文章编号:1001 - 9081(2005)01 - 0011 - 03

基于 N 元语言模型的文本分类方法

周新栋,王 挺
(国防科技大学 计算机学院,湖南 长沙 410073)
(zhouxindong @163. com)

摘 要:分类是近年来自然语言处理领域的一个研究热点。在分析了传统的分类模型后,文中提出了用 N 元语言模型作为中文文本分类模型。该模型不以传统的"词袋"(bag of words)方法表示文档,而将文档视为词的随机观察序列。根据该方法,设计并实现一个基于词的 2 元语言模型分类器。通过 N 元语言模型与传统分类模型(向量空间模型和 Naive Bayes 模型)的实验对比,结果表明:N 元模型分类器具有更好的分类性能。

关键词:文本分类;N 元语言模型;参数平滑
中图分类号: TP391 **文献标识码**:A

Text classification based on N-gram language model

ZHOU Xin-dong, WANG Ting
(*School of Computer Science, National University of Defense Technology, Changsha Hunan* 410073, *China*)

Abstract: Text classification has become a research focus in the field of natural language processing. After the review of traditional text classification models, a method using N-gram language models to classify Chinese text was presented. This model doesn't present documents with bag of words, but regards documents as random observation sequences. With the bi-gram model, a text classifier based on word level was implemented. The performance of the N-gram model classifier was compared with that of the traditional models (Vector Space Model and Naive Bayes Model). Experiment result shows that the accuracy and the stability of the N-gram model classifier are better than others.

Key words: text classification; N-gram language model; parameter smoothing

0 引言

随着 Internet 上可用的信息日益增多,对那些能够更好地发现、过滤和管理这些资源的工具的需求也日益迫切。文本自动分类器,一种将文档根据其内容分派到预先定义的类别中去的工具,对于许多信息管理任务来说是一个十分重要、不可或缺的部件[1]。目前,已有不少统计分类法[2]和机器学习方法[3]被应用到文本分类中,且都取得了相当好的效果,这其中包括:回归模型[4]、向量空间模型[5,6]、决策树模型[1]、Naive Bayes 模型[5,7]、支撑向量机[8]和神经网络[1,9]等。

许多基于统计的自然语言处理方法都采用 N 元语言模型[10],它在词典编纂[11]、语音识别[12]、词性标注[13]、机器翻译[14]等应用上取得了相当的成功。在中文文本分类模型中结合 N 元语言模型,之前的一些工作仅仅局限于以 Bi-gram 或 Tri-gram 作为中文文档的索引单元[15,16],来弥补中文分词的不足。Peng[17]提出了 CAN(Chain Augmented Naive Bayes)模型,它是 TAN(Tree Augmented Naive Bayes)[18]模型的一种变形,该模型辅以一条局部 Markov 链作为 Naive Bayes 模型的独立性假设的一种增强,允许文档中的词以该 Markov 链形式相关。本文则提出了直接用 N 元语言模型作为分类模型这一方法,将文档本身视为一个词(字)的随机观察序列,文档归属哪个类别由在该类别中观察到此随机序列的概率值大小决定。本文在同一训练和测试集上对 N 元模型分类器和两种传统分类模型(向量空间模型和 Naive Bayes 模型)的分类器分别进行了参数训练和分类测试。实验结果显示:在分类准确率及其稳定性上,N 元模型分类器要好于后两种分类器。

1 任务描述

对于文本分类可描述如下:假定有一个文档类别集 C 和一个已知文档类别的训练文档集 D,存在着一个映射 T,对于 D 中的每个文档 d 有:$T(d) \in C$。问题在于我们并不知道这个 T 的确切形式,只知道它属于由这个训练数据集所构成的假设空间 H。分类任务就是利用训练集中的信息在 H 中找出一个模型或假设 h,使之 $h(d) \in C$ 尽可能地接近 $T(d) \in C$,更重要的是这个 h 还可以用来对新文档进行类别判别。

2 N 元模型分类器

2.1 问题提出

传统分类模型的一个共同之处在于先将文档由词的有序排列变成一个无序的"词袋",这虽然简化了处理过程,使得机器易于学习,但在这种变换过程中信息的损失是无法避免的,这也是这些分类模型的准确率无法达到更高的重要原因之一。

现在我们这样考虑:如果把文档看成词的序列(或字的序列,以下略同),那么这些词的出现与否及其出现的次序可以看成是一种语言结合模式,这些结合模式信息完全可以被用来进行文档类别判别,也就是说不同类别的文档,其语言结合模式是不同的。这其实就是统计语言模型[10,19]。对于一个文

收稿日期:2004 - 07 - 05;**修订日期**:2004 - 11 - 24 **基金项目**:国家 863 计划资助项目(2001AA114110)
作者简介:周新栋(1972 -),男,江苏启东人,工程师,硕士研究生,主要研究方向:自然语言处理、信息检索; 王挺(1970 -),男,湖南长沙人,副教授,博士,主要研究方向:自然语言处理、计算机软件.

