第28卷 第5期 计 算 机 工 程 2002年5月
Vol.28 № 5 Computer Engineering May 2002
·人工智能与识别技术· 文章编号：1000—3428(2002)05—00137—02 文献标识码：A 中图分类号：TP 391.4

基于语义的模糊匹配在模糊汉字辨认中的应用

周拥峰，张 彪，夏宽理
（复旦大学计算机系，上海 200433）

摘 要：模糊汉字的辨认在文本处理中是一个有待解决的问题，提出了一种基于语义的模糊匹配算法，该方法把语义的理解和模糊模式匹配相结合，从而来解决模糊字的辨认问题。

关键词：模式识别；模糊匹配；匹配度

Blurry Match Based on Semantic for the Recognition of Blurry Chinese Words

ZHOU Yongfeng，ZHANG Biao，XIA Kuanli
(Department of Computer Science,Fudan University,Shanghai 200433)

【Abstract】 The recognition of blurry Chinese words should be settled in the processing of the Chinese words. A blurry match algorithm based on semantic strategy is presented in this paper. This method combined the recognition of the Chinese semantic with blurry pattern matching to solve the recognition of blurry Chinese words.

【Key words】 Pattern recognition；Blurry match；Matching degree

现实生活中经常会碰到辨认不清的文字，尤其是在考察古代文献的时候，有些字由于年代久远，已经变得模糊不清了。这时辨认起来就有些困难，因此，就产生了模糊字的自动辨认这个问题。本文提出了基于语义的模糊匹配算法，可以在模糊字的辨认中发挥一定的作用。此算法的基本思路很简单：特定时期人们的用语中词语的搭配有一定的规律，每个字可以在什么字前面或者后面是有一定的范围的，根据这个性质，可以缩小该模糊字可能代表的字的范围，然后利用匹配算法，就可以确定文献中的模糊字应该是什么字。

1 模式识别

模式识别，也是用计算机模拟人的智能行为，就是要研究如何用计算机分析各种模式，并对未知模式给出分类和结构描述。

模式识别问题是已知事物的各种类别，然后来判断给定的对象是属于哪一个类别的问题，"模式"是指标准的模板。实际生活中，有些事物的类别（即模式）是明确、清晰和肯定的，但也有很多事物的模式带有不同程度的模糊性，对这些具有模糊性的模式借助于模糊理论来刻画。具有"模糊模式"的模式识别问题，可以用"模糊模式识别"方法来处理[1]。

设U为给定的待识别对象的全体的集合，U中的每一对象u有p个特性指标u_1，…，u_p。每个特性指标刻画了对象u的某个特性，由p个特性指标确定的对象u可记成特性向量$u=(u_1,\cdots,u_p)$。设识别对象集合U可分成n个类别，每个类别均为U上的一个模糊集。记作A_1，…，A_n，则称它们为模糊模式。模糊模式识别就是把对象$u=(u_1,\cdots,u_p)$划归到与其最相似的一个类别A_I中（1<=I<=n）去。当一个识别算法作用于对象u时，就将产生一组隶属度$pA_1(u)$，…，$pA_n(u)$，它们分别表示对象u隶属于类别A_1，…，A_n的程度。建立了模糊模式的隶属函数组之后，就可以按照某种隶属原则对对象u进行判断，指出它应归属哪个类别。

本文所提到的基于语义的模糊匹配算法在匹配时候就是以模糊匹配为理论基础来设计的，其目的就是得到各文字和所辨认模糊字的匹配度。

2 基于语义的模糊匹配算法

2.1 字的连接字属性向量

一般说来，在一定的时期，人们的用词有一定的规律。某一个特定的字前面或者后面跟的字基本上有一定的范围。我们把特定字前面出现的字叫该字的前连接字，后面出现的字叫该字的后连接字。

