中 文 信 息 学 报
第 18 卷 第 1 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol.18 No.1

文章编号:1003-0077(2004)07-0048-07

信息抽取模式自动生成方法的研究*

郑家恒,王兴义,李 飞

(山西大学 计算机科学系,山西太原 030006)

摘要:模式匹配是信息抽取系统通常使用的方法,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。通过计算模式实例间的相似度,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。

关键词:人工智能;自然语言处理;信息抽取;模式匹配;信息抽取模式

中图分类号:TP391 **文献标识码**:A

Research on Automatic Generation of Extraction Patterns

ZHENG Jia-heng, WANG Xing-yi, LI Fei

(Computer Science of Shanxi University, Shanxi Taiyuan 030006)

Abstract: Most information extraction (IE) systems adopt a pattern-matching approach. As a result, how to generate extraction patterns has become an essential step. As the cost of man-made patterns is very high, we propose a method to generate extraction patterns automatically by clustering. Calculating the similarity between pattern examples and Using single-link clustering, examples of patterns can be clustered into various categories, each of which represents a pattern. We applied the method to Chinese agricultural texts. After clustering, the rate of wrong classification and rate of miss classification are 0.21% and 1.07%, respectively. The patterns obtained from merging include 24 types of the information that belong to the 25 types proposed by manual analysis.

Key words: artificial intelligence; natural language processing; information extraction; pattern-matching; extraction patterns

1 引言

模式匹配(pattern-matching)是信息抽取系统普遍采用的方法。信息抽取模式是指可以传递特定领域中关系或事件信息的语言表达式。因此,如何获取信息抽取模式就成为研究的重点。目前,大多采用的信息抽取方法都能适应各种领域的不同信息抽取任务,唯独信息抽取模式是针对特定任务的,当转向一个新的信息抽取任务时,就必须重新创建一套模式。但是模式的人工创建不仅耗时费力,而且需要既熟悉信息抽取模式格式又精通应用领域的专家,因而极大地限制了信息抽取的应用。另外,信息抽取模式的分布不均匀,存在少数出现频率高的模式以及大量的低频模式。高频模式的生成相对容易,但要覆盖数目庞大的低频模式往往变得非常困难。

* 收稿日期:2003-08-06
基金项目:国家 863 资助项目(2001AA114031)
作者简介:郑家恒(1948—),女,教授,硕士生导师,主要研究方向为中文信息处理.

实现信息抽取模式的自动生成,将在很大程度上克服上述两个障碍。一方面,在转向新的信息抽取任务时,系统可以完全自动或只在少量人工干预下,快速创建信息抽取系统所需的模式资源。另一方面,通过提供给系统更多的训练语料,可以获取尽可能多的抽取模式,从而尽可能覆盖更多语言现象,提高信息抽取系统的性能。

国外,AutoSlog[1]、CRYSTAL[2]等系统从标注了领域语义信息的语料中获取模式。而AutoSlog-TS[3]系统命名用预分类语料,各语料只需标明是否与当前信息抽取任务相关。Ellen Riloff[4]与 Roman Yangarber[5]等则分别尝试从完全未标注的语料中,自动生成信息抽取模式的不同方法。

中文信息抽取模式的自动生成技术研究发展较晚,在国内还鲜有报道。中文信息抽取模式的生成是以中文信息处理研究为基础的。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。鉴于中文信息抽取模式生成所面临的困难,我们根据中文自身的特点,将信息抽取模式看作项的序列,同时借鉴已有的面向英文文本的信息抽取系统的成功经验,提出一种基于聚类的信息抽取模式自动生成方法,从中文文本中自动获取模式。通过计算模式实例间的相似度,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集。将同一模式集中的模式实例进行合并,就可以得到最终的信息抽取模式。我们以农作物信息文本作为实验语料,实验证明,该方法在生成抽取模式时取得很好效果。

2 相关概念

2.1 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,每个项对应于一个词(或者词组)的集合。每个集合中的词(或词组)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;而"50 公分"、"三十厘米"、"8 米"等一些数量词表示了长度概念。

