基于时间序列分析的动态分布平滑方法

黄永文[1], 何中市[1], 王海燕[2]
(1. 重庆大学计算机学院, 重庆 400044; 2. 四川美术学院美术学系, 重庆 400052)

摘 要: 统计语言模型在实际应用中显示出了不俗的效果, 但由于语言的灵活性, 模型的数据稀疏问题始终不能避免, 现有的平滑方法只考虑了模型中元素出现的频数, 没有考虑到语言的使用是随着时间变化的. 本文分析了模型中词语随着时间的变化而出现的频数变化情况, 利用时间序列模型分析中的预测方法获得下一个阶段的数据来估计模型的参数, 提出了一种对在时间线上频数增加的词语增加概率值, 对频数减少的则降低概率值的动态分布平滑方法. 实验数据显示, 本平滑方法具有一定的优越性.

关键词: 自然语言处理; 统计语言模型; 数据稀疏; 时间序列分析; 动态分布
中图分类号: TP391 文献标识码: A 文章编号: 0372-2112 (2008) 12A-147-05

The Dynamic Distribution Smoothing Technique Based on Time Series Analysis

HUANG Yong-wen[1], HE Zhong-shi[1], WANG Hai-yan[2]
(1. *College of Computer Science, Chongqing University, Chongqing 400044 China;*
2. *Fine Arts Department, Sichuan Fine Arts Institute, Chongqing 400052 China*)

Abstract: This thesis analyzes the changes of the word occurrence frequency in models with time, uses the prediction technique in the time series model analysis to obtain the next data and thus estimate the model parameter, and get a new smoothing method, which increases the probability if the word's frequency is increasing, reduces the probability if the word's occurrence frequency is decreasing on temporal dimension. The experimental data show that this new smoothing method is superior to the others.

Key words: natural language process; statistical language model; data sparseness; time series analysis; dynamic distribution

1 引言

在统计语言模型中, 自然语言被看作是一个随机过程[1], 其中每一个语言单位包括字、词、句、段落和篇章等都被看作是有一定概率分布的随机变量. 假定自然语言句子 S 由最小的结构单位词 $w_1 \cdots w_n$ 组成, 概率值 $p(S)$ 的计算是利用离散概率的乘法定律将 $p(S)$ 分解为条件概率的乘积, 见式(1):

$$\begin{aligned} p(S) &= p(w_1, w_2, \cdots, w_n) \\ &= p(w_1)p(w_2|w_1)\cdots p(w_n|w_1, w_2, \cdots, w_{n-1}) \\ &= \prod_{i=1\cdots n} p(w_i|w_1, w_2, \cdots, w_{i-1}) \end{aligned} \quad (1)$$

在实际应用中, 由于当前词 w_i 只和前面若干个词相关, 同时由于统计语言模型特有的数据稀疏问题, 所以通常只考虑一定范围内的上下文, 这就是常用的 N-gram 模型. N-gram 模型实质是利用马尔可夫过程减少参数估计的维数, 见式(2):

$$\begin{aligned} p(S) &= p(w_1, w_2, \cdots, w_n) \\ &= \prod_{i=1\cdots n} p(w_i|w_{i-n+1}, w_{i-n+2}, \cdots, w_{i-1}) \end{aligned} \quad (2)$$

模型中 n 的大小要考虑估计有效性和描述能力的折衷. n 值越大, 其描述能力越强, 但是估计的有效性反而越差. N-gram 模型中的一个主要问题是模型的参数空间随着 n 值呈指数性增长, 从而极大限制了 n 值的大小. 由于自然语言遵循 Zip' f Law, 使得大量的语言现象不能出现在训练语料中, 从而导致数据稀疏问题[1].

解决 N-gram 模型中数据稀疏的方法就是平滑. 平滑是对采用最大似然原理的概率估计模型参数进行调整, 从而消除模型参数为零的数据稀疏现象. 主要的平滑算法有: 加法平滑[2]; Good-Turing 平滑算法[3]; Jelinek-Mercer 平滑算法[4]; Katz 平滑算法[5]. 以及还有一些根据上面的平滑方法进行改进的方法, 在众多已经实用的平滑算法中, Katz 平滑实现起来简单, 性能表现也很不错.