档 d,其词序形式为:$d = w_1 w_2 \cdots w_n$,现在要做的是计算该文档属于不同的文档类别的概率值 $P_c(d = w_1 w_2 \cdots w_n)$ 的大小,C 表示不同文档类别。从语言学的角度来看,$P_c(d = w_1 w_2 \cdots w_n)$ 反映了这个词序列在文档类别 C 中被使用的情况。统计语言模型实际上就是一个概率分布,它给出了某个文档类别中所有可能的文本的出现概率。在统计语言模型看来,对于一个文档类别,任何一个文本(即任何一个词序列的组合)都是可接受的,只是接受的可能性(概率)不同而已。

2.2 统计模型的构造

常用的统计语言模型有 N 元语言(N-gram)模型[20]、隐 Markov 模型(HMM)[20]、概率上下文无关语法(PCFG)[20]和概率链语法(Probabilistic Link Grammar)[21]。对于文本分类这一任务,我们选择了N元语言模型。考虑文档 d,$d = w_1 w_2 \cdots w_n$,根据条件概率的定义,有:

$$P_c(d) = P_c(w_1 w_2 \cdots w_n) = \prod_{i=1}^{n} P_c(w_n \mid w_1 \cdots w_{n-1})$$

其中 $P_c(w_n \mid w_1 w_2 \cdots w_{n-1})$ 表示在给定历史信息 $w_1 w_2 \cdots w_{n-1}$ 的条件下,词 w_n 在类别 C 中出现的概率。所有这些信息构成了一条 Markov 链,也就是说 N 元语言模型就是 N - 1 阶的 Markov 模型。$P_c(w_n \mid w_1 \cdots w_{n-1})$ 的计算量是非常巨大的,尤其当 n 取值较大时,在实际应用中,为简化计算,往往只考虑一个或两个历史信息,形成 2 元语言模型和 3 元语言模型。表面上看,N 取值越大,计算出来的概率的准确性越高。但是,这种准确性的提高是以计算量的级数上升为代价的,同时,高阶模型的数据稀疏问题也比低阶模型的要严重得多,从而会降低估计值的可靠性,这会对分类的性能起到负面的影响。下面简要介绍一下实验中用到的 2 元模型。

在 2 元模型中,一个词的出现概率仅与其前面的那个词相关,即只考虑了一个历史信息,这时,

$$P_c(d) = P_c(w_1 w_2 \cdots w_n) = P_c(w_1) \prod_{i=2}^{n} P_c(w_i \mid w_{i-1})$$

2.3 参数估计与平滑

对于 N 元模型中条件概率的估计主要是利用训练样本中词同现的相对频率,采用极大似然估计法(MLE),如下式:

$$P_c(w_i \mid w_1 w_2 \cdots w_{i-1}) = \frac{N_c(w_1 w_2 \cdots w_i)}{N_c(w_1 \cdots w_{i-1})}$$

其中 N_c 表示相应的 n 元词串在 C 类的训练样本中出现的频次。问题的关键在于如何对待那些没有在训练样本中出现的词串。如果武断地认定它们的频次为 0 是不可取的,因为这样即使某词串中包含有其他具有较高概率的子串,其整个词串的最终计算概率也为 0,无法体现出不同词串对于类别的适合程度。简单地靠通过增加训练样本量也无济于事,因为有限的训练样本只能包含有限的语言现象,而语言的生成能力是无限的。解决方法就是在参数估计中引入平滑技术[10,20],平滑的原则为:适当减少训练样本中出现了的 n 元词串的概率,将减少的这部分概率分派给那些没有在训练样本中出现的 n 元词串,并保证最终满足概率的归一性约束。在设计 N 元语言模型分类器时,我们分别使用了三种平滑方法:Laplace 平滑、Lidstone 平滑和 Good-Turing 平滑[20]。

