文章编号:1001 - 9081(2007)01 - 0221 - 04

# 移动环境下的垃圾短信过滤系统的研究

邓维维,彭 宏
(华南理工大学 计算机科学与工程学院,广东 广州 510640)
(denvy@ tom. com)

**摘 要**:提出了一种分布式的垃圾短信过滤系统,它适合于移动网络,具有自学习能力,能够及时发现垃圾信息源,有效的过滤垃圾短信。在传统以词为属性的贝叶斯过滤算法的基础上,加入了规则和长度信息,利用互信息减小单词属性的个数。实验表明,它在短信过滤方面具有空间占用小和性能更好的特点,适合在移动电话上使用。同时还提出了一种垃圾短信发送者的可能性排名的方法。

**关键词**:移动计算; 垃圾短信; 短信过滤; 朴素贝叶斯
**中图分类号**: TN918.91;TP393.09 **文献标识码**:A

# Research on junk SMS filtering system on mobile environment

DENG Wei-wei, PENG Hong
(*School of Computer Science & Engineering, South China University of Technology, Guangzhou Guangdong* 510640, *China*)

**Abstract:** This paper introduced a distributed Short Message Service (SMS) filtering system which was applicable on mobile network. This system has self-learning and knowledge updating capability and it can find junk SMS sender with a proper high credibility. The main algorithm used in this system is the Nave Bayesian classification algorithm. Some attributes such as the length of the SMS and rules found by statistics are added to attribute set, and experiments show that it results in a better performance than the traditional word based Bayesian approach. This paper also provided an approach to rank the suspicious SMS senders on their probabilities to be real junk SMS senders according to some measures.

**Key words:** mobile computing; junk short message; SMS(Short Message Service) filtering; Naïve bayesian classification

## 0 引言

手机短信与邮件一样存在着令人苦恼的垃圾信息问题,来自不同渠道的各种垃圾短信充斥手机短信,给用户带来了很多的烦恼。手机垃圾短信是指未经请求或允许而收到的,对接收者来说无用的短信,例如未经短信接收人请求或允许而发送的商业广告。垃圾短信的常见内容包括广告信息、色情信息、假中奖信息、欺诈信息、恶作剧等。

目前,垃圾信息过滤的研究主要集中在邮件方面。垃圾信息过滤可以看作是文本分类技术的应用[1]。黑白名单、基于规则的方法、贝叶斯和 SVM 等,都常用于文本分类。主要的垃圾邮件过滤系统,比如 SpamAssassin 和 Brigntmail 是基于规则的,这些规则可以是关键词,假的邮件头,特殊的文本格式或者特定的单词。使用 SpamAssassin 方法可以过滤大约 90% 的垃圾信息。但是,规则是静态的,因此,可以比较容易的绕过。文献[2]提出了使用贝叶斯方法来进行分类。文中的实验结果表明,它能过滤 92% 的垃圾,大约 1.16% 的合法邮件被分类为垃圾,显然这还不是很实用。文献[1,4]提出了基于贝叶斯的混合过滤方法,将准确率提高到了 99.5%,只有 0.03% 的合法邮件被错误的标记为垃圾。在短信过滤方面,这些方法是否还会有效呢?

和垃圾邮件过滤比较,短信过滤有以下的特点:

(1) 短信只包含文本和发送者的号码,没有其他的信息,比如附件,图片,链接。这将大大减少可以利用的规则的数量;

(2) 短消息的长度是有限制的,一般为 140 个英文字符长度,也就是 70 个中文字符。短文本提供的可用文字信息更少,需要其他的可用特征来补充;

(3) 对垃圾短信的处理有两种技术:一种是在 SMSC(短信服务中心)进行处理,另一种方法是直接在手机上用编制的内嵌程序实现。若在短信中心过滤可能导致被错误分类的信息无法到达用户手中。而且,有些信息,对于不同的人,它可能是垃圾,也可能正是接收者所需要的,比如彩票信息,因此过滤需要有个性化的选择,这在短信中心是比较难实现的。同时,过滤系统还应该能够实时处理,在短时间处理完,接收者不可能较长时间等你判断某条消息是否是垃圾。和个人电脑相比,手机的计算速度和空间都是有限的。因此,过滤方法的速度和空间消耗是一个需要考虑的问题;

