计算机科学2000Vol. 27№. 4

可视化中文文本挖掘模型*)

Visualization Model for Chinese Text Mining

林鸿飞[1] 贡大跃[2] 张 跃[1] 姚天顺[1]

(东北大学计算机系 沈阳110006)[1] (辽宁铁岭征稽处)[2]

Abstract This paper briefly describes the background of text mining and the main difficulties in Chinese text mining, presents a visual model for Chinese text mining and puts forward the method of text categories based on concept, the method of text summary based on statistics and the method of identifying Chinese name.

Keywords Text mining, Text clustering, Text summary, Chinese name identification, Text feature extraction, Fuzzy semantic representation

面对今天浩如烟海的信息，如何帮助人们有效地收集和选择所感兴趣的信息，更关键的是如何帮助用户在日益增多的信息中自动发现新的概念并自动分析它们之间的关系，使之能够真正地做到信息处理的自动化，这已成为信息技术领域的热点问题。在这样的需求驱动下，文本挖掘得到了长足的发展，并取得了相当的成功。由于目前在因特网上大多数的信息表现形式为文本形式，只有通过文本挖掘才能充分地利用信息资源。

对于中文文本的文本挖掘其难度较大，体现为汉语分词问题，建立完整的汉语概念体系的困难和汉语语法、语义和语用分析的困难。我们在CETRAN[1]的词典、概念词典和汉语分析器的基础上，建立了一个中文文本挖掘模型。其基本思想是借鉴数据挖掘的思想，首先将文本按照内容的相似程度划分成若干类别；抽取每类的特征，作为该类的标记信息。然后对每个文本进行文本结构分析，将文本分解为层次状的结构单元，抽取各个结构单元的特征，并生成文本摘要。最终形成文本结构树，每个树结点代表一个文本结构单元，通过单元的标记信息进行导航，发现新的概念和获取相应的关系。

一、基于概念的文本聚类

目前的文本聚类分析大多建立在词频的基础上。然而，人们在表达相同概念时，使用的词汇具有很大的不同，如个人的喜好，有人愿意用"电脑"一词，而其它人喜欢用"计算机"一词；也可能因文章修辞的缘故，用词要求比较简洁，经常出现同义替换的现象，以避免单调重复；或者词汇表述的概念层次有所不同。因此，仅仅依靠特征词的重复而产生的频率信息是完全不够的。虽然选用的词汇可能不同，但表述的概念却是一致的。如果将特征项映射至概念级，无疑将有助于加强同一类别文本的聚合能力。

1 概念映射

输入文本经过分词处理和停用词处理后，获取文本的特征项信息，这里主要获得文本的项集特征量，经过概念映射后，得到概念集特征向量。具体算法如下：

设文本T的项集特征向量 $P_t=((t_1,d_1,f_1),(t_2,d_2,f_2),\cdots,(t_m,d_m,f_m))$。其中：$t_i$ 为特征项，d_i 为分词时从词典中获取的 t_i 的概念码，f_i 为 t_i 的频率。

概念词典是层次结构的语义组织，不同的层次表明其抽象的程度不同。层次越高，概括性越强，包含的下位概念可能越多。下位概念往往是上位概念的属性、特征、部分或说明；上位概念常常是下位概念的抽象、概括或整体表示。因此，在概念映射中将特征项映射至概念体系的哪个层次上是值得关注的。映射的层次太高，则容易造成主题过于笼统，失去层次划分的意义。

定义概念映射 $\Phi(P,\lambda):P\xrightarrow{\Phi}Q$。其中 P 为项集特征向量，Q 为概念集特征向量。

$Q=((c_1,g_1),(c_2,g_2),\cdots,(c_i,g_i))$，$c_i$ 是 λ 层的概念结点的代码，g_i 是 c_i 的概念密度。

2 概念密度

$g(c)=\sum_{t\in S}f(t)/K^{d-1}$。它表示概念 c 在文本中的

*)国家自然科学基金、国家教委博士点基金资助项目。

集聚程度。其中集合 S 是项集特征向量 P 中概念 c 的所有下位概念的项的集合。t 是属于集合 S 的特征项，$f(t)$ 是 t 的频率，d 是 t 的概念结点到概念 c 的最短路径长度，K 是常数（$K>1$，如 $K=\sqrt{2}$，$K=2$等）。

