一种介词-动词模式的获取方法*)

吴昱明[1,2]　曹存根[2]

(首都师范大学计算机联合研究院　北京 100037)[1]　(中国科学院计算技术研究所　北京 100080)[2]

摘　要　基于模式的知识获取方法研究是当前文本知识获取的重点研究之一，如何获得文本知识模式是该研究中的一个重要研究内容。提出一种新的基于介词和动词模式(称为 PV 模式)的获取方法。首先构造出一个候选的动词介词组合(称为 PV 组合)，使用统计方法对其进行过滤。度量 PV 组合好坏有两个标准：一个是模式词的表示能力，另一个是模式词与概念词之间及多个概念词之间的相关性。依据这两个标准构造了 6 个数值特征，通过训练产生了 3 个分类器，采用交叉验证的方式估计出 3 个分类器的精度分别达到 0.853，0.862 和 0.856。这些分类器为从 PV 组合中自动挑选 PV 模式提供依据。

关键词　文本知识获取，文本模式获取，模式分类

Method of Preposition-verb Pattern Acquisition

WU Yu-ming[1,2]　CAO Cun-gen[2]

(Joint Faculty of Computer Scientific Research, Capital Normal University, Beijing 100037, China) [1]

(Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China)[2]

Abstract　Pattern-based knowledge acquisition is an important research area in the research of knowledge acquisition from text (KAT). One topic of this research is how to harvest textual knowledge patterns. A novel method on acquisition of preposition-verb patterns (PV Patterns) was proposed. First, candidate preposition-verb pairs (PV pairs) were generated, and filtered by a combination of a rule-based method and statistical methods. Designed two criteria to evaluate PV patterns: coverage on instances of semantic relations and relevance among the concept words and pattern words, which lead us to construct six numeric features for PV patterns. Three classifiers were trained using these six features. The precision rates, which are estimated via cross-validation, of three classifiers are up to 0.853, 0.862 and 0.856, respectively. These classifiers provide a solid basis for automatically selecting PV patterns from PV pairs.

Keywords　Knowledge acquisition from tex, Text pattern acquisition, Pattern classification

1　引言

知识获取是与自然语言处理、人工智能、数理统计学、认知科学等学科紧密相关的研究领域，是典型的多学科研究领域。知识获取被认为是一项非常困难和耗时耗力的任务，几十年来一直是阻碍智能系统等研究和开发的瓶颈问题。文本知识获取(Knowledge Acquisition from Text，简称 KAT)是指将自然语言描述的文本知识变为计算机可理解的知识形式的过程。

本体是一种能在语义知识的层次上描述系统的概念模型。它由一个概念术语集和这些概念之间的关系组成。目前有很多自动概念获取方法取得了不错的结果[1,2]，但是现实世界中的每个概念所表示的类别或者实体都不是孤立的，这些类别(实体)之间存在着复杂的语义关系。如何获取概念之间的语义关系，成为构建大规模知识库的过程中一个非常重要的问题。

基于模式的知识获取方法在知识获取系统中被普遍使用[3-7]。虽然人们在使用自然语言表述知识的时候有着很大的随意性和不确定性，但是自然语言在表述的时候，依然遵循着一定的规则，并且这些规则被相对简单而又固定的模式表示。如果能获取这些模式，就可以使用模式找到包含感兴趣的知识的文本，进而获取我们感兴趣的知识，如概念或实体的上下位关系、整体部分关系和同义关系等。基于模式的知识获取方式不仅可以用于构建海量知识库，而且会在知识库更新的过程中发挥重要的作用。

最早使用模式的思想去获取语义关系开始于 1992 年，Hearst 通过一些暗示语义关系的模式词来获取大量的上下位关系，她使用了 4 个句型："such as"，"(and/or other)"，"including"和"especially" 来获取上下位关系。她认为不需要对语料做出细致的理解，就能获取语料中的上下位关系。实验中使用非标记的语料，但是在获取流程中需要较多的人的参与，她使用了 WordNet 作为评价获取到的上下位关系的一种工具，但是她没有把这种方法成功地运用于其它类型的关系的获取[3,4]。刘磊等人使用基于中文的"是一个"的模式从大规模中文语料库中获取上下位关系，并引入了概念验证和上下位关系验证的机制[6]。田国刚等人使用预定义的同指关系模式和多特征约束的方法从大规模中文语料库中获取同指关系[7]。但是上述这些模式获取语义关系方法所用的模式都是手工整理的，模式数量少，能表示的语义关系类型单一并且获取的语义关系召回率低。为了解决这些问题，需要一种自动