(4) 邮件地址是很容易伪造的,但是短信的号码是很难伪造的,因此,黑白名单方法在短信系统中是很有效的;

(5) 短信相对于互联网来说,需要精确计费,因此发送和接收一般是集中处理,可以很方便地在短信中心阻止黑名单用户。

本文提出了一些新的分类属性,比如短信长度属性和规则匹配属性。实验表明,它能提高分类的准确度。为了能够防止发送者绕过过滤器的过滤模型,过滤器必须不断的更新。我们提出了一种分布式的短信过滤系统,系统将学习部分放

**收稿日期**:2006 - 07 - 13;**修订日期**:2006 - 09 - 30
**基金项目**:国家自然科学基金资助项目(60574078);广东省自然科学基金资助项目(31454)
**作者简介**:邓维维(1978 - ),男,湖南岳阳人,博士研究生,主要研究方向:移动计算、数据流挖掘; 彭宏(1956 - ),男,重庆人,教授,博士生导师,博士,主要研究方向:数据仓库、数据挖掘.

置在短信中心,过滤器放置在手机上。过滤中心从手机上收集被错误分类的短信(由用户选择出),然后对它们进行学习。学习后将产生增量更新信息,然后手机从中心下载这些信息,这样,手机上的过滤器将不会变得形同虚设。垃圾信息发送者相对于正常的手机信息发送者,会有些不同的特征。我们根据这些特征,提出了一种对可能的垃圾短信发送者进行排名的方法,能够更准确的找到垃圾发送者,从而在短信处理中心将它们过滤掉。

## 1 贝叶斯分类

贝叶斯分类是统计学分类方法,它可以预测类成员关系的可能性,如给定样本属于某个特定类的概率。朴素贝叶斯分类算法可以和判定树和神经网络分类算法媲美。贝叶斯分类假定一个属性值对应给定类的影响独立于其他的属性值。朴素贝叶斯分类可以描述如下:

(1) 设样本可以用一个 $n$ 维特征向量 $X=(x_1,x_2,\cdots,x_n)$ 表示,其中 $x_i$ 分别表示对 $n$ 个属性 $A_1,A_2,\cdots,A_n$ 的度量。

(2) 假定有 $m$ 个类 $C_1,C_2,\cdots,C_m$。给定未知数据样本 $X$,根据贝叶斯公式,样本属于某个类 $C_l$ 的概率为:

$$P(C_l \mid X)=\frac{P(X \mid C_l)P(C_l)}{P(X)}=\frac{P(X \mid C_l)P(C_l)}{\sum_{k=1}^{m}P(C_k)P(X \mid C_k)} \tag{1}$$

其中类的先验概率可以用 $P(C_l)=s_l/s$ 估计,$s_l$ 是类 $C_l$ 中训练样本数,而 $s$ 是样本总数。

(3) 当特征项很多时,计算 $P(X \mid C_l)$ 很复杂。为了简化计算,可以作特征项间条件独立的朴素假定。即属性间不存在依赖关系(大量的研究表明,这种假设很有效[5,6])。由此可得:

$$P(X \mid C_l)=\prod_{k=1}^{n}P(x_k \mid C_l) \tag{2}$$

其中 $P(x_k \mid C_l)$ 由样本估计得来,所以我们可以得到:

$$P(C_l \mid X)=\frac{P(C_l)\prod_{k=1}^{n}P(x_k \mid C_l)}{\sum_{h=1}^{m}\left[P(C_h)\prod_{k=1}^{n}P(x_k \mid C_h)\right]} \tag{3}$$

为了对某个未知的短信 $X$ 进行分类,需要对每个分类 $C_l$ 计算他的 $P(C_l \mid X)$。短信 $X$ 属于某个类 $C_l$,当前仅当:

$P(C_l \mid X) > P(C_j \mid X), 1 \leqslant j \leqslant m; j \neq l$

## 2 特征提取

一般的基于贝叶斯的文本过滤,只考虑采用分词后的词作为属性;我们考虑了短信的特殊情况,在将词作为特征项的同时,加入了短信长度以及规则作为新的特征项。实验表明,这些属性的加入提高了垃圾短信的识别效果。

### 2.1 短信样本预处理

样本中的信息,可以用向量来表示。短信 $m$ 可以表示为 $n$ 维的特征向量 $X_m=(x_1,x_2,\cdots,x_n)$,对于文本分类,特征项一般采用对文本进行分词后的单词。算法采用以下步骤进行样本预处理。

