郑家恒，王兴义，李飞 《信息抽取模式自动生成方法研究》2004 【农作物信息】

山西大学，计算机科学系，山西太原 030006
中文信息学报，第 18 卷，第 1 期，收稿日期：2003 年 8 月 6 日
摘要：模式匹配是信息抽取系统常用的方法，如何生成信息抽取模式就成为信息抽取的关键问题。由于手工编写模式的代价太大，本文尝试采用***聚类方法***自动生成针对中文文本的信息抽取模式。通过计算实例间的相似度，采用单链法聚类，将模式实例划分为不同的类别，每个类别对应一个模式，将同一类别中的模式实例进行合并就可以得到最终的信息抽取模式。以农作物信息文本为实验语料，进行了聚类测试，错分率与漏分率分别为 0.21%和 1.07%，合并后的模式覆盖了人工分析提出的 25 类中的 24 类。
目前中文自动分词和句法分析还存在一定不足，尤其在对未登录词的切分时错误较多，影响信息抽取模式的自动生成。
信息抽取模式：信息抽取模式被看作是由项组成的有序序列，每个项对于一个词（词组）的集合。每个集合中的词（词组）在当前信息抽取领域内具有**相同或相近**含义。以农作物信息文本为例，“播种时间”、“播种期”等词组都表示一类信息，即农作物的播种时间。

设信息抽取模式为 P，则 $P = Item_1, Item_2, \ldots, Item_n$，其中 $Item_i = \{W_{i1}, W_{i2}, , W_{it}\}(1 \leqslant i \leqslant n, W_{ij}(1 \leqslant j \leqslant t)$为词或词组。

例如：部分水稻文本信息抽取模式

<“该”><“品种”>[“株高”](CENTIMETER)<“左右”>
<“平均”><“每”>[“穗”]<“总”><“粒数”>(NUMBER)<“左右”>
[“结实率”](PERCENT)<“左右”,“以上”>
[“千粒重”](GRAM)<“左右”>
[“糙米率”](PERCENT)<“左右”>
<“整”>[“精米率”](PERCENT)
<“作”><“晚季稻”,“早季稻”,“麦茬稻”><“全”>[“生育期”](DAY)<“左右”>
[“成穗率”](PERCENT)<“以上”>
[“蛋白质”]<“含量”>(PERCENT)

其中,“平均”表示词;NUMBER、PERCENT 等表示数目、百分比等含义;[“蛋白质”]表示特征项;<“含量”>表示可选项;(PERCENT)表示抽取项。

考虑公共子序列：给定两个序列A{a1...am}和B{b1...bn}，若存在单调增的整数序列i1<i2...<il和j1<j2<...<jl，满足aik = bjk = ck，k=1,2...l，则C{c1,c2...cl}称A和B的公共子序列。
考虑公共子序列中的**连续元素**时的公共子序列分值公式：Score(C) = |C|+p*delta；|C|是C的长度，p是连续元素对数，delta是连续元素的奖惩值；
模式实例相似度计算：Sim(Ei,Ej) = max(Score(C(Ei,Ej)))/f(|Ei|,|Ej|)，|Ei|表示Ei的长度，f(|Ei|,|Ej|)=min(|Ei|,|Ej|)*(1+delta)-delta【没有详细介绍这个函数】。
创建模式实例
文本分析：从不同格式的文档中分离格式标记，提取纯文本；
文本分割：利用标点符号等分割标记，将文本划分成具有独立意义的字段；
专有特征项识别：对文本中由数词和各种特征次组成的特殊项，如日期、长度、百分比等；
自动分词：将字段分成由基本语言单位（字、词、词组）组成的字段。

模式实例聚类

在向量空间模型中，文档是同一空间中的点，任意文档间的相似度都可由距离表示。只要相似度满足阈值条件，就认为是同一类。

单链法聚类输入数据是模式实例集的相似度矩阵，根据设定的相似度阈值将矩阵转换为无向图，模式实例是顶点，相似度是边的权重。**采用BFS对图遍历，每次搜索得到一个连通分量就是一个模式实例类别。**对全图遍历结束，就生成了模式实例集的类别集。