*) 工作得到国家自然基金(60496326，60573063，60573064 和 60773059)和 863 课题(2007AA01Z325)的资助。**吴昱明**　硕士研究生，主要研究方向为知识获取；**曹存根**　博士生导师，主要研究方向为人工智能。

的模式获取方法。

Surdeanu 等使用了一种混合的方法获取信息抽取模式，他们使用 Co-Training 方法，把文档中出现的词汇和文档中出现的模式作为两种条件独立的视角，从这两个视角出发训练出两个一致的文档分类器。当两个分类器都收敛时，模式的获取过程结束[8]。但是目前的模式获取方法都是获取某些指定类型的关系的模式。

本文提出一种介词-动词模式（Preposition-verb 模式，简称 PV 模式）获取方法，PV 模式不局限于表达某个单一类型的语义关系。在 PV 模式的获取过程中，首先生成候选的 PV 模式；然后基于模式表示能力和概念词与模式词之间的相关性构造出 PV 模式的数值特征，最后用监督学习的办法对候选模式进行分类。

为简化模式获取过程中的复杂性，我们在 PV 模式中引入了疑问词作为约束。我们的依据有两点。第一，我们感兴趣的是知识型 PV 模式，而“介词-疑问-动词”类的 Web 网页问句中体现出知识型 PV 模式的特点。第二，人们关注的知识型 PV 模式应该是有限的，在海量 Web 网页上应该可以找到。

2 介词-动词模式及其特征

2.1 介词-动词模式的特征

介词-动词模式（Preposition-Verb Pattern）是一种特殊的语言模式。通过对语料的观察发现，中文在表述两个术语语义关系的时候经常使用如下两种结构之一：

〈? C1:类型〉介词〈? C2:类型〉动词

介词〈? C1:类型〉动词〈? C1:类型〉

例如，从句子“中共十七大在北京召开”和“计算机由中央处理器、内存、硬盘等部件组成”中，我们很容易获得以下两个常见 PV 模式，其中“在”、“召开”、“由”、“组成”称为模式词。

①〈? C1:会议名〉在〈? C2:地名〉召开

②〈? C1:实体〉由〈? C2:实体〉组成

最后需要说明两点。

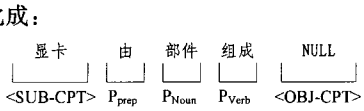
第一，在形如“中共十七大在北京召开”和“计算机由中央处理器、内存、硬盘等部件组成”的句子中，本文将“中共十七大”和“北京”等称为个体词，它们指代个体；将“计算机”、“中央处理器”、“内存”和“硬盘”等称为概念词，它们指代概念。在不致混淆的情况下，我们将两者都称为概念词。

第二，在本文中，我们不考虑介词“被”和“把”所构成的 PV 模式，因为这种模式中往往不含有我们所关心的知识，或者其中所含的知识可以通过别的模式获得。

2.2 介词-动词模式的特征

一个 PV 模式的好坏需要从多个方面考虑。例如，是否与自然语言中的表述习惯相符？模式是否适用于多个概念？模式词和概念词之间及概念词和概念词之间是否有一定相关性等等。

下面，我们从定量分析的角度给出刻画一个 PV 模式的特性，以便可以通过学习的方法获得 PV 模型。语料中的句子将被转化成〈SUB - CPT〉P_{Prep} P_{Noun} P_{Verb}〈OBJ - CPT〉的结构，其中〈SUB-CPT 〉表示主语部分的概念；〈OBJ-CPT〉表示宾语部分的概念。例如：“主板和显卡分别都是由什么部件组成的”被转化成：



只保留离 PV 模式最近的概念，如果主语部分或宾语部分没出现概念，用 NULL 占位。

2.2.1 简单表示能力度量

把 PV 模式表示的主语部分概念和宾语部分概念组成一个二元组〈SUB-CPT，OBJ-CPT〉，把一个模式在 Google 返回的语料中能表示不同的二元组个数作为该模式的度量。

