文章编号：2095-2163(2021)07-0202-04 中图分类号：TP391.1 文献标志码：A

# 基于统计自然语言分析的九宫格输入法最优键盘布局

周 迪[1]，周晓聪[2]，候 婷[2]
（1 河海大学 物联网工程学院，江苏 常州 213022；2 河海大学 理学院，南京 211100）

**摘 要**：九宫格输入法是手机端常用的中文输入法之一，可将 26 个英文字母按顺序布局在 8 个数字键上，每个键上有 3~4 个字母。然而顺序布局显然不是最优的。本文采用统计自然语言处理计算键盘布局的平均击键次数，并采用模拟退火算法优化，在搜寻键盘数据时，对其实行哈希计算，避免重复搜索，最终找到九宫格输入法的最优键盘布局方案。结果显示，本文的最优键盘布局方案比顺序布局的输入效率明显提升，可以提高生活的便利程度和工作效率。

**关键词**：九宫格输入法；最优键盘布局；击键次数；统计自然语言处理；模拟退火

## Optimal keyboard layout of nine-grid input method based on statistical natural language analysis

ZHOU Di[1], ZHOU Xiaocong[2], HOU Ting[2]
(1 College of Internet of Things Engineering, Hohai University, Changzhou Jiangsu 213022, China;
2 College of Science, Hohai University, Nanjing 211100, China)

**【Abstract】** Nine-grid input method is one of the Chinese input methods commonly used in mobile phone terminal. It lays out the 26 English letters in order on eight numeric keys, each with 3~4 letters. However, sequential layout is clearly not optimal. In this paper, statistical natural language processing is used to calculate the average keystroke times of keyboard layout, Simulated Annealing algorithm is used to optimize the keyboard data, and hash calculation is implemented to avoid repeated search, so as to find the optimal keyboard layout scheme of the nine-grid input method. The results show that the input efficiency of the optimal keyboard layout in this paper is significantly higher than that of the sequential layout, which can improve the convenience of life and work efficiency.

**【Key words】** nine-grid input method; optimal keyboard layout; keystroke times; statistical natural language processing; Simulated Annealing

## 0 引 言

现有的手机中文文本输入法，以拼音输入法为主，键盘主要采用九宫格键盘和 26 键全键盘两种。其中，九宫格输入法通常把 26 个字母顺序放置在 2~8 这 8 个数字键上，每个键上有 3~4 个字母，数字键 0 和 1 作为它用。由于手机屏幕大小的限制，26 键全键盘使用率不如九宫格键盘高。以苹果 IOS 输入法为例，对中文输入不太友好，经过改进后采用了九宫格输入法。

出于历史原因，PC 端键盘的 26 个字母并不是按顺序排列的。26 键全键盘是根据电脑键盘来布局的，九宫格键盘输入法是按字母表顺序排列的。但这种按英文字母顺序布局的键盘是不合理的，并不适合中文输入，没有考虑汉字的频率分布特征，对汉字拼音输入的速度具有一定的限制作用。例如汉语拼音中，字母 s 的使用频率很高，但却跟 p、q、r 共用一个数字键，导致选候选词时击键次数太多。且还要选拼音，当输入 7426 的时候，可能的拼音有 pian、piao、qian、qiao、shan、shao 等。

在手机中文输入法的改进方面，国内外的一些研究者进行了相关的研究。如 Lin 和 Sears 等人以笔画输入法为基础，研究了手机键盘的中文输入效率，研究结论表明：只需重新设计手机按键上的图标，就能提高手机键盘的中文输入效率[1-2]。王晓龙等人[3]发明了数字键盘智能拼音汉字输入方法，自动处理汉字输入过程中的数字键位歧义、拼音组合歧义和同音多字歧义。用户只需输入对应汉字拼音的数字键，系统便根据上下文在整个语句范围内调整相应的汉字，保证汉字语句的正确。在《手机键盘文本输入法研究综述》中，何灿群等人[4]从手机键盘文本输入法的改进研究、中文文本输入法的

作者简介：周 迪(2002-)，女，本科生，主要研究方向：机器学习、集成电路设计、芯片；周晓聪(2001-)，女，本科生，主要研究方向：优化算法、计算科学；候 婷(2000-)，女，本科生，主要研究方向：优化算法、分布式计算与处理。
通讯作者：周晓聪 Email：sailinwei@hhu.edu.cn

收稿日期：2021-03-15

研究、模型预测与评价等多个角度归纳了国内外有关手机键盘文本输入法的研究动态。在此基础上，指出了目前研究存在以下不足：基于西方文字设计的手机键盘不适合中文输入；新的中文输入法在应用上存在诸多不足；大多数手机的键盘改进没有考虑用户的操作特点。此外，又提出了今后的研究发展方向：根据用户操作特征以及中文输入特点来优化现有中文手机键盘的设计，对提高中文文本输入绩效具有很高的应用价值和较强的可操作性。

