汉语词性标注 自然语言处理 知识库 (27)

计算机科学2000Vol 27№. 7

71-75

汉语词性标注方法的研究*)

The Study of Speech Tagging of Chinese Texts

TP391

魏 欧 孙玉芳

(中国科学院软件研究所 北京 100080)

Abstract With the development of computer technology and more large corpus available, the techniques of statistics-based natural language processing, especially speech tagging technique, become one of the most actively researched project in computational linguistics. In the paper, we studied the statistics-based methods applied to Chinese part-of-speech tagging, analyzed the distinguishing features of Chinese speech ambiguity phenomena, discussed the standard of determining Chinese part-of-speech, and the related topics of selecting a tag set, introduced the n-gram model used in statistical methods and studied the related algorithmes.

Keywords Corpus, Part-of-speech tagging, Speech ambiguity phenomena, N-gram, Chinese

1 引言

自然语言中,表达意义的符号(词)往往在各个层面上有歧义。在句法层面上,一个词可以兼好几种词性;在语义层面上,一个词可能有多个义项,词性歧义是由语言中的兼类词,即具有不止一个词性特征的词所引起的,只有在一定的上下文语境关系中,词所表现的特征才能被唯一地确定。

词性标注(Part-of-Speech Tagging)的作用就是通过采取适当的方法,根据上下文语境关系,消除句子中词的语法兼类,使得无论一个词兼有几种词性,在特定的场合下只保留其中最合适的一种。

词性标注在许多领域,如语音合成、语音识别、OCR、语料库加工、机器翻译、大规模文本数据的信息检索等中都有重要的作用,对词性自动标注的研究和讨论具有重要的实用意义和理论意义。

本文概述了基于规则和基于统计的两种词性标注方法,介绍了选择词性标记集的原则,并给出我们进行词性标记的统计语言学模型。

2 词性标注方法

目前词性标注方法主要分为基于规则和基于统计两种。

2.1 基于规则的方法

基于规则的标注方法首先要获取能表达一定语言上下文关系及相关语境的规则库,规则知识库是基于规则的处理的基础,它的构造则需要考虑两个基本问题:规则对语言现象的覆盖率和规则处理的正确率。一般而言,对于一条规则,这两种性能往往显示反比关系。因此,一个好的规则库的获取是比较困难的,必须综合考虑两方面的因素,合理安排不同规则的分布,使规则处理的整体效果达到最佳。

国外对词性标注方法的研究最早是从基于规则方法开始的,1971年,文[1]采用基于规则的方法开发了TAGGIT系统用于Brown语料库的辅助词性标注工作。Brown语料库约含一百万词,其中35%的词属于兼类词,TAGGIT对Brown语料库的标注正确率约为77%左右。后来基于规则的标注工作还有文[2],使用错误驱动的基于变换的方法利用Penn Treebank中已标注的约100万语料,自动构造规则库,对15万词的标注测试,正确率达到96%,此外,进行这方面研究的还有文[3,4]等。

国内山西大学[5]、北京大学计算语言学研究所[6]等针对汉语开展了有关研究工作。

2.2 基于统计的方法

近年来,随着经验主义方法在计算语言学研究中

*)本文得到国家重点科技攻关项目《开放系统中文API框架与多平台联接系统》(96-B08-1-3)和《计算机中文信息处理技术及产品开发》(98-779-01-02)的资助。魏 欧 硕士生,主要研究方向为中文信息处理。孙玉芳 研究员,博士导师,主要研究方向为中文信息处理、系统软件、大型数据库与网络工程。

的逐渐流行，基于统计的词性标注方法开始得到应用。基于统计的方法基本上是用马尔可夫语言模型，即n元语法(n-gram)模型。实验证明[7]，如果选择具有最大概率的标记串，即使采用最简单的一元语法模型，正确率也可达到85%以上。这说明，在处理词性标注方面，基于统计的模型的性能要远远优于基于规则的系统。

对于基于统计的词性标注的研究主要包括两个方面：一是概率参数的获取；二是如何应用所获得的概率参数对文本进行词性自动标注。

1)概率参数的获取　目前，对于如何通过对语料进行训练而获得有关的参数，主要有两种方法：一是监督方式，利用已标注过的语料库作为训练语料，从中统计出相关的参数值，即相对频率训练，RFT，文[8]利用已标注的Brown语料库进行训练，得到的标注正确率在95%以上。