3 概念消歧

在分词和概念标注中，会出现未登录词、没有概念标注的词和一词具有多个概念标注的情况。对于前两种情况，利用如下算法确定其概念标注。在含有词 w 的段落中，统计共现词频数 $f_w(t)=l$，l 为 w 与 t 共同出现的句子数。获取频数量大者 t 的概念结点 c，将 w 的概念标注定义为 c 的子结点，c 为其父结点。

对于第三种情况，假设词典中 w 有 m 个概念标注 $c_1,c_2,\cdots,c_m$，在含有词 w 的段落中，统计共现概念函数 $h_w(c_i)=\frac{1}{D}\sum_{t\in T}f(c_i,t)$，$D$ 为 c_i 的子结点数，T 是一棵以 c_i 为顶点的子树，$f(c_i,t)$ 是 t 在段落中的频率。取共现概念函数最大者 c 为 w 的概念标注。

经过如上处理，获得各段的概念集特征向量。采用这种方法可以增强相似文本之间的相似程度，而且在某种程度上减少向量中各个分量间的依赖情况，即"斜交"现象，降低了向量的维数。因此，可以提高向量空间模型应用的效率。

4 基于概念的文本聚类

假定文本集为 D，共分为 n 类。采用示例文本集作为各类的表示，$D=D'\cup D''$，其中 D' 是训练文本集，D'' 是待分类的文本。该聚类方法的基本思想是将待分类的文本与每个类别的文本重心相比较，以确定与之最相似的类别。这里文本重心按如下计算：

假设第 k 类的文本重心为 $W=(w_1,w_2,\cdots,w_m)$，L 代表其训练集文本数，训练集 $D_k=(T_1,T_2,\cdots,T_L)$，其中 $T_i=(w_{i1},w_{i2},\cdots,w_{im})$，$w_{ij}$ 是概念密度，它表明概念在文本中的集聚程度。则有 $w_j=\frac{1}{L}\sum_{p=1}^{L}\sum_{q=1}^{m}w_{pq}(j=1,2,\cdots,m)$。设待分类文本为 $T=(a_1,a_2,\cdots,a_m)$，计算相似程度 $Sim(T,W)=\frac{1}{\|T\|\|W\|}\sum_{j=1}^{m}a_jw_j$，取最大者的类别为其所属，这里不允许兼类。

选择重心方法的主要目的在于这种算法的响应速度快，计算简便。由于采用概念密度作为权重，减少了分量之间的依赖关系，与单纯的词频相比精度较高。

二、文本特征的提取

文本特征项包括两个部分：一是一般特征项，即由一般名词导出的概念；二是由专有名词，包括人名和数量信息构成的专有特征项。

1 一般特征项的抽取

一般特征项根据阈值，将其权重大于阈值的特征项列出。特征项的权重函数定义如下：

$$f_w(t_i)=\frac{f_w(t_i)\log_2(1+f_s(t_i))'}{\sqrt{\sum_{j=1}^{m}(f_w(t_j)\log_2(1+f_s(t_j))')^2}}$$

其中：$f_w(t_i)$ 表示特征项 t_i 的权重函数；$f_w(t_i)$ 表示特征项在文本内的频数；$f_s(t_i)$ 表示特征项 t_i 的段落频率，即包含 t_i 的段落数/文本总段落数；l 表示特征项 t_i 的长度。

这个公式实质上是著名的权重公式 $tf*idf$ 的扩展，权重函数的设计基于如下的事实：

特征项的段落频率越高，表明该特征项反映文本主题的能力越强，因此应赋予较大的权重。另外，短词具有较高的频率，更多的含义，是面向功能的；而长词的频率较低，是面向内容的，加大长词的权重，增强词汇的区分度，也可以减轻单个汉字成词的不稳定性。

标题、副标题以及关键字表中出现的词汇和短语是当然的特征项。

2 专有特征项的抽取

（1）中文姓名识别　根据文[3]的中文姓名统计结果的提示，应建立姓氏用字表（First Name List）和名字用字表（Last Name List）和常用姓名表（Common Name List），检测可能的姓名用字。

为了方便表述和处理，为文本中出现的字和词赋予相应的属性函数值 ATTRIBUTE(x)，x 为字串，x=c1,c2,…,cn,c1,…,cn 为单字。标点符号的属性值为 Sign。

