基于结构描述的汉字字形相似度计算

林民[1,3]，宋柔[1,2]

（1. 北京工业大学计算机学院，北京 100022；2. 北京语言大学信息科学学院，北京 100083；3. 内蒙古师范大学计算机与信息工程学院，呼和浩特 010022）

摘　要： 汉语是一种大字符集语言，汉字数量繁多、结构复杂。汉字字形的相似度计算是汉语信息处理的一项基础研究，对于计算机辅助的汉语文章校对和汉字教学有重要作用。本文从图形相似角度改进了汉字结构的描述方法，给出了一种基于结构描述的汉字字形相似度计算方法，并完成了 GB2312 的 6763 个汉字的相似字表。这一相似字表用于计算机辅助校对系统中的修改提示，初步显示出效用。

关键词： 汉字字形，结构描述，相似度

The Similarity Calculation of Chinese Character Glyph Based on Structure Description

Lin Min[1,3], Song Rou[2]

(1.College of Computer Science & Technology, Beijing University of Technology ,Beijing 100022; 2.College of Information Sciences, Beijing Language and Culture University, Beijing 100083;3.College of Computer & Information Engineering, Inner Mongolia Normal University, Hohhot 010022)

Abstract: Chinese is a large character set language. There are many Chinese characters with complex structure. The similarity calculation of Chinese character glyph is a fundamental research in Chinese information processing. It is also important in computer aided proof reading of Chinese articles and Chinese characters teaching. This paper improved the description of Chinese character structure from the perspective of graphic, provided a method to calculate the glyph similarity of Chinese characters based on their structure descriptions, and completed the list of similar characters in the GB2312 standard which includes 6763 Chinese characters. The application of the similar character list in the computer aided proof reading system improved the modification guide for users.

Keywords: Chinese character glyph, structure description, similarity

1　引言

汉语是一种大字符集语言，常用汉字大约 3 千多个，国家标准 GB2312 中有 6763 个汉字，国际标准的大字符集 ISO10646 中的中 CJK 汉字有 20902 个，加上扩充 A、扩充 B 已达到 6 万多个。古籍整理使用的汉字有 10 万左右。从汉字字形上看，汉字的笔画和部件的种类都比较多，汉字的基本部件就有 560 个[2]。而且笔画的组合方式和部件的组合方式也都很多，使得汉字的结构十分复杂。

对于数量如此巨大、结构复杂的汉字字符集，采用类似拼音文字的处理方法，把汉字整体看作一个记号，仅仅列出每个汉字的字形是不够的。人们容易认错、写错、录入错，计算机自动识别系统也会识别错，因此就需要

基金资助：本项研究得到国家自然科学基金项目（60272055，60572159）的资助。

作者简介：林民（1969-），男，山东日照，副教授，在读博士，linmin@blcu.edu.cn

对字形进行形式化描述，进而进行字形相似度的计算，并对字符集中的每个汉字，把它的相似字都找出来，按照相似度排序。这样，对因人认错或 OCR 识别软件识别错而造成的录入错误，就可以估计字形错误的可能性，从而给出较准确的修改建议。此外，教中小学生和外国人学习汉字，也需要根据字形的相似度进行字形辨析的教学。

关于汉字字形的研究，过去的工作主要集中在语言学界，研究汉字字形的演化。从应用的角度进行研究，有上海交通大学出版的汉字信息字典[1]，详细描述了 7785 个汉字的结构和部件。在此基础上，国家于 1997 年颁发了汉字部件标准[2]，穷尽式地列出了国家标准通用多八位编码字符集（UCS）中 20902 个字的部件表，并按部件对这些汉字进行了逐个拆分。还有一些学者把汉字描述成部件为操作数、部件间位置关系为运算符的数学表达式形式[4]，或定义了一种语言采用笔画、部件分层次对汉字进行描述[3]，这些工作为汉字字形相似度计算提供了基础，但是，从字形相似度计算的应用需求看有下面一些不足。[1]中给出的结构描述和[2]中的汉字部件拆分标准主要是从汉字字形的演变过程的角度（即字理的角度）出发的，这种角度同软件或人识别的角度（即认字的角度）并不完全相同；[3]中使用平面坐标来描述笔画、部件在汉字中的位置，过于精细、刻板，掺入了一些非本质的位置关系特征，使字形结构的同一性很难判定；[4]中划分的结构类型比较粗糙，包容结构只有“左下、左上、右上、全包”四种类型，把“区”和“凶”这样结构不同的字作为同一类型对待，计算机进行相似性计算时容易误判。这些字形描述研究都未给出字形相似度计算的方法。