近年来，国内的研究人员也在这方面进行了探索，清华大学人工智能国家实验室利用手工标注的语料库进行训练，开放测试的标注正确率在93%以上[7]，此外，还有北京大学计算语言学研究所利用基于统计和规则的方法，在约三十万词的汉语语料基础上进行训练，开放测试的标注正确率达到了95.8%左右[9]。

二是非监督方式，利用未标注的语料来进行训练。此种情况与已标注的语料库不同，每一个词的词性是未知的。所使用的模型一般称为隐型马尔可夫模型HMM，模型中的参数可以使用Baum-Welch(也称forward-backward)算法来进行估计[9]，文[10]在五十万词的语料库上，用这种方法实现了一个英语词性自动标注器，据称标注正确率达到96%，但是目前国内还没有关于汉语非监督方法方面的报告。

2)对文本进行自动标注　在对文本进行标注时，也有两种方法，一是基于词的，给句中的每一个词选择一个最合适的标记，即：

$$\Phi(S)_i=T_T^{\max} p(C_i=T^i|S)=T_T^{\max}\sum_{C_i=T} p(S,C_i)$$

这种方法被称为极大似然标注(Maximum Likelihood (ML) Tagging)。二是基于句子的标注方法，为每个句子选取一个最可能的标记串，即：

$$\Phi(S)=C_S^{\max} p(C_S|S)=C_S^{\max} p(S,C_S)$$

这种方法称为Viterbi标注。

3　汉语兼类词的特点及词性标记集的选择

了解汉语中兼类词的特点和词性划分的依据对于我们在运用统计方法进行自动标注中采取正确的策略有重要的启示意义。

3.1　汉语兼类词的特点

对于汉语来说，由于对词类划分和词性确定的标准尚有较多的争议，因此对兼类词的统计分析还没有权威的结果，但一般来讲，汉语中的兼类词有以下几个特点：

·兼类词只占汉语词汇的很小一部分，据文[11]统计为5.86%，而文[6]统计为4.28%。

·常用词兼类现象严重，往往越是常用词，不同用法越多，所以兼类词的使用频度并不低。

·兼类现象纷繁，覆盖面广，但分布很不均匀，一些常见的词类歧义组合占很大比例，据文[11]统计，仅(名词，动词)就占兼类总条数的49.8%，而有些兼类，如(形容词，连词)所包含的全部词条寥寥无几。

这说明：(1)兼类词问题在语言处理中所占的"份量"不轻。(2)处理时应该分清主次，根据其分布规律在自动标注中采取正确的策略。

3.2　词性标记集的选择

3.2.1　划分词类的依据　为了对语词进行词性标注，即对语词进行分类，首先要确定词类划分的依据，词类是在研究词的组合关系时根据词的聚合关系而形成的类别，它是词的聚合关系的产物，又是以词的组合关系为前提的[12]。

对于汉语来说，词类问题相当复杂、相当困难。从理论上讲，划分词类的本质依据只是词的语法功能，这是由词类的含义与分类的目的所决定的。所谓词的语法功能，概括地讲是指词在句法结构中的位置与分布，具体地讲是指以下两类功能：1)词在句法结构中充当句法成分的能力；2)词与特意选择的某类词或某些词组合成短语的能力，词的意义不能作为划分词类的依据，如果按词汇的意义进行分类，所得到的只能是概念的类或者说是语义的类，而不是服务于语法研究的类，因为表示同类概念的词的语法并不一定相同。例如"忽然"和"突然"是同义词，"忽然"是副词，"突然"却可以是形容词，如"忽然起了一阵风"，"这件事情很突然"，"战争"与"打仗"是指的同一个概念，但"战争"是名词，"打仗"是动词。

由于汉语词类与其担任的句法成分之间不存在简单的一一对应关系，因此汉语词类的多功能现象特别突出，这就容易混淆词类的界线。但是随着语法研究的深入，发现汉语的某些词类不仅具有多种功能，而且这多种功能的分布概率是不相同的。词类功能的概率分布有两方面的含义：1)是指某词类虽然有担任多种句法成分的潜在可能性，但是在大规模的真实文本中，实际成为各种句法成分的概率是不相等的。2)是指句法成分对词类的选择也存在不同的概率分布，例如主谓结构中的主语多数由名词来担任，谓语则主要由动词、形容词、状态词担任，宾语主要由名词担任等。各个词类的语法功能的优势分布成为各个词类的语法特点。