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。

2.2 模式实例

模式实例(pattern example)是信息抽取模式在文本中的具体表现形式。

以农作物文本为例，关于植株高度的信息在各文本中有不同的表达方式，如“株高 100－110 厘米”、“株高 85 厘米左右”、“该品种株高 105 厘米左右”等等。这些语言表达式都是描述株高的信息抽取模式的具体表现形式，即模式的不同实例。将这些字段进行分词，就组成了模式实例。

2.3 公共子序列分值

给定两个序列 $A=\{a_1,a_2,\cdots,a_m\}$ 和 $B=\{b_1,b_2,\cdots,b_n\}$，若存在单调增的整数序列：$i_1<i_2<\cdots<i_l$ 和 $j_1<j_2<\cdots<j_l$，满足 $a_{i_k}=b_{j_k}=c_k,k=1,2,\cdots,l$，则记这个子序列为 $C=\{c_1,c_2,\cdots,c_l\}$，称序列 C 是 A 和 B 的公共子序列，用符号 $CS(A,B)$ 表示。

若公共子序列 C 中的两个相邻元素 c_s,c_t，并且 $s+1=t$，$c_s=a_{i_s}=b_{j_s}$，$c_t=a_{i_t}=b_{j_t}$ 满足条件 $i_s+1=i_t$ 与 $j_s+1=j_t$ 时，称 c_s 和 c_t 为公共子序列 C 中的一对连续元素。

考虑到在公共子序列中连续元素的特殊作用，定义公共子序列的分值公式为：$Score(C)=|C|+p.delta$。其中，$|C|$ 为公共子序列 C 的长度，p 是 C 中连续元素的对数，$delta$ 是为每对连续元素设定的奖惩值。

2.4 模式实例相似度

模式实例可以看作是项的有序序列，项是模式实例含有的基本语言单位（字、词或词组）。因此，根据两个模式实例对应项序列的公共子序列，定义模式实例相似度。设有模式实例 E_i,E_j，二者的相似度为 $Sim(E_i,E_j)$：

$$Sim(E_i,E_j)=\frac{\max(Score(CS(E_i,E_j)))}{f(|E_i|,|E_j|)};$$

其中，$CS(E_i,E_j)$ 是模式实例 E_i 和 E_j 的一个公共子序列，$Score(CS(E_i,E_j))$ 表示公共子序列 $CS(E_i,E_j)$ 的分值；$|E_i|$、$|E_j|$ 分别表示 E_i 和 E_j 的长度，$f(|E_i|,|E_j|)$ 则为模式实例 E_i 与 E_j 长度的函数。

2.5 模式实例长度函数

设定相似度最大值为“1”，并且当两个实例完全相同，或者当其中一个模式实例是另一个实例中一个连续的子序列时，定义这两个模式实例的相似度为最大值“1”。

在模式实例相似度计算公式中，定义模式实例长度函数 f 为：

$$f(|E_i|,|E_j|)=\min(|E_i|,|E_j|)\times(1+delta)\text{-}delta。$$

2.6 模式生成流程

由于相同的信息抽取模式的各实例间具有相似性，而不同模式的实例不相似。因此，模式实例可以依据相互间的相似性进行分类，每个分类对应一个类别，将各个类别中的模式实例进行合并就得到信息抽取模式。

整个方法流程大致可分三步 模式实例创建、模式实例聚类和模式合并，如图 1 所示。

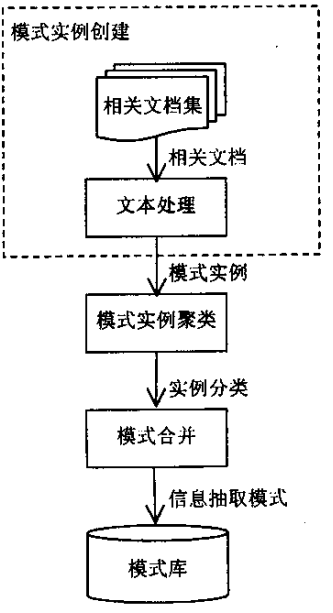

图 1 信息抽取模式自动生成流程

3 模式实例的创建与聚类

3.1 创建模式实例

模式实例（pattern example）是信息抽取模式在文本中的具体表现形式，对应文本中具有

独立意义的字段。在模式实例创建过程中,系统将对相关文档进行一定程度的文本处理,收集所有字段,生成模式实例集。

文本处理过程包括:文本分析、文本分割、专有特征项识别和自动分词4个部分。其中,文本分析的目的是从不同格式的文档中分离格式标记,进而提取纯文本。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。专有特征项识别主要针对文本中一些由数词和各种特征词组成的特殊项,如日期、长度、百分比等,这些专有特征相的识别对信息抽取往往非常重要。自动分词是将字段分成由基本语言单位(字、词或词组)组成的字段。