此外，汉字自动识别的方法是利用大量手写或印刷汉字的图形实例，提取汉字图形的特征，给出了汉字间的相似度，作为识别的依据[5]。但是，这种相似度往往对不同人群、不同印刷特征的字形普遍适应性不够，是一种依赖于图形实例的汉字相似度计算方法。

事实上，软件或人在认字的时候，把汉字看成平面图形，只要图形相似就有可能看错，同字理并无必然联系。我们根据人的图形相似性感知改进了[1]的汉字结构描述体系，对 GB2312 中的每一个汉字给出了结构描述；设计了一种不依赖于图形实例的汉字字形相似度计算算法；对 GB2312 字符集内的 6763 个汉字，给出每一个字在集内的字形相似字，并对相似字按相似度排序；将这样的相似字表用于汉语文章校对系统的别字修改提示，取得了实效。本文介绍了这几个方面的工作。

2 汉字字形的结构描述

按照[1]的体系，汉字分为独体字和复合字。复合字都按二分法切分，包括左右结构、上下结构、包容结构、被包容结构和嵌套结构 5 大类。每一类结构中的每一个成分又可以看作独体部件或这 5 种结构之一的复合字。如此递归切分，直至二分法的每个成分都是不可分的独体字（部件）。下面分类举例说明。

（1）独体字。包括不能拆分的字（如“一”、“了”、“儿”、“口”、“中”、“牛”、“车”等）和许多非字的偏旁部首部件（如“匚”、“刂”、“冂”、“亻”、“灬”等）。

（2）左右结构复合字。如“代”，左部和右部分别是“亻”和“弋”，表示作（亻 | 弋），其中“|”是左右结构的表示符。对于左中右结构，一般是先在“左”和“中右”之间切分，例如“淅”，表示为（氵 | （木 | 斤））；有些“左中”是一个完整字的，则先在“左中”和“右”之间切分，如“彬”表示为（（木 | 木） | 彡）。

（3）上下结构复合字。如“早”，表示作（日 / 十），其中“/”是上下结构的表示符。同样，上下多层结构的复合字，如“享”，[1]按“上”和“中下”方式切分成（亠 / （口 / 子））；“案”则按字理从“安”和“木”切分为（（宀 / 女） / 木）。

（4）包容结构复合字。如“超”，表示作（走 > 召），或再分解“召”，得到（走 >（刀 / 口））。

（5）被包容结构复合字。如“道”，表示作（首 < 辶），或再分解“首”，得到（（𠂒 / 口）< 辶）。包容结构与被包容结构的区别是：包容结构先书写“包”的部件，后书写“被包”的部件，而被包容结构的书写顺序相反。

（6）嵌套结构复合字。如“团”是“口”中嵌入“才”，表示作（口 @ 才）。“裹”是“衣”中嵌入“果”，表示作（衣 @ 果）。

这样的结构分类方法从总体上看确实反映了汉字的平面图形结构，但也有涉及字源、书写顺序的非图形因素，对于图形相似性分析起了干扰的作用。为此，我们对这一体系作了修改，使它只考虑汉字图形的几何结构，排除非图形因素，从而更彻底地支持汉字字形的相似性计算。我们的修改有以下几方面：

（1）合并包容结构和被包容结构，统称包容结构。即不考虑部件书写的先后顺序，只看是否一个部件从某几个方位包容另外一个（或几个）部件。如“超”和“迢”都是包容结构，分别表示作（走 > 召）和（辶 > 召）。容易看出它们是字形相似的。