本文认为，不同词类语法功能的优势分布与劣势分布的语法特点是基于统计的方法应用于词性标注并取得显著成效的一个重要原因，基于统计的方法实际上就是通过对大规模的真实语料的调查统计，使知识在统计的意义上被解释，词类语法功能的分布特点通过相应的概率参数得到体现，这些数据从语料库中自动学习获得。反过来，在用于词性自动标注时，利用这些参数，对应词类语法功能不同的概率分布，通过上下文语境所提供的信息，比较不同情况下的概率值的大小而选出最有可能的词性。因此，在利用统计模型进行词性标注时，选择标记集也应该考虑到这一点，使得基于该标记集进行训练后所得到的概率参数值能真实、精确地反映出词类语法功能的分布特点，文[7]中曾提出了选择标记集的四条原则：

1）完备性准则：分类应当覆盖所有的语言现象，保证文本中的每一个词都可标上词性标记。

2）确定性准则：同一个词的同一语言现象不能落入不同的范畴，即在一语言现象 X 中对词 W 的特定出现，不能同时隶属于范畴 C_1 和 C_2（C_1 和 C_2 没有包含关系），这样可以使模型在进行统计训练时不会造成混乱，从而影响统计的精确性和标注系统的性能。

3）交叉最小准则：即最少兼类准测。对范畴 C_1 和 C_2，二者没有包含关系，$C_1 \cap C_2$ 应尽可能小，保证语词在总体上有较少的平均词性歧义。

4）分布性准则：具有不同范畴的词在句法结构上应具有不同的分布。

我们认为，除了这四条准则外，在进行词性标注集的选择时，还应注意"相关性原则"，即根据上述的词类的语法功能，特别是第二种特点，在选择标记集时，注意使不同词性标记之间有较强的分布特征，一个词性的出现应对其它词性的出现有一定的预测能力。这也是对第四条准则的补充，例如，如果标注集中含有量词的子类名量词，那么我们也应把名词分成可受名量词修饰的一般名词和不可受名量词修饰的无量名词；如果进一步将名量词划分成个体量词、种类量词等，那么也应将可受名量词修饰的一般名词对应地细分成受个体量词修饰的可数个体名词和受种类量词修饰的种类名词等，使得训练后得到的概率参数值能更好、更充分地反映出不同词性间的组合关系特征。

3.2.2 汉语词性标记集　根据上述的指导思想以及系统的实际需要，以《现代汉语语法信息词典》[12]对词语属性的描述和文[7]中的语料库词性标注标记集为基础，我们制订了一个包含26个大类、82个子类、25个标点和符号、总共107个标记的词标注标记集[13]。

一个合适的标记集对于自动标注的影响很大，同规则一样，一般来说，标注集分得越细，其颗粒度越小，不同词类间语法功能的特点就表现得越明显，但这也会增加模型的复杂度，需要更多的语料才能获得比较可靠的统计参数；如果分类过粗，会使词类的界限过于模糊而不能反映出词类的语法功能特征，也不利于提高标注正确率。必须通过对大规模的真实语料的调查统计分析，才能把握住不同词类语法功能的分布特点，制定出更加符合语言规律的词性标注标记集合。

4　词性标注的统计语言学模型

基于统计的自然语言处理方法是近年来兴起的一种新方法，目前，自然语言的统计模型在处理语言中的歧义消除和句法分析等方面的工作中得到越来越广泛的应用。本节介绍自然语言的 n 元语法模型，其基本思想是：假设自然语言文本的产生服从马尔可夫链，把语言中的某个语法单位（单个词、词性等）看成马尔可夫过程的状态间的一个"转移"，并假设第 n 个语法单位只与紧挨着它的前面的很少的 n－1（n>1）个词有关，利用这些语法单位间的同现概率作为状态间的转移来进行词性标注，通过适当的语料进行训练来提取各种概率参数，根据这些参数来计算出一个给定的词串可能对应的标记串的概率，然后按一定的标准选择适当的标记串作为输出，最后分析利用该模型进行词性标注的算法。

