计算机科学2005Vol. 32№. 3

自然语言信息抽取中的机器学习方法研究

周俊生[1,2] 戴新宇[1] 尹存燕[1] 陈家骏[1]
(南京大学计算机软件新技术国家重点实验室 计算机科学与技术系 南京210093)[1]
(南京师范大学计算机科学系 南京210097)[2]

摘 要 信息抽取是一种用于处理各种类型文本文档的非常有效的方法,然而建立一个文本信息抽取系统却是非常困难和耗费时间的。近年来,基于统计的机器学习方法在信息抽取领域的研究受到了广泛关注。本文深入探讨了当前自然语言信息抽取领域广泛采用的几种非常有效的统计学习方法,比较分析了各种方法的统计推断过程和学习算法及其优缺点,讨论了各种统计学习方法所面临的训练语料匮乏问题的主要解决方法,并指出了今后进一步研究的方向。

关键词 自然语言,信息抽取,统计学习,命名实体

The Methods of Machine Learning for Natural Language Information Extraction

ZHOU Jun-Sheng[1,2] DAI Xin-Yu[1] YIN Cun-Yan[1] CHEN Jia-Jun[1]
(State Key Laboratory for Novel Software Technology, Department of Computer Science Technology, Nanjing University, Nanjing 210093)[1]
(Department of Computer Science, Nanjing Normal University, Nanjing 210097)[2]

Abstract Information extraction is an effective way of processing various texts, but building a information extraction system is very difficult and time-consuming. In recent years, statistical learning methods for information extraction receive more attention. This paper first deeply analyses the main statistical learning methods employed by researchers in the filed of information extraction and compares the advantages and shortness of various methods. It then explores the solutions to the problem of lack of the training corpus. It also points out the direction for further research in the future.

Keywords Natural language, Information extraction, Statistical learning, Named entity

1 引言

在当今信息爆炸的时代,大量的信息存在于自然语言形式的文档中。如果要使得这些文档能够被自动地处理和分析,这些文档首先必须要被转化为一种结构化的形式,才能使得包含于文档中的各种"事实"信息可以被方便地访问和处理,从而给信息使用者提供有效的支持。信息抽取研究正是在这种背景下产生的[1,2]。信息抽取技术具有非常广泛的应用领域,如可以将信息抽取应用于传统的信息检索系统和 Web 搜索引擎之中,在信息检索之后对相关的文本进行指定信息的抽取,使单纯的信息查找过程进一步变成信息理解(匹配)过程,从而把传统的信息检索系统变成智能系统,以用户更满意的方式输出信息。除强烈的应用需求外,近几年来正在推动信息抽取研究进一步发展的动力,则主要来自美国国家标准技术研究所(NIST)组织的自动内容抽取(ACE)评测会议。这项评测从1999年7月开始酝酿,2000年12月正式开始启动,迄今已经举办过四次评测,研究的主要内容是自动抽取新闻语料中出现的实体、关系、事件等内容,即对新闻语料中实体、关系、事件的识别与描述[3]。

信息抽取虽然是一种用于处理各种类型文本文档的非常有效的方法,然而建立一个文本信息抽取系统却是非常费时费力的。早期出现的信息抽取系统往往依赖于人们手工建立的抽取规则或模式[4],而由人建立的规则很难保证具有整体的系统性和逻辑性,并且这些规则一般具有高度的领域相关性和较差的可移植性。因此,迫切需要寻找更加有效的方法来自动学习信息抽取的规则,这种形势使得机器学习在信息抽取系统中的应用研究显得尤为重要和迫切。近几年来在国外,机器学习方法在信息抽取领域的应用研究受到了广泛的关注[5~8],特别是对各种基于统计的机器学习方法的研究更是热点。本文深入探讨了当前自然语言信息抽取领域广泛采用的几种非常有效的统计学习方法,分析比较了各种方法的统计推断过程和学习算法及其优缺点,并讨论了各种统计学习方法所面临的标注语料匮乏问题的解决方法,最后指出了我们今后进一步发展的方向。

2 统计机器学习的基本问题

2.1 建模

在建立模型时,有两个相互影响的问题需要着重考虑:怎样参数化一个模型和怎样估计模型的参数值。如果我们构造的模型有太多的参数而太复杂,会导致模型过分依赖于训练数据集,而不能较好地预测将来的其它实例,这种现象称为"过配"(overfitting)。相反,如果模型过于泛化,也会存在问题,如一个过于泛化的语法模型所包含的规则可能生成任何可能的字符串,这种现象称为"低配"(underfitting)。除了过配与低配问题外,另一个建模的基本问题是选择"产生式"(generative)模型还是"判别式"(discriminative)模型。产生式模型的学习过程就是估计隐变量的分布和描述其相互关系的参数辨识的过程。通常产生式模型具有清晰的分层结构,而且学习