定义1　姓氏用字表中的字称为姓氏用字。"王"，"林"等等。属性值为 Surname。

定义2　姓名中不使用的字或极少使用的单字称为名字禁用字。如"死"、"奸"、"吧"、"呢"等等。属性值为 Stop。

定义3　出现在常用姓名表中的词汇称为姓名用词。如"王雪松"。属性值为 Name。

定义4　姓名中出现的词汇称为姓名用词，如"高尚"、"方圆"。属性值为 Pass。

定义5　对于符合下列条件的双字词组称为普通用词。属性值为 Common。否则称为非普通用词，属性值为 None。

条件1：出现在分词字典中的非姓名用词。如"翻阅"、"浏览"。

条件2：首字为动词，尾字为虚词。如"笑了"、"跑着"。

条件3：首字为虚字，尾字为动词。如"亦是"、"也算"。

条件4：首字为动词，尾字为方位词。如"翻过"、"跳

上"。

条件5：首字为数词，尾字为量词。如"三台"、"八张"。

条件6：不是多字词的前两个字。如"才华横溢"、"张牙舞爪"。

定义6 有一些词或符号经常出现在姓名的左右，包括表示称谓的名词和指界动词。如"先生"、"省长"、"经理"、"指出"、"说"、"授予"等等。出现在姓名左边的称为前称谓词，属性值为Left；出现在姓名右边的称为后称谓词，属性值为Right。

定义7 若x在文本T中确认为姓氏用字，则First_Name(T,x)为真，否则为假。

定义8 若y在文本T中确认为名字用字，则Last_Name(T,y)为真，否则为假。

定义9 姓名解析表达式Name_Describe(x,m,n)=e[0]f[m]…f[2]f[1]xb[1]b[2]…b[n]e[1]，其中ATTRIBUTE(x)=Surname，ATTRIBUTE(e[0])=Stop OR Sign，ATTRIBUTE(e[1]=Stop OR Sign，m>=0，n>=0。实际上是将含有姓氏用字的句子分段后的结果。

姓氏识别算法：

```
PROCEDURE Identification_Surname(Text T, Name_Describe(x,m,n))
BEGIN
    IF ATTRIBUTE(f[2],f[1])=Left THEN First_Name(T,x)=TRUE
    IF ATTRIBUTE(b[1],b[2])=Right OR ATTRIBUTE(b[2],b[3])=Right OR
       ATTRIBUTE(b[3],b[4])=Right THEN First_Name(T,x)=True
    IF ATTRIBUTE(f[1],x)=None AND ATTRIBUTE(x,b[1])=None
       THEN First_Name(T,x)=True ELSE
    IF ATTRIBUTE(f[1],x)=Common THEN
       IF AND ATTRIBUTE(f[2],f[1])=None THEN First_Name(T,x)=False ELSE
     IF ATTRIBUTE(f[3],f[2])=None THEN First_Name(T,x)=True ELSE
       MessageBox"Can't identify the Surname!"
END.
```

名字识别算法：

```
PROCEDURE Identification_Name(Text T,Name_Describe(x,m,n))
BEGIN
    IF n=1 THEN Last_Name(T,x,b[1])=True ELSE
    IF n=2 THEN Last_Name(T,x,b[1],b[2])=True ELSE
    IF ATTRIBUTE(b[1],b[2])=Right THEN Last_Name(T,b[1],b[2])=True ELSE
    IF ATTRIBUTE(b[3],b[4])=Right THEN Last_Name(T,b[1],b[2])=True ELSE
    IF ATTRIBUTE(b[2],b[3])=Right THEN Last_Name(T,b[1])=True ELSE
    IF ATTRIBUTE(b[2],b[3])=Pass THEN Last_Name(T,b[1],b[2])=True ELSE
    IF ATTRIBUTE(x,b[1],b[2])=Name THEN Last_Name(T,b[1],b[2])=True ELSE
    IF ATTRIBUTE(b[3],b[4])=Common THEN Last_Name(T,b[1],b[2])=True ELSE
    IF ATTRIBUTE(b[3])=Stop THEN Last_Name(T,b[1],b[2])=True ELSE
    IF ATTRIBUTE(x,b[1])=Name THEN Last_Name(T,b[1])=True ELSE
    IF ATTRIBUTE(b[2],b[3])=Common THEN Last_Name(T,b[1])=True ELSE
    IF ATTRIBUTE(b[2])=Stop THEN Last_Name(T,b[1])=True ELSE
       Last_Name(T,b[1],b[2])=True
END
```