根据特定词库资料来统计特定字w的各前连接字频率，从大到小选取m个字，w_i(i: 1…m)作为其前连接字列表，根据其出现的统计频率设定其频率值$p(w_i,w)$(i: 1…m)。同理可以得到w的后连接字列表q_i(i: 1...m)，频率值$p(w,q_i)$(i: 1…m)。由此，可以得到一个关于w的2×m维向量f(w):

$$f(w) = (p(w_1,w),...p(w_i,w),...p(w_m,w)\ p(w,q_1),...p(w,q_i),...p(w,q_m)).$$

正规化得到字的前、后连接字属性向量

$$f0(w) = |\ p(w1,\ w)/\sum_{i=1}^{m} p(wi,\ w)\ ...p(w,\ q1)/\sum_{i=1}^{m} p(w,\ qi)\ ...\ |$$

2.2 有限向量的相似度

单个元素相似度：元素x,y的相似度，可以根据其接近程度给出一定的数值。越是接近，相似度越高。当两个元素相同时相似度为1，完全不同时候为0，其余在0和1之间。

等长向量的相似度：向量$P = (p_1,\cdots,p_m)$,向量$Q = (q_1,\cdots,p_m)$；P跟Q的相似度等于向量所有相对应元素的相似度之和，假设各单个元素相似度为$f(p_i,q_j)$,向量P、Q的相似度为f(P,Q)，即得

作者简介：周拥峰(1977～)，男，硕士生，研究方向：软件与理论；张 彪，硕士生；夏宽理，教授

收稿日期：2001-07-060

$$f(P,Q)=\sum_{i=1}^{m} f(pi,qi)$$

不等长向量间的相似度：设有向量P=（p_1，…，p_m），Q =（q_1，…，p_n),其中m>=n。把Q跟P去匹配，Q向量的任何位置都可以插入空格,如插入一个空格得到Q' =(q_1',..., q_i',...q'_{n+1}), q_i是空格，当j<i时候q_j'是q_j,当j>i时候是q_{i-1}。要求插入的空格数目加上n等于m。然后计算向量P和Q的相似度，把所有可能的插入方法都计算相似度，其中最大的那个相似度数值就是这两个不等长向量的相似度。

2.3 字的匹配度

对于汉字P, P字按照汉语的笔画顺序把每一笔作为一个元素形成一个笔画向量，如‘三’的笔画向量为（‘一’,‘一’,‘一’）。所以‘三’可以根据规则生成向量P=(p_1, p_2,p_3)。 p_1=‘一’，p_2 =‘一’，p_3 =‘一’。

对于两个字P，Q,按照上面的规则可以生成两个向量P =(p_1,..., p_m)，Q =(q_1,..., q_n)。向量P，Q的相似度记为f(P,Q)，那么字P、Q的匹配度就是f(P,Q)。

汉字笔画的匹配应该根据汉字的特殊性质，设定各笔画之间的相似度。譬如：‘一’和‘|’可以认为相似度为0；‘一’和‘一’可以认为相似度为0.5。

通过设定，形成一个笔画相似度矩阵（如图1），然后就可以进行汉字笔画向量的匹配。

$$SR=\begin{array}{c} X_1\ X_2 \cdots\cdots X_n \\ \left|\begin{array}{l} R_{11}\ R_{12} \cdots\cdots R_{1n} \\ R_{21} \cdots\cdots \\ \vdots \\ R_{n1} \cdots\cdots R_{nn} \end{array}\right| \end{array} \begin{array}{c} X_1 \\ X_2 \\ \vdots \\ X_n \end{array}$$

图1 汉字笔画相似度矩阵

该矩阵中，X_i代表汉字的各种笔画（包括空格），R_{ij}代表笔画X_i和X_j的相似度。有以下特性：

(1)相似度的值在0和1之间，1表示两个笔画完全相似，0则表示两者完全不相似。(2)每个笔画同自己的相似度为1。(3)不同的笔画之间的相似度是小于等于1，大于等于0的数。(4)相似度对称，即R_{ij}等于R_{ji}。