例如，“由什么成分组成”和“用什么材料制作”分别表示表 1 所示的二元组。

表 1 PV 模式表示概念示例

PV 模式	句子	二元组
由什么成分组成	保鲜膜是由什么成分组成的	〈保鲜膜，NULL〉
	花岗岩由什么成分组成	〈花岗岩，NULL〉
	究竟蜡是由什么成分组成的	〈蜡，NULL〉
用什么材料制作	用什么材料制作琴弦	〈NULL，琴弦〉

则“由什么成分组成”的简单表示能力度量是 3，“用什么材料制作”的简单表示能力度量是 1。

给定一个 PV 模式，它的简单表示能力度量定义为它所能表示的二元组集合的基数，即

$$M_{Simple}(P)=|\{\langle ?\ C1,?\ C2\rangle|\langle ?\ C1,?\ C2\rangle \text{被模式 } P \text{ 表示}\}| \quad (1)$$

直观地，M_{Simple} 越大，越说明模式 P 是一个好模式。

2.2.2 概念词集合投影度量

$TF_{Google}(String)$ 表示用 Google 搜索引擎搜索 $String$ 得到的词频，例如 TF_{Google} 表示用 Google 搜索引擎搜索“”用 * 构造””（参数半角引号是查询项的一部分）返回的词频，下同。针对概念识别程序识别出来的每一个概念词，在其前面添加“和”，在后面添加“等”构造查询项，使用 Google 搜索引擎进行查询，把返回来的词频取自然对数作为该概念的权重。

例如，概念词为“计算机”，对查询项“”和计算机等””（查询项中包含半角引号），Google 返回的词频是 118000，则置概念词“计算机”的权重 $w_{计算机}=\ln(118000)=11.67844$，依此类推。

以权重高低保留主语部分的前 3500 个概念词，记为 S^{Sub}，宾语部分保留前 3000 个，记为 S^{Obj}。S^{Sub} 和 S^{Obj} 用于计算 PV 模式的各项度量。

PV 模式在主语部分概念词的投影矩阵 S^{Sub} 的构造方法如下：

$$A_{ij}^{Sub}=k_{ij}^{Sub}\times w_j^{Sub} \quad (2)$$

其中，k_{ij}^{Sub} 是第 j 个概念词在第 i 个模式所表示二元组的主语部分出现的次数，w_i^{Sub} 是 S^{Sub} 中第 j 个概念词的权重。模式 P_i 主语部分概念词集合投影度量定义为：

$$M_{project}^{Sub}(P_i)=\sum_j A_{ij}^{Sub} \quad (3)$$

类似地可定义 $M_{Project}^{Obj}(P_i)$，并定义模式 P_i 的概念词投影度量为：

$$M_{Project}(P_i)=M_{Project}^{Sub}(P_i)+M_{Project}^{Obj}(P_i) \quad (4)$$

简单表示能力度量并没有将概念词和非概念词区别开来。概念词集合投影度量考虑不同概念的权重，减小权重小的概念词（很有可能是噪声数据）对度量值的影响，从理论上提高了度量的稳定性。

2.2.3 潜层语义度量

潜层语义核（Latent Semantic Kernel）[9]方法的基本思想是，把带 n 个分类项在 p 个特征维上投影得到投影矩阵 A，A 是一个 n 行 p 列的矩阵，该矩阵的每一行是对一个待分类项

的数值表示。把矩阵 A 进行 SVD 分解成如下形式：

$$A_{n\times p}=U_{n\times n}S_{n\times p}V^T_{p\times p} \tag{5}$$

其中 $U^TU=I_{n\times n}$，$V^TV=I_{p\times p}$，保留 $S_{n\times p}$ 中前 k 个特征值，其余的特征值置零得到 $S_{n\times p}{}'$，然后计算核矩阵：

$$\hat{K}=U_{n\times n}S_{n\times p}{}'V^T_{p\times p}\times(U_{n\times n}S_{n\times p}{}'V^T_{p\times p})^T=U_{n\times n}\times S_{n\times p}{}'\times(S_{n\times p}{}')^T\times U^T_{n\times n} \tag{6}$$