（2）合并后的包容结构按照“包”的方位分成7种类型：

结构名称	表示符	例字
左上包	lu> (left-upper)	者、压、居、病、友、存
左下包	ld> (left-down)	翘、勉、题、魅、飓、赶、这、延
右上包	ru> (right-upper)	句、可、氧、岛、虱、武
右下包	rd> (right-down)	头
上左右包（上三包）	ulr> (upper-left-right)	问、向、同、风、咸、肉
左上下包（左三包）	lud> (left-upper-down)	区、匣、巨
下左右包（下三包）	dlr> (down-left-right)	凶、画、函、幽、幽

如表中“者”为左上包结构，表示作“(夕 lu> 日)”；而“延”为左下包结构，不考虑部件书写顺序，表示作“(廴 ld> 　)”；“凶”为下三包结构，不考虑书写顺序，表示作“(凵 dlr> 乂)”。在[1]中“延”和“凶”属于同一结构类型，容易被计算机误认为字形相似。现在它们分属不同的结构类型，一般来说不再会被误判相似。

（3）修改了一些字的结构类型和结构分解顺序，使其更符合人的感知。如[1]中把“夺”字称为包容结构，包和被包成分分别是“大”和“寸”。但从汉字图形角度看，“夺”与上下结构的“奇、奈”等字很相似，因此我们把“夺”分解为“上下结构”，表示为“(大 / 寸)”，类似的字还有“分佘冬尽参卷泰春昼基”。又如“幽”，[1]中认为“山”是外框，左右结构的两个“幺”是内嵌成分，也称为嵌套结构，表示作“(山 @ (幺 | 幺))。事实上，“幽”与全封闭包围的其它字外形差别都较大。我们根据“幽”与“函”外形更相似而把“幽”归为“下三包”包容结构，表示作“(山 dlr> (幺 | 幺))”。类似的字还有“幽”。又如[1]把“案”分解成((宀 / 女) / 木)，我们则按“上”和“中下”方式切分为“(宀 / (女 / 木))”，使“案”与“察”的相似度高。再如“我”在[1]中是独体字，但我们描述成左右结构，左边类似于牛字旁“牜”，右边是“戈”，这样处理就使它同“找”很相似，符合感知事实。

3 汉字字形相似度计算的算法

汉字字形相似度计算需要考虑以下因素：

（1）独体字（部件）的相似性

独体字（部件）个数不多，有些差别较大，有些十分相似，很容易被认错。我们按外形相似程度分级并划组。第一级是基本相同的，如“日”、“曰”，“七”、“匕”；第二级是虽然不同但十分相似，如“干”、“于”、“千”，“人”、“入”、“八”，“讠”、“氵”，“礻”、“衤”；第三级是基本相似的，如“开”、“升”、“井”，“东”、“车”。其余的字是互不相似的。

（2）复合字结构的相似性

一般来说，复合字同独体字是不会相似的，首层结构不同的复合字也不会相似。两个首层结构相同的复合字有了相似的基础，最后的相似度则还要看两对对应成分的相似度。越到结构的深层，其相似程度对整体相似程度的影响越小。比如，“椅”同“猗、倚、掎、犄”的相似度很高，因为它们都是左右结构，右边完全相同，左边十分相似。“椅”同“琦”的相似程度就略差一些，因为它们的左边木字旁同斜王旁差别较大。“椅”同“掩”相似程度更差，因为它们的左部不同，右部都是上下结构支持了相似性，但右下部“可”和“电”差别很大。因此，我们设计了复合字按结构层次进行相似性比较的递归算法。

以上情况也存在例外，例如“真”、“其”、“具”都是上下结构的复合字，但它们与独体字“甚”字形基本相似，按外形相似程度应划属第二级相似字。再如“血”、“四”都是复合字，并且“血”为上下结构，“四”是嵌套结构，虽然结构不同，但从感知角度来看是字形基本相似的，也应划属第三级相似字，对这些特殊复合字我们按独体字的处理方法把它们和相似的独体字同列在相似字表中。

再如“勺”、“句”、“勾”、“司”既是第三级相似字，又都是右上包（首层结构相同）的复合字，类似的还如“左”、“右”、“石”，既是第二级相似字，也都是左上包复合字，最后相似度应取按相似等级和结构层次比较得到的两种计算结果中较大的值。

