Advances in Applied Mathematics 应用数学进展, 2016, 5(2), 159-165
Published Online May 2016 in Hans. http://www.hanspub.org/journal/aam
http://dx.doi.org/10.12677/aam.2016.52021

Hans汉斯

Algorithm Design of Restoring Two-Way Single/Double-Sized Shredded Documents

Chen Zhang, Shiyun Wang

Science Department, Shenyang Aerospace University, Shenyang Liaoning
Email: wsy0902@163.com

Received: Apr. 14th, 2016; accepted: May 2nd, 2016; published: May 5th, 2016

Abstract

This paper designs an algorithm to restore English shredded documents no matter they are single-sized or double-sized text files which are cut both vertically and horizontally. Firstly, we cluster the fragments which were located in the same line in original text files according to the structural features of English letters and the row spacing. Then, using l_1 norm difference model, we attach the fragments in the same class. By this way, the scraps of paper in the same line can be restored as a whole crosscutting shredded document. Finally, we should splice the crosscutting shredded documents into a complete image. In the numerical test part, taking the 2013 national mathematics model contest problem as examples, our algorithm restores 209 pieces of English shredded documents. Numerical results show that the correct rate of clustering is over 93% which demonstrates the efficiency of the algorithm.

Keywords

Peak Weight, Row Spacing Weight, Clustering Credibility, Jffreys & Matusita Distance, l_1 Norm

双向切割单/双面英文碎纸片拼接复原算法设计

张　晨，王诗云

沈阳航空航天大学理学院，辽宁 沈阳
Email: wsy0902@163.com

收稿日期：2016年4月14日；录用日期：2016年5月2日；发布日期：2016年5月5日

文章引用：张晨，王诗云．双向切割单/双面英文碎纸片拼接复原算法设计[J]．应用数学进展, 2016, 5(2): 159-165.
http://dx.doi.org/10.12677/aam.2016.52021

摘　要

针对单/双面英文文本文件，经过双向(横向 + 纵向)切割后形成的碎纸片，本文通过设计拼接算法将其还原。首先，利用“英文字母的结构特征”和“空白行间距”这两个几何特征将原图中同行的碎纸片按行聚类。在此基础上，我们利用向量的l_1范数差异度模型对每类碎片进行列拼接，以形成一个横切碎片，最后再对所有的横切碎片进行行拼接即可。在算法的数值检验部分，我们以2013年全国大学生数学建模赛题为例，对横纵切后形成的209块单/双面英文碎纸片进行拼接复原。数值复原结果证实了该算法实现简单，且聚类成功率高，其中聚类部分的正确率可以达到93%以上。

关键词

峰值权数，行间距权数，聚类可信度，Jffreys & Matusita距离，l_1范数

1. 引言

破碎文件的拼接在司法物证复原、历史文献修复以及军事情报获取等领域都有着重要的应用。传统上，拼接复原工作需由人工完成，虽然准确率较高，但效率很低。特别是在碎片数量巨大的情况下，人工拼接很难在短时间内完成任务。

如今，人们则期望结合计算机强大的计算能力与合理的数学模型，获得精确高效的碎纸片自动拼接复原技术。目前，在拼接形状规则的矩形碎纸片时，最为常规的拼接复原算法大多基于碎纸片的行列距离，以及碎纸片边缘的几何特征。通过特定的算法对所有碎片按行进行聚类，然后再根据碎片边缘黑色像素的灰度特征，建立碎片之间的相似度，最后根据相似度的大小筛选出邻接碎片。但是无论筛选的条件有多精确，还是会出现匹配错误，这时就会要求引入人工干预。

为了提高拼接复原的准确度，国内外众多学者从不同的角度出发，进行了诸多的尝试，也得到了很多优秀的拼接复原算法。2009 年，维也纳科技大学的 Mattias Prandtstetter 以及 Gunther R. Raidl 提出了基于变邻域查询(VNS)和蚁群优化算法(ACO) [1]，这两个算法可以分别提高拼接准确度和拼接复原时间。2012 年，弗吉尼亚理工大学计算机科学系与发现分析中心的 Patrick Butler 等人则采用一种可视化的分析方法来复原碎纸片文件[2]。2014 年，中国政法大学的鲁嘉琪发表了基于文字信息的碎纸片拼接复原算法，该算法根据中文文字的字体稳定性与英文字母的形状固定性，分别采用聚类分析或模式识别算法来进行拼接复原[3]。