(1) 对短信进行分词,分词是将短信分割成一个个有意义的单词。像英文等以词为最小语言单位不同,中文的最小语言元素是字,字再构成词,而且词之间没有分隔标记(比如空格),所以中文的分词相对来说复杂很多;

(2) 对分类没有用的词,主要是分词后形成的单个的字,以及叹词,语气助词,代词等。通过查阅助词,代词表等方式除掉;

(3) 单词抽象化。将单词提供的信息抽象到更高的一个层次,例如:

数字:比如 123,456 等;

电话号码:139*******等标记为手机号码;020***标记为固定电话;

URL 地址:如 http://www.scut.edu.cn/等;

金钱:如 10 元。

以上这些可以通过正则匹配来识别。

### 2.2 分词特征提取

由于短信由很多不同的词组成,如果把这些词都作为特征项,则特征项的维度过大[1]。不难发现,有些单词对区分正常或者垃圾信息所起的贡献很小,完全可以忽略,因此相应的维数可以减小。一般可以根据词的信息熵增益或互信息决定特征的选取,我们采用互信息。计算每个候选属性 $X$ 出现与否和某个分类的互信息 $MI(X;C)$,这里我们只有两个分类 $junk$, $legit$

$$\sum_{x\in\{0,1\},c\in\{junk,legit\}}P(X=x,C=c)\cdot\log\frac{P(X=x,C=c)}{P(X=x)P(C=c)}$$

然后从中选出具有最高互信息的属性作为分词后的特征。$P(X,C)$, $P(X)$ 和 $P(C)$ 可以从样本中统计得来。

### 2.3 长度特征提取

单个短信最大长度为 70 个中文字符,我们对短信长度按照中文字符个数作了统计,其中英文字符按半个中文字符计算。统计结果如图 1。

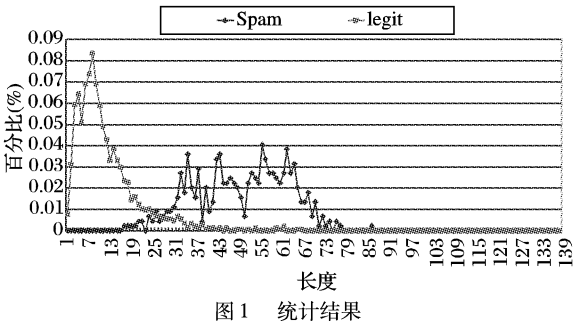


图 1 统计结果

我们可以看到,垃圾和非垃圾短信在长度上有很明显的区别,垃圾短信普遍具有更长的长度,也就是说具有更多的信息。如果短信 $X$ 的长度 $L$ 是 $b$, 那么 $P(b \mid C_{junk})$ 和 $P(b \mid C_{legit})$ 可以从样本中统计得来。很少有样本超过 70 个中文字符,所以根据样本得来的统计误差比较大。当 $L>70$, 我们假设 $P(b \mid C_{junk})$, $P(b \mid C_{legit})$ 具有相同的值 $\tau$, 这样,长度这个属性将对短信 $X$ 分类没有影响。

### 2.4 规则特征提取

我们统计了一些规则的匹配频率,如表 1 所示。其中 $P(R_i \mid C_{junk})$ 为规则 $R_i$ 在垃圾短信中规则匹配频率;$P(R_i \mid C_{legit})$ 为规则 $R_i$ 正常短信中规则匹配频率。

假设各个规则间独立,并将这些规则也当作分类属性。$P(R_i \mid C_{junk})$,$P(R_i \mid C_{legit})$ 均可从样本中得来。

表 1　规则的匹配频率

| 规则 | 描述 | $P(R_i \mid C_{junk})$ | $P(R_i \mid C_{legit})$ |
|---|---|---|---|
| $R_1$ | 包含有电话号码 | 20.3% | 5.3% |
| $R_2$ | 包含有 URL | 3.1% | 1.1% |
| $R_3$ | 包含有钱信息 | 5.2% | 1.2% |
| $R_4$ | 包含在白名单列表 | 0% | 73.1% |
| $R_5$ | 包含在黑名单列表 | 23.3% | 0% |
| $R_6$ | 发送者的号码是一个有效的固话(小灵通)或者移动电话号码 | 65% | 100% |