（3）笔画对相似性的影响

不在相似字表中的两个独体字或一个独体字一个复合字，以及不在相似字表中且结构不相同的复合字，虽然完全不相似，但其不相似的程度也需要定量地给出，因为它们可能是另外两个相似字的成分。一般说来，可以按照两个字笔画数的差异来衡量它们的不相似程度。比如，“椅”同“朴”的右部就很不相似，不仅因为它们一个是复合字一个是独体字，而且因为这两个成分的笔画数差很多。

根据以上三方面的考虑，设计的字形相似度计算算法如下：

汉字 A 和 B 的相似度记为 sim(A,B,lev)，该值越大两字越相似。

其中参数 lev 用来表示当前比较所在的结构层次，结构层次越深 lev 越小。

设 A 和 B 的笔画数分别为 a 和 b，mix(a,b)表示 A 和 B 笔画数的均值；i 表示 A、B 两字的相似等级，i=1，2，3；A1、A2 表示 A 按首层结构分解得到的两个成分，B1、B2 表示 B 按首层结构分解得到的两个成分。

1．如果 A 和 B 是 i 级相似字，则

1）如果 A 和 B 都是复合字并且结构类型相同，则

```
sim(A,B,lev)=max(mix(a,b)+(4-i)*lev,
                 lev+sim(A1,B1,lev/2)+sim(A2,B2,lev/2));
```

2）否则，sim(A,B,lev)=mix(a,b)+(4-i)*lev;

2．否则，即 A 和 B 是 i 级相似以外的字，

1）如果 A 和 B 都是复合字并且结构类型相同，则

```
sim(A,B,lev)=lev+sim(A1,B1,lev/2)+sim(A2,B2,lev/2);
```

2）否则，sim(A,B,lev)=-|a-b|*lev。

4 实验和结果

我们目前研究的是 GB2312 中的 6763 个汉字。首先确定了独体字和复合字的集合，并分出了以独体字为主体的一级相似字 18 组、二级相似字 120 组、三级相似字 90 组，对每个复合字进行了结构拆分。在此基础上，按算法对每个汉字算出它同其余所有字的相似度，并按相似度大小对这些字排序，舍掉排在 100 以后的字。

相似度算法中使用了参数 lev，lev 值不同会影响相似汉字的排序，因此要对 lev 取值进行优化。根据经验，lev 的取值范围限制为 32、16、8、4，从中选优。具体做法是：随机取 10 个汉字，用不同的 lev 分别求出各字的前 100 个相似字，然后把根据人的直觉手工调整顺序得到的字序列作为评判标准，计算不同 lev 下自动得到的各字相似字序列与评判标准序列的接近程度，这里，两字符序列接程近度 SIM 的定义如下：

设 $X=x_1,x_2,...x_n$ 为任一由 n 个字符组成的序列，定义集合 $X_i=\{x_k \mid 1\le k\le i\}$，$1\le i\le n$，则对任意两个长为 n 的字符序列 A 和 B，接近度为 $SIM(A,B)=\sum_{i=1}^{n}\frac{|A_i\cap B_i|}{i}\times w_i$，

其中权重 w_i 体现序列中排在越靠前的字符对整个序列的接近度影响越大，并且应满足 $\sum_{i=1}^{n} w_i=1$，计算公式为 $w_i=n-i+1\Big/\sum_{k=1}^{n}k=2(n-i+1)/n(n+1)$。按以上公式计算得到的结果如下表所示：

参数	各字的相似字序列与评判标准序列的接近度（SIM 值）										
lev	稿	梁	陈	蔼	闷	恋	货	椅	钻	察	平均值
4	0.89	0.96	0.99	0.94	0.97	0.94	0.98	0.91	0.99	0.90	0.946
8	0.96	0.98	0.94	0.96	0.97	0.94	1.00	0.98	1.00	0.97	0.969
16	0.91	0.96	0.93	0.92	0.97	0.94	1.00	0.97	1.00	0.94	0.955
32	0.86	0.96	0.93	0.90	0.97	0.94	1.00	0.97	1.00	0.89	0.942