平滑后模型的性能是通过测试模型在测试集中的困惑度来评价的. 困惑度(perplexity)代表了一给定语言

收稿日期: 2008-05-07; 修回日期: 2008-10-01
基金项目: 国家自然科学基金(No. 60173060); 重庆市自然科学基金(No. 2007BB2134); 重庆市高等教育教学改革研究项目(No. 0635207)

模型处理语料的困难程度和不确定成分的程度, 其定义如下:

$$PP = \prod_{i=1}^{m} [p(w_i | w_{i-n+1}^{i-1})]^{-\frac{1}{m}} \quad (3)$$

其中 m 为测试语料总词序列的长度, n 是模型的阶数(如 $n=2$ 时称为 Bigram).

根据信息论, PP 值反映了信源的不可知程度. 直观上, 困惑度可以理解为在给定的语言模型中某个词后面可能接的词的平均数量[5]. 显然, 困惑度越小, 语言模型对上下文的约束能力就越强, 模型性能则越好. 因而语言模型的困惑度是评价语言模型好坏的一个重要标准[6~8].

2 语言的时间特性

语言的使用具有变迁性, 人们在使用时总会随着时间的推移而变化其风格, 尤其是时效性很强的新闻用语. 现有根据语料库建立的模型虽然反映了语料中的用词情况, 但不能反映出随着时间的变化而出现的词语变化情况. 为了考察语言运用变化的情况, 把从 2003 年到 2007 年 5 月共 500 多兆的人民日报和光明日报新闻语料按时间先后进行等分后, 分别统计其中词语的使用情况.

随着时间的变化, 一些常用词语会变得不常用, 如图 1.

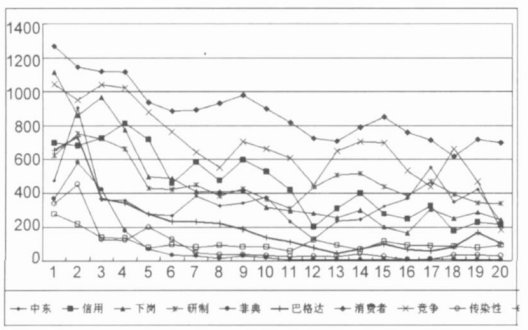


图1 词频降低

一些不常用的词语会频繁使用, 如图 2

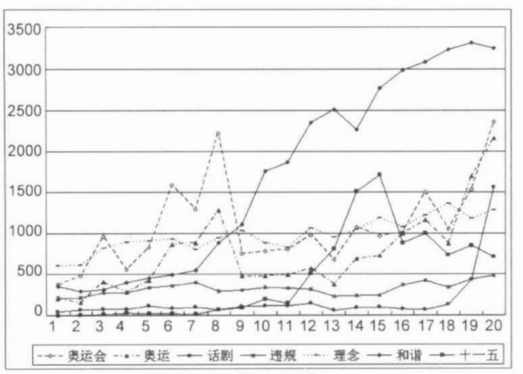


图2 词频增高

也会有一些词在一段时间内常用, 但在其它时间的使用频率很低, 如图 3.

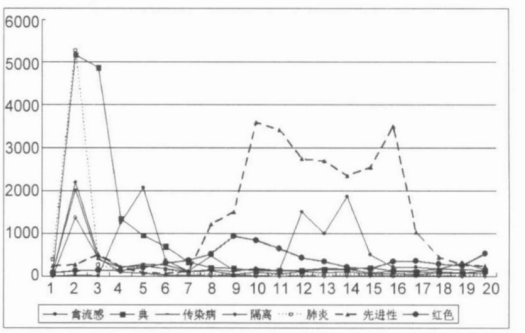


图3 瞬时常用的词

从以上所列图表中可以看出, 语言的使用随着时间推移有很大的变化. 通过分析发现词语会随着政治、经济以及其他重大事件而变得常用或短时间常用, 而一些词会因为脱离社会实际而变得越来越不常用. 故如果仅仅从静态的角度来看模型就会丢失很多有用的信息, 从而造成模型中统计到的信息有很大的偏差. 故本文特针对语料库中用词随着时间的变化而进行相应处理的平滑方法进行研究.