2.3.1 Laplace 平滑

这是最早被采用的一种平滑技术:[20]

$$P_c(w_i \mid w_1 w_2 \cdots w_{i-1}) = \frac{N_c(w_1 w_2 \cdots w_i) + 1}{N_c(w_1 \cdots w_{i-1}) + B}$$

它在极大似然估计的基础上分子加了 1,这样可以保证即使 n 元词串 $w_1 w_2 \cdots w_n$ 没有在训练样本中出现过,其条件概率也不会为 0。B 为词表大小相当的量。

2.3.2 Lidstone 平滑

有实验报告称 Laplace 平滑给未在训练样本中出现的 n 元词串分配的概率过多了,因而现在更多采用的是 Lidstone 平滑[20]:

$$P_c(w_i \mid w_1 w_2 \cdots w_{i-1}) = \frac{N_c(w_1 w_2 \cdots w_i) + \lambda}{N_c(w_1 \cdots w_{i-1}) + \lambda B}$$

λ 为 0 ~ 1 之间的参数,通过它可以调整分配给未在训练样本中出现的 n 元词串的概率空间的大小。

2.3.3 Good-Turing 平滑

Good-Turing 平滑[20]适用于那些服从于二项分布的随机变量的参数估计,而我们知道文本中的词近似服从于 Zipf 分布[22],但在训练样本集规模很大时,词的出现也可以认为近似服从于二项分布。

设 N 为训练样本集中所有 n 元词串出现的总频次,n 元词串 w_i 在训练样本集中出现的次数为 r,p_i 表示 w_i 的出现概率。任务是对 p_i 进行估算,首先利用折扣频次将 r 修改为:

$$r^* = \frac{(r+1) * n_{r+1}}{n_r}$$

其中 n_r 表示在训练样本中出现且仅出现 r 次的 n 元词串的数目,然后再利用极大似然估计法求 p_i,这时 p_i 的 Good-Turing 估算公式为:

$$\hat{p}_i = \frac{R^*}{n} = \frac{(r+1) * n_{r+1}}{N * n_r}$$

对于那些没有在训练样本中出现的 n 元词串,即 $r = 0$ 的 n 元词串,则将通过折扣频次减少的那部分概率空间平均分派给其中的每个元素:

$$\hat{p} = \frac{1 - \sum_{r=1}^{m} n_r \frac{r^*}{N}}{n_0} \approx \frac{n_1}{n_0 * N}$$

其中 m 为在训练样本中所有 n 元词串出现的最大频次数,N 为训练样本中 n 元词串的总频次。

2.4 基于 N 元语言模型的文本分类

在设计基于 N 元语言模型的中文文本分类器时,首要的一个问题就是选择基于字级还是基于词级的 N 元语言模型。在汉语中有实验表明[23] 在字级时,$N = 6$ 对是汉语使用的一个较优逼近。然而设计一个 6 元语言模型分类器,其计算量是十分惊人的,以 GB2312 为准,需要评估的参数量为 $6\ 763^6$,这会带来极其严重的数据稀疏问题。因此选择基于词级的 N 元语言模型是比较适合的。在语言学界,词被认为是最小的、能独立活动的、有意义的语言成分,其 2 元模型的表示能力相当于字级的 5 元模型,而且相比后者还避免了大量垃圾字串的出现,无论对于分类性能还是分类速度,其优势是显而易见的。设词典的条目数为 100 000,则需要估计的参数量为 $1\ 000\ 00^2$,远远小于 $6\ 763^5$。本文采取了基于词的 2 元语言模型作为分类器的模型。

用 N 元语言模型进行文本分类要注意两个问题:一是要有足够的训练数据来提取“好”的模型参数;二是统计数据稀疏问题,即用平滑技术来处理小概率事件,防止过度拟合(over-fitting)现象。

新文本的分类过程:对待分类的新文本 d,先预处理成连续的汉字串集合。对每个连续的汉字串在进行分词、停用词和非实词过滤后所得到的一个词串 W_i,分别计算其属于各个文

档类别 C 的概率:

$$P_c(W_i) = P_c(w_1 w_2 \cdots w_m) = P_c(w_1)\prod_{i=2}^{m} p_c(w_i \mid w_{i-1}) \propto \log p_c(w_1) + \sum_{i=2}^{m} \log p_c(w_i \mid w_{i-1})$$

