语法模型概率词典

1.1 关键词屏蔽的技术难点

汉语作为一种语言本身有其复杂性,而词语关键词屏蔽又具有其本身特殊性,一个词是否被作为屏蔽关键词处理是需要考虑到其所在的语境和句子成分的,不能简单地根据这个词的存在与否来判断。比如在一些娱乐网站上要对一些政治词汇进行屏蔽,而在政府网站上需要对一些娱乐词汇进行屏蔽,也就是说不同的时间、不同的环境所需要建立的屏蔽词库都是不一样的,因此要建立一个灵活的关键词屏蔽机制。

对一个汉语句子进行处理,首先要对其进行分词操作,将句子分成一个个的简单词,同时进行关键词的识别和屏蔽。目前汉字分词中存在的主要难点包括切分时的分歧和识别词典中的未记录词语,切分时的分歧和词典中的未记录词语降低了汉字分词的准确率,同时增加了关键词的判定难度,如果遇见未登录词往往和切分歧义交织在一起,进一步增加了关键词屏蔽的难度。如:在“还不走私下回来了”中,“还不走”是一个词典中没有收录的网络人名,“走私”是一个屏蔽词,实际切分的时候,“走”与“私”,“下”与“回”往往会粘在一起,导致错误的切分结果“还/不/走私/下回/来了”,这样就导致错误屏蔽关键词,也就是说关键词屏蔽在具体句子的语法分析中也存在难题。

1.2 创建基于概率的词典

1.2.1 二元语法模型

在汉语分析中,首先是由字构成一个个词,然后由词构成句子,而字构成词可以通过统计得到,即在大量预料中统计两个字相邻并且构成一个词的概率,这里先介绍基于统计学的二元语法模型[8]。

N-gram 模型是在词语分词中应用较为广泛的语法统计模型。用 s 代表由一系列连续组合的词 $w_1,w_2,\cdots,w_n$ 结合而成的句子,计算机对该语句的识别就是计算语句 s 在整个句子中存在的几率。假设某个词 w 存在的概率使用 $P(w)$ 来表示,根据条件概率公式可知,s 语句在句子中存在的几率就是组成 s 的词语相乘的概率,于是 $P(s)$ 可展开为:

$$P(s)=P(w_1)P(w_2\mid w_1)P(w_3\mid w_1w_2)\cdots P(w_n\mid w_1w_2\cdots w_{n-1}) \tag{1}$$

其中,$P(w_1)$ 是代表词语 w_1 在句子中存在的概率;$P(w_2\mid w_1)$ 是在词 w_1 存在的前提下,词 w_2 存在的概率;不难得到,词 w_n 在句子中存在的概率是由它前面所有词语所决定的。从实际计算考虑,存在的情况太多,难以计算。因此文中假定任意词语 w_i 存在的概率只和它前面 N 个词语有关联,假设当 $N=1$ 时,那么这时词语 w_i 存在的概率只同它前面一个词语 w_{i-1} 有关,这就是所讨论的二元语法模型。由此,s 存在的概率为:

$$P(s)=P(w_1)P(w_2\mid w_1)P(w_3\mid w_2)\cdots P(w_n\mid w_{n-1}) \tag{2}$$

1.2.2 建立基于概率的词典

那么如何估算 $P(w_i\mid w_{i-1})$ 的值呢?在进行大量语料库统计和学习后,可以得到统计后的取值,假设该词语 (w_{i-1},w_i) 根据统计后的结果,其在文本中出现的次数为 $C_{i-1,i}$,同样 w_{i-1} 在文本中存在时与该词相邻的次数为 C_{i-1},则可以得到:

$$P(w_i\mid w_{i-1})\approx C_{i-1,i}/C_{i-1} \tag{3}$$

在大规模语料库训练的基础上,根据大数定理,即在大样本统计的前提下,样本的频率接近其概率值,所以 $P(w_i\mid w_{i-1})$ 就可以作为普通词或关键词组成概率存入词库。通过这种方法,可以创建一个基于概率的词典,分别得到普通词和屏蔽关键词的概率词典。如普通词词典中有“吃饭 85%”,即“吃饭”出现在分词结果中,准确率为 85%;同理屏蔽关键词词典中有“贩毒 65%”,分词结果中如出现“贩毒”,它是一个屏蔽词的概率为 65%。使用二元语法模型所建立的词典结构如图 1 所示。