周俊生 博士研究生,主要从事自然语言处理、信息抽取等方面的研究。**戴新宇** 博士研究生,主要从事自然语言处理、机器翻译等方面的研究。**尹存燕** 博士研究生,主要从事自然语言处理等方面的研究。**陈家骏** 教授,博士生导师,主要从事自然语言处理、机器翻译、软件工程等方面的研究。

得到的模型很容易满足模型解释要求。而如果以识别为学习的目的，学习得到的模型需要尽量从样本数据中抽取共有的特征，以得到正确的分类边界，这样的模型通常属于判别式模型，它并不包含单一样本的具体特性。这个选择依赖于是否我们在考虑建立一个能够生成语言的装置或语言的部分已经给定(可被观察)的情形，在后一种情况下，建模的任务将是在供选择的多个结构中进行判别。

2.2 特征选择

主要有四种策略用于统计学习中的特征选择。第一种策略称为"包装器方法"(wrapper approach)[9]，它的思想是先生成不同特征子集，然后通过执行学习算法和测量结果分类器的准确性对各个子集进行评估。各个特征子集一般通过前向选择或后向删除方法来生成；第二种策略是将所有可能的特征包含到模型中，但对模型中的参数值引入一个惩罚值，这将导致与无用特征相关的参数将变得非常小，甚至可能为0[10]；第三个策略是计算特征的某种相关性，然后删除相关性低的特征。测度特征相关性一个最简单的方法是计算一个特征和某类别的互信息[11]。不过，这种相关性测度方法却不能捕捉特征之间的交互性。另外几种方法已经被提出用于确定这种特征间的交互性，如 RELIEFF[12]，马尔可夫毯(Markov blankets)[13]等方法；第四个策略是先拟合一个简单的模型，然后分析这个被拟合的模型以确定相关的特征。如 Chow 描述了一个高效的算法用于对一个数据集拟合一个树结构的贝叶斯网络[14]，这个网络可以被用来分析以删除对类别影响较小的特征。

3 信息抽取的统计学习方法比较

3.1 最大熵方法

最大熵原理其实就是遵循这样一个原则："对已知的建模，对未知的不做任何假设"。Della Pietra 等人于1996 年首次将它应用于自然语言处理的语言模型建立中[15]。近年来，最大熵方法开始被广泛地应用于命名实体识别等自然语言信息抽取研究中[16,17]。

命名实体可细分为不同的类型，一般主要涉及到5种：人物、地点、机构组织、时间、货币数量，另加上非实体标志，共6个元素构成标注集合。最大熵方法将对命名实体的提取过程转化为在一定的上下文 x 条件下对文本中词序列的标记过程。即对给定的一个自然语言文本输入序列 $w_1^N=w_1\cdots w_n\cdots w_N$，确定一个对应的命名实体标注序列 $c_1^N=c_1\cdots n\cdots c_N$，通过最大熵模型可以在所有可能的命名实体标注序列中选择一个具有最大概率的标注序列：

$$C_1^N=\text{argmax}\{\Pr(c_1^N|w_1^N)\}$$

为求出后验概率 $\Pr(c_1^N|w_1^N)$，需对其进行分解以对输入序列中的每一个词确定对应的命名实体标注。一般为减小计算的复杂性和考虑到实际的语言规律，可将每一个词的上下文约定为围绕当前词 w_n 的一个受限窗口 w_{n-2}^{n+2}和其前面的两个标注符号，这样，可将后验概率 $\Pr(c_1^N|w_1^N)$转化为下面的二阶模型：

$$\Pr(c_1^N|w_1^N)=\prod_{n=1}^{N}\Pr(c_n|c_1^{n-1},w_1^N)\underset{\text{mod }el}{=}\prod p(c_n|c_{n-2}^{n-1},w_{n-2}^{n+2})$$

而最大熵模型是一种用于对后验概率 $p(c_n|c_{n-2}^{n-1},w_{n-2}^{n+2})$进行建模的成熟模型。最大熵模型不依赖语言模型，独立于特定的任务，并且由于最大熵方法善于将各种不同的知识结合起来，因此最大熵模型比一般的统计模型能获取到更丰富的不受限文本特征，诸如可以灵活地把一些跨距离的特征加入到模型中去，能达到较好的识别效果，比较适合于信息抽取中分类问题的解决。但最大熵模型的一个明显缺点是计算量巨大，同时它也可能出现数据稀疏问题，需要进行平滑处理。