新文档 d 属于类别 C 的概率为 $\sum_i P_c(w_i)$,拥有最大概率的那个类别被判别为文档 d 所属的类别,即。$\arg\max_c \sum_i P_c(w_i)$

3　分类测试及实验结果

实验使用的训练语料和测试语料均来自于复旦大学的一个中文分类语料库。该语料库共有近 20 000 篇文档,20 个文档类别,其分类原则为中图分类法,文档大多来自于网站。由于不同类别间的文档数量分布不均匀,根据类别的代表性和样本文档的多少,笔者从中挑选了 8 个类别,训练语料和测试语料按照 2∶1(400 篇∶200 篇)的比例组成。向量空间模型(VSM)分类器的实现技术方案为:词作为索引单元,信息增益作为特征子集选取的评估函数,使用 TFIDF 加权策略,以各样本向量的数值均值作为中心趋势度量方式产生类原型向量,夹角余弦作为向量间相似度量。Naive Bayes 模型分类器的实现技术方案为:词作为索引单元,类文档频次作为特征子集选取的评估函数。实验同时也比较了不同特征维数对 VSM 分类器和 Naive Bayes 分类器的性能影响及 N 元模型分类器中不同平滑技术的优劣。

表 1　三种分类器开放测试结果(各 200 篇)

特征维数	VSM 分类器			Naive Bayes 分类器			N 元模型分类器(N=2)		
(平滑技术)	50 维	100 维	500 维	50 维	100 维	500 维	Laplace 平滑	Lidstone 平滑(0.1)	Good-Turing 平滑
Agriculture	86.5%	87.5%	92%	54%	54.5%	61.5%	83.5%	82.5%	77%
Art	91.5%	96%	95%	90%	92.5%	91.5%	93%	93.5%	88%
Computer	88%	91%	92.5%	98%	98%	98%	100%	100%	96.5%
Economy	93.5%	93%	88.5%	82%	82.5%	78%	78.5%	79%	78.5%
Environment	79%	72.5%	79.5%	90%	88.5%	86%	91%	91%	91.5%
Politics	92.5%	90.5%	92%	95%	94.5%	96%	97%	96.5%	92.5%
Space	72%	68%	67%	80%	80.5%	80%	74.5%	74.5%	72.5%
Sports	66%	71.5%	77.5%	92.5%	91%	92%	95.5%	96.5%	97.5%
总.计	83.6%	83.8%	85.5%	85.19%	85.25%	85.38%	89.13%	89.2%	86.75%
标准方差	10.2%	11.2%	9.8%	13.9%	13.7%	12%	9.2%	9.3%	9.5%

从分类测试的实验结果看:

1)从整体上讲,2 元模型分类器(89.2 %)在分类准确率上要好于 VSM(85.5 %)和 Naive Bayes 分类器(85.38 %),其分类准确率上升了近 4 %。VSM 分类器的准确率略低(83.6 %～85.5 %),但稳定性较好(9.8 %～11.2 %);Naive Bayes 分类器的准确率稍高(85.19 %～85.38 %),但稳定性较差(12 %～13.9 %);而 2 元模型分类器无论是在准确率(86.75 %～89.2 %)还是在稳定性上(9.2 %～9.5 %)的表现在三者中都是最优。

2)2 元模型分类器在对各个类别的文档上表现出了更好的适应性。各个文档类别之间分类准确率的差距较小(72.5 %～100 %),表明它具有更好的通用性和稳定性。而对于 VSM(66 %～96 %)和 Naive Bayes 分类器(54 %～98 %),特别是后者,它们在不同文档类别的分类准确率上的差别就十分明显。究其原因主要在于,特征子集选取的优劣对于 VSM 和 Naive Bayes 分类器的性能有着至关重要的影响,而且同一特征评估函数对于不同类别的文档的适用性也不一样,不同的特征维数在分类准确率上对于不同类别的文档其效果也不尽相同,基本上无规律可循,难以在诸多因素和各类别之间达到优化平衡。这些不定因素造成了 VSM 和 Naive Bayes 分类器在不同类别文档上分类性能的较大差异。形成对比的是,这些左右分类性能的因素在 N 元模型分类器中均不存在,其单一的实现方式保证了在分类性能上的稳定性,对于 N 元模型分类器来说,所需考虑的只有一点:训练语料对于语言现象的覆盖程度,覆盖度越高,模型参数就越精确,分类性能就越好。因此 N 元模型分类器的提高和改进也相对容易一些。