用式(2)中的 A^{Sub} 作为投影矩阵，并计算 $\hat{K}^{Sub}$，PV 模式在主语部分的 SVD 度量定义如下：

$$M^{Sub}_{SVD}(P_i)=\hat{K}^{Sub}_{ii} \tag{7}$$

其中 $\hat{K}^{Sub}_{ii}$ 是 PV 模式在主语部分的 3500 个概念上投影并计算的核矩阵。与 M^{Sub}_{SVD} 类似地定义 PV 模式在宾语位置的 SVD 度量 M^{Obj}_{SVD}，PV 模式的 SVD 度量定义如下：

$$M_{SVD}(P_i)=M^{Sub}_{SVD}(P_i)+M^{Obj}_{SVD}(P_i) \tag{8}$$

获取概念词准确的权重是一个很难处理的问题，而且这个权重是和具体的度量标准相关的。为此，我们把模式词和概念词综合起来考虑。对投影矩阵做 SVD 分解后，$S_{n\times p}$ 中接近零或者小于零的特征值对应的模式词和概念词是不好的概率更大，把这些特征值置零的目的在于让好坏模式在数值上更容易区分。

把简单表示能力度量、概念词集合投影度量和潜层语义度量统成为表示能力度量。

2.2.4 概念词之间句子级互信息度量

可以用互信息的方法度量两个概念词之间的相关性：

$$MI(\langle ?\ C1\rangle,\langle ?\ C2\rangle)=\log\frac{p(\langle ?\ C1\rangle,\langle ?\ C2\rangle)}{p(\langle ?\ C1\rangle)p(\langle ?\ C2\rangle)} \tag{9}$$

句子级互信息用于度量在句子级别两个概念词的相关程度，在计算中使用词频和共现频率估计 $p(\langle ?\ C1\rangle,\langle ?\ C2\rangle)$，$p(\langle ?\ C1\rangle)$和 $p(\langle ?\ C2\rangle)$。本文中的方法使用 Google 搜索引擎搜索查询项返回词频，使用词频去近似共现率。查询项构造方法如下：在两个概念中间分别添加 1 到 3 个“*”，并把其放在两个双引号中间。采用附加引号构造的查询可以指定两个词在文档中出现的距离，一个“*”表示两个概念相隔一个词，两个“*”表示相隔两个词，依此类推。使用如下的公式估计 $p(\langle ?\ C1\rangle,\langle ?\ C2\rangle)$，$p(\langle ?\ C1\rangle)$和 $p(\langle ?\ C2\rangle)$：

$$p(\langle ?\ C1\rangle)\approx\hat{p}(\langle ?\ C1\rangle)=TF_{Google}(\langle ?\ C1\rangle)/N \tag{10}$$

$$p(\langle ?\ C1\rangle,\langle ?\ C2\rangle)\approx\hat{p}(\langle ?\ C1\rangle,\langle ?\ C2\rangle)=CO^{Sen}_{Google}(\langle ?\ C1\rangle,\langle ?\ C2\rangle)/N \tag{11}$$

CO^{Sen}_{Google} 为 Google 搜索引擎返回的两个概念在同一个句子中共同出现的次数，用〈i asterisks〉表示连续的 i 个星号(*)(下同)：

$$CO^{Sen}_{Google}(\langle ?\ C1\rangle,\langle ?\ C2\rangle)\approx\sum_{i=1}^{3}TF_{Google}(\text{“}\langle ?\ C1\rangle\langle \text{i asterisks}\rangle\langle ?\ C2\rangle\text{”}) \tag{12}$$

则句子级互信息用如下公式近似：

$$\begin{aligned}&MI_{Sen}(\langle ?\ C1\rangle,\langle ?\ C2\rangle)\\&\approx\log\frac{N\times CO^{Sen}_{Google}(\langle ?\ C1\rangle,\langle ?\ C2\rangle)}{TF_{Google}(\langle ?\ C1\rangle)TF_{Google}(\langle ?\ C2\rangle)}\\&\approx\log\frac{N\sum_{i=1}^{3}TF_{Google}(\text{“}\langle ?\ C1\rangle\langle \text{i asterisks}\rangle\langle ?\ C2\rangle\text{”})}{TF_{Google}(\langle ?\ C1\rangle)TF_{Google}(\langle ?\ C2\rangle)}\\&=\log\frac{\sum_{i=1}^{3}TF_{Google}(\text{“}\langle ?\ C1\rangle\langle \text{i asterisks}\rangle\langle ?\ C2\rangle\text{”})}{TF_{Google}(\langle ?\ C1\rangle)TF_{Google}(\langle ?\ C2\rangle)}+\log(N)\end{aligned} \tag{13}$$