对《人民日报》1994年800余篇语料处理结果，表明查全率和正确率分别为98.62%和81.94%，而且误判率大于漏判率。

(2)日期、时间、数字和货币特征的识别 对于数词而言，按照语义分类则分成系数词和位数词。系数词是数字的名字，位数词是数字所处位置的指称。位数词又可以分成层位数词和子位数词，为了叙述简便，本文限定层位数词最高为"万亿"。假定数字只有一层时，即小于一万的数字其层位数词为Ω(其字符值为""，其数值为10^0)，小于十的数字的子位数词也为Ω。则数词结构定义如下：

数词::={子数词块+层位数词}*

子数词块::={子系数词+子位数词}

层位数词::={万亿、亿、万、Ω}

子位数词::={千、百、十、Ω}

子系数词::={零、一、二、三、四、五、六、七、八、九}

下面讨论汉语数字转化的规则。首先将层位数词、子位数词和子系数词转化为相应的十进制数字，特别地将Ω转化为10^0。根据层位数词分段，得到如下的数字特征解析式：$Digit(x)=(x_{n3}y_3+x_{n2}y_2+x_{n1}y_1+x_{n0}y_0)w_n+\cdots+(x_{03}y_3+x_{02}y_2+x_{01}y_1+x_{00}y_0)w_0$，其中$n$称为层数，$y_3$=千，$y_2$=百，$y_1$=十，$y_0$=""，$x_{ij}(i=n,\cdots,0;j=3,2,1,0)$为子系数词，$w_i(i=n,n-1,\cdots,0)$为层位数词。则转化的数字为：

$$Digit(x)=\sum_{i=0}^{n}\left(\sum_{j=0}^{3}x_{ij}10^j\right)10^i.$$

如果有小数，通过"点"或其它标志分段，对于小数部分单独处理后，再与整数部分相加，小数部分数字解析式(不包括小数点)，$Digit(x)=x_1x_2\cdots x_m$，$x_i(i=1,\cdots,m)$是子系数词。则转化的数字为：

$$Digit(x)=\sum_{i=1}^{m}x_i10^{-i}.$$

识别出的数字还应考虑相应的种类特征，即分成日期、时间、数字和货币等加以处理。这些种类的主要特点是相应的词组一般由数词和各种特征词构成，如年、月、日、元、角、美元，马克等；数词表现方式比较复杂：有汉字，有阿拉伯数字，数字间可能存在其它字。如二十八岁、50马克、五月六日、1234.01、四分之三、百分之四十五等等。

对于日期特征，存在三种日期形式：一是绝对日

期，如一九九九年六月八日；另一种是相对日期，即相对于某日期原点的日期，若给定日期原点值，则可以转化成绝对日期，如三年前；最后一种是泛指日期，无法转化为绝对日期，如几天前，若干年后。为了统一日期表示，便于检索处理和统计需求，将各类日期转化为如下格式：(日期原点为 d_origin)

一九九九年六月八日→1999-06-08
五月八日→d_origin+0000-05-08
二十四日→d_origin+0000-00-24
三年前→d_origin-0003-00-00
几天前→[d_origin-0000-00-09, d_origin-0000-00-01]

对于时间特征，往往与日期同时出现，这里在日期的基础上，单独考虑时间因素。它同样存在三种时间形式：一是绝对时间，如三点五十分三十一秒；另一种是相对时间，即相对于某时间原点的时间，若给定原点值，则可以转化成绝对时间，如二小时前；最后一种是泛指时间，无法转化为绝对时间，如几分钟，几小时后。相仿日期表示，将各类时间转化为如下格式：(时间原点为 t_origin)

三点五十分三十一秒→03:50:31
十八时五分→18:05:00
十点整→10:00:00
三小时前→t_origin-03:00:00
六分钟后→t_origin+00:06:00
几小时前→[t_origin-12:00:00, t_origin-01:00:00]