3)不同的平滑技术对 N 元模型分类器有不同的影响。Laplace 平滑和 Lidstone 平滑同出一源,所以在分类性能上相差不明显,而 Good-Turing 平滑的适用前提是训练语料量较大时,词的分布近似满足二项分布。由于在文档预处理时进行了非实词过滤,而有研究表明[10],实词(content word)在语料库中的分布不能很好地符合二项分布规律,因为实词倾向于"突发性"地出现,因此在分类准确率上略有下降。从上面的实验数据可以看出,基于 N 元语言模型的文本分类器在分类准确率上比之当前常用的文本分类器较优,而且实现方式更统一,影响模型的因素较少,对不同类别的文档适应性也更好。另外,对于 VSM 和 Naive Bayes 等分类模型来说,特征的选取是与语种相关的。由于没有特征选取过程,N 元语言模型作为分类模型其本身是独立于语种的,其移植性更好,只要提供相应语种的训练语料就能够实现不同语种的文本分类器。

4　结语

随着计算机技术的飞速发展,基于统计的方法日益显现其重要性。N 元语言模型在自然语言处理的许多领域上取得了相当的成功。然而以其作为文本分类模型的却很少见,之前的一些工作仅限于用 N-gram 生成文本的索引项[15,16]。本文将 N 元语言模型作为文本分类模型,通过构建一个 N 元模型分类器(N = 2),进行了分类实验,同时以相同的训练样本集构建了 VSM 分类器和 Naive Bayes 分类器,对三者进行了实验对比。实验表明:在分类准确率及其稳定性上,N 元模型分类器要优于后两种分类器。

但是 N 元语言模型分类器也有不利因素,其模型参数的存储空间需求较大,且训练过程也相对复杂、缓慢。下一步我们将对 N 元模型分类器做进一步的研究,重点在于考查各种因素对该模型的影响,如不同N元数、基于字和基于词、选用

(下转第 16 页)

表 2 SDTF * PDF 算法进行小篇幅短文优化的实验结果

词汇	权重	词汇	权重	词汇	权重	词汇	权重
中国	306.63	抵制	547.35	比赛	274.32	欧洲杯	432.1
台湾	653.24	日货	299.6	战胜	399.3		
日本	587.48	足球	449.86	葡萄牙	273.76		

3.3 词性语义相关性对结果的影响

在本组实验中,采用 SDTF * PDF 算法计算词汇权重,在实验 3.2 的基础上,进一步考虑词汇语义相关度的影响(相似度增益因子 p = 1),实验结果如表 3 所示:

表 3 SDTF * PDF 算法考虑词汇语义相关度的实验结果

词汇	权重	词汇	权重	词汇	权重	词汇	权重
中国	321.6	抵制	587.32	比赛	344.6	欧洲杯	463.57
台湾	698.32	日货	312.51	战胜	720.6		
日本	607.36	足球	469.3	击败	720.6		

从结果中看出,考虑到词性语义相关性后,表 2 中的关键词"葡萄牙"被表 3 中的关键词"击败"所代替,"战胜"和"击败"两词的权重相等,"战胜"的权重大幅提高,其他词的权重也有不同程度的提高。这是因为"战胜"和"击败"两词的语义相关度为 1,即它们是同义词。考虑词汇语义相关度后识别出了原先遗漏的关键词,这与我们的设计意图相符。

4 结语

本文介绍了一个词汇权重计算算法的设计和实现,它适用于多信源、信源流量不均衡、短文篇幅小的汉语短文话题提取系统。从初步的模拟实验结果看,算法达到了预期的目标。对系统的完善和进行实际环境中的完整测试还需要做大量的工作,此外对小篇幅短文的优化、对文本数据的并行处理仍是值得研究的问题。

参考文献:

[1] WAYNE CL. Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation[A]. Language Resources and Evaluation Conference (LREC) 2000[C],2000.1487 - 1494.

[2] 李保利,俞士汶. 话题识别与跟踪研究[J]. 计算机工程与应用,2003,39(17).

[3] SALTON G, BUCKLEY C. Term-Weighting Approaches in Automatic Text Retrieval[J]. Information Processing and Management, 1989,4(5):513 - 523.

[4] BUN KK, ISHIZUKA M. Emerging Topic Tracking System[A]. Proceedings of Web Intelligent (WI 2001)[C], LNAI 2198 (Springer), Maebashi, Japan 2001.125 - 130.