Google 搜索引擎索引的网页数每天都发生变化，无法确切具体的 N，令

$$MI_{Sen}{}'(\langle ?\ C1\rangle,\langle ?\ C2\rangle)=MI_{Sen}(\langle ?\ C1\rangle,\langle ?\ C2\rangle)-\log(N) \tag{14}$$

令 $\overline{MI^{Sub}_{Sen}}$ 为所有主语部分概念〈SUB-CPT〉和模式中概念 P_{Noun} 计算出的 $MI_{Sen}{}'$ 的均值，模式 P 相对于主语部分概念的句子级互信息度量定义为

$$M^{Sub}_{Sen-MI}(P)=\sum_{\langle SUB-CPT\rangle\in S^{sub}_P}(MI_{Sen}{}'(SUB-CPT),P_{Noun}-\overline{MI^{Sub'}_{Sen}})^2 \tag{15}$$

其中 S^{Sub}_P 是 S^{Sub} 中在 P 的主语部分出现过的概念的集合。类似的方式定义相对于宾语部分概念的句子级互信息度量 $M^{Obj}_{Sen-MI}(P)$。模式 P 的句子级互信息度量定义为：

$$M_{Sen-MI}(P)=M^{Sub}_{Sen-MI}(P)+M^{Obj}_{SNE-MI}(P) \tag{16}$$

一般来说，两个概念词相关是两个概念词具有某种语意关系的必要条件，但是反过来不一定成立。直观地，句子级互信息是一个比较严格的度量标准，可能会出现两个语义相关的概念词句子互信息很低的情况，为此需要引入文档级互信息作为补充。

2.2.5 概念词之间文档级互信息度量

文档级互信息是度量在文档级别两个概念词的相关程度，估计方法和句子级互信息类似，区别在于用如下的公式估计 $p(\langle ?\ C1\rangle,\langle ?\ C2\rangle)$：

$$\begin{aligned}p(\langle ?\ C1\rangle,\langle ?\ C2\rangle)&\approx\hat{p}(\langle ?\ C1\rangle,\langle ?\ C2\rangle)\\&=CO^{Doc}_{Google}(\langle ?\ C1\rangle,\langle ?\ C2\rangle)/N\\&=TF_{Google}(\langle ?\ C1\rangle*\langle ?\ C2\rangle)/N\end{aligned} \tag{17}$$

其中 $CO^{Doc\,Google}$ 为 Google 搜索引擎返回的两个概念在同一个文档中共同出现的次数。将式(17)替换式(11)，可依次计算出 MI_{Doc}，$MI_{Doc}{}'$，M^{Sub}_{Doc-MI} 和 M^{Sub}_{Doc-MI}。

$$M_{Doc-MI}=M^{Sub}_{Doc-MI}+M^{Obj}_{Doc-MI} \tag{18}$$

句子级互信息和文档级互信息的计算公式中只有 CO_{Google} 的计算方法不同。句子级共现次数总是小于文档级共现次数，即 $CO^{Sen}_{Google}\leqslant CO^{Doc}_{Google}$。文档级互信息可以看作是句子级互信息的补充。为了表述的方便，我们把句子级互信息度量和文档级互信息度量统称为互信息度量。

2.2.6 概念词和模式词之间相关性度量

概念词和 PV 模式的相关性通过概念词和模式词共同出现的次数来度量：

$$CO^{Sen}_{Google}(\langle SUB-CPT\rangle,P)\approx\sum_{j=1}^{3}TF_{Google}(\text{“}\langle SUB-CPT\rangle P_{prep}\langle \text{i asterisks}\rangle P_{Verb}\text{”}) \tag{19}$$

S^{Sub}_P 是 S^{Sub} 中 P 在的主语部分出现过的概念的集合，则主语部分概念〈SUB－CPT〉和模式词的相关性度量定义如下：

$$M^{Sub}_{Relatness}(P)=\sum_{\langle SUB-CPT\rangle\in S^{Sub}_P}\log(CO^{Sen}_{Google}(\langle SUB-CPT\rangle,P)) \tag{20}$$