本文用统计自然语言方法考虑了汉语的词频，将 26 个字母重新布局到数字键上。采用模拟退火优化，找到了最优键盘布局。显著提高了中文输入法效率。

## 1　统计自然语言分析

### 1.1　统计自然语言分析

自然语言[5]，是日常生活中使用的语言类型，包括汉语、日语和英语等。通过计算机技术对自然语言加以处理和运用，整体上可归属于人工智能和语言领域的分支学科。自然语言充当语料库与统计学研究领域的主要方向，自然语言处理技术则旨在完成人类和计算机之间的交互[6]。对于语料库的信息处理和语言学习，可以将以统计学为基础的自然语言处理技术作为重要方式，从而获得信息数据的来源，提取主要语料库信息，得到多种知识。

通过搜集不同的文本对汉语语料库进行统计，为九宫格最优键盘布局研究提供强有力的数据支撑。想要利用统计自然语言分析，设计出最优的九宫格键盘布局方案，就要对语料库进行清洗和统计词频。词频统计[7]是数据与信息处理、知识挖掘与传播中的中心和基础性工作，只有比较准确地在文章中统计出词及其词频，才能进行下一步的工作。

### 1.2　词频统计

利用 Python 编程，对语料库进行清洗，剔除符号并将文本进行分词，统计每个词的频率，再将词频表导出为表格文件。此次搜集的汉语语料库共有 857 276 个词，通过词频统计后，最终可得 42 535 个不同的词，在考虑词频的基础上得到每个词平均字数为 1.582 个。

获取 GB2312 国标码中一级常用汉字和二级不常用汉字的拼音，进而生成每个词的拼音。得到完整的统计文档后，将候选词按照词频降序排列，统计自然语言处理到此完成。以初始键盘为例，通过统计自然语言处理的文本片段见表 1。

**表 1　候选词的文本表**

**Tab. 1　Text table of candidate words**

| 候选词 | 词频 | 拼音 | 击键数字 |
|---|---|---|---|
| 及时 | 2 901 | jishi | 43633 |
| 历史 | 2 290 | lishi | 43633 |
| 即使 | 1 431 | jishi | 43633 |
| 即时 | 222 | jishi | 43633 |
| 忌食 | 220 | jishi | 43633 |
| 既是 | 145 | jishi | 43633 |
| 计时 | 106 | jishi | 43633 |
| 技师 | 83 | jishi | 43633 |
| 集市 | 51 | jishi | 43633 |
| 历时 | 51 | lishi | 43633 |
| 即食 | 47 | jishi | 43633 |
| 离世 | 47 | lishi | 43633 |

### 1.3　平均击键次数计算

对给定的键盘布局，查询每个词中汉字的拼音，再将每个字母转成数字键，得到每个词的击键数字序列。将击键数字序列相同的词作为一组，计算候选词排布。排布方式为每页 4 行，每行不超过 8 个汉字。每个词的击键次数为数字序列长度+页码+1。将所有词的击键次数与词频相乘再求和，就是平均击键次数。

## 2　模拟退火和哈希映射

### 2.1　模拟退火算法

模拟退火算法（Simulated Annealing，SA）[8] 思想是在 1997 年由 Steinbrunn 等人首次提出。这是一种基于 Monte－Carlo 迭代求解策略的随机寻优算法，在局部最优解的情况下能概率性地跳出，并最终趋于全局最优。

模拟退火算法（SA）包含 2 个部分，即：Metropolis 算法和退火过程。其中，Metropolis 算法就是如何在局部最优解的情况下让其跳出来，是退火的基础。1953 年，Metropolis 提出重要性采样方法，即以概率来接受新状态，而不是使用完全确定的规则，称为 Metropolis 准则，计算量较低。

假设前一个状态为 $X(n)$，状态变为 $X(n+1)$ 时，同时系统的能量（平均击键次数）由 $Y(n)$ 变为 $Y(n+1)$，定义系统由 $Y(n)$ 变为 $Y(n+1)$ 的接受概率 $P$ 为：

$$P=\begin{cases}1, & Y(n+1)<Y(n)\\ e^{-\frac{Y(n+1)-Y(n)}{T}}, & Y(n+1)\geqslant Y(n)\end{cases} \tag{1}$$

从式(1)可以看到,如果平均击键次数减小,那么新解状态就被接受(概率为 1),如果平均击键次数增大,就说明系统偏离全局最优值位置更远,此时算法不会立刻将其抛弃,而是进行概率操作:首先在区间[0,1]产生一个均匀分布的随机数 $t$, 如果 $t < P$, 则新解状态接受,否则拒绝新解状态。虽然这种算法对于较差的键盘布局存在一定的接受概率,但是爬坡能力较强,不会轻易陷入局部最优,可以从局部最优的情况下跳出来。当退火模拟曲线震荡幅度逐渐减小,趋于平稳时,跳出循环,得到最优布局。