3 时间序列分析

时间序列分析就是通过研究某段时间内所观察到数据的统计关系, 来揭示系统的动态结构特征及其发展变化规律, 是一种重要的现代统计分析方法[9,10]. 一个本质特征就是相邻观察值之间的相互依赖性, 人们根据这种依赖性对时间序列数据生成随机动态模型, 并将这种模型应用于不同领域的分析预测[11~13]. 用时间序列进行预测时, 一般总是要依据所观察的时间序列建立预测模型, 然后用趋势外推法进行预测.

平稳时间序列分析有三种重要的形式, 即 AR 序列、MA 序列、ARMA 序列. 对非平稳时间序列主要用 ARIMA 序列来刻画. 因为词语的使用都有一个惯性的过渡, 故词语序列的变化不会是很剧烈的, 故可以把词语的ARIMA 序列进行差分处理得到平稳的 ARMA 序列.

ARIMA 模型是建立在马尔可夫随机过程上, 反映了动态的特点, 既吸取了回归分析的优点, 又发扬了移动平均的长处[16]. 假设所观测到的样本数据序列为 $\{Z_t, t=1,2\cdots\}$, 则其 ARIMA 序列为:

$$\Phi(B)\nabla^d Z_t = \Theta(B)\alpha_t \quad (4)$$

其中 $\Phi(B)$ 和 $\Theta(B)$ 是两个分别次数为 p 和 q 的特征多项式, p 和 q 都是正整数, 表达式分别为:

$$\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p \quad (5)$$

$$\Theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_p B^q \quad (6)$$

$\phi_1, \phi_2, \cdots, \phi_p$ 为自回归模型的参数, $\theta_1, \theta_2, \cdots, \theta_q$ 为移动平均模型的参数. B 是后移算子, 即:

$$BZ_t = Z_{t-1} \quad (7)$$

∇为向后差分算子, 即: $\nabla_{Z_t} = Z_t - Z_t t - 1 = (1 - B) Z_t$, d 为差分的次数, 则:

$$\nabla^d Z_t = (1 - B)^d Z_t \quad (8)$$

α_t 为高斯白噪声序列, 服从正态分布.

对满足ARIMA(p, d, q)模型的样本数据序列$\{Z_t, t = 1, 2 \cdots\}$进行 d 次差分后(一般 d 不超过 2)就可得到平稳ARMA(p, q)序列, 数据平稳化后, 可以用ARMA 模型的参数估计方法对处理后的数据进行建模.

4　基于时间序列的动态分布平滑方法

从前面的分析可以看出, 词的使用随着时间的推移有很大的变化, 直接从语料库中统计获得的数据并不能真正反映出词语在下一个时间段的使用情况, 故我们可以针对词频的变化改变相应词语的概率分布来对模型进行平滑.

4.1　参数估计

利用时间序列分析处理数据主要有模型识别、参数估计及预测三个阶段, 通过模型识别确定模型类型, 再对模型的参数进行估计, 得到模型的参数后就可以进行预测, 获得序列的下一个值.

4.1.1　模型识别

根据原始数据$\{Z_t\}$的趋势, 对$\{Z_t\}$做相应的变换, 如果是指数形的可以作对数变换将指数趋势转化为线形趋势, 再进行差分消除线形趋势, 结合自相关系数, 通过一阶差分后的序列是近似平稳的.

4.1.2　模型的估计和诊断

针对模型的各种可能排列判定模型的阶数, 即确定ARIMA(p, d, q)中的(p, d, q)的值, 主要通过序列的自相关系数和偏自相关系数是否快速落入随机区间确定 d 值, 再通过其截尾和拖尾情况确定 p 及 q 的值, 得到确定的ARIMA(p, d, q)模型.