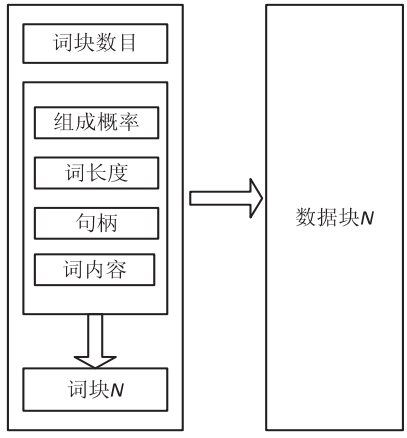

图 1 词典结构

1.2.3 数据平滑

受限于训练模型的语料库规模,在语料库中未出现的词并不代表其真实概率等于 0,而出现次数相同的词其概率也并不一定一样,因此,引入数据平滑技术来应对二元语法的数据稀疏现象,调整最大似然估计的概率值,使零概率增加,非零概率下降,以提高模型的整体正确率。目前常用的数据平滑算法[9-11]有折扣法、回退法、插值法等,文中使用的是文献[12]介绍的 Additive Smoothing[12],具有一定的先进性。

2 分词与关键词屏蔽相结合

2.1 改进的层叠隐马尔可夫模型

隐马模型(Hidden Markov Model,HMM)在很多领域都得到了广泛的应用,同样在自然语言处理上隐马模型也十分有效。然而,与一般领域相比,自然语言无

论是在层次上还是在内容上都更为复杂,因此文中使用的是一种多层的隐马尔可夫模型,称为层叠隐马尔可夫模型(Cascaded Hidden Markov Model, CHMM)。和传统的隐马模型相比,层叠隐马模型是由多个隐马模型组合而成,各个层次间使用以下几种方式关联,构成紧凑的耦合关系:各层间使用同样的切分词图作为它们公共区域的数据结构;所有层次模型策略全部使用 *N*—Best 策略,每层得到的最好结果作为该层的最终结果汇总到更高层次使用;低层次的模型在向高层输送数据时,同时也记录这些数据的相关参数。

针对关键词屏蔽和分词[13]中各个层面的处理对象及问题特点,引入 CHMM 统一建模,该模型包含原子切分、未登录词识别、屏蔽关键词识别、基于类的隐马切分、词类标注共 5 个层面的隐马模型(见图 2)。

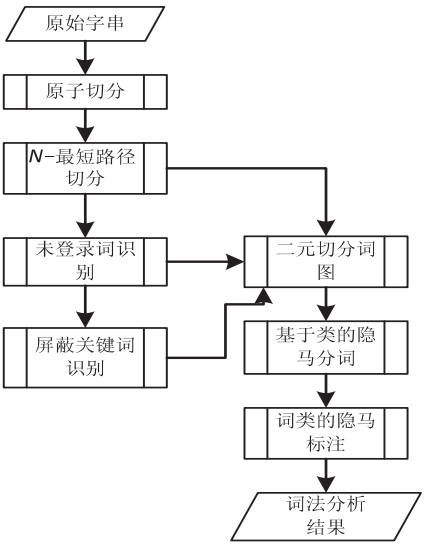

图 2　基于 CHMM 的关键词屏蔽分析框架

基于类的隐马分词[14]算法,把所有的词分为以下 3 种类型:

(1)词典中的普通词;

(2)词典中的屏蔽关键词;

(3)未登录词。

对未登录词识别可以使用角色标注的方法。

2.2　关键词屏蔽概率计算

设中文句子 S 由字序列组成,即 $S=\{C_1, C_2, \cdots, C_n\}$,字以不同的方式组合在一起成为词。经过隐马可夫分词会得到一个句子的多种切分方案,S 的第 j 种切分称为 W_j,m_j 为 W_j 中词的个数,第 k 个词为 $W_{j,k}$($0<k<m_j$)。如果所有词确定,则句子唯一确定,对于第 j 种切分有:

$$P(W_j)=P(W_{j,1})P(W_{j,2})\cdots P(W_{j,m_j}) \tag{4}$$

假设一个句子共得到 a 种切分方案,其中 $1\sim b$ 种为含有屏蔽关键词的切分方案,$b+1\sim a$ 种为不含有关键词的切分方案,则这个句子含有屏蔽关键词的概率为:

$$P(\text{shield})=\sum_{i=1}^{b}P(W_i)/\sum_{k=1}^{a}P(W_k) \tag{5}$$

通过设置标准值 q_1, q_2, q_3 来设置屏蔽度,可以用来针对不同的环境需求。

2.3　训练与分词屏蔽流程

模型处理过程主要由训练、分词和屏蔽三部分构成:首先通过大量语料的训练得到贝叶斯先验分布、二元语法信息,然后使用隐马可夫模型结合二元语法进行分词,最后对分词结果计算其屏蔽概率。屏蔽过程如图 3 所示。