4.1　n-元语法模型

我们可以采用最小错误率的方法来判断在给定输入的情况下所产生的某一个输出的正确率。设 W 是词汇集，T 是词性标记集，对于一个给定的词串 $S=S_1S_2\cdots S_i\cdots S_M$（$S_i \in W$），我们要根据一定的策略，从 S 产生的所有可能的标记串中找出最适合该特定词串的一个标记序列 $C_S=C_1C_2\cdots C_i\cdots C_M$（$C_i \in T$），记 $P(C_S|S)$ 为在给定输入词串 S 的条件下所产生的输出标记串 C_S 的后验概率，根据贝叶斯公式，有

$$P(C_S|S)=P(C_1,C_2\cdots C_i\cdots C_M|S_1,S_2\cdots S_i\cdots S_M)=\frac{P(C_S)P(S|C_S)}{P(S)}$$

其中，$P(C_S)$ 是标记串 C_S 的先验概率，$P(S|C_S)$ 是在标记串 C_S 已知情况下词串 S 产生的条件概率，$P(S)$ 是词串 S 的非条件概率。

对标记串 C_S，根据条件概率的分布，有

$$P(C_S)=P(C_1)P(C_2|C_1)P(C_3|C_1C_2)\cdots P(C_i|C_1C_2\cdots C_{i-1})\cdots P(C_M|C_1C_2\cdots C_{M-1})=PC_1\prod_{i=2}^{M}P(C_i|C_1C_2\cdots C_{i-1}) \quad (1)$$

$$P(S|C_S)=P(S_1|C_1)P(S_2|C_1,C_2,S_1)\cdots P(S_i|C_1,C_2,\cdots C_i,S_1,S_2\cdots S_{i-1})\cdots P(S_M|C_1,C_2\cdots C_M,S_1,$$

$$S_2\cdots S_{M-1})=P(S_1|C_1)\prod_{i=2}^{M}P(S_i|C_1C_2\cdots C_i,S_2,S_2\cdots S_{i-1}) \quad (2)$$

在词串已知的情况下，$P(S)$对于所有可能的标记串都是相同的，因此可以不考虑$P(S)$，而用$P(C_S)$与$P(S|C_S)$的乘积来表示每个标记串C_S的后验概率，

$$P(C_S)P(S|C_S)=P(C_1)\prod_{i=2}^{M}P(C_i|C_1C_2\cdots C_{i-1})P(S_i|C_1)\prod_{i=2}^{M}P(S_i|C_1,C_2,\cdots,C_i,S_1,S_2\cdots,S_{i-1}) \quad (3)$$

我们的目的就是要找到这样的标记串C_S，使得

$$P(C_S|S)=\max_{C_S}P(C_S|S) \quad (4)$$

从(1)式中可以看出，要计算$P(C_i|C_1C_2\cdots C_{i-1})$，就需要对语料中由$t$个标记组成的向量进行统计计算。假设词性标记集$T$中共有$N_T$个标记，对于长度为$t$的向量，就必须估计$N(T)^t$个参数。随着$t$的增加，计算量将呈指数增大；另外，当$t$较大时，有一部分向量的出现概率极小。这样，随着$t$的增大，参数空间将会呈指数级增长变得庞大无比，并且在训练集规模不变时，参数会变得越来越不可靠，从而将导致数据稀疏问题。

为了解决上述的问题，可以通过两个简单的假设来减少参数空间的规模。

首先，从(2)式中可以看出，要计算$P(S|C_S)$，对于句子中的每一个词S_i，不仅要考虑它前面的所有词，也要考虑它前面的所有词的词性标记。为了使问题简化，我们可以假设S_i的出现只与其自身的词性C_i相关，而与前$t-1$个词无关，从而有

$$P(S_i|C_1,C_2,\cdots C_i,S_1,S_2\cdots S_{i-1})\approx P(S_i|C_i) \quad (5)$$

另外，对于(1)式，可以假设局部的上下文信息对于C_i的出现是足够的，认为C_i的出现只与紧接着第t个词的前面的很少的$(n-1)(n>1)$个词的词性相关，这样就可以大大减少需要估计的概率参数的规模以及用于估计这些参数的数据的数量，使问题得以简化，这样的模型称为n元语法模型，这实际上是一个$n-1$阶的马尔可夫过程。如果取$n=2$，那么在处理第t个词的时候，只需考虑前面第$t-1$个词出现的情况和这个词本身的特性，而不考虑其它词的影响，这时采用的就是二元语法模型。此外，常用的还有一元语法和三元语法模型。对二元语法模型，(3)式可以重新写成

$$P(C_S)P(S|C_S)=P(C_1)P(C_1|C_i)\prod_{i=2}^{M}P(C_i|C_{i-1})P(S_i|C_i) \quad (3')$$