### 2.2 模拟退火算法的设计

(1)初始化。给定初始温度 $T$(充分大),产生初始键盘布局(初始解状态 $n$),同时计算当前键盘布局平均击键次数为 $Y(n)$。

(2)判断迭代次数是否达到要求:是,转(7);否则转(3)。

(3)产生新解 $n'$。随机选择 2 个数字键,从中各选一个字母,交换(如图 1 右所示)。或者将其中一个字母移动到另一个数字键上(如图 1 左所示),保证移动后每个键上的字母数在 2~5 之间。

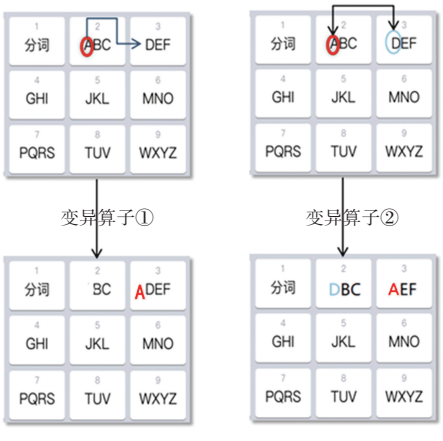


**图 1 键盘布局变换**

**Fig. 1 Transpose of keyboard layout**

(4)模拟退火算法计算是否接受。对于新布局方案 $n'$,计算其平均击键次数 $Y(n')$,利用公式(1)判断接受、还是拒绝该键盘布局。对于一个给定的键盘布局方案 $n$,计算每个词的拼音输入和候选词选择的击键次数。再根据词频加权计算所有词的平均击键次数,得到 $Y(n)$ 函数,每次产生新解 $n'$ 后,通过 $\Delta T = Y(n') - Y(n)$,计算 $\Delta T$ 的大小。

(5) 温度 $T$ 逐渐减少。

(6)转(2)。

(7)退出程序,打印最优键盘布局。

至此,研究中给出了算法程序流程如图 2 所示。

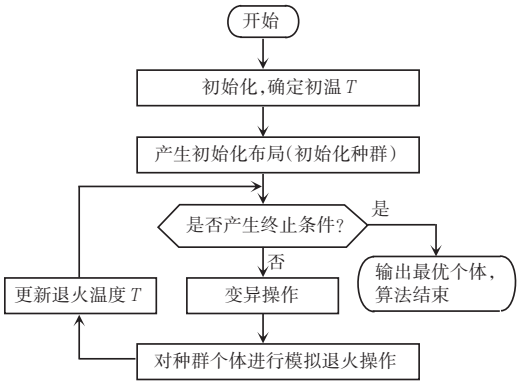


**图 2 模拟退火算法流程图**

**Fig. 2 Flow chart of Simulated Annealing algorithm**

### 2.3 哈希映射

在模拟退火过程中,为避免搜索键盘布局重复,采用哈希映射(Hash Map)的方法进行判重。键盘布局与键内字母顺序无关,也与数字顺序无关。因此,先将每个数字键上的字符串排序,再将 8 个字符串排序后拼接成一个长度为 26 的字符串,最后求该字符串的哈希值。

定义一个集合容器来存储搜索过的键盘布局哈希值。对于一个新的键盘布局,先计算其哈希值,然后在集合中查找是否已有。如果已有,则继续产生一个新的键盘布局;否则将该哈希值放入集合,并用模拟退火算法判断是否接受该键盘布局。

## 3 布局比较及结果分析

利用模拟退火算法,迭代 100 万次,得到最优键盘布局,平均击键次数下降曲线如图 3 所示。

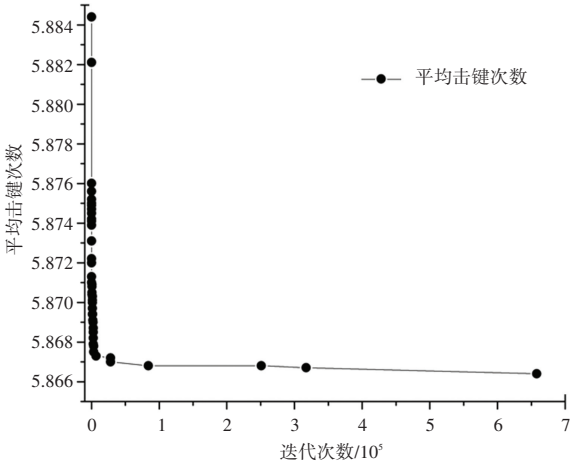


**图 3 模拟退火曲线图**

**Fig. 3 Simulated Annealing result**

键盘布局1是目前使用的顺序布局,在搜集的

(下转第 211 页)