碎纸片的按行聚类作为整个拼接复原过程的重要组成部分，同样也吸引众多国内外学者对其进行研究。2013 年，尹玉萍等人提出了基于碎纸片特征向量的动态行聚类算法进行行初步聚类，再根据文字特征及行距对初步聚类进行调整修改，才确定最终的行分类[4]。Azzam Sleit 提出了另一种聚类思想，他认为聚类应该作为整个拼接复原过程其的一部分，而不是预处理步骤，这样的改变可以使得结果更加精确[5]。

本文的研究重点也聚焦在聚类过程上，本文在聚类过程中引入了“行间距权数”，“模板峰值权数”，以及“模板更新”的概念，以实现碎纸片的更好聚类。在本文中，我们假设所有的碎纸片无破损，且只有黑白双色(灰度值介于 0~255 之间)，切割整齐均匀。

2. 单面英文双向切割的碎纸片的拼接复原算法

本节针对双向切割的单面英文碎纸片的情况，设计拼接复原算法。设单/双面文本文件被横向平均分割成 M 份，纵向平均切割成 N 份，则碎纸片的总个数为 MN 份。

2.1. 预处理

这些碎纸片在 Matlab 软件下，皆有唯一对应的矩阵 $A_i\,(i=1,\cdots,MN)$。为了使得矩阵的计算存储快捷，我们将数值矩阵转换成相应的稀疏矩阵，即令 $A_i=255-A_i$。此时，矩阵中的元素“0”代表“白色”，“255”代表“黑色”。本文中的矩阵 A_i 皆作该处理。

Principle 1：任意两张列相邻的碎纸片，其相邻的两个向量中，黑色像素灰度值以及分布位置相近。例如，给定碎片 A_i，其最右边的列向量为 a_i；其余纸片为 $A_k\,(k\neq i)$，其最左边的列向量为 a_k。记 $k_0=\arg\min\limits_{k\neq i}\|a_i-a_k\|$，则模板碎片与 A_{k_0} 左右相邻的可能性最大。

Principle 2：处于同一行的碎纸片，具有相似的空白行间距特征，构成英文字母的黑色像素具有分布特征。

Principle 3：任意两张行相邻的横切碎纸片，也具有两种相似特征。具体的，若给定模板碎片 B_i 的最后一行向量为零向量(图像中的表现为空白行)，且该碎片底端的零向量共有 j 行，而与之相邻的碎纸片的顶端前 j' 行也为零向量，根据空白行间距 d 固定的特征，$j+j'=d$；另一方面，若 B_i 的最后一行 b_i 不为零向量，其余纸片为 $B_k\,(k\neq i)$ 及其最上边的行向量为 b_k，若 $k_0=\arg\min\limits_{k\neq i}\|b_i-b_k\|$，则指定的横切碎片与待拼接的横切碎片 B_{k_0} 上下相邻的可能性最大。

2.2. 同行碎纸片的聚类

本节，我们将 MN 块纸片，先按行聚类，即聚成 M 类，每类包含 N 块纸片。通过观察，我们发现同行的英文碎纸片具有两种几何特征。首先构成文字的黑色像素的规律分布。同行英文字母的黑色像素，在该行的某些固定行高位置处出现频率较高。其次是基于空白的行间距，同行碎纸片的行间距所占的高度范围是基本相同的。结合这两种几何特征，我们来设计英文碎纸片的聚类算法。

2.2.1. 聚类原则

首先对碎纸片矩阵 A_i 作横向求和，得到的向量称为横向求和向量 s，横向求和向量中的第 k 个元素 $s(k)$ 表示矩阵 A_i 的第 k 行的元素之和。我们发现如果两个碎纸片位于同一行，那么碎纸片对应的两个横向求和向量中峰值出现的位置重合程度就越高(见图 1 与图 2)，图 1 与图 2 所对应的碎纸片来自于 2013 年全国大学生数学建模 B 题提供的附件 4 的 081.bmp 和 077.bmp。