模式合并

> 模式合并算法如下：
>
> (1)计算候选模式集中任意候选模式间的相似度。
>
> (2)寻找相似度最大的两个候选模式，如 P_i 和 P_j，若 P_i 和 P_j 的相似度大于相似度阈值 t，则转向(3)；否则，转向(4)。
>
> (3)将 P_i 与 P_j 合并成新的候选模式 P_k，将 P_k 加入候选模式集，并删除 P_i 和 P_j。若当前候选模式集中只有候选模式 P_k，则转向(4)；否则，计算 P_k 与其他候选模式的相似度，并转向第(2)步。
>
> (4)模式合并算法结束，输出所有候选模式。

候选模式相似度计算：

候选模式相似度的计算不考虑两个连续项的特殊性，只需计算两个候选模式的最长公共子序列，即：

$$Sim(P_i,P_j)=\frac{|LCS(P_i,P_j)|}{min(|P_i|,|P_j|)}$$

其中，$|LCS(P_i,P_j)|$ 是 P_i 与 P_j 的最长公共子序列 $LCS(P_i,P_j)$ 的长度；$|P_i|$，$|P_j|$ 分别是 P_i 与 P_j 的长度。

模式合并

设有候选模式 P_s 和 P_t，对应的最长公共子序列为 C：

$P_s=PS_1PS_2\ PS_3\cdots PS_n$；$P_t=PT_1\ PT_2\ PT_3\cdots PT_m$；$C=I_1\ I_2\cdots I_l$；其中 $I_1=PS_{i1}=PT_{j1}$，$I_2=PS_{i2}=PT_{j2}$，…，$I_l=PS_{il}=PT_{jl}$；且 $i_1<i_2<\cdots<i_l$ 和 $j_1<j_2<\cdots<j_l$。

将候选模式 Ps，Pt 中与最长公共子序列中对应项一一对齐，则有：

$C=I_1I_2\cdots I_l$

$P_s=PS_1\cdots PS_{(i1-1)}PS_{i1}PS_{(i1+1)}\cdots PS_{(i2-1)}PS_{i2}\cdots PS_{il}\cdots PS_n$

$P_t=PT_1\cdots PT_{(j1-1)}PT_{j1}PT_{(j1+1)}\cdots PT_{(j2-1)}PT_{j2}\cdots PT_{jl}\cdots PT_m$

P_s 与 P_t 被划分为 $l+1$ 组对应的片断，对每组片断分别进行合并，得到新的候选模式。

候选模式片断合并的基本操作有两种：交换与忽略。

设候选模式片断为 P_1，P_2，P_3：

(1)交换(exchange)

$P_1=ABC$，$P_2=ADC$；

P_1，$P_2\rightarrow P_3=A\ exchC$，其中 $exch=B/D$。

(2)忽略(ignore)

$P_1=ABC$，$P_2=AB$；

P_1，$P_2\rightarrow P_3=A\ B\ ignr$，其中 $ignr=C$。

分别将每个候选模式集中的候选模式进行合并，就得到全部的信息抽取模式。

实验结果：

数据来源：中科院、Internet，水稻文本 39 篇，模式实例 1518 个；

模式聚类：

通过分析，设定 $t=0.70$，$delta=1.4$ 对水稻文本中的 1，518 个模式实例进行聚类，共得到 421 个类别，错分率与漏分率分别为 0.21％和 1.07％。

模式合并（421 类别中对 57 类别进行合并）：

在 421 个类别中，有 364 个只包含 3 个以下（含 3 个）的模式实例。它们可以分为非信息抽取模式实例、非常用的信息抽取模式实例、特殊表达方式的模式实例等几种情况。这些模式实例对模式的生成贡献不大，所以不进行模式合并操作。

对其余 57 个类别进行合并，经人工审查，最后得到 42 个针对水稻品种的信息抽取模式，涉及农作物的株高、穗长、成穗率、结实率、生育期等方面的信息。它们覆盖人工分析提出的 25 类信息中的 24 类。