3.2 隐马尔可夫模型方法(HMMs)

隐马尔可夫模型可以看成是有穷状态自动机，它通过定义观察序列和标号序列的联合概率对生成过程进行建模。每一个观察序列可以看成是由一个状态转移序列生成，状态转移过程是从某一初始状态开始，当到达某一预先指定的结束状态为止，在每一个状态将随机产生一个观察序列的一个元素。用 HMMs 来解决信息提取的一般途径是：每个域(待提取的每个语义项称之为域)对应一个或多个状态，原始文本中的符号作为状态的输出符号，如果模型给定，那么信息提取过程就是搜索最可能创建符号序列的状态序列[18,19]。这个问题可以由 Viterbi 算法通过动态规划解决[20]。

尽管 HMMs 被广泛使用，但它和其它产生性模型一样并不是用于标注序列化数据的最佳模型。产生性模型定义了标号序列和观察序列的联合概率，定义这样一个联合概率意味着所有可能的观察序列都应该被枚举出来，然而如果观察元素间具有长距离依赖性，这个任务将是很困难的。因此，对于产生性模型而言，为了保证推导的正确性，应该作出严格的独立性假设。事实上，大多数序列数据都不能被表示成一系列独立的元素，往往在观察元素之间存在长距离依赖性，这样的数据更适宜于被允许这种依赖性的模型所表示，从而使观察序列被表示成一系列的非独立的、重叠的特征。

3.3 最大熵隐马尔可夫模型方法(MEMMs)

最大熵马尔可夫模型是对 HMMs 的一种改进[21]，它试图克服 HMMs 的上述缺点。在 MEMMs 中，传统 HMMs 的转换函数和观察函数被单个函数 $P(s|s',o)$所替代，这个函数给出了在给定以前的状态 s'和当前的观察 o 的条件下转移到当前状态 s 的概率。MEMMs 从训练数据中学习 $P(s|s',o)$，它是通过使用最大熵方法来使得该模型最大可能地与训练数据中的特征约束保持一致，这使得 $P(s|s',o)$具有如下的指数形式：

$$P_{s'}(s|o)=\frac{1}{Z(o,s')}\exp\left(\sum_a\lambda_a f_a(o,s)\right)$$

其中，λ_a 是需要被学习的参数，$Z(o,s')$是一个归一化因子。每一个 f_a 是一个布尔特征值，它依赖于状态 s 和输入观察序列 s 的任何特征，如"o 以一个数字开始"、"o 以一个问号结束"。

由于最大熵马尔可夫模型结合了隐马尔可夫模型和最大熵模型的优点，它允许状态转移可以基于输入序列中非独立性特征。因而使用 MEMMs 模型处理自然语言的信息抽取任务时，性能明显优于 HMMs 和无状态的最大熵模型[21]。但是 MEMMs 模型和其它判别性有限状态模型一样，在特定情形下都存在一个共同的问题—标注偏置问题(label bias problem)：离开一个给定状态的转移仅仅彼此竞争，而不会与模型中的其它转移竞争，按概率术语，转移值是在给定当前状态和观察序列条件下的条件概率。每个状态的转移概率值的归一化意味着一种"转移概率值总量的守恒"，由此，到达一个状态的总量应该在所有可能的后继状态之间分配。一个观察值可能影响哪一个目标状态获取转移概率总量，但并不会影响传递多少总量，这就会引起一个向着带有更少分支转移的状态的偏置。

3.4 条件随机场方法(Conditional Random Fields)

针对 MEMMs 的标注偏置问题，Lafferty 等人提出了一个条件随机场(CRFs)的概率模型来克服标注偏置问题[22]。条件随机场是一种用于在给定了指定的输入结点值时计算指定的输出结点值的概率的无向图模型。若 O 是一个"输入"随