本文将同行碎纸片中位置重合频率较高处的峰值称为必然峰值，而每行英文字母基本都有五个必然峰值。本文选取图 1 与图 2 所对应的两张碎纸片图像，并在图中标出每行字母中五个必然峰值所在位置(见图 3)。图 3 也证实了同行碎片中必然峰值的重合性。本文将横向求和向量 s 中剩余的峰值称为“偶然峰值”，它们在同行碎纸片中各横向求和向量 s 中的出现具有偶然性。必然峰值在其形状特征上，表现为该峰值很大且很凸出；而偶然峰值的形状特征则是很小且很平缓。为了区别对待两种峰值情况，我们引入“峰值权数”的概念。这里，我们所说的“峰值”，取横向求和向量元素中的“极值”之意，本文假设“峰值”元素不小于其前三个和后三个元素。

定义 1：设某模板碎片有 n 个峰值，F 表示峰值权数向量，记 $F=\big(F(1),\cdots,F(n)\big)$。其中 $F(k)$ 表示第 k 个峰值的峰值权数：

$$F(k)=\begin{cases}3*s(D_k)-s(D_k+1)-s(D_k+2)-s(D_k+3), & k=1\\ 6*s(D_k)-\sum\limits_{\substack{t=-3\\ t\neq 0}}^{t=3}s(D_k+t), & 1<k<n\\ 3*s(D_k)-s(D_k-1)-s(D_k-2)-s(D_k-3), & k=n\end{cases}\tag{1}$$

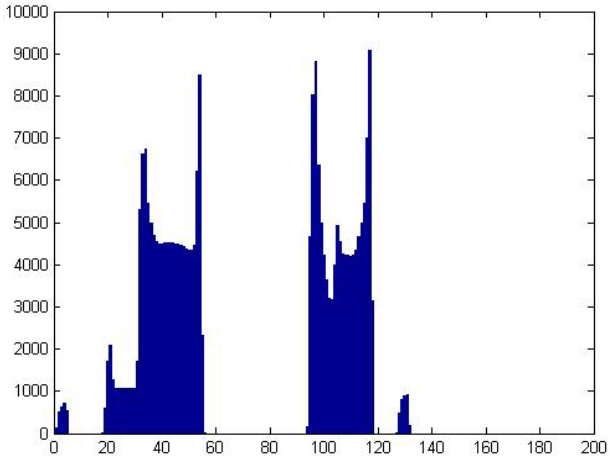

Figure 1. The horizontal sum of figure 081.bmp
图 1. 图片 081.bmp 的横向求和图像

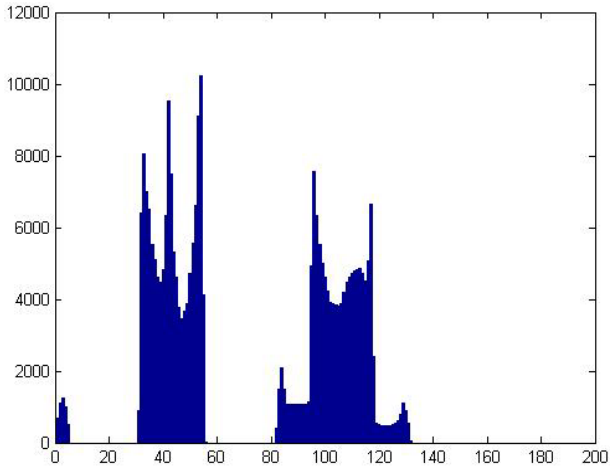

Figure 2. The horizontal sum of figure 077.bmp
图 2. 图片 077.bmp 的横向求和图像

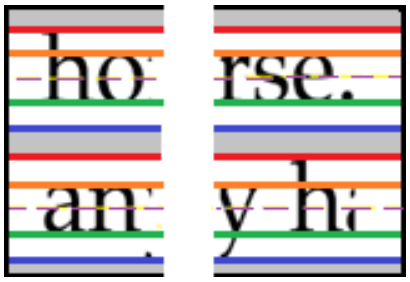

Figure 3. The positions of five inevitable peaks (081.bmp, 077.bmp)
图 3. 五个必然峰值出现的位置(081.bmp, 077.bmp)

其中，s 表示模板碎片的横向求和向量，而第 k 个峰值在横向求和向量 s 中的位置用 D_k 表示，即 $s(D_k)$ 表示第 k 个峰值的大小，同时称 $D=(D_1,\cdots,D_n)$ 为峰值位置向量。