其中$P(C_i|C_{i-1})$只与相邻词的词性有关，我们称为词性概率参数，$P(S_i|C_i)$既与词性相关，又与词本身相关，我们称此项为词汇概率参数。

假设T中共有N_T个标记，W中共有N_W个词汇，那么所有的词性概率参数组成一个$N_T\times N_T$的二维矩阵$A_{N_T\times N_T}$，其中任一元素$a_{ij}(1\leqslant i,j\leqslant N_T)$表示词性标记$T_i$到$T_j$的转移概率$P(T_j|T_i)$的值，对任一$i$$(1\leqslant i\leqslant N_T)$，满足$\sum_{j=1}^{N_T}a_{ij}=1$。所有的词汇概率参数组成一个$N_T\times N_W$的二维矩阵$B_{N_T\times N_W}$，其中任一元素$b_{jk}(1\leqslant j\leqslant N_T,1\leqslant k\leqslant N_W)$表示在出现词性标记$T_j$时产生词汇$W_k$的概率$P(W_k|T_j)$的值，对任一$j(1\leqslant j\leqslant N_T)$，满足$\sum_{k=1}^{N_W}b_{jk}=1$。

利用大规模语料对二元语法模型进行训练，就是要获得$B_{N_T\times N_W}$的参数值，然后使用适当的标注算法就可以对文本进行自动词性标注。文[14]和[15]分别介绍了采用基于监督与非监督模式从训练集中获取参数的方法及实验结果的分析。

4.2 自动标注方法

从前面的讨论可以看出，为了对一给定的词串$S=S_1S_2\cdots S_i\cdots S_M$进行词性标注，我们要找出一个标记串$C_S$，使得

$$P(C_S|S)=\max_{C_S}P(C_S|S)=\max_{C_S}P(C_1)P(S_1|C_1)\prod_{i=2}^{M}P(C_i|C_{i-1})P(S_i|C_i) \quad (4')$$

寻找满足条件的C_i的过程也是标记选择的过程。在实际出现的标注系统中，我们选择S_1和S_M为词性唯一的词或标点符号所包括的词汇序列作为标注单位，那么其余的每个词S_i最多有N_T个可能的词性，从S_1到S_M的所有可能的标记路径就形成一个有向图，如图1所示。可以看出，从起点到终点的路径总数是$(N_T)^{M-2}$，如果直接采用计算每条标记路径的概率值从中选取最大值的办法，算法的复杂度将是指数级的，显然不可行，

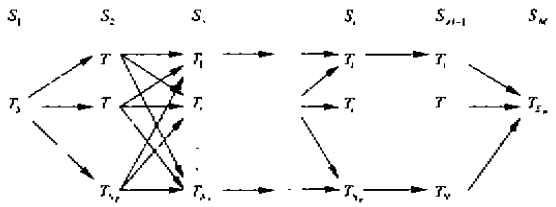

图1 在S_1、S_M的词性唯一的条件下词串S的标记路径有向图

通过对标记路径有向图的结构和二元语法模型的

特点的分析，在本系统中，我们采用基于动态规划的Viterbi算法[11]来进行最优标记串的选择，其基本思想是把求解整个问题的最佳解归结为求解其子问题的最佳解。对于我们的模型，首先从起始点开始，求出起始点到其后继结点的最佳路径并记下此路径，若从起点到某一结点的所有直接前趋结点的最佳路径已求出，则可利用前趋结点到此结点的转移的权值和起始点到前趋结点的最佳路径的累积权值计算出起始点到此结点的最佳路径，当计算出到终点的最佳路径时，路径上每一结点的标记就组成了最佳选择的标记串。