机变量的集合，且它们的值可以被观察，S 是一个“输出”随机变量的集合，它们的值是要求模型能够预测的。这些随机变量之间通过指示依赖关系的无向边所连接，让 $C(O,S)$ 表示这个图中的团(cliques)的集合，根据 Hammersley-Clifford 定理，CRFs 将在给定一系列输入随机变量值的情况下，一系列输出随机变量值的条件概率定义为与无向图中各个团的势函数(potential function) 的乘积成正比：

$$P_A(s|o)=\frac{1}{Z_o}\prod_{c\in C(s,o)}\Phi_c(S_c,O_c)$$

其中，$\Phi_c(s_c,o_c)$ 表示是团 C 的势函数，一般定义为团的所有特征的带权和的指数形式，$\Phi_c(S_c,O_c)=\exp(\sum_{k=1}^{K}\lambda_k f_k(s_c,o_c))$，$Z_o$ 是一个归一化因子。

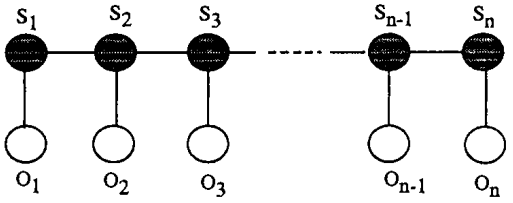

图1 链状 CRFs 的图形结构(非阴影节点表示的观察值并不是由模型产生)

在图形模型中的各指定输出结点被边连接成一条线性链的特殊情形下(如图1)，CRFs 假设在各个输出结点之间存在一阶马尔可夫独立性，这种 CRFs 可以被理解为条件训练的有限状态机(FSMs)。由于这种类型的 CRFs 是对 MEMMs 的一种全局归一化扩展，因此，它较好地解决了 MEMMs 中存在的标注偏置问题。

若让 $o=(o_1,o_2,\cdots,o_T)$ 表示被观察的输入数据序列，如文本文档中的词序列。S 表示一个 FSM 的状态集合，每一个状态均与一个标号相关，让 $s=(s_1,s_2,\cdots,s_T)$ 表示一个状态序列。图中的各个团现在被限制为仅包含序列中相邻的状态对 (s_{t-1},s_t)；但在输入结点之间 o 的连接并不受限制。这样，在一个输入序列给定的情况下，线性链的 CRFs 定义状态序列的条件概率为：

$$P_A(s|0)=\frac{1}{Z_o}\exp(\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(s_{t-1},s_t,o,t))$$

其中，$f_k(s_{t-1},s_t,o,t)$ 是一个任意的特征函数，λ_k(从 $-\infty$ 到 $+\infty$ 变化)是一个需要被学习的对应每个特征函数的权值。一般而言，特征函数可以对输入序列提出任意的问题，包括询问前面的词、后面的词以及它们的联合。

CRFs 模型与 MEMMs 模型的主要区别在于：在一个最大熵马尔可夫模型中，对每个状态均定义一个指数模型作为在给定当前状态时下一状态的条件概率；而一个条件随机场模型仅使用一个指数模型作为在给定观察值序列的条件下整个标号序列的联合概率，因此，在条件随机场模型中不同状态的不同特征的权值可以彼此平衡。McCallum 等人将条件随机场模型应用于命名实体识别、文本浅层分析等信息抽取任务的实验[23,24]，实验结果显示该模型具有良好的性能。

3.5 核(kernel)的方法

前面的几种机器学习方法均是从训练数据中抽象和构造出一个模型，它的各个参数都需要从训练数据中估计，因此可以说这些不同的实例化模型其实是对训练数据的一个总结；另外，前面各种学习方法均依赖于对象的特征表示，一个对象通常被转换成一系列特征 $f_1,\cdots,f_N$，而在许多情况下，数据并不便于通过特征来表达。

核方法则是一种完全不同的方法[25]，它的模型可以通过一组关键的训练样本来确定。在核方法中样本保留了它们的原始表示形式，在算法中仅仅通过计算一对样本对象间的核函数的方式使用样本对象。一个核函数是一个满足一定特性的相似函数，更准确地说，一个在对象空间 X 上的核函数 K 是一个二元函数 $K:X\times X\to[0,\infty]$，它映射一对对象 $x,y\in X$ 到它们的相似值 $K(x,y)$，一个核函数要求必须是对称的和半正定的(positive semidefinite)。任何核函数均在高维的特征空间中隐式地计算对象的特征向量的点积，也就是说，若存在特征 $f(\cdot)=(f_1(\cdot),f_2(\cdot),\cdots)$，$f_i:X\to R$，则 $K(x,y)=\langle f(x),f(y)\rangle$。在许多情况下，不用枚举出所有的特征也可以计算出某些特征的点积。在自然语言处理中典型的例子有关于子序列核[26]和解析树核[27]的例子。如在子序列核的例子中，对象是字符串，核函数计算在两个字符串中存在的公共字符子序列的个数。尽管特征的数量是指数级的，但子序列核的计算仍可以在多项式时间内完成。因此在核方法中可以充分利用字符串中的长距离特征，而不需要明确地枚举出这些特征。Zelenko 提出了一个基于核方法的机器学习方法用于信息抽取中的关系抽取[28]，他首先在文本的浅层解析表示的基础上定义了核，并设计了一个用于计算核的高效的动态规划算法。然后分别应用支持向量机(SVM)和表决感知器(Voted Perceptron)算法实现信息抽取，实验显示这种核方法导致了非常好的性能。