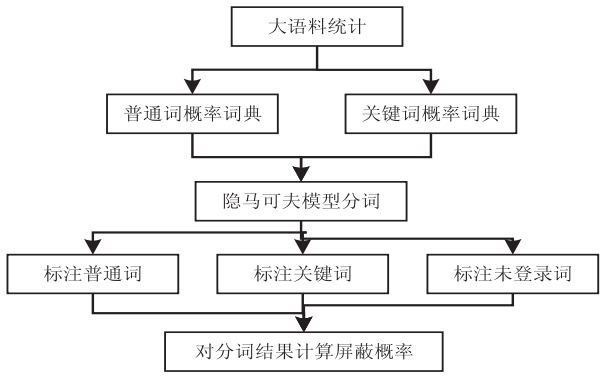

图 3　屏蔽过程

3　结束语

文中提出了一种新的关键词屏蔽技术,可以计算出某句话中含有屏蔽关键词的概率,在一定程度上提高了关键词屏蔽的概率和准确性。但是还有一定不足之处,对如何确定一个科学的值作为关键词屏蔽的阈值,即如果计算结果大于它,则确定为需要进行关键词屏蔽,这个值需要通过对具体的语料进行分析统计得到,是未来这个方面研究的一个方向。

参考文献:

[1]　龙树全,赵正文,唐　华. 中文分词算法概述[J]. 电脑知识与技术:学术交流,2009,5(4):2605-2607.

[2]　Emerson T. The second international Chinese word segmentation bakeoff[C]//Proceeding of fourth SIGHAN workshop on Chinese language processing. Korea:[s. n.],2005.

[3]　刘　群,张华平,俞鸿魁,等. 基于层叠隐马模型的汉语词法分析[J]. 计算机研究与发展,2004,41(8):1421-1429.

[4]　张华平,刘　群. 基于 N-最短路径方法的中文词语粗分模型[J]. 中文信息学报,2002,16(5):1-7.

[5]　刘　挺,马金山. 汉语自动句法分析的理论与方法[J]. 当代语言学,2009(2):100-112.

[6]　王晓龙,王开铸. 最少分词问题及其解法[J]. 科学通报,1989,34(13):1030-1032.

[7]　苑春法,陈　刚,黄昌宁. 基于词性和语义知识的汉语句法

(下转第 174 页)

器件,应有较高鲁棒裕度,亦应分层予以规定。前述对于网电空域仿生防护所进行的分级、分层规划不仅仅是一种标准化的措施,更是策略上的选择和战略上的部署。在复杂的仿生系统中,可根据外来干扰的时频特性及功率强度等,设计不同的阵列冗余及鲁棒特性。此时亦必然存在着电子系统与生物系统的鲁棒性研究与对等标定的问题,即完成整体鲁棒性与局部鲁棒性之间的差异、仿生系统鲁棒性与被仿系统鲁棒性之间的差异研究,及层面、结构中鲁棒裕度问题的确定。

总之,分级冗余研究旨在解决系统受损后程度评估、修复与否、时机选择和可供使用资源的问题,而鲁棒裕度研究则是决定受损系统修复和恢复的程度问题。因此,这两项指标将为“模块化设计”赋予更新、更深的含义。

5 结束语

在借助生物进化的概念和仿生规划的基础上,通过采用全新的思路、方法深入进行了网电空域的定义及其仿生防护等方面的理论研究。特别是从耗散结构、自律机制和鲁棒特性这三个方面,初步论述了网电空域中电磁防护仿生的原理和实现方法。进而,有望建立一种新型的防护研究模式,对传统干扰的抗扰方式进行强化、补充与完善,使得网电空域最终能够满足在不同层面上的多种安全、稳定的运行要求。可以看出,电磁防护仿生是尝试使用新理论、新器件解决网电空域中传统问题的一个研究切入点,又是将网电概念、生物技术、电磁防护与微电子技术等诸多学科结合之后诞生的一个技术增长点,甚至可能作为奠定网电空域仿生发展方向的一个长远立足点。因此,该类研究的提出和开展,既能满足网电空域发展的需要,又能符合科技发展的潮流。亦可望从整体拓宽电磁防护的科研领域与项目规模,将网电空域的安全与稳定研究推向一个新的阶段。

参考文献:

[1] 郑 箇,史燕中. 计算机网络电磁安全的脆弱性分析及防护对策[J]. 保密科学技术,2011(9):57-60.