## 3　学习和分类

假设在属性抽取完成后,得到 $u$ 个词属性和 $v$ 个规则属性。短信 $X$ 可以表示为 $(w_1,w_2,\cdots,w_u,b,r_1,r_2,\cdots,r_v)$。学习时计算每个属性 $x_i$ 的 $P(x_i \mid C_{junk})$ 和 $P(x_i \mid C_{legit})$。在分类时,根据公式(3)计算 $P(C_{junk} \mid X)$ 和 $P(C_{legit} \mid X)$。短信过滤是两类分类,存在两种分类错误:将垃圾短信判别为非垃圾短信或者将非垃圾短信判别为垃圾短信。对我们来说,第二种错误是更严重的。我们定义一个变量 $\alpha = \dfrac{P(C_{junk} \mid X)}{P(C_{legit} \mid X)}$,并给定一个常量 $\lambda$,如果 $\alpha > \lambda$,则分类器预测 $X$ 是一个垃圾短信,否则认为它是个合法短信。

## 4　系统设计

系统为一个分布式的结构。它包含一个短信处理中心和若干个短信过滤代理者。结构如图 2 所示。

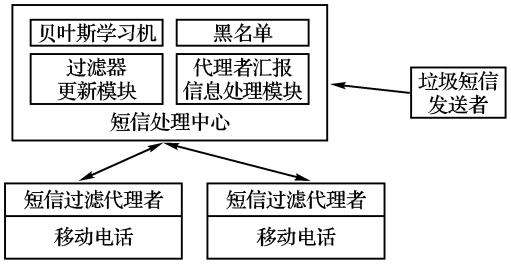


图 2　系统结构

### 4.1　短信过滤代理

短信代理运行在用户的手机上,它具有以下功能:

(1) 当短信到达时,它判断短信是否属于垃圾短信;

(2) 向短信处理中心汇报被错误分类的短信(由用户指出),主要包含以下两种:

1) 将垃圾短信识别为非垃圾;

2) 将非垃圾短信识别为垃圾。

(3) 定期向短信中心汇报被错误正确分类的垃圾短信。包含以下信息:

发送者号码 $m$,发送时间 $t$,接收者号码 $n$。可以表示为向量 $E_i(m_i,t_i,n_i)$;

(4) 定期通过 GPRS 从短信中心下载模型更改和规则的更改。

所有的信息以 GPRS 的方式发送或者接受。GPRS 能提供 9.6kbps 的连接速度。如果用户每天收到 10 条垃圾短信,其中有一条被错误分类,则它需要汇报的字节数最多为:$L_{\text{最大短信长度}} + 10 \times (N_{\text{保存电话号码所需的字节数}} + T_{\text{保存时间所需的字节数}}) = 140 + 10 \times (18 + 2) = 340\text{B}$。如果有 300 个单词存在属性列表中,则需要最多 $300 \times (W_{\text{保存单词所需的平均字节数}} + 2 \times S_{\text{保存条件概率所需的字节数}}) = 300 \times (5 + 2 \times 4) = 3900\text{B}$ 来保存过滤模型。在增量更新的情况下,需要的字节数更少。因此,每天手机使用者仅需要少于 1s 的时间就能够汇报错误并下载更新。信息汇报存在隐私问题,用户会担心隐私泄漏。其实,手机用户现在发送的短信都要经过短信处理中心,因此,用户是默认相信短信中心会保障用户的隐私权的,不会将这些短信泄露给第三方。短信处理中心服务器运行在短信中心,也就是说在手机用户的信任范围内。同时,用户可以选择不汇报错误分类的短信。

### 4.2　短信处理中心

(1) 从短信代理收集错误分类的短信,进行增量学习,每隔一段时间形成增量更新。增量更新包含以下信息:

1) 新增的单词特征项及其在垃圾短信和正常短信中出现的条件概率;

2) 删除的单词特征项;

3) 修改单词特征项在垃圾短信和正常短信中出现的条件概率。

(2) 根据代理者汇报的信息和垃圾短信发送者的特点,识别出真正的垃圾短信发送者。

垃圾信息发送者一般具有以下特点:

- 发送的消息基本上都是垃圾短消息;
- 发送频率比较高,即单位时间内发送的数量大;
- 一般每条短信发送给不同的人;
- 发送没有时间特征,即任何等长时间段发送短信的数目相同。

