

网络舆情监控中新词识别问题的研究

唐籍涛¹, 李 飞², 郭昌松¹

(1. 成都信息工程学院 计算机系, 四川 成都 610225;

2. 成都信息工程学院 网络工程系, 四川 成都 610225)

摘 要:在网络舆情监控中,由于事件的突发性和网络词汇的泛滥,各种各样的新兴词汇以及新的字符串大量涌现,而有穷的分词词典对新词的识别基本上无能为力,这些无法识别的字符串将被现有的分词系统分为零散的碎片,这将极大地影响热点词和主题词提取的准确性,成为网络舆情监控系统性能提升的瓶颈。文中分析了当前主要的几种分词技术的优缺点,利用网络舆情监控中未被词典收录的主题词的局部高频这一特性,通过计算异常分词与周围分词之间的粘结度,从而识别出未被词典收录的主题词。实验结果表明:所提出的分词算法能识别出未被词典收录的主题词,相比传统的分词算法,更加适合于网络舆情监控。

关键词:网络舆情监控;新词识别;分词词典

中图分类号:TP31

文献标识码:A

文章编号:1673-629X(2012)01-0119-03

Research of New Word Pattern Recognition in Network Monitoring Public Opinion

TANG Ji-tao¹, LI Fei², GUO Chang-song¹

(1. Department of Computer Science, Chengdu University of Information Technology, Chengdu 610225, China;

2. Department of Network Engineering, Chengdu University of Information Technology, Chengdu 610225, China)

Abstract: With rapid development and deepen evolution of internet public opinion in the internet, a variety of new vocabulary and new string comes out due to the sudden of matters and the high frequency of new words occur on network, therefore, the current method of sub-dictionary has no effect on them in a large extent. The most important and most deadly is that those rare appear strings are divided into scattered fragments by the existing segmentation system, which will greatly affect the accuracy in extracting out the hot words and the keywords. Know that the situation will become the bottleneck of improving performance in network monitoring system. It analyzes the major advantages and disadvantages of several word segmentation and draw out the characteristics, using the local high-frequency of the keyword not included into dictionary in the monitoring public opinion, then calculating the anomalous bond between the abnormal words and its around words, finally, to identify the keywords not edit. The experiment shows: compared to the traditional segmentation algorithm, this segmentation algorithm can identify the keywords better and is more suitable for network monitoring public opinion.

Key words: network monitoring public opinion; new word pattern recognition; dictionary

0 引言

中文分词是网络舆情监控系统中的基础,它的性能直接影响聚类的效果,而新词识别一直都是中文分词的难点。目前主要的分词技术有三种:基于统计的分词、基于词典的分词和基于理解的分词^[1]。基于统计的分词能够识别出高频的新词,对于网络中低频的新词不能识别,同时它还需要大量的训练数据用于建立模型,计算量非常大,分词精度与训练数据的选取有

很大关系。基于词典的分词切分简单,但不能识别新词^[2]。基于理解的分词又名知识分词,它是一种理想的分词方式,它利用句法以及句子中的语义信息或者从大量语料中找出汉字组词的结合特点来进行评价,从而找到最贴近于原句语义的分词结果,目前还不存在可以很好工作的基于理解的分词算法,它只作为一个概念而存在^[3]。这三种方式都不能满足网络舆情监控对于分词的需求。文中结合网络舆情监控的特点,改进了目前的基于词典的分词技术,提出自动扩充词典机制,使它更适合于网络舆情监控。

1 中文分词技术

中文分词技术属于语言处理技术的范畴,对于文

收稿日期:2011-05-31;修回日期:2011-09-08

基金项目:四川省教育科研项目(川教函[2011]210号)