从机器学习系统设计的观点看，核方法是将焦点从特征选择问题转移到核的构造问题，由于在一个核学习系统中核是唯一与领域相关的构件，因此设计一个能充分封装用于预测的各种信息的核是非常关键的。另一方面，由于在核计算中长距离依赖性的使用，使得基于核的算法比基于特征的算法能够搜索更大的空间。但使用核方法还需要进行数值优化，当前用于支持向量机(SVMs)的各种估计方法不能具有较好的伸缩性，因此，恰当的训练对于某些应用而言可能还不可行。

4 多种机器学习方法的集成

在过去的十年中有一个重要发现，即如果将多种不同的学习模型组合成一个集成系统，则系统的性能经常会得到明显改善[29]，当前对各种集成技术(如 boosting，bagging 和 stacking 等)的研究非常热门[32,33]。它们的基本思想是所有的学习模型都在某些方面有所偏置，而通过对多个不同的模型的平均，可以有效地消除这些偏置。

在自然语言处理领域，集成方法已经应用到词性标注、语法解析、文本分类和信息抽取等多个领域[30]。Florian 利用多分类器组合的方法设计了一个命名实体识别系统[31]，在该集成系统中包含了一个隐马尔可夫模型的分类器、一个最大熵分类器和一个基于规则的分类器(基于转换的学习分类器)等四个分类器。各个模型均通过对文本中的每个词赋予一个该词在一个命名实体中的位置标注来确定命名实体，多个模型可以通过对当前正在处理的分类问题以投票的方式进行组合，另外也可以对每一个组成模型按照其对测试集的性能赋以一定的权值。一般，若给定 n 个分类器，各分类器的组合框架可以被定义为如下形式的组合概率分布：

$$P(C|w,C_1^n)=f((P_i(C|w,C_i))_{i=1,\cdots,n})$$

其中，C_i 表示第 i 个分类器的分类输出，f 是一个组合函数。当前广泛使用的组合方案是对各分类器的类别概率分布进行线性插值：

$$P(C|w,C_1^n)=\sum_{i=1}^{n}P(C|W,i,C_i)\cdot P(i|w)$$

$$=\sum_{i=1}^{n} P_i(C|w,C_i)\cdot\lambda(w)$$

权值 $\lambda(w)$ 表征了对于词 w 的上下文，第 i 个分类器在组合中的重要程度。$P_i(C|w,C_i)$ 是在给定第 i 个分类器对词 w 的输出是 C_i 的情况下，正确分类结果是 C 的概率的估计。Florian 采用了五种不同的方法对各插值参数进行估计[31]，实验结果表明：几乎在各种情况下，集成方法都产生了最好的查准率和查全率。但代价是集成方法极大地增加了由于参数估计所带来的计算负担；另外，集成模型将系统的复杂性提高到极点，使其很难被解释。现有研究成果也表明，当多学习模型集成中的个体学习模型差异较大时，集成的效果会较好。

5 弱指导学习

统计学习方法在信息抽取中比基于规则的学习方法具有优越性，但当前各种统计学习方法均面临一个困境，即需要大量的标注语料的支持，而创建新的标注语料库资源是十分高代价的。因而，近年来有许多研究聚焦于如何从现存的小规模已标注语料通过自扩展(bootstrapping)方法生成大规模的标注语料库。Blum 提出了一种将相对少量的手工标注语料与大量的未标注语料组合的方法[34]，称之为互助训练(Co-Training)。Co-Training 方法可以非形式化地描述为：首先为一个分类问题选择两个或更多的视图(Views)，然后为每一个视图建立一个独立的模型，并基于少量的标注数据训练每一个模型；接下来从未标注数据集中选择被每一个模型独立地以高可信度标注的数据样本，并将这些样本看成是有用的训练样本，不断地迭代这个过程直至整个未标注数据集为空时结束。在 Co-Training 方法中，各个模型所学到的特征是相互独立的，同时各个模型在学习过程中相互帮助，把各自学到的东西交给对方，使各自的学习成果进一步提高。Collins 将 Co-Training 方法应用于命名实体的识别[35]，他提出基于 Co-training 思想的 Adaboost 算法来解决命名实体的识别问题，称之为 Co-boosting，将标注样本的特征空间分成构造特征(Spelling)和上下文特征(Context)。尽管类似于 Co-training 的学习方法在一定程度上提高了定义在不同特征空间的弱分类器的分类准确性。但是它还是存在着一些缺陷，如要求必须满足特征空间的冗余性[36]等。