对于数字统一转化为＃＃＃，＃＃＃，＃＃＃，＃＃＃、＃＃格式，加上相应的量词或货币名称构成数字词组，表示单纯数量特征。

一万两千三百四十美元→12,340.00美元
三十六吨→36.00吨
六百八十多元→[680—690]元
高于1600转/分→[1600—∞]转/分

值得指出的是将泛指日期，泛指时间和不定数量词的模糊语义数量化，转化为相应的区段，即模糊区间数，在文本挖掘中具有重大意义。采用的基本思想是设 U 是线性有序的论域，这里 U 可以是所有合法日期的集合、所有时间的集合和数字的集合，$[a,b]/p$ 表示一个模糊区间数，$a,b\in U$，$0<p\leqslant 1$，$[a,b]$称为区间，p 称为可能度。区间数表示某个模糊数落在区间$[a,b]$的可能度。在日期特征中，指某个具体日期落在日期区段的可能性；对于数字，表示数字落在该数字区间的可能性。对于出现的伪日期、伪时间和不定数量词，将其区间加大到足够大范围，使其能够包括通常意义下的所有可能取值，此时 p 的值为1，因此对于常见的伪日期、伪时间和其它不定数量的词汇说明相应的区间。对于确切的日期和数字认为它们也是的模糊区间数，$[a,b]/1$，$a=b$。在检索操作时，两个模糊数之间的语义距离为：

$$S(x,y)=\sqrt{w_1|a_1-a_2|^r+w_2|b_2-b_2|^r+w_3|p_1-p_2|^r},$$
$$x=[a_1,b_1]/p_1,\ y=[a_2,b_2]/p_2$$

这里 $p_1=p_2=1$，$r=2$，$w_1+w_2+w_3=1$，$w_1=w_2=0.5$，$w_3=0$。所以

$$S(x,y)=\sqrt{0.5*|a_1-a_2|^2+0.5*|b_1-b_2|^2}$$

如果 S(x,y)小于指定的阈值，则可以认为符合检索条件。

三、文本摘要

文本挖掘中的文本摘要是为了用户对文本的内容有一个比较全面的认识，以决定是否深入了解该文本。基于统计的文本摘要自动生成方法的基本思想是原文中与主题密切相关的句子挑选出来，这样的句子往往位于比较特殊的部分或含有较强的提示，而且含有较多的特征项。因此设计句子权重函数，以评价各个句子的重要性。句子的权重函数为：

$$f_i(s)=a+\frac{\beta}{Cl(s)k(s)}\sum_{j=1}^{n}f_w(t_j)$$

其中：$f_i(s)$表示句子的权重函数，$f_w(t_j)$表示特征项的权重函数，l 表示句子的长度，$k(s)$表示句子所包含的分句个数，C 表示比例因子，$a=1$，$\beta=0$当句子为标题、子标题或者句子含有线索词，如"综上所述"、"本文论述了"等；$a=0$，$\beta=1.5$当句子位于段首；$a=0$，$\beta=1.1$当句子位于段尾。

句子权重函数的设计基于如下事实：结论性的句子，包含更多的特征项，句子长度较短，具有很少的分句，因此应该赋予较高的权重。不言而喻，标题，子标题应该赋予较高的权重。对于包含诸如"综上所述"、"总而言之"等线索词的句子，由于它们往往是结论性的句子，所以也应获得较高的权重。大量统计资料表明，每个段落的开头和结尾都是需要考虑的地方。美国学者 P. E. Baxendale 进行过统计，反映主题的句子，85%出现在段首，7%出现在段尾。尤其对于新闻语料，Searchable Lead[6]系统仅仅从文章开头部分抽取给定长度的一部分形成摘要，就达到87%—96%的可接受率。因此对首句和尾句也给予相应的权重。

设文本 $T,s_1,s_2,\cdots,s_n$ 是其句子集合 S，并假定已经按照权重排好序，即：$f_i(s_{i_1})\geqslant f_i(s_{i_2})\geqslant\cdots\geqslant f_i(s_{i_n})$。*Length* 是长度函数，$r$ 是摘要压缩率，*Text_Summary* 是文本 T 的摘要信息。

摘要生成算法：

```
PROCEDURE Generate_Text_Summary(T,S,n,r)
BEGIN
    Temp_Summary="";p=1
    DO WHILE Length(Temp_Summary)+length(s[i_p])
       <=Length(T)*r
    BEGIN
        Temp_Summary=Temp_Summary+s[i_p];t[i_p]=
            True;p=p+1
```

```
    END
    Text_Summary=""
    FOR I=1 To n DO IF t[I] THEN Text_Summary=
        Text_Summary+s[I]
END.
```