作者简介:唐籍涛(1986-),男,四川成都人,硕士研究生,研究方向为网络舆情监控;李 飞,教授,硕士生导师,研究方向为计算机应用与信息安全。

本中的一段文字,人可以识别出哪些是一个词汇,哪些不是一个词汇,应用计算机来模拟人的这一过程的方法就是分词算法。目前的中文分词算法主要有三类:基于词典的分词方法、基于理解的分词方法和基于统计的分词方法。国内有许多学者在这三种基本方法基础上做了一些改进。

1.1 基于词典的分词方法

基于词典的分词方法又叫做基于规则的分词方法,它的核心是一个能够被机器所识别的词典,该词典囊括了大多数的词汇,它分词的基本方法是待分析的字符串与词典中的词条进行匹配,若匹配成功,那么就识别出一个词汇。按照扫描方向的不同,该方法又分为正向匹配和逆向匹配;按照匹配长度的不同,可以分为最大匹配和最小匹配。按照是否标注词性,又可以分为单纯分词方法和分词与标注相结合的一体化方法^[4]。

常见的几种分词方法如下:

- 1) 逆向最大匹配法。
- 2) 正向最大匹配法。
- 3) 最少切分法(使每句话切出的词数最小)。

目前基于词典的分词占主流地位的是正向匹配法和逆向匹配法。一般来说,逆向匹配的精度要高于正向匹配的精度。有关统计结果表明,正向匹配的错误率为1/169,逆向匹配的错误率为1/245^[5]。产生错误的原因主要为某些词汇(如地名、人名以及一些生僻词和新兴词汇)并未被词典所收录,对于一些对这些未被收录的词汇不敏感的系统来说,这种精度能够达到其要求,但是这种精度很难达到网络舆情监控的要求。

基于词典的分词方法,原理简单,实用性强,但是该方法的最大缺点在于准确率很大程度上依赖于有穷的词典^[6]。据相关文献统计,用一个含有70000个词的词典去切分15000个词的语料库,仍然有30%以上的词没有被切分出来^[7]。

1.2 基于统计的分词方法

基于统计的方法的基本原理是:通过计算汉字同时出现的概率,通过对大量语料库无监督的学习,得到一种语言的语言模型。该方法的优点在于可以识别一些新词汇,但是仅限于在训练语料库中出现的高频新词汇,对于低频新词汇也无能为力,缺点是该方法的准确率依赖于语料库的选择,要得出一个较好的语言模型,需要较大的花费,非常困难^[8]。

1.3 基于理解的分词方法

基于理解的中文分词又称之为知识分词,知识分词是一种理解的分词方法,它是一种理想的分词方法,它通常包括三个部分:分词子系统、句法语义子系统、

总控制部分。在总控制部分的协调下,分词子系统模拟人对句子的理解过程,它需要使用大量的语言知识和信息,由于汉语言的知识的笼统性、复杂性,难以将各种语言信息组织成机器可直接读取的形式,这类分词方法算法复杂度太高,其有效性与可行性有待在实际工作中得到进一步的验证。知识分词利用有关的词、句子等句法和语义信息或者从大量语料中找出汉字组词的规律特点来进行评价,以找到最贴近于原语的分词结构^[9]。

2 网络舆情监控中分词问题分析

中文分词是网络舆情监控的基础,爬虫程序所采集的文本要使用向量表示,基本单位都是单个的词汇,这些词就是通过分词算法来取得,网络舆情监控的一个基本任务就是从大量的网络语料数据中快速地识别新话题、热点话题、突发事件等^[10]。而这些新话题、热点话题、突发事件的主题词往往都是一些新词汇,它常常未被有穷的机器字典所收录,这些未收录的词汇不能被识别,它们被分词系统分为零散的片段,而这些主题词是网络舆情监控系统聚类的重要依据之一,如果不能很好地识别这些关键的词汇,常常导致聚类效果不佳、产生的结果集过大等问题。如果采用基于统计的分词方法的话,那么需要大量训练数据建立模型,分词精度很大程度取决于训练数据的选取,即便能够达到很高分词精度,也只能识别高频的词汇,对于那些新出现的新词汇的识别也无能为力^[11]。