4.1.3　预测

在对序列进行拟合后, 获得序列的模型参数, 也就可以对序列进行预测, 估计在下一个时间段出现的数据.

4.2　模型平滑

对模型进行平滑时不能脱离模型进行, 因此本文的平滑方法是以模型原有的参数为基础, 以时间序列分析预测的数据作为调整参数, 对模型中相应元素的参数值进行调整, 具体平滑方法如下:

以预测值乘上序列值个数作为参考值, 与模型中原始数据进行比较, 如果大于 1, 则以此词为后词的二元对的概率值增加, 否则降低.

设模型中词的出现频数为 $f1$, 预测值乘上序列值个数为 $f2$, 则平滑系数为:

$$\lambda = f2/f1 \quad (9)$$

模型中二元对平滑后的概率值:

$$p_{s1}(w_2 | w_1) = \lambda * p(w_2 | w_1) \quad (10)$$

如果词语 w_2 越来越常用, 预测结果值 $f2$ 就高, λ 的值就会大于 1, 平滑后的概率值就会比原来的值大, 即对此二元对进行补偿, 否则就进行折扣.

如果是模型中未出现的二元对, 则回退到一元模型, 即:

$$p_{s2}(w_2 | w_1) = \mu * p(w_2) \quad (11)$$

其中 μ 为补偿系数.

平滑不仅是调整各个二元对的概率值, 也必须满足概率归一性, 即对每一个二元对的前词 $w1$ 来说, 必须满足:

$$\sum_{w_2} p(w_2 | w_1) = 1 \quad (12)$$

即:

$$\sum_{w_2} \alpha * p_{s1}(w_2 | w_1) + p_{s2}(w_2 | w_1) = 1 \quad (13)$$

其中 α 为归一化系数, 其作用在于调整折扣和补偿的值, 不使补偿值高于折扣值, 另也留适当的折扣值给予在模型中不可见的二元对.

在以上的公式中, 只有 α, μ 两个参数是未知的, 确定好每个词的两个参数之后就可以实现基于动态分布的平滑方法.

5　实验步骤及结果分析

实验用训练语料数据为 2003-2007 年 5 月总计 500M 的人民日报和光明日报的新闻语料, 测试数据为 2007 年 6-8 月份的人民日报和光明日报语料, 分词后进行日期、数字、英文单词归类处理.

5.1　实验步骤

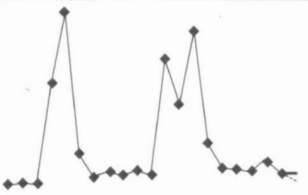
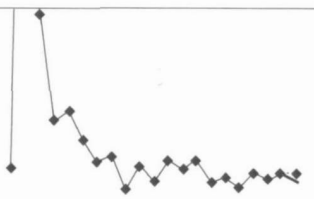
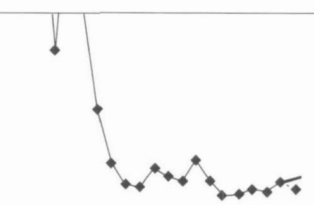
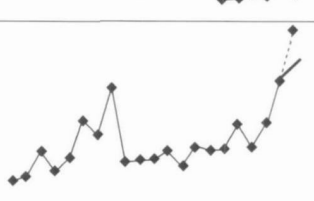
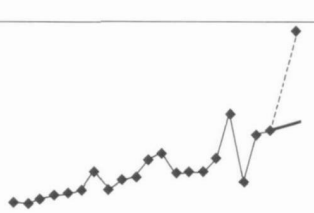
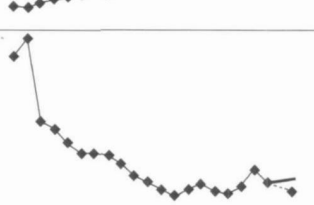
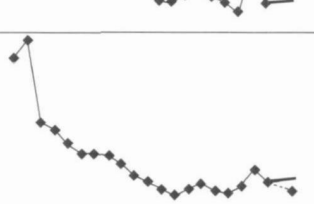
(1)把训练语料按时间先后顺序分成词语数量相近的二十等份, 分别统计出各部分中词语的频数, 建立词语的时间序列.