使相似字序列最接近标准的 lev=8。此时得到的部分相似字序列例举如下：

蔼 霭蕴落藩蒎莎薄藻蒲菘蒎菏菠菹萍蓖葆萨蒋茫茫荡莅簿茨荐蘅藐葙蒌薪莉箔荷菠篌

闷 闰闱闲闼闾闵闶闫闭问闯闸闹阅闪闺闻阔闽闾阀阁阂闩阑阄阆阈阊阅阌阁阁阙阐

恋 挛栾娈孪弯鸾变脔蛮峦恁恐恕恙恕恣恶恩息恳恶銮怎怒思怠急怨怼恿悉悠患您恙悬

货 贷货赁贤贽资贸贺华赘赞责贪贯员贡贫贲贯贵费柴背贞负贾页赛贺肾览垒冀恁柴毖

椅 犄掎倚猗琦畸绮骑崎捺掩榛椿楱栲棒橙桥栓桧稽枪嵇松檎棕棱桷梅梢梭梯检楔棺楂

钻 钻铅铝铬铭锆镕惦拈沾阽钙错锾钪铎铛铞钵铃铮铯销锁锃锅铧镗锉锐锑锌镂镇枯桔

察 蔡寮祭禀寄箐寥茶寡褰赛蹇寰宽富宣寓寨攀甯蜜赛塞寅赛宰害宴宵容宽宾密寒案官

我们使用这样的相似字表对 OCR 软件识别出的 1 万多字文章进行别字修改提示的测试。把相似字序列用作候选字符集，再利用语言模型进行筛选，前十个修改提示的平均准确率在 80%以上。如果不采用相似字表，仅使用语言模型，就要计算几千个转移概率连乘积，再找出最大的几个作为修改提示。现在只需在几十个相似字中选择，计算量大大减少，可见这一方法是有用的。

5 结束语

汉字字形的相似度计算，对于汉字教学、错别字校对等有重要意义。本文在[1]中汉字结构描述方法的基础上，从汉字图形相似性角度出发，对[1]中方法进行了 4 方面的改进：

1） 把[1]中“包容”和“被包容”结构的复合字合并为一类称为“包容结构”；

2） 把[1]中“包容”和“被包容”结构又细分成 7 种类型；

3） 对[1]中没有描述结构的独体字和一部分结构类型不同但容易混淆、认错或写错的复合字，按外形相似程度划分出一、二、三级相似字若干组；

4） 给出了一种不依赖汉字图形实例的计算字形相似度的算法。

并在此基础上，对 GB2312 的每个汉字给出了相似字表，试验结果初步显示出了实用效果。

但是由于汉字字形结构的复杂性，[1]中给出的汉字结构表示法中还存在一些问题本文没有解决。如一些独体字的相似性比较，上下结构的层次划分深度，嵌套结构的分类，多层左右结构或多层上下结构的相似度计算等。

如“帝”字[1]中二分为“(/ 巾)”，上偏旁“ ”作为一个整体不再细分，但对字形相似的“帚、带”字，[1]切分为“((⺕ / 冖) / 巾)” 和“((/ 冖) / 巾)”，上偏旁都进一步细分为两部分。从保持细分层次一致性的角度，“ ”也应该细分，但应该二分为“(⺊ / (丷 / 冖))”还是“((⺊ / 丷) / 冖)”则很难确定。又如[1]中“乘”是独体字，“乘”则是嵌套结构（禾 @ 北），结构完全不同，如何表示其相似性还需进一步研究。

综上所述，汉字字形结构规律非常复杂，找出一种比较完善的汉字字形描述方法是一项很困难的工作，我们今后需要进一步完善汉字的结构表示法，进一步改进相似度的计算方法。

参考文献：

[1] 上海交通大学汉字编码组. 汉字信息字典. 北京：科学出版社，1988

[2] 国家语言文字工作委员会. GF3001-1997 信息处理用 GB13000.1 字符集汉字部件规范. 北京：语文出版社，1997.12.1 发布，1998.5.1 实施

[3] Richard Cook. A Specification for CDL(Character Description Language):an extract of [PhD Dissertation]. UC Berkeley,Dept.of Linguistics,2003

[4] 孙星明，殷建平，陈火旺等.汉字的数学表达式研究[J].计算机研究与发展, 2002,39(6): 707-711

[5] Hsi-Jian Lee,Hung-Chi Hsu. A hierarchical model-guided generation of Chinese characters. Proc of the 12th International conf on Pattern Recognition, 256-260, Jerusalem,Israel,1994.10