类似地可定义模式词和宾语部分概念〈OBJ－CPT〉的相关性度量 $M^{Sub}_{Relatness}(P)$。概念和模式词相关性度量定义为：

$$M_{Relatness}(P)=M^{Sub}_{Relatness}(P)+M^{Obj}_{Relatness}(P) \tag{21}$$

和互信息度量的目的类似，概念词和模式词相关是该模式能指示语义信息的一个必要条件。如果模式相应的概念词相关性很低，模式词可能根本起不到指示概念词之间的语义关系的作用。

3 PV 模式获取的方法

PV 模式的获取分为 6 个步骤。

Step1 词汇、语法过滤

手工整理介词(Preposition)和动词(Verb)的词典,并以两个词典的笛卡尔积作为候选的介词动词组合(PV 组合),但这些 PV 组合中有效的部分只有很小的比例,其中大部分不符合自然语言表述的习惯。例如,"按 * 伴读"、"按 * 暴乱"和"像 * 滥伐"。这些组合很少出现或根本不出现在语料中,可以通过如下方法过滤掉这些 没有意义的介词加动词的组合:

$$\alpha P(prep|verb)+(1-\alpha)P(verb|prep)\geqslant Threshhold \tag{22}$$

其中

$$P(verb|prep)\approx\frac{TF_{Google}(\text{“}prep * verb\text{”})}{TF_{Google}(prep)} \tag{23}$$

$$P(prep|verb)\approx\frac{TF_{Google}(\text{“}prep * verb\text{”})}{TF_{Google}(verb)} \tag{24}$$

通过该规则对条件概率很低的 PV 组合进行过滤,保留剩余的 PV 组合 1486 个。

Step2 使用疑问词构造候选模式

使用获取到的 PV 组合,在 P 和 V 中间添加"什么 * "构造查询项,用 Google 搜索引擎进行查询,取 Google 返回的前 100 个结果,生成〈 P_{Prep}〉〈 P_{Noun}〉〈 P_{Verb}〉的结构,其中 P_{Prep} 表示 PV 模式中的介词;P_{Verb} 表示 PV 模式中出现的动词;P_{Noun} 表示 PV 模式中出现的概念词(在 P_{Prep} 与 P_{Verb} 中间的名词)。

例如用 PV 组合"由…组成"构造如下的查询项"由什么 * 组成",从 Google 返回的前 100 个结果中,得到如下的候选模式:

由什么材料组成→〈由〉〈材料〉〈组成〉

由什么元素组成→〈由〉〈元素〉〈组成〉

……

由什么原料组成→〈由〉〈原料〉〈组成〉

Step3 获取语料

针对每个候选模式,提交 Google 搜索引擎进行查询,并将返回的前 100 个结果中包含当前候选模式的句子作为语料。

Step4 概念识别

使用王石等人[1]提出的概念识别算法识别出句子中离候选模式最近的概念。将语料中的句子转化成〈 SUB-CPT 〉$P_{Prep}P_{Noun}P_{Verb}$〈 OBJ-CPT 〉的结构。

Step5 构造数值特征

从 PV 模式的表示能力和模式词、概念词之间相关程度定义了 6 种度量标准,分别是简单能力表示度量 M_{Simple}、概念词集合投影度量 $M_{Probject}$、潜层语义度量 M_{SVD}、句子级互信息度量 M_{Sen-MI}、文档级互信息度量 M_{Doc-MI} 和模式词概念词相关性度量 $M_{Relatness}$。

Step6 分类

标记一部分候选模式集,以 STEP5 构造的特征项为基础,用监督学习的方法对未标记模式进行分类。

4 实验结果及分析

手工标记了 836 个候选模式,其中正样本 414 个,负样本 422 个,以简单能力表示度量 M_{Simple}、概念词集合投影度量 $M_{Project}$、浅层语义度量 M_{SVD}、句子级互信息度量 M_{Sen-MI}、文档级互信息度量 M_{Doc-MI} 和模式词概念词相关性度量 $M_{Relatness}$ 为输入特征,分别采用了 LibSVM(径向基核)、J48 和 AdaboostM1[10]对数据进行训练,并用 10 折交叉验证(10-Fold Cross-Validation)对数据进行误差估计。实验结果如表 2 所示。