2.4基于语义的模糊匹配

任意3个顺序字x，y，z，假设y模糊不清，只能看到一部分。要判断y到底是什么字。x的后连接字列表向量P=(p_1, …, p_m)，z的前连接字列表向量Q =(q_1,…, q_m)，得到x的后连接字属性向量f0(P)；z的前连接字属性向量 f0(Q)。假设里面不重复的文字一共k个，计算所有这k个字的符合度f(w_1, ..., w_k) =(f(w_1),…, f(w_k))。其中f(w_i)的数值为w_i在 f0(P)，f0(Q) 中的相对应的数值之和，如果w_i在P中不出现，以0来代替其在P中数值，同理应用于Q。然后在这k个文字中找出符合度最高的d 个文字作为候选的文字。模糊辨认字y，按照初步辨认写出该字的可见的笔画向量，当然，此向量肯定会比原来的真实的字的笔画向量缺少几个元素或则错误几个笔画。然后把这个笔画向量跟前面获得的d个候选文字进行字的匹配，得到各字的匹配度。最后根据前面得到的符合度和匹配度来权衡得到该字到底是哪个字。

2.5 该算法的可行性

汉字的数量是一定的，而且语言文字的应用也有一定的规律，所以可以统计整理生成一个汉字的前连接字和后连接字的信息库。可以选择频率比较高的一定数量的文字，譬如可以选择500个。组成一个汉字的笔画数量是有限的，甚至可以说是一个很小的数字，所以笔画向量的长度也不大，这就不会产生无穷计算问题。各种笔画的数量是一定的，这就使得对各个笔画之间的相似度的规定完全可以实现，当然这也需要语言学家的帮助。这些事实保证了算法的有效实行。

3 小结

当前，在自然语言处理中，模糊字的辨认是一个比较困难的事情，因此迫切需要一种高效率的自动的辨认方法。本文提出了一种基于语义的模糊匹配算法，能够很好地解决这个问题，而且具有实际应用的可能。

参考文献

1 王述亭,刘家瑾模糊数学在煤层层位识别中的应用[J].测井技术, 1985(2): 66-71

2 Voutilainen A. Asyntax Basedpart of Speech Analyzer.Dublin：In: Proceedings of 7th EACL,1995

3 边肇祺. 模式识别. 北京：清华大学出版社，1988

★★

（上接第134页）

时间连续投票，Cookie和数据库记录联合控制一个IP地址在某个时段以及整个投票项目过程中的选票数量。这些都在相当程度上解决了"一人多票"问题，但对于"有组织投票"还很难找到有效的控制方式，因此对于"可信度"要求很高的调查项目，应该尽量降低匿名用户选票的权值，鼓励用户以注册账户投票，甚至可以只接受一级用户选票，以满足调查任务的"可信度"要求。

4 结束语

文中针对当前投票系统存在的"信用"问题进行了分析，提出了一套比较有效的"可信度"解决方案，实现了一个可满足不同可信度要求的受限调查平台—"校务公开"通用调查平台。该平台比较真实地反映了用户意愿，为校务决策提供了有效的参照依据。同时该平台不仅提高了"向上"的可信度，也很好地解决了"向下"的可信度问题，平台只记录哪些用户对哪些调查项目已经发表了意见，对用户与其选票之间的对应关系不作记录，而且投票结果采取直接累加的方式进行记录，用户与选票之间不可能找到任何对应关系。

平台针对受限调查任务设计，但通过部门的延伸，可将其设计思想拓展到社会性调查问题。同时，平台对于Web环境下其它类型的可信度等问题的解决也有一定的借鉴意义。

参考文献

1 周 渡. 网上新闻调查初探. 镇江师专学报(社会科学版), 2000(1): 98-100

2 梁新明,李怀民. 计算机网络应用于企业统计调查的若干问题探析. 山西统计,2000(7):18-20

3 叶 青,刘向民.在线调查系统的建立与应用. 现代计算机,1998(4)

4 Hall M, 宋 文,钟向群译.Web 编程指南北京清华大学出版社, 1999

5 BuczekLU G,李 博,于 骞译. ASP 应用开发指南. 北京：科学出版社, 2000

6 Howard M,Levy M, Waymire R. Designing Secure Web-based Applications for Microsoft Windows 2000.Washington:Microsoft Press,2000