[2] 廖晓阳,陈徐飞,于鑫刚,等. 赛博空间技术的军事应用研究及对策[J]. 电子科技,2011,24(11):147-149.

[3] 吴 巍. 赛博空间技术发展现状与通信网络安全问题[J]. 无线电通信技术,2012,38(3):1-4.

[4] 李 昊,龙晓波. 赛博行动与电子战[J]. 中国电子科学研究院学报,2011,6(3):240-242.

[5] 林 峰,舒少龙. 赛博物理系统发展综述[J]. 同济大学学报(自然科学版),2010,38(8):1243-1248.

[6] 吕信明. 关于网络电磁空间战的思考[J]. 国防科技,2012,33(4):1-7.

[7] 路甬祥. 仿生学的科学意义与前沿[J]. 科学中国人,2004(4):22-24.

[8] 刘尚合,原 亮,褚 杰. 电磁仿生学—电磁防护研究的新领域[J]. 自然杂志,2009,31(1):1-7.

[9] 原 亮,巨政权,满梦华,等. 仿生层级标定与电磁仿生模型建立[J]. 高技术通讯,2012,22(6):631-637.

[10] 何传启. 第六次科技革命的中国战略机遇[J]. 决策与信息,2012(6):20-22.

[11] Wang Zhongqiang, Xu Haiyang, Li Xinghua, et al. Synaptic learning and memory functions achieved using oxygen ion migration/diffusion in an amorphous InGaZnO memristor[J]. Advanced Functional Materials,2012,22(13):2759-2765.

[12] 巨政权,原 亮,满梦华,等. 电子系统的层次分解及建模[J]. 现代电子技术,2011,34(5):18-20.

[13] 满梦华. 嵌入式异构冗余容错计算系统的研究与实现[D]. 石家庄:军械工程学院,2009.

[14] 吕大刚,宋鹏彦,崔双双,等. 结构鲁棒性及其评价指标[J]. 建筑结构学报,2011,32(11):44-54.

[15] 崔新风. 基于神经网络结构模型的数字电路演化设计与实现[D]. 石家庄:军械工程学院,2010.

[16] 黄 涛,夏 志,汪清祥,等. NMDA 受体在运动易化学习记忆突触机制-LTP 中的作用[J]. 广州体育学院学报,2008,28(4):100-104.

(上接第 169 页)

规则学习[J]. 中文信息学报,2001,15(3):1-8.

[8] 吴应良,韦 岗,李海洲. 一种基于 N-gram 模型和机器学习的汉语分词算法[J]. 电子与信息学报,2001,23(11):1148-1153.

[9] 黄建中,王肖雷. Katz 平滑算法在中文分词系统中的应用[J]. 计算机工程,2004,30(B12):371-372.

[10] Chen S F, Goodman J. An empirical study of smoothing techniques for language modeling[C]//Proc of the 34th annual meeting on association for computational linguistics. Stroudsburg: Association for Computational Linguistics, 1996: 310-318.

[11] Gale W A, Sampson G. Good turing frequency estimation without tears[J]. Journal of Quantitative Linguistics, 1995, 2(3): 217-237.

[12] 翟海保,程浩忠,吕干云,等. 基于模式记忆并行蚁群算法的输电网规划[J]. 中国电机工程学报,2005,25(9):17-22.

[13] Kit C, Pan Haihua, Chen Hongbiao. Learning case-based knowledge for disambiguating Chinese word segmentation: a preliminary study[C]//Proc of first SIGHAN workshop attached with the 19th COLING. Taipei: [s. n.], 2002.

[14] 张华平,刘 群. 基于角色标注的中国人名自动识别研究[J]. 计算机学报,2004,27(1):85-91.

基于层叠隐马模型的屏蔽关键词研究

作者： 陶非凡， TAO Fei-fan
作者单位： 上海海事大学 信息工程学院,上海,201300
刊名： 计算机技术与发展 ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2014(11)