表 2 实验结果

		真正率 (TP Rate)	假正率 (FP Rate)	精度 (Precision)	召回率 (Recall)	F 度量 (F-Measure)
AdaBoostM1	正样本	0.86	0.154	0.846	0.86	0.853
	负样本	0.846	0.14	0.86	0.846	0.853
	合计	—	—	0.853	—	—
J48	正样本	0.87	0.145	0.855	0.87	0.862
	负样本	0.855	0.13	0.87	0.855	0.863
	合计	—	—	0.862	—	—
LibSVM (径向基核)	正样本	0.877	0.164	0.84	0.877	0.858
	负样本	0.836	0.123	0.874	0.836	0.855
	合计	—	—	0.856	—	—

由表 2 可以看出,该方法的精度、召回率及 F 度量最高分别可以达到 86.2%,87.7%及 0.863;平均精度、召回率及 F 度量也都超过了 85%。剩余的 14%左右的错误预测可以被分成如下两类:

第一是正样本被预测错误。

例如:"以什么作品获得"是一个好的 PV 模式,但是各项度量都很低。"高行健是以什么作品获得诺贝尔文学奖的"是和该模式相关的一个句子,宾语部分的概念是"诺贝尔文学奖",因为"诺贝尔文学奖"未包含于 S^{Obj} 中(如果 S^{Obj} 过大会引入更多的噪声)。

第二是负样本被预测错误。

例如:"到什么错误提示"不是一个好的 PV 模式,这种错误产生的是"错误"和"提示"在一起组合成了一个新的名词短语。

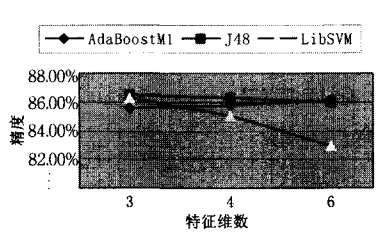


图 1 特征维数/精度变化曲线

为了观察不同的特征对精度的影响,我们分别选取所有特征集合的两个子集和全部特征作为输入。子集之一不包含

互信息度量和概念词模式词相关性度量，特征维数是3；子集之二仅不包含互信息度量，特征维数是4。实验结果如图1所示。

如图1所示，在仅使用概念表示能力度量的情况下，3种分类器的精度都已达到83%以上。但是J48和LibSVM的精度随着输入维数增加不但没有增长反而降低。出现这种现象的原因是：LibSVM是把输入的特征数据映射到高维空间进行分类，J48是以决策树的方式分类。随着维数的增加，LibSVM在训练时，样本在高维空间变得非常稀疏；而J48训练出的决策树随着输入维数的增加决策树的复杂性上升，产生过拟合的现象。所以，在我们保持原有训练样本不变的情况下，随着维数的增加，精度出现不同程度的下滑。和LibSVM与J48不同，AdaBoostM1训练出的模型是一些弱分类器的线性组合，对训练样本数要求不高。从AdaBoostM1算法精度的变化可以看出，在分别增加了概念词和模式词相关性度量和互信息度量之后，精度约有1.3%的提升。

结束语 本文提出一种介词-动词结构的模式获取方法。通过介词集合和动词集合的笛卡尔积运算构造候选模式，并通过"什么"疑问词构造查询项，获取用于评价PV模式的句子。综合考虑PV模式的表示能力和模式词概念词之间的相关性，构造了6个数值特征，并采用监督学习的方法对PV模式进行预测。用交叉验证的方法估计出精度、召回率和F度量均在0.85左右。

在获取的PV模的精度达到一定程度的基础上，还需要在以下几个方面做进一步的研究：

第一，从深度的角度来说，还需要增加对主语部分概念〈SUB-CPT〉或者宾语部分概念〈OBJ-CPT〉的语义限制，从而提高用PV模式获取语义关系的精度。

第二，从广度的角度考虑，虽然自然语言中的很多知识能通过PV模式进行表述，但是除了PV模式以外还存在很多其他结构的模式，这就需要考虑生成其他结构的模式候选集合的方法。

参 考 文 献

[1] Wang Shi, Cao Yanan, Cao Xiny, et al. Learning Concepts from Text Based on the Inner-constructive Model // Proceedings of 2nd International Conference on Knowledge Science Engineering and Management. Melbourne, Australia, 2007