另一种广泛使用的弱指导学习方法是主动学习(Active Learning)[37]。主动学习的核心思路就是在机器学习中考虑到不同样本对最后分类器的作用其实是不一样的，我们称这种作用为样本的信息量，样本含有的信息量越大，对分类结果的确定越重要。主动学习算法主动在未标注样本集中选择测试例子，并将这些实例以一定的方式加入到训练集中。主动学习明显不同于 Co-training 的是：当对某个例子两个(或更多视图)都预测产生不同的标注时，则将那个例子提交给人进行标注。Muslea 提出了一种基于两个视图的有效的主动学习方法，称之为 naïve co-testing 方法[38]。在该方法中，分别对应于两个视图的两个分类器首先在可利用的标注数据上分别进行训练，然后，将它们运行于未标注数据，这样就产生了一个实例的不确定集合(contention set)，位于这个集合中的实例将被随机地抽取以提交给人进行标注。在两个分类器将保持不变的情况下，这个过程将不断重复。Jones 在一个用于抽取各语义类的名词短语的信息抽取系统中对 naïve co-testing 方法进行了改进[39]，他结合 Cotraining 的思想，使用标注数据和未标注数据来共同建立分类器；另外，Jones 还尝试使用不同的策略从不确定集合选择最好的实例。

结束语 当前，机器学习方法在信息抽取领域的应用研究受到了广泛的关注，特别是对各种基于统计的机器学习方法的研究更是研究热点。统计学习方法在信息抽取中具有优越性，一些实验数据表明，基于各种统计学习方法的系统的查准率和查全率一般都达到或超过了基于规则的系统所能达到的水平。但统计学习方法也存在不足，本文认为它存在下列发展趋势：

首先它的模型、算法还需要不断改善。统计学习归根到底是一个优化问题，只能在人预先规定的范围内选择一个最优解，或近似最优解，如何将人工规则加入到统计模型中，特别是如何将各种语义约束规则加入到模型中是需要进一步研究的内容。

再者，当前统计学习方法主要是有指导的学习方法，因而都面临着标注语料的匮乏问题。而语料库的人工标注是一件很费时费力的工作，尤其是针对汉语语料库的标注工作，迄今为止可利用的汉语语料库资源又很有限，大规模语料的获取与加工成为统计学习技术面临的最大困境。主动学习方法是目前用于减小语料标注代价的一种有效方法，但目前的各种主动学习方法均是基于单个学习模型的，如果将这种基于单个模型的主动学习方法扩展为基于集成(ensemble)的主动学习，一定会进一步减少语料标注的代价。

此外，虽然当前各种实验数据均表明使用多学习器集成的方法能够比使用单个学习器的系统具有更好的性能，并且近年来提出了针对各种广泛使用的集成方法的有效性的理论解释，但目前依然还缺乏一个支持各种集成方法的统一理论框架，如果能为多学习器的集成建立一个统一的理论框架，不仅可以为集成技术的理论研究提供方便，还将有利于促进其应用层面的发展。

参 考 文 献

1 Gaizauskas R, Wilks Y. Information Extraction: Beyond Document Retrieval. Journal of Documentation, 1997

2 Appelt D E, Israel D J. Introduction to Information Extraction Technology IJCAI-99

3 The ACE 2003 Evaluation Plan. http://www.nist.gov/speech/tests/ace/ace03/, Site visited on August 30th, 2003

4 Aone C, Halverson L, Hampton T, Ramos-Santacruz M. SRA: Description of the IE2 system used for MUC-7. In: Proc. of MUC-7, 1998

5 Miller S, Crystal M, Fox H, Ramshaw L, Schwartz R, Stone R, Weischedel R. Algorithms that learn to extract information - BBN: Description of the SIFT system as used for MUC-7. In: Proc. of MUC-7, 1998

6 Freitag D. Machine Learning for Information Extraction in Informal Domains: [PhD thesis]. Carnegie Mellon University, 1998