我们定义一些度量来定量描述这些特征的大小。

设 $E_{1r},E_{2r},\cdots,E_{qr}$ 为发送号码 $r$ 在某段时间 $(T_a,T_b)$ 之间发送的垃圾短信(由 Agent 汇报),

发送频率:

$$f_{1r} = \frac{q}{T_b - T_a}$$

构造数列:

$$L_r = \{l_1,l_2,\cdots,l_{q-1}\}$$

其中 $l_i = t_{i+1} - t_i; 1 \leqslant i \leqslant q-1$。

短信间的分隔程度我们用度量来表示:

$$f_{2r} = \frac{\frac{1}{q}\sqrt{\sum_{i=1}^{q-1}\left(l_i - \frac{T_b - T_a}{q}\right)^2}}{\frac{T_b - T_a}{q}} = \frac{\sqrt{\sum_{i=1}^{q-1}\left(l_i - \frac{T_b - T_a}{q}\right)^2}}{T_b - T_a},\ 1 \leqslant i \leqslant q-1$$

$f_{2r}$ 值越低,表示间隔越平均;

设 $p$ 为 $n_1,n_2,\cdots,n_q$ 中不同的电话号码的数目,则每个号码发送的消息数 $f_{3r} = \dfrac{q}{p}$,通过对 $f_r = k_1 f_{1r} + k_2 f_{2r} + k_3 f_{3r}$ 来比较衡量 $r$ 为垃圾短信发送者的可能性,$f_r$ 越大,可能性越大。其中 $k_1,k_2,k_3$ 为经验常数。

(3) 在获得人工确认后,封锁该短信号码。

## 5　试验

### 5.1　识别效果测试

为了评价垃圾短信过滤效果的好坏,我们使用两个评价指标:$SP$(垃圾短信识别准确率)和 $SR$(垃圾短信识别查全率),用 $n_{junk\to junk}$ 为正确识别出的垃圾短信数,$n_{legit\to junk}$ 为正常短信识别误判为垃圾短信数,$n_{junk}$ 为垃圾短信总数,试验中

$n_{junk}$ 取 1 500：

$$SP = \frac{n_{junk\to junk}}{n_{junk\to junk} + n_{legit\to junk}}$$

$$SR = \frac{n_{junk\to junk}}{n_{junk}}$$

表 2 采用 10 次交叉验证方法得到的结果

| 属性个数 | SP | SR | $S_{junk\to junk}$ | $S_{legit\to junk}$ |
|---|---|---|---|---|
| 100 | 0.933 | 0.788 | 1 182 | 85 |
| 200 | 0.966 | 0.814 | 1 221 | 43 |
| 500 | 0.965 | 0.818 | 1 227 | 45 |

表 3 加入规则后得到的结果

| 属性个数 | SP | SR | $S_{junk\to junk}$ | $S_{legit\to junk}$ |
|---|---|---|---|---|
| 100 | 0.97 | 0.889 | 1 333 | 37 |
| 200 | 0.985 | 0.939 | 1 409 | 22 |
| 500 | 0.985 | 0.933 | 1 400 | 21 |

表 4 加入长度判断后得到的结果

| 属性个数 | SP | SR | $S_{junk\to junk}$ | $S_{legit\to junk}$ |
|---|---|---|---|---|
| 100 | 0.988 | 0.954 | 1 431 | 18 |
| 200 | 0.991 | 0.963 | 1 445 | 13 |
| 500 | 0.991 | 0.964 | 1 446 | 13 |

### 5.2 垃圾短信发送者识别

我们共从代理者收到垃圾短信号码汇报的个数为 35 个，其中确认垃圾发送者 34 个，而且这 1 个非垃圾发送者 $f$ 值最小，这说明公式 $f$ 对识别垃圾发送者有帮助。

### 5.3 性能测试

表 5 系统的性能分析

| 属性个数 | 分词时间/s | 分类时间/s | 内存消耗/MB |
|---|---|---|---|
| 100 | 2.3 | 0.01 | 2.7 |
| 300 | 2.4 | 0.01 | 2.8 |
| 500 | 2.3 | 0.01 | 2.9 |