[2] 余蕾，曹存根. 基于Web语料的概念获取系统的研究与实现. 计算机科学，2007，34(2)：161-165，195

[3] Hearst M. Automatic acquisition of hyponyms from large text corpora // Proceedings of the 14th Conference on Computational Linguistics. Nantes, France, 1992

[4] Hearst M. Automated discovery of wordnet relations // Fellbaum C, ed. WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press, 1998: 131-151

[5] Riloff E. Automatically generating extraction patterns from untagged text // Proceedings of the 13rd National Conference on Artificial Intelligence. Oregon, USA, 1996

[6] 刘磊，曹存根，王海涛，等. 一种基于"是一个"模式的下位概念获取方法. 计算机科学，2006，33(9)：161-165

[7] Tian Guogang, Cao Cungen, Liu Lie, et al. MFC: A Method of Co-referent Relation Acquisition from Large-scale Chinese Corpora // Proceedings of Third International Conference on Fuzzy Systems and Knowledge Discovery. Xi'an, China, 2006

[8] Surdeanu M, Turmo J, Ageno A. A Hybrid Approach for the Acquisition of Information Extraction Patterns // Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics. Trento, Italy, 2006

[9] Nello C, John S-T, Huma L. Latent Semantic Kernels // Proceedings of the 18th International Conference on Machine Learning. MA, USA, 2001

[10] Ian W, Eibe F. Data Mining: Practical Machine Learning Tools and Techniques. Second Edition. Burlington, MA: Morgan Kaufmann, 2005

(上接第134页)

现、调用、组合以及自动优化也是今后工作的重点。

参 考 文 献

[1] Kim S M, Rosu M C. A survey of public Web services // Feldman S I, Uretsky M, Najork M, et al. eds. Proc. of the 13th Int'l Conf. on World Wide Web (WWW 2004). New York: ACM Press, 2004. 312-313

[2] Tian M, Gramm A, Ritter H, et al. Efficient selection and monitoring of QoS-aware Web services with the WS-QoS framework // Liu J, Cercone N, eds. Proc. of the IEEE Int'l Conf. on Web Intelligence (WI 2004). New York: IEEE Press, 2004: 152-158

[3] Bryson J, Martin D, McIlraith S, et al. Toward behavioral intelligence in the semantic Web. IEEE Computer, 2002, 25(11): 48-54

[4] Cardoso J, Sheth A, Miller J, et al. Quality of service for workflows and Web service process. Journal of Web Semantic, 2004, 13: 281-308

[5] Zeng L Z, Benatallah B, Ngu A H H, et al. QoS-aware middleware for Web services composition. IEEE Transactions on Software Engineering, 2004, 30(5): 311-327

[6] 代鲸，杨雷，张斌，等. 支持组合服务选取的QoS模型及优化求解[J]. 计算机学报，2006，29(7)：1167-1178

[7] Jensen K. An introduction to the Theoretical Aspects of Colored Petri Nets[J]. Lecture Notes in Computer Science, 1994, 803: 230-272

[8] Tan Z, Lin C, Yin H, et al. Approximate Performance Analysis of Web Sercices Flow Using Stochastic Petri Net[J]. Lecture Notes in Computer Science, 2004, 3251: 192-200

[9] Narayanan S, McIlraith S. Ananlysis and simulation of Web Services[J]. Computer Networks, 2003, 42(5): 675-693

[10] 杨胜文，史美林. 一种支持QoS约束的Web服务发现模型[J]. 计算机学报，2005，28(4)：589-594

[11] 胡建强，邹鹏，王怀民，等. Web服务描述语言QWSDL和服务匹配模型研究[J]. 计算机学报，2005，28(4)：505-513

[12] Gao Xiang, Yang Jian, Papazoglou M P, et al. The capability matching of Web services // Proceedings of the IEEE Four International Symposium on Multimedia Software Engineering (MSE'02). California, USA, 2002: 56-63

[13] Ashemian V H S. A Graph2Based Approach to Web Services Composition // Proceedings of the 2005 Symposium on Applications and the Internet Society. 2003: 191-200

[14] Milanovic N, Malek M. Current Solutions for Web Service Composition[J]. IEEE Internet Computing, 2004, 8(6): 51-59

[15] Hamadi R, Benatallah B. A Petri Net - Based Model for Web Service Composition [C]// Proceedings of the 14th Australasian Database Conference. Adelaide: Australia, Australian Computer