由于 $s(D_k)$ 为峰值，由(1)式，$F(k)\geq 0$，因此，该定义有意义。下面，我们以模板碎纸片的峰值权数为参照，计算其余待聚类碎片与模板碎纸片的峰值位置重合处的峰值权数之和。

定义 2：如果待聚类的碎片 j 与模板碎片 i 在某些相同位置处皆出现了峰值，则重合的位置所形成的重合位置集合记为 D^{ij}，称 $H_j=\sum_{k\in D^{ij}}F(k)$ 为待聚类碎片 j 的重合峰值权数之和。

由定义 2 可知，衡量待聚类碎片 j 与模板碎片 i 位于同一类的可能性的大小，就是衡量基于峰值重合位置集合 D^{ij} 所对应的峰值权数之和的大小，若 $i_0=\arg\max_{j\neq i}H_j$，则表示待聚类碎片 i_0 与模板碎片处于同一行可能性最大。

此外，处于同一行的碎纸片，还具有相同的空白行间距特征。为了利用这一几何特征，我们引入了行间距权数 λ。其定义如下。

定义 3：令 h 为模板碎片中空白间距的像素行之和，待聚类碎片 j 本身在模板碎片 i 的空白间距范围内的空白像素行之和为 h_j，则待聚类碎片 j 的行间距权数定义为

$$\lambda_j=\frac{h_j}{h} \tag{2}$$

从定义 3 可以看出，$\lambda_j\in[0,1]$，且 λ_j 的值越大，则待聚类碎片 j 的空白行特征与模板碎片的空白行特征的相似性越好。

定义 4：待聚类碎片 j 与模板碎片 i 位于同一类的可信度大小定义为 $K_j=\lambda_jH_j$。

显然，综合了重合峰值权数之和 H_j 和行间距权数 λ_j 所得到的聚类可信度 K_j 越大，则该碎片与模板碎片是同一类的可能性就越高。由于 H_j 和 λ_j 的数量级不同，差别很大，因此，我们以 H_j 和 λ_j 相乘的方式来可信度。

2.2.2. 模板碎片的更新

在聚类过程中，搜索拼接一个新碎片，往往要依赖于一个指定的模板碎片。如果模板碎片选取不当(横向求和向量中信息量较少)，筛选出的结果就会差强人意，此外还会引起连锁反应，导致后续筛选的碎片都出错。反过来想，如果一个模板中包含的碎片个数越多，则其所体现的峰值信息与行间距信息也就越全面，聚类过程中产生误差的可能性也越小。因此，我们所采取的改进方法是，指定模板碎片 A_i 后，根据聚类可信度 K_j 筛选出碎片 A_{i_0}，并且还要将 A_{i_0} 纳入到模板碎片中，即将$[A_i,A_{i_0}]$作为新模板碎片，这样就可以使得下一次聚类时，新模板的峰值权数和行间距权数更加精确，再次聚类筛选所得的结果也更加准确。我们可将聚类算法归纳如下。

聚类算法：

Step 0. 指定该聚类过程的初始模板碎片 A_0。

Step 1. 计算出模板碎片的峰值权数向量 F 和峰值位置向量 D，以及行间距宽度 h 和范围。

Step 2. 用待聚类碎片与模板碎片进行匹配，求解待聚类碎片 j 的 H_j 和 λ_j。

Step 3. 求解待聚类碎片 j 的聚类可信度 K_j。

Step 4. 筛选出剩余碎片中聚类可信度最大的碎片 i_0，并将其纳入模板碎片中，即

$$A_0=\left[A_0,A_{i_0}\right]。$$

Step 5. 若模板碎片中碎片的数量 $k\geq 19$，则聚类结束；否则跳转到 **step 2**，再次聚类。

2.3. 类碎片序列的列拼接

当 $M*N$ 个碎纸片都被聚类之后，就形成了 M 个类碎片序列，每个序列中有 N 个碎纸片。接下来，本

文会将每个类碎片序列进行列拼接复原，拼接复原完成后的每类的碎片序列就可以作为一个完整的横切碎片。

若两碎纸片左右是相邻的，则前一个碎纸片右边一列的像素与后一个碎纸片左边一列的像素向量，在相邻行高位置处的黑色像素灰度值是渐进变化的。这个几何特点在数值上表现为，切缝边缘的两个列向量在相同高度上的非零数值元素(即黑色像素)的大小变化是“渐进的”。为此本文引入边缘差异度 c，其定义如下：