3 改进策略

3.1 改进思路

由前面的分析,我们知道目前应用比较广泛的基于词典的分词方法和基于统计的分词方法都不能满足网络舆情监控的需求,主要原因在于,新话题、热点话题、突发事件等的主题词往往都是一些新词汇,如果采用的分词方法不能对这些新词识别,那么舆情监控的性能将大打折扣,从本质上来说,对于新词识别属于分词系统中对未登录词识别的范畴,目前对未登录识别也存在一些算法,但是这些算法并不能解决舆情监控中未登录词的识别,因为这些算法对全局低频的新词识别性能都很差,而舆情监控中的主题词往往具有局部高频性和全局低频性的特点^[12]。所谓局部高频性是指在这个主题词出现的这篇文章中这个词出现的频率比较高,所谓全局低频性是指这个主题词在整个语料库中出现的频率是比较低的,基于一篇文章的主题词多次在该篇文章中出现的这个特点,文中提出可扩展词典的方法来解决这一问题,通过观察,目前基于词典的分词方法对于不能识别的词汇往往都会分为单个

的字,除去常见的单字,如“是”、“的”、“得”等,那么剩下的这些被分出的单个的字称之为异常分词,例如对文本“药家鑫是大学生”进行分词处理,分词结果是:药/家/鑫/是/大学生,异常分词便是:“药”、“家”、“鑫”,识别出异常分词后将计算异常分词与前面一个词和后面一个词之间的粘结度(关于粘结度的计算将在3.2节中介绍),如果粘结度高于一个阈值,那么将其合并成为一个词,并将该词选入待选新词集,在以后的分词过程中统计待选新词的出现的数目,如果该待选新词出现的次数超过一个阈值,那么将该词存入词典中,采用这种机制能够大大地提高分词的效率。

3.2 模型建立

对于文章 M 中的 AB , A 代表一个字, B 代表一个字,定义 A 与 B 之间的粘接度为 $T(A,B)$ 。

$$T(A,B) = \log_2 \frac{P(A,B)}{P(A)P(B)} \quad (1)$$

其中 $P(AB)$ 为 AB 在文章 M 中出现的概率, $P(A)$ 为 A 在文章 M 中出现的概率, $P(B)$ 为 B 在文章 M 中出现的概率,其中 AB 、 A 、 B 在文章 M 中出现的次数分别为 $n(AB)$ 、 $n(A)$ 、 $n(B)$,其总数 $n = n(AB) + n(A) + n(B)$,则

$$P(A,B) = \frac{n(AB)}{n} \quad (2)$$

$$P(A) = \frac{n(A)}{n} \quad (3)$$

$$P(B) = \frac{n(B)}{n} \quad (4)$$

粘结度体现了 A 与 B 之间结合关系的紧密程度,当 $T(A,B)$ 大于一个预先设定的一个阈值 F 时,即可认为 AB 为粘结在一起概率很高,很有可能 AB 就是一个新词的一部分。当 $T(A,B) > 0$ 时, A 与 B 是正相关的, $T(A,B)$ 越大,则 A 与 B 的相关性越大,则 AB 越可能是一个词语;当 $T(A,B) \approx 0$ 时, A 与 B 不相关;当 $T(A,B) < 0$ 时, A 与 B 互斥,则 AB 基本不可能结合成词。在网络舆情监控中,阈值 F 如何选取,直接决定分词算法对新词的识别能力,若阈值 F 选取过大,那么将导致一些新词无法识别,若阈值 F 选取过小,那么将导致误将某些字的组合当成新词,在实际应用中,一般选取阈值 F 介于8到12之间,能达到比较理想的新词识别效果,能够满足网络舆情监控中对新词识别的要求。

3.3 算法描述

设 C_m 为句子,其中 m 为正整数,那么待处理文章 $M = \{C_1, C_2, \dots, C_n\}$,令 k 的初始值为1。

改进后的算法描述如下:

1) 若 $k > n$,则结束,否则将句子 C_k 用基于词典的分词方法分词,按逆向匹配分词,分词后的结果为 D ,

$= \{D_1, D_2, \dots, D_t\}$, D_i (t 为正整数)代表一个分词。

2) 除去 D 中常见的单个字分词,如“的”、“像”、“是”等,寻找剩下的集合中的异常分词,即寻找剩下集合中的单个字分词,若找到 D_i 为异常分词,且 D_i 不为已经处理过的异常分词,那么转入第三步,若 D 中不存在异常分词或者所有的异常分词都被处理完毕,则 $k++$,转入第一步继续进行。

3) 计算 D_i 与 D_i 周围字或者词的粘结度 T ,若 $T > F$ (F 为预先定义的阈值),则将 D_i 与 D_i 周围字或者词合并为一个分词,更新 D ,并将该分词选入待选新词集,在以后的分词过程中统计待选新词的出现的数目,如果该待选新词出现的次数超过一个阈值,那么将该词存入词典中;若 $T \leq F$,则不将 D_i 与 D_i 周围的字或者词合并。最后转入第二步,继续寻找 D 中的异常分词。

4 实验

采用nutch爬取了新浪网新闻版块2011年5月1日到2011年5月31日的新闻文本。采用上述改进算法,选取阈值 $F=9$,得到实验结果见表1,由于新增词汇较多,表1仅列出具有代表性的一些:

表1 实验结果

词典中新 增词汇	是否是主题词	能否被传统分词 方法识别
朝鲜半岛	是	不能
真维斯楼	是	不能
药家鑫	是	不能
本·拉登	是	不能
黑砖窑	是	不能
联合利华	是	不能
毒玩具	是	不能
富士康	是	不能

5 结束语

文中结合网络舆情监控的特点,利用了舆情监控中主题词在文章中的高频性,通过计算异常分词与周围单元之间的粘结度,主动发现未被收录的新词汇,提出了一种更适合于网络舆情监控的分词方法。经测试,该方法能够识别一些未被词典收录的词汇,对新闻文本中未被词典收录的主题词的识别准确率高达92%,初步证明了该方法在网络舆情监控中的有效性。该算的缺陷在于,寻找异常分词的方法单一,仅仅是通过单个字分词来判断异常,在寻找异常分词的方法上有待进一步研究。

(下转第125页)

的匹配效率。

实验将测得的数据拟合为阈值敏感性波动图,如图4所示:

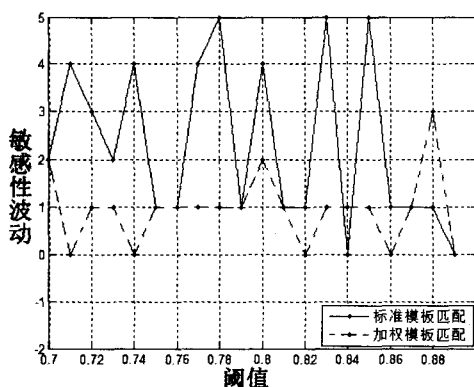


图4 两种模板匹配的阈值敏感性波动对比

通过对比两条波动曲线,进一步说明了加权后的模板匹配算法对二值化阈值的不敏感性要明显优于标准模板匹配算法,更直观地说明了加权的模板匹配算法对二值化阈值的不敏感性更好。

4 结束语

根据数字图像在二值化过程中对阈值选取的较高要求,文中提出了基于加权的模板匹配算法,并且通过对比标准模板匹配算法和加权模板匹配算法对散点图像中点的匹配结果,通过分析实验结果,对比两种模板匹配算法对数字图像选取的二值化阈值的敏感性波动情况,得出了当目标图像经二值化处理时,阈值的变化严重影响了标准模板匹配的匹配结果,匹配结果对阈值十分敏感,使得匹配效率和准确率大大降低。而当采用加权的模板匹配算法时,在一定的阈值选取范围内可以快速得到较优的匹配结果,匹配结果对阈值