从(4')式可以看出，在计算标记路径的概率权值时，需要对值在[0,1]之间的数据进行连乘运算，可能会产生数据下溢。为此，我们采用取对数值的方法来处理。

对(4')式两边取对数，有

$$\log(P(C_S|S)=\max_{C_S}(\log(P(C_1))+\log(P(S_1|C_1))+\sum_{i=2}^{M}(\log(C_i|C_{i-1}))+\log(P(S_i|C_i))) \tag{6}$$

相应地，算法中

$$\mathrm{fai}[t,j]=\max_{1\le i\le N_T}(\mathrm{fai}[t-1,i]\times a_{ij})\times b_j(S_t)$$

也被改写成：

$$\mathrm{fai}[t,j]=\max_{1\le i\le N_T}(\mathrm{fai}[t-1,i]+\log 10(a_{ij}))\times\log 10(b_j(S_t))$$

算法的空间复杂度为$O(N_T\max(M,N_W))$，对于除S_1和S_M外的每个词，所需要进行的运算次数为$N_T\times N_T$，对于长度为M的词串，总运算次数为$M\times N_T\times N_T$，算法的时间复杂度为$O(MN_T^2)$(事实上，实际标注时，算法的时间复杂度可以比这还小)。

小结 随着语料库建设规模的扩大和数量的增长，统计技术的不断完善以及计算机技术的发展，基于统计的方法在自动词性标注技术中占了主导地位。

本文认为，不同词类语法功能的优势分布的语法特点是基于统计的方法应用于词性标注并取得显著成效的一个重要原因。本文提出，除了完备性、确定性、交叉最小性、分布性等四条选择标记集的原则之外，还应有"相关性"准则，这样可以使得训练后得到的概率参数数值能更好、更充分地反映不同词性间的组合关系特征。根据上述指导思想和系统的实际需要，以《现代汉语语法信息词典》对词语属性的描述和清华大学语言学家给出的语料库词性标注标记集为基础，我们制订了新的词性标注集。文中详细讨论了词性标注的统计语言学n-元语法模型，特别是分析了在我们构造的实际系统中所采用的相关算法，我们的自动词性标注系统应用于Search '97系统中，取得了较好的效果，表明我们制订的新的词性标注集和相关算法是成功的。当然，随着技术的进步和我们研究工作的进一步深入，我们还将把新的技术和方法应用到实际系统中。

参考文献

1 Greene B B, Rubin G M. Automated Grammatical Tagging of English. Brown University, 1971

2 Brill E. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. Computational Linguistics, 1995, 21(4): 543～565

3 Benny B. Problems with tagging and a solution. Nordic Journal of Linguistics, 1982: 93～116

4 Paulussen H, Martin W. Dilemma-2: A lemmetizer-tagger for medical abstracts. In: Proc Third Conf. on Applied Language Processing. Trento, Italy, 1992. 141～146

5 周莉娜，郑家恒，刘开瑛. 汉语词类标注规则的获取技术. 计算机研究与运用. 北京语言学院出版社，1993. 120～135

6 周强. 规则与统计相结合的汉语词类标注方法. 中文信息学报，1995, 9(2): 1～10

7 白拴虎，夏莹，黄昌宁. 汉语语料库词性标注方法研究. 机器翻译研究进展，1992: 408～418

8 Church K. A stochastic parts program and noun phrase parser for unrestricted text. In: Proc of ICASSP-89. Glasgow, Scotland. 1989. 695～698

9 Rabiner R. A tutorial on hidden markov Models and selected applications in speech recognition. Proceedings of the IEEE, 1989, 77(2): 257～285

10 Kupiec J. Robust part-of-speech tagging using a hidden Markov model. Computer Speech and Language, 1992, 6: 225～242

11 孙茂松，黄昌宁. 汉语中的兼类词、同形词类组及其处理策略. 中国人工智能学会自然语言理解学会第三次学术讨论会，1988

12 俞士汶，朱学峰，王惠，张芸芸. 现代汉语语法信息词典详解. 清华大学出版社、广西科学技术出版社，1998

13 魏欧. 基于统计的汉语词性标注方法的研究与实现. [硕士学位论文]. 中科院软件所，1998. 6

14 魏欧，吴健，孙玉芳. 基于统计的汉语性标注方法的分析与改进. 软件学报，2000, 11(4)

15 魏欧，孙玉芳. 汉语词性标注与基于非监督训练方法的实验与分析. 计算机研究与发展，1999. 2