(2)建立时间序列模型, 通过自相关和偏自相关系数是否落入随机区间确定 d 值的大小, 再通过其截尾和拖尾确定 p 及 q 的值, 得到序列的 ARIMA(p, d, q)的阶数.

(3)根据上一个步骤得到序列的阶数, 对时间序列进行拟合并预测, 获得参考值, 通过式(9)计算出每个词的折扣或补偿系数, 统计出每个词语的折扣和补偿个数, 适当调整归一化系数 α 值. 根据式(13)计算出每个词的补偿系数 μ.

绝大多数词的归一化系数 α 为 1, 如果遇到无折扣(所有后接词都要补偿)、欠折扣(折扣概率低, 补偿概

表 1 预测值与实际值的比较

词语	后接数据	预测值	粗线为预测值,虚线为实际值
禽流感	99	110	
隔离	84	60	
肺炎	17	25	
奥运会	3389	2885	
节能	3266	1575	
奥运	3574	2783	
理念	1682	1354	
巴格达	67	145	
消费者	900	705	

率过高)、过折扣(折扣概率高,补偿概率过低)三种情况则需要调整 α 系数相应进行补偿.

这样获得每个词的平滑参数 α, μ, 根据式(9)~(13)把此平滑方法进行应用.

5.2 实验结果及分析

为了直观的展示序列预测的准确性,选择了一些词在测试集中的值与预测值进行了比较,比较结果见表 1. 在表 1 中,后接数据为生成模型的语料库下一个阶段出现的数据,可以作为实际值来看,预测值是以 ARIMA 模型进行预测得到的结果值;表 1 图中的纵坐标为对应词语出现频数,横坐标为词语序列对应时间段,曲线显示了词语随着时间变化而变化的情况.

从表 1 中可以看出,预测值和实际值有一定的差距,毕竟语言的使用是很复杂的,但是预测值与数据的变化趋势还是吻合的,说明应用时间序列分析模型处理词语变化是可行的.

为了对平滑后的模型进行评价,特把此平滑方法和其他平滑方法进行了比较.

表 2 平滑方法困惑度比较

平滑方法	Katz	J-M	W-B	DD
困惑度	215	207	197	189

表中 Katz 表示使用 Katz 平滑方法, J-M 表示 Jelinek-Mercer, W-B 表示 Witterr Bell, DD(Dynamic Distribution)表示动态分布平滑方法.

本文提出的方法在测试数据上的困惑度低于其它平滑方法,也就是说本文提出的对模型进行平滑处理后的模型更适合于测试数据,在实际应用中就会有更佳的表现.

表 3 是从分词消歧角度对平滑结果进行评价,在对 2007 年 6-8 月的人民日报语料处理后通过消除测试语料中的交集性歧义词语来评价平滑模型.

表 3 消歧比较

平滑方法	KATZ	J-M	W-B	DD
正确率%	94.3	94.5	95.1	96.3

从表 3 可以看出,用动态分布平滑处理的模型在歧义消除方面分词正确率相对其他平滑有所提高.

6 总结与展望

本文在分析了模型中随着时间的变化用词的变化情况后,提出了基于动态分布的平滑方法,应用时间序列分析模型对序列进行预测,以预测值来对模型参数进行调整.与常用的 Katz 平滑方法进行比较后可发现本平滑方法的效果更好.

由于语料库的规模有限,本文只对二元模型进行了分析与测试,此动态分布的平滑方法可以用于更高阶的模型,如对于三元模型来说,既可以考虑到词语的分布情况,也可以考虑到二元对的分布情况,这样处理后的模型更能够反映实际情况,因此应用效果会更好.