7 Ciravegna F. Adaptive information extraction from text by rule induction and generalisation. In: Proc. of the Seventeenth Intl. Joint Conf. on Artificial Intelligence, 2001

8 Califf M E, Mooney R J. Relational learning of pattern-match rules for information extraction. In: Proc. of the Sixteenth National Conf. on Artificial Intelligence, 1999. 328～334

9 Kohavi R, John G H. Wrappers for feature subset selection. Artificial Intelligence, 1997, 97(1-2): 273～324

10 Weigend A S, Rumelhart D E, Huberman B A. Generalization by weight-elimination with application to forecasting. Adv. Neural Inf. Proc. Sys. Morgan Kaufmann, 1991, 3: 875～882

11 Quinlan J R. C4.5: Programs for machine learning. Morgan Kaufmann, 1993

(下转第199页)

名称位置为空，表示没有采用该本体中的技术，其数据特点是当前节点的数据特征。当处理完生成树同一层次中最后一个叶节点时，表明在该阶段的所有可能的技术组合都已经考虑，可以进行下一阶段操作。既考虑子类集合中的下一个子类，直至将集合中的所有子类都遍历之后，我们也将生成相应的生成树了。

```
输入：只有根节点的树 T，子类的有序集合{C1,C2,…,Cn}
输出：有效 DM 过程的生成树
过程：
for (i=1,i<=n,i++)
{ leaves=getleaf(T);//得到 T 的叶子节点集合；
  for each r in leaves
  { for each ontology Oj in Ci
    {if ( r.数据特征= Oj.前提 and r.排斥() Oj.名称)
       new(t,Oj.名称,Oj.效果,Oj.排斥);
     //生成新节点 t,名称为 Oj 所代表的技术,数据特征为执行该技
           术后的结果 add(t,r);//将 t 作为 r 的子节点加入 T 中
    }
    new(t,null,r.数据特征,r.排斥);
    //生成一个没有名称的新节点，表示没有采用 Ci 中的任何技术。
    add(t,r);
  }
}
```

我们对 T 进行遍历生成所有最长路径，一条最长路径上的所有节点，即为一个可执行计划方案中所有细节。

总结 数据挖掘是一个由多个阶段组成的知识发现过程，在每个阶段都有很多的相关技术。随着数据挖掘技术在商业领域中的日益普及，越来越多的新技术被提了出来，此时，不论是数据挖掘领域的专家还是新手，都可能会忽略有用的技术。为此，我们提出为数据挖掘方法建立本体，来解决上述问题，并初步建立数据挖掘方法的本体以及相关算法。更重要的是，数据挖掘方法本体的建立还可以为数据挖掘工作者之间共享信息提供平台，使他们的工作彼此不再独立。

本文中，我们在概念上探讨将本体引入数据挖掘方法中，并对数据挖掘方法本体和其相关算法进行了初步设计，目的在于帮助数据挖掘工作者在工作过程中，面对如何选择数据挖掘技术时不再困惑。目前，我们已经在动手建立部分本体，以实现本文中所提出的算法，同时在着手设计对得到的方案计划按用户需求进行排列的算法。

下一步的目标是考虑如何更好地共享知识发现成果，实现所谓的网络外延性，我们将基于课题项目设计一个原型系统。

参 考 文 献

1 陆汝泠. 世纪之交的知识工程与知识科学. 清华大学出版社，2001
2 周肖彬，曹存根. 基于本体的医学知识获取. 计算机科学，2003，30(10)：35～39
3 Uschold M，Gruninger M. ONTOLOGIES：Principles，methods and applications. Knowledge Engineering Review，1996，11(2)：93～155
4 Guarino N. Formal ontology and information system. In：Guarino N ed. Formal Ontology in Information System. Trento：IOS Press，1998. 6～8
5 Han Jiawei，Kambr M. Data Mining Concepts and Techniques. 高等教育出版社，2001
6 Bernstein A，Provost F，Hill S. An Intelligent Assistant for the Knowledge Discovery Process：An Ontology-based Approach. 2002. Working Paper of the Center for Digital Economy Research，New York University - Leonard Stern School of Business，CeDER Working Paper # IS-01-01
7 Keim D A. Information Visualization and Visual Data Mining. IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS，JANUARY-MARCH，2002，7(1)：100～107