贝叶斯分类计算速度很快，影响处理速度的是分词算法。我们在 dopod 535 移动电话上进行了测试。dopod 535 系统运行 Windows Mobile phone。它的 CPU 为 200MHz 的 ARM 微处理器，32MB内存。当系统发现某短信为垃圾短信时，会在短信内容前标记“垃圾”两个字。表 5 给出了系统的性能分析。它显示属性的个数对性能影响很小，一般手机的运算速度和内存完全能满足贝叶斯过滤算法的需要。

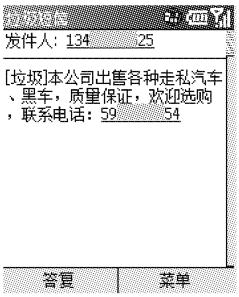


图 3 测试界面

## 6 结语

基于贝叶斯的分布式短信过滤系统具有自学习和更新能力，因此它能克服传统过滤器容易过时的问题。文章还提供了一种对垃圾短信发送者进行排名的方法。实验表明，我们提供的系统能够以较高的准确率识别垃圾信息，并且更适合于移动环境。

随着彩信的普及，基于彩信的垃圾信息也出现了，如何有效的过滤彩信将是我们下一步的工作。

**参考文献：**

[1] PAUL G. A plan for spam[EB/OL]. Http://www.Paulgraham.com/spam.html, 2002-12-10.

[2] PATRICK P, DEKAN GL. Spam Cop: A spam classification & organization program[A]. Proceedings of AAAI Workshop on Learning for Text Categorization[C]. 1998.

[3] MEHRAN S, SUSAN D, DAVID H. A Bayesian approach to filtering junk Email[A]. Proceedings of AAAI Workshop on Learning for Text Categorization[C]. 1998.

[4] PAUL G. Better Bayesian filtering[EB/OL]. Http://www.paulgraham.com/better.html, 2003-12-10.

[5] DOMINGOS P, PAZZANI M. Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier[A]. Proceeding of the 13th International Conference on Machine Learning[C]. 1996.

[6] LANGLEY P, WAYNE I, THOMPSON K. An Analysis of Bayesian Classifiers[A]. Proceeding of the 10th National Conference[C]. 1992.

(上接第 220 页)

多个实例使用递归不断插入在 $A$ 和 $C$ 中。

4）不具有先验运行时知识的多实例模式的解决办法

这与具有先验运行时知识的多实例模式的解决办法很相似，不同的是这种模式中实例化的次数预先不可知，在需要时就创建新实例，直到不需要时为止。由此，不要使用名字 *start* 和 *run*，可得到如下的 $\pi$ 演算形式化表示：

$$A = \tau_A.A_1(c)$$

$$A_1(x) = (\upsilon y)\bar{b} < y >.y < x >.A_1(y) + \bar{x}.0$$

$$B = !b(y).y(x).\tau_B.y.\bar{x}.0$$

$$C = c.\tau_C.C'$$

## 3 结语

多实例工作流模式是一类重要的工作流模式。本文利用 $\pi$ 演算对多实例工作流模式进行了详细的描述，这种方法是完全形式化的，具有较强的语义表达能力，同时使过程模型的语义更加精确。运用 $\pi$ 演算作为理论基础，形式化描述业务过程，能够方便地描述和分析工作流。

**参考文献：**

[1] MILNER R. The polyadic π - Calculus: A tutorial[A]. BAUER FL, BRAUER W, SCHWICHTENBERG H, eds. Logic and Algebra of Specification[C]. Berlin: Springer-Verlag, 1993. 203-246.

[2] MILNER R. Communicating and Mobile Systems: The π-calculus[M]. Cambridge: Cambridge University Press, 1999.

[3] SMITH H, FINGAR P. Business Process Management - The Third Wave[M]. Tampa: Meghan-Kiffer Press, 2002.

[4] VAN DER AALST WMP. Pi calculus versus petri nets: Let us eat "humble pie" rather than further inflate the "pi hype"[EB/OL]. Http://is.tm.tue.nl/research/patterns/download/pi-hype.pdf, 2005-05-31.

[5] VAN DER ALAST WMP. Ter Hofstede AHM. Workflow Patterns[EB/OL]. Http://www.tm.tue.nl/it/research/patterns, 2003-03-10.

[6] LI CY, GOU J, WU HF, *et al.* A Process Meta-Model Supporting Domain Reuse[A]. 2005 International software process workshop[C]. 2005. 459-461.