参考文献:

[1] 林亚平, 刘云中, 等. 基于最大熵的隐马尔可夫模型文本信息抽取[J]. 电子学报, 2005, 2(2): 236- 240.
Lin Ya-ping, Liu Yun-zhong, et al. Using hidden markov model for text information extraction based on maximum entropy[J]. Acta Electronica Sinica, 2005, 2(2): 236- 240. (in Chinese)

[2] F Jelinek. Self-Organized Language Modeling for Speech Recognition[M]. Readings in speech recognition, San Mateo: Morgan Kaufmann Publishers, 1990. 450- 506.

[3] G J Lidstone. Note on the general case of the Bayes-Laplace formula for inductive or a posteriori probabilities[J]. Transactions of the Faculty of Actuaries, 1920(8): 182- 192.

[4] I J Good. The population frequencies of species and the estimation of population parameters[J]. Biometrika, 1953, 12(40): 237- 264.

[5] 黄萱菁, 吴立德, 等. 现代汉语熵的计算及语言模型中稀疏事件的概率估计[J]. 电子学报, 2000, 8(8): 110- 112.
Huang Xuan-jing, Wu Li-de, et al. Computation of the entropy of modern Chinese and the probability estimation of sparse event in statistical language model[J]. Acta Electronica Sinica, 2000, 8(8): 110- 112. (in Chinese)

[6] F Jelinek, R L Mercer. Interpolated estimation of markov source parameters from sparse data[A]. Proceedings of Workshop on Pattern Recognition in Practice[C]. Amsterdam: ICEIS Press, 1980. 381- 397.

[7] Slava M Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer[J]. IEEE Transactions on Acoustics, Speech and Signal Processing, March 1987, ASSP- 35(3): 400- 401.

[8] 吴军, 王作英. 汉语信息熵和语言模型的复杂度[J]. 电子学报, 1996, 10(10): 69- 71.
Wu Jun, Wang Zuo-ying. The entropy of Chinese and the perplexity of the language models[J]. Acta Electronica Sinica, 1996, 10(10): 69- 71. (in Chinese)

[9] 王军, 彭喜元, 彭宇. 一种新型复杂时间序列实时预测模型研究[J]. 电子学报, 2006, 34(2): 2391- 2394.
Wang Jun, Peng Xi-yuan, Peng Yu. A novel real time predictor for complex time series[J]. Acta Electronica Sinica, 2006, 34(12): 2391- 2394. (in Chinese)

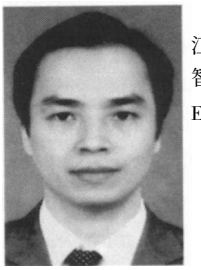
[10] 王永利, 周景华, 徐宏炳, 等. 时间序列数据流的自适应预测[J]. 自动化学报, 2007, 2: 197- 201.
Wang Yong-li, Zhou Jing-hua, Xu Hong-bing, et al. An adaptive forecasting method for time-series data streams[J]. Acta Automatica Sinica, 2007, 2: 197- 201. (in Chinese)

[11] George E P Box, Gwilym M Jenkins, Gregory C Reinsel. Time Series Analysis Forecasting and Control[M]. Prentice-Hall, 3rd edition, 1994.

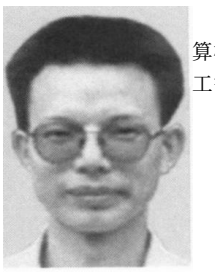
[12] Nancy Tran, Daniel A Reed. Automatic ARIMA time series modeling for adaptive I/O prefetching[J]. IEEE Trans. Parallel Distrib. Syst, 2004, 15(4): 362- 377.

[13] I Herraiz, J M González-Barahona, G Robles. Forecasting the number of changes in eclipse using time series analysis[A]. MSR2007: ICSE Workshops[C]. Washington, USA, 2007. 32- 33.

作者简介:



黄永文　男, 1970 年 10 月出生于四川开江. 重庆大学博士研究生, 主要研究领域为人工智能、自然语言处理、智能搜索.
Email: lanf@tom.com



何中市　男, 1965 年生, 博士, 重庆大学计算机学院教授, 博士生导师, 主要研究领域为人工智能、自然语言处理、数据挖掘与机器学习.