（上接第189页）

12 Kononenko I，Simec E，Robnik-Sikonja M. Overcoming the myopic of inductive learning algorithms with RELIEFF. Applied Intelligence，1997，7(1)：39～55
13 Koller D，Sahami M. Toward optimal feature selection. In：Proc. 13th Int. Conf. Machine Learning，Morgan Kaufmann，1996. 284～292
14 Chow C，Liu C. Approximating discrete probability distributions with dependence trees. IEEE Transactions on Information Theory，1968，14：462～467
15 Berger A L，Della Pietra S A，Della Pietra V J. A maximum entropy approach to natural language processing. Computational Linguistics，1996，22(1)：39～72
16 Bender O，Josef Och F，Ney H. Maximum Entropy Models for Named Entity Recognition. In：Proc. of the Seventh CoNLL conf. Edmonton，May-June 2003
17 Chieu H L，Ng H T. Named Entity Recognition with a Maximum Entropy Approach. In：Proc. of the Seventh CoNLL conf. Edmonton，May-June 2003. 160～163
18 Freitag D，McCallum A. Information Extraction with HMM Structures Learned by Stochastic Optimization. In：Proc. of AAAI-2000
19 Freitag D，MaCallum A K. Information Extraction with HMMs and Shrinkage. AAAI99
20 Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. of the IEEE，1989，77(2)
21 McCallum A，Freitag D，Pereira F. Maximum entropy Markov models for information extraction and segmentation. In：Proc. ICML，Stanford，California，2000. 591～598
22 Lafferty J，McCallum A，Pereira F. Conditional random fields：Probabilistic models for segmenting and labeling sequence data. In：Proc. ICML
23 McCallum A，Li Wei. Early Results for Named Entity Recognition with Conditional Random Fields，Feature Induction and Web-Enhanced Lexicons. In：Proc. of the Seventh CoNLL conf. Edmonton，May-June 2003
24 Sha F，Pereira F. Shallow parsing with conditional random fields. In：Proc. of Human Language Technology，NAACL
25 Furey T，Cristianini N，Duffy N，Bednarski D，Schummer M，Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression. Bioinformatics，2000
26 Lodhi H，Saunders C，Shawe-Taylor J，Cristianini N，Watkins C. Text classification using string kernels. Journal of Machine Learning Research，2002
27 Collins M，Duffy N. Convolution kernels for natural language. In：Proc. of NIPS-2001，2001
28 Zelenko D，Aone C，Richardella A. Kernel methods for relation extraction. Journal of Machine Learning Research，2003
29 Dietterich T G. An experimental comparison of three methods for constructing ensembles of decision trees：Bagging，boosting，and randomization. Machine Learning，2000，40(2)：139～157
30 Ghani R. Using error-correcting codes for text classification. In：P. Langley，ed. Proc. of ICML-00，17th Intl. Conf. on Machine Learning，Stanford，US：Morgan Kaufmann Publishers，San Francisco，US，2000. 303～310
31 Florian R，Ittycheriah A. Named Entity Recognition through Classifier Combination. In：Proc. of the Seventh CoNLL Conf. Edmonton，May-June 2003. 168～171
32 Kleinberg E M. A Mathematically Rigorous Foundation for Supervised Learning. In：J. Kittler，F. Roli，eds. Multiple Classifier Systems. First International Workshop，MCS 2000，Cagliari，Italy，volume 1857 of Lecture Notes in Computer Science，Springer-Verlag，2000. 67～76
33 Allwein E L，Schapire R E，Singer Y. Reducing multiclass to binary：a unifying approach for margin classifiers. Journal of Machine Learning Research，2000，1：113～141
34 Avrim B，Mitchell T. Combining labeled and unlabeled data with co-training. In：Proc. of the Workshop on Computational Learning Theory. Morgan Kaufmann，1998
35 Collins M，Yoram S. Unsupervised models for named entity classification. In：Proc. of the 1999 Conf. on Empirical Methods in Natural Language Processing，1999
36 Nigam K，Ghani R. Analyzing the effectiveness and applicability of co-training[A]. In：Proc. of Ninth Intl Conf. on Information and Knowledge Management (CIKM)[C]，2000
37 Thompson C A，Califf M E，Mooney R J. Active Learning for Natural Language Parsingand Information Extraction. In：Proc. of the Sixteenth Intl. Machine Learning Conf. Bled，Slovenia，June 1999. 406～414
38 Muslea I，Minton S，Knoblock C A. elective sampling with redundant views. AAAI/IAAI
39 Jones R，Ghani R，Mitchell T，Riloff E. Active Learning with Multiple View Feature Sets. ECML 2003 Workshop on Adaptive Text Extraction and Mining