结束语 本文提出的文本挖掘模型是建立在文本特征提取基础上的，面对大量的电子文本，良好的响应能力和不依赖具体领域的特性是首要关注的方面。其基本结构如图1。

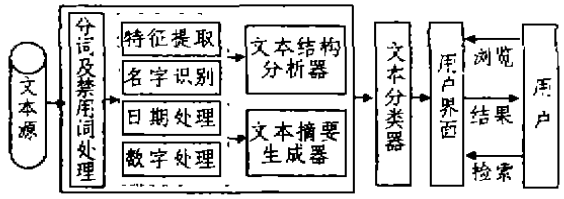

图1 中文文本挖掘模型结构示意图

在文本挖掘模型中，本文提出了基于概念的文本分类方法，初步解决了概念的影射和概念消歧问题。在文本特征提取方面，提出了一般特征项的筛选方法和中文姓名的识别算法以及基于模糊语义的数字特征（日期、时间、数字和货币等）表示方法，在文本摘要方面，提出了基于统计的摘要生成算法。面对海量文本，采用统计方法有着较强的适应性和良好的反映能力，不依赖于具体领域知识，但是，随着需求的深入，引入基于自然语言理解的语法、语义、语用分析势在必行，以便挖掘更深入的知识。但如何协调适应性和精确性的关系，文本的来源多样化与领域知识库的关系是关键问题，也是下一步重点。

参考文献

1 姚天顺，等. 自然语言理解. 清华大学出版社，1995
2 麻志毅，林鸿飞，姚天顺 基于情境的文本中时间信息分析. 东北大学学报，1999，13(6)
3 吴立德，等. 大规模中文文本处理. 复旦大学出版社，1997
4 刘开瑛，等. 中文文本中抽取特征信息的区域与技术，中文信息学报，1998，12(2)
5 Udo Klemens, Schnattinger. Deep Knowledge Discovery from Natural Language Texts. In Proc of the 3rd Conf on Knowledge Discovery and Data Mining. 1997，175～178
6 Salton G, et al. Automatic Structuring and Retrieval of Large Text Files. Communications of the ACM，1993，37(2)：97～108

（上接第31页）

internet and multimedia repositories, doctoral dissertation. University of Simon Fraser，1999
3 Cooley R, et al. Web Mining: Information and Pattern Discovery on the World Wide Web. In: Proc. of Intl. Conf. on Tools with Artificial Intelligenc, Newport Beach, IEEE，1997
4 Cooley R, et al. Grouping Web Page References into Transactions for Mining World Wide Web Browsing Patterns. In: Proc. of Knowledge and Data Engineering Workshop. Newport Beach, CA, IEEE. 1997
5 Cooley R, et al. Data Preparation for Mining World Wide Web Browsing Patterns. Knowledge and Information Systems，1999，1(1)
6 Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. In: 7th Int. Conf. WWW. Brisbane, Australia, April 1998
7 Chakrabarti S, et al. Experiments in topic distillation. In: ACM SIGIR Workshop on Hypertext Information Retrieval on the Web. Melbourne, Australia，1998
8 Luotonen A. The common log file format. Available at: http: // www.w3.org/pub/WWW/，1995
9 World Wide Web Consortium. Available at: http: //www.w3.org/XML/，1998
10 Perkowitz M, Etzioni O. Adaptive Web Sites: Automatically Synthesizing Web Pages. In: Proc. of AAAI-98. 1998
11 Perkowitz M, Etzioni O. Adaptive web sites: an AI challenge. In: Proc. 15th Int. Joint Conf AI. 1997a
12 Perkowitz M, Etzioni O. Adaptive web sites: Automatically learning from user access patterns. In: Proc. of the Sixth Int. WWW Conference. 5 1997b
13 Spiliopoulou M. The laborious way from data mining to web mining. Int. Journal of Comp. Sys., Sci. & Eng., Special Issue on "Semantics of the Web", Mar. 1999
14 Spiliopoulou M, Faulstich L C. WUM: A Web Utilization Miner. 1998
15 Craven M, et al. Learning to Extract Symbolic Knowledge from the World Wide Web. Same to [10]
16 Buchner A. et al. Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining. SIGMOD Record，1998，27(4)
17 Fink J, et al. User-oriented Adaptivity and Adaptability in the AVANTI Project. In Designing for the Web: Empirical Studies. 1996
18 Wexelblat A, Maes P. Footprints: History-rich web browsing. In: Proc. Conf. Computer-Assisted Information Retrieval(RIAO). 1997. 75～84
19 Chen M S, et al. Data mining for path traversal patterns in a web environment. In ICDCS, 1996. 385～392