3.2 模式实例聚类

由于无法预先确定信息抽取模式的类别,所以对模式实例集采取无指导的分类,即聚类的方法。

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。而模式实例间的相似度是针对不同模式实例间公共子序列而定义的。为了避免相似的模式实例被划分到不同的类别中,使同一类别中相似模式实例的数目过少,导致模式合并无法进行,这里采用单链法聚类。因而,只要模式实例与类别中任一实例的相似度满足阈值条件,就将其加入类中,避免遗漏模式实例。

单链法聚类的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可以将相似度矩阵转换为无向图,每个模式实例对应图中的一个顶点,两个模式实例间的相似度代表对应顶点间边的权重。采用广度优先搜索(breadth-first search)对图进行遍历,每次搜索得到一个连通分量,即一个模式实例类别。在对全图遍历结束后,可以生成模式实例集的类别集。

设模式实例集中包含 n 个模式实例 $\{E_1, E_2, \cdots, En\}$,计算任两个模式实例间相似度,生成相似度矩阵 S:

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \Lambda & s_{nn} \end{bmatrix}$$

其中,$s_{ij} = Sim(E_i, E_j) = Sim(E_j, E_i)$,$(1 \leqslant i, j \leqslant n)$。

设定相似度阈值 t,则相似度矩阵转换算法为:

(1)创建空图,设置 $i = 1$。

(2)若 $i \leqslant n$,则在图中增加一个顶点 i,并作 $j = 1$,转向(3)。否则,转向(5)。

(3)若 $j < i$,则转向(4)。否则,作 $i = i + 1$,转向(2)。

(4)若 $Sim(E_i, E_j) \geqslant t$,在图中顶点 i 与 j 之间加一条边,并作 $j = j + 1$,转向(3)。否则,作 $j = j + 1$,转向(3)。

(5)算法结束,输出图结构。

从图中顶点 v_0 出发,广度优先搜索算法如下:

(1)初始化空队列 Q。

(2)输出顶点 v_0,并将 v_0 加入队列 Q 中。

(3)若队列不为空,则转向(4)。否则,转向(8)。

(4)取出队列中的头元素 v,寻找 v 的邻接点 w。

(5)若顶点 w 存在,则转向(6)。否则转向(3)。

(6)若顶点 w 未被遍历,则输出顶点 w,并将 w 加入队列 Q 中,并转向(7)。否则,直接

转向(7)。

(7)取 v 的下一个邻接点 w,并转向(5)。

(8)搜索算法结束。

4 模式合并

4.1 合并算法

我们把每个模式实例的类别定义为候选模式集,模式实例类别中的模式实例定义为候选模式。模式合并就是通过不断地将候选模式集中两个候选模式进行合并,得到最终的信息抽取模式。

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。

4.2 候选模式相似度计算

候选模式相似度的计算同样是以公共子序列为基础的,与模式实例相似度计算的不同处在于 相似度计算方法不同。

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。

4.3 候选模式合并操作

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。

5 实验结果与分析

5.1 语料介绍

测试语料是农业领域的农作物信息文本，是各类农作物(如小麦、水稻等)的不同品种的描述，包括：品种来源、特征特性、栽培的技术要点等几大类信息。相关文本由中科院计算所提供，以及从 Internet 中查找。从全部语料中，挑选水稻文本 39 篇，作为实验语料。抽取模式实例 1518 个。

5.2 模式聚类实验

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%。

由于公共子序列奖惩值 $delta$ 比较高，分类的错分率相对较低。错误主要出自：1. 表达方式的不统一。例如，关于米质的描述通常为“米质优”、“米质优良”、“米质较优”等，但存在个别情况‘米质优于津稻 1187”，导致将模式实例‘津稻 1187”归入同一类别中。2. 相近概念不易区分。例如，作物的抗病虫害能力与病虫害防治是两个不同但又相互联系的，在表达中的一些共同词语导致聚类程序将两者合并，像‘中抗稻瘟病和白叶枯病”与‘注意防治稻瘟病和白叶枯病等病虫害”等。