[5] BUN KK, ISHIZUKA M. Information Area Tracking and Changes Summarizing in WWW[A]. Proc of WebNet 2001[C], International Conf on WWW and Internet, Orlando, Florida 2001.680 - 685.

[6] BUN KK, ISHIZUKAM. Topic Extraction from News Archive Using TF * PDF Algorithm[A]. Proceedings of the 3rd International Conference on Web Information Systems Engineering, 2002.

[7] 王永恒,贾焰,杨树强. 面向汉语短文的话题识别系统研究[A]. NDBC2004[C]. 福建厦门,2004.

[8] 刘群,李素建. 基于《知网》的词汇语义相似度计算[J]. Computational Linguistics and Chinese Language Processing,2002,7(2):59 - 76.

(上接第 13 页)

更为复杂的平滑技术(如折扣平滑、back-off 平滑和内插值平滑等)及训练容量的大小等等。另外,对巨大的模型参数空间进行无损的数据压缩,以提高分类的速度,满足实时处理的需求,也是一个值得深入研究的问题。

参考文献:

[1] AAS K, EIKVIL L. Text Categorization: A Survey[R]. Technical Report # 941, Norwegian Computing Center, 1999.

[2] DUDA RO, HART PE, STORK DG. Patten Classification[M]. 2nd Edition. America: John Wiley & Sons Inc, 2001.

[3] MITCHELL TM. Machine Learning [M]. America: McGraw - Hill Companies, Inc, 1997.

[4] YANG Y, PEDERSEN JP. Feature Selection in Statistical Learning of Text Categorization[A]. Tthe 14th Inc Conf, On Machine learning, 1997. 412 - 420.

[5] JOACHIM T. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization[A]. Processing of ICML-97, 14th International Conference on Machine Learning[C], 1996.143 - 151.

[6] 庞剑锋,卜东波,白硕. 基于向量空间模型的文本自动分类系统的研究与实现[J]. 计算机应用研究,2001,18(9):23 - 26.

[7] 石洪波,王志海,黄厚宽,等. 一种限定性的双层贝叶斯分类模型[J]. 软件学报,2004,15(2):193 - 199.

[8] JOACHIMS T. Text categorization with support vector machines: learning with many relevant features[A]. Proceedings of ECML-98, 10th European Conference on Machine Learning[C], 1998.137 - 142.

[9] 都去琪,肖诗斌. 基于支持向量机的中文文本自动分类研究[J]. 计算机工程,2002,28(1):137 - 139.

[10] 周强. 基于语料库和面向统计学的自然语言处理技术介绍[J]. 计算机科学,1995,22(4):36 - 40.

[11] 黄科,马少平. 基于统计分词的中文网页分类[J]. 中文信息学报,2003,16(6):25 - 31.

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

[13] 朱靖波,姚天顺. 一种基于 NA 假设的训练数据自动构造方法[J]. 东北大学学报(自然科学版),1999,20(4):366 - 368.

[14] 刘群. 统计机器翻译综述[J]. 中文信息学报,2003,17(4):1 - 12.

[15] 何浩,杨海棠. 一种基于 N-Gram 技术的中文文献自动分类方法[J]. 情报学报,2002,21(4):421 - 427.

[16] 俞红奇. 中文文本分类研究[D]. 上海:复旦大学,2000.

[17] PENG F, SCHUURMANS D, WANG S. Augmenting Naive Bayes Classifiers with Statistical Language Models[J]. Information Retrieval, 2004, 7(3 - 4):317 - 345.

[18] FRIEDMAN N, GEIGER D, GOLDSZMIDT M. Bayesian network classifiers[J]. Machine Learning, 1997,29(2 - 3):131 - 163.

[19] 姚天顺,朱靖波. 自然语言理解[M]. 第 2 版. 北京:清华大学出版社,2002.

[20] 王小捷,常宝宝. 自然语言处理技术基础[M]. 北京:北京邮电大学出版社,2002.

[21] SLEATOR D, TEMPERLEY D. Parsing English with a Link Grammar [R]. Carnegie Mellon University Computer Science technical report CMU-CS-91-196, October 1991.

[22] RIJSBERGEN CV. Information Retrieval (2nd Edition) [M]. London: Butterworths, 1979.

[23] 张树武,黄泰翼. 汉语统计语言模型的 N 值分析 [J]. 中文信息学报,1998,22(1):35 - 41.