的变化不敏感,有效地提高了对目标图像匹配的准确率和效率。对比结果表明,加权模板匹配算法有较好的不敏感性。

参考文献:

- [1] 梁路宏,艾海舟,肖习攀.基于模板匹配与支持矢量机的人脸检测[J].计算机学报,2002,25(1):22-29.
- [2] 罗 帅,毛奇凰,董玉德.一种选取工程图纸图像二值化阈值的新方法[J].微型机与应用,2003(8):52-54.
- [3] 费俊琳,俞王新,王志中.一种改进的基于模板匹配眼睛特征点定位算法[J].计算机工程与应用,2007,43(32):207-209.
- [4] 崔 政,李 壮.两种改进的模板匹配识别算法[J].计算机工程与设计,2006,27(6):1083-1085.
- [5] 唐 琰,李 青.一种快速的模板匹配算法[J].计算机应用,2010,30(6):1549-1561.
- [6] Starck J L, Candese J, Donoho D L. The curvelet transform for image de-noising[J]. IEEE Transactions on Image Processing, 2002, 11(6): 670-684.
- [7] 王 欣,殷肖川,周翔翔.一种改进的模板匹配识别算法[J].计算机工程与应用,2007(36):238-240.
- [8] Shi Hongchi. Two image-template operations for binary image processing[J]. Journal of Mathematical Imaging and Vision, 1997, 7(3): 269-274.
- [9] 李晓东,李志强,雷晓平,等.彩色数字仪表图像二值化技术研究[J].计算机技术与发展,2010,20(4):120-123.
- [10] Gavrilu D M. Multi-feature hierarchical template matching using distance transforms[J]. Pattern Recognition, 1998(1): 439-444.
- [11] Gonzalez R C, Woods R E. 数字图像处理[M].北京:电子工业出版社,2006:224-272.
- [12] 罗军辉,冯 平,哈力旦·A. MATLAB7.0在图像处理中的应用[M].北京:机械工业出版社,2005:76-79.

(上接第121页)

参考文献:

- [1] 罗桂琼,费洪晓,戴 戈.基于反序词典的中文分词技术研究[J].计算机技术与发展,2008,18(1):80-83.
- [2] 陈 平,刘晓霞,李亚军.基于字典和统计的分词方法[J].计算机工程与应用,2008(10):144-146.
- [3] 梁卓明,陈炬桦.基于专有名词优先的快速中文分词[J].计算机技术与发展,2008,18(3):24-27.
- [4] 王永景,刘功申,李申红,等.用于文本校对的分词与词性标注一体化算法[J].计算机技术与发展,2008,18(8):1-3.
- [5] 唐培利,胡 明,张 勇.基于中文文本主题提取的分词方法研究[J].吉林工程技术师范学院学报,2005(2):34-36.
- [6] Gao Jianfeng, Li Mu, Wu Andi, et al. Chinese Word Segmentation: A Pragmatic Approach[M]. [s. l.]: [s. n.], 2004.
- [7] Chen K J, Ma W Y. Unknown word extraction for Chinese documents[C]//The 19th COLING 2002. [s. l.]: [s. n.], 2002.
- [8] 傅赛香,袁鼎荣,黄柏雄,等.基于统计的无词典分词方法[J].广西科学院学报,2002(4):252-255.
- [9] Palmer D D. A Trainable Rule-Based Algorithm for Word Segmentation[C]//Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics. [s. l.]: [s. n.], 1997:321-328.
- [10] 王 伟,许 鑫.基于聚类的网络舆情热点发现与分析[J].现代图书情报技术,2009(3):74-79.
- [11] 柳 虹,徐金华.网络舆情热点发现研究[J].科技通报,2011(3):421-425.
- [12] 魏莎莎.一种中文未登录词识别及词典设计新方法[D].重庆:西南大学,2011.