漏分率相对较高，主要是因为相似度阈值 t 的设置比较大。由于一些模式实例的长度较长，而相互间公共子序列的长度又比较短，导致相似度值低，从而被错误地分开。例如，描述作物品种来源的模式实例，“由天津市农科院作物所选育而成”、“由福建省漳州市农科所选育而成”、“由江苏省泗阳棉花原种场用泗稻 8 号×中丹 3 号选育而成”等。

5.3 模式合并实验

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。

6 结束语

针对一些特定的领域文本，采用项序列来表示信息抽取模式具有一定应用价值，使用聚类方法自动生成这类模式是一种有效的方法，并且在对农作物信息文本的应用中表现出一定成效。但是，也有几点需要改进之处：一、信息抽取模式生成是一个渐进的过程，模式实例聚类算法应进一步完善以适应不断增加的相关文本；二、模式合并过程中，对两个模式实例的合并操作过于简单，需要增加语义属性等信息以提高合并的正确性。

参 考 文 献：

[1] Ellen Riloff. Automatically Constructing a Dictionary for Information Extraction Tasks[C]. In：Proceedings of the Eleventh National Conference on Artificial Intelligence，811－816. AAAI Press/The MIT Press，1993.

[2] Stephen Soderland，David Fisher，Jonathan Aseltine，and Wendy Lehnert. CRYSTAL：Inducing a conceptual dictionary[C]. In：Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence，1314－1319，1995.

[3] Ellen Riloff. Automatically Generating Extraction Patterns from Untagged Text[C]. In：Proceedings of Thirteenth National Conference on Artificial Intelligence（AAAI－96），1044－1049. 1996.

[4] Ellen Riloff，Rosie Jones. Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping[C]. In：Proceedings of the Sixteenth National Conference on Artificial Intelligence（AAAI－99），Orlando FL. 1999.

[5] Roman Yangarber，Ralph Grishman，Pasi Tapanainen and Silja Huttunen. Unsupervised Discovery of Scenario-Level Patterns for Information Extraction[C]. In：Proceedings of Sixth Applied Natural Language Processing Conference（ANLP－2000），282－289，Seattle WA. 2000.

[会议消息]

第 2 届学生计算语言学研讨会（SWCL　2004）

“学生计算语言学研讨会”是由中国中文信息学会发起的系列学术会议，其目的旨在培养青年计算语言学科技工作者，其特色是全部活动完全由学生自己组织。继‘第 1 届学生计算语言学研讨会（SWCL2002）’于 2002 年 8 月在北京大学计算语言学研究所成功召开之后，“第 2 届学生计算语言学研讨会（SWCL2004）”将由北京语言大学信息科学学院承办，拟于 2004 年暑假期间在北京语言大学举行。

研讨会将围绕语言信息处理技术与语言知识库两大主题，安排专家讲座和专题培训，组织论文报告，讨论和学术参观。对计算语言学以及相关学科的在读博士生，硕士生，大学生，免收会议费。

面向新世纪的生力军，面向未来的社会发展，学生计算语言学研讨会将成为计算语言学和相关专业学生学习和交流的生动课堂。

主要议题：（但不局限于此）语料库语言学与语料库建设；句法分析与语义研究；文本分析与生成；语义 Wed 及 Ontology；信息检索与信息提取；机器翻译与机器翻译评价；计算词典学与机器词典的构建；术语学研究与术语标准化；语言模型在 OCR、语音识别与合成上的应用；机器学习方法；面向计算的语言学研究；应用语言学。

论文：中文或英文。论文必须有中文和英文的题目与摘要；建议报告所用的 PowerPoint 或投影片尽量使用英文；论文提交形式要求电子版，使用 pdf 格式或 rtf 格式；会议将评选优秀论文并给予奖励，同时向相关的核心学术刊物推荐正式发表。

重要日期：论文提交截止日期： 2004 年 5 月 31 日　论文录用通知日期： 2004 年 6 月 30 日

来搞请发至：edxun@blcu.edu.cn

信息抽取模式自动生成方法的研究

作者： 郑家恒， 王兴义， 李飞
作者单位： 山西大学,计算机科学系,山西太原,030006
刊名： 中文信息学报 ISTIC PKU
英文刊名： JOURNAL OF CHINESE INFORMATION PROCESSING
年，卷(期)： 2004,18(1)
被引用次数： 15次

参考文献(5条)

1. Ellen Riloff Automatically Constructing a Dictionary for Information Extraction Tasks 1993

2. Stephen Soderland;David Fisher;Jonathan Aseltine;Wendy Lehnert CRYSTAL:Inducing a conceptual dictionary 1995

3. Ellen Riloff Automatically Generating Extraction Patterns from Untagged Text 1996

4. Ellen Riloff Rosie Jones. Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping 1999

5. Roman Yangarber;Ralph Grishman;Pasi Tapanainen;Silja Huttunen Unsupervised Discovery of Scenario-Level Patterns for Information Extraction 2000

本文读者也读过(5条)

1. 于江德. 王立新. 樊孝忠. YU Jiang-de. WANG Li-xin. FAN Xiao-zhong 基于自扩展的信息抽取模式自动获取[期刊论文]-小型微型计算机系统2009,30(5)

2. 钟秀琴. 符红光. 丁盘苹. ZHONG Xiu-Qin. FU Hong-Guang. DING Pan-Ping 基于本体与Prolog的平面几何定理证明[期刊论文]-电子科技大学学报2011,40(3)

3. 史旗凯. 郭菊娥. SHI Qi-kai. GUO Ju-e 基于SMA信息抽取的主题诊断研究[期刊论文]-管理工程学报2010,24(1)

4. 王波. 姚敏 基于信息抽取的匿名用户兴趣描述[期刊论文]-华南理工大学学报（自然科学版）2004,32(z1)

5. 梁海华. 朱淼良. LIANG Hai-hua. ZHU Miao-liang 一种用于多Agent系统的领域工程方法[期刊论文]-计算机工程2008,34(11)

引证文献(16条)

1. 许威. ZHAO Ke. 亿珍珍 一个确定汉语句子主干的递归模型[期刊论文]-航空计算技术 2008(4)

2. 奚斌. 钱龙华. 周国栋. 朱巧明. 钱培德 语言学组合特征在语义关系抽取中的应用[期刊论文]-中文信息学报 2008(3)

3. 吕国英. 冯艳. 李茹 基于CFN的教材内容提要信息抽取研究[期刊论文]-山西大学学报(自然科学版)2010(1)

4. 郑家恒. 菅小艳 农作物信息抽取系统的设计与实现[期刊论文]-计算机工程 2006(7)

5. 贾美英. 杨炳儒. 郑德权. 曹鸿强. 杨靖. 张练 基于模式匹配的军事演习情报信息抽取[期刊论文]-现代图书情报技术 2009(9)

6. 屈赟. 杨捧. 张文静 基于信息粒度的主题相似性信息检索[期刊论文]-河北农业大学学报 2011(1)

7. 郭俊荣. 杨捧. 王紫薇 一种基于信息粒度的信息检索优化方法[期刊论文]-计算机仿真 2010(8)

8. 袁毓林 用动词的论元结构跟事件模板相匹配--一种由动词驱动的信息抽取方法[期刊论文]-中文信息学报 2005(5)

9. 屈赟. 杨捧. 张文静 基于信息粒度的主题相似性信息检索[期刊论文]-河北农业大学学报 2011(1)

10. 曹冬林. 廖祥文. 许洪波. 白硕 基于网页格式信息量的博客文章和评论抽取模型[期刊论文]-软件学报 2009(5)

11. 陈西选 基于机器学习的中医病案信息抽取系统的研究[学位论文]硕士 2006

12. 曹冬林. 廖祥文. 许洪波. 白硕 基于网页格式信息量的博客文章和评论抽取模型[期刊论文]-软件学报 2009(5)

13. 徐超 基于种子自扩展的命名实体关系抽取方法的研究[学位论文]硕士 2006

14. 叶娜 面向信息抽取的文本预处理和规则自动学习技术研究[学位论文]硕士 2004

15. 李跃进 基于Internet的信息抽取技术研究[学位论文]硕士 2005

16. 王照亮 基于XML的数据抽取的研究与应用[学位论文]硕士 2007