$$c_k = \|y_k - y_i\|_1 \ (k \neq i)$$

其中 $\|\cdot\|_1$ 为向量的 l_1 范数，y_i 为指定碎片，y_k 是待聚类碎片，而 c_k 则表示待拼接碎片 k 与指定碎片 i 的边缘差异度。而待聚类碎片中边缘相似度最大的碎片，即边缘差异度 c_j 最小的碎片。所以指定碎片 y_i 的邻接碎片 i_0，定义如下：

$$i_0 = \arg\min_{k \neq i} \|y_k - y_i\|_1$$

通过该 1-范数最小差异模型就可以得到指定碎片 i 的邻接碎片 i_0。为了放大误差的效果，本文还对 1-范数最小差异模型进行了优化，即引入 Jffreys & Matusita 距离[6]，此时向量 x, y 的边缘差异度 C 定义如下：

$$C(x,y) = \left[\sum_{i=1}^{n}\left(\sqrt{x_i} - \sqrt{y_i}\right)^2\right]^{1/2}$$

类碎片序列的列拼接算法归纳如下。

类碎片序列的列拼接算法：

Step 0 导入同类的 N 个碎纸片所对应的数值矩阵 $A_i\ (i = 1, 2, \cdots, N)$。

Step 1 寻找原图中该行碎片最左边的碎纸片，且令其为 A_{k_0}；

Step 2 利用 1-范数最小差异模型找到 A_{k_0} 的邻接碎片 A_{i_0}；

Step 3 若拼接完的碎片个数小于 N，则令 $k_0 = i_0$，转至 **Step 2**；否则转至 **Step 4**。

Step 4 将稀疏矩阵进行还原，输出复原图像。

2.4. 横切碎片的行拼接

同行的类碎片序列按序拼接复原后就形成了一个完整的横切碎片 L_i。所以在整个拼接复原流程中，对横切碎片的拼接就成为了最后一步。横切碎片的拼接复原基于两个准则，其一就是 1-范数最小相似模型，其二则是行间距固定的几何特征。如果指定横切碎片 L_i 的底端有黑色像素，则可以使用优化后的 1-范数最小相似模型搜索拼接邻接碎片。如果横切碎片 L_i 的底端为空白行，则利用该碎片底端空白行距 d_1 与相邻碎片顶端空白行距 d_2 之和与行间距 d 相等的几何特征，即 $|d_1 + d_2 - d| = 0$ 来完成横切碎片的拼接复原。该算法归纳如下。

列拼接算法：

Step 0 导入同类的 N 个碎纸片所对应的数值矩阵 $A_i\ (i = 1, 2, \cdots, N)$。

Step 1 寻找原图中该行碎片最左边的碎纸片，且令其为 A_{k_0}；

Step 2 利用 1-范数最小差异模型找到 A_{k_0} 的邻接碎片 A_{i_0}；

Step 3 若拼接完的碎片个数小于 N，则令 $k_0 = i_0$，转至 **Step 2**；否则转至 **Step 4**。

Step 4 将稀疏矩阵进行还原，输出复原图像。

Table 1. The results of row-clustering of single-sized text file
表 1. 单面行聚类序列统计表

T	1	2	3	4	5	6	7	8	9	10	11
初始模板	002	006	005	019	001	015	007	008	003	009	000
p	16	19	17	17	19	18	18	18	18	18	19
q	84%	100%	89%	89%	100%	94%	94%	94%	94%	94%	100%

3. 数值算例

我们以 2013 年全国大学生数学建模竞赛 B 题的附件 4 为例，该附件中 $M=11, N=19$。在表 1 中，我们用 T 代表聚类顺序； p 代表该类碎片中正确的个数； q 代表该类碎片的聚类正确率。根据表 1 的结果可知，单面碎纸片的平均聚类正确率为 94.3%。

4. 小结

以往的算法针对英文碎片进行聚类时，由于英文字母的高度不一致导致算法聚类时正确率较低，从而间接的增加人工干预的次数。本文提出的模板峰值权数向量和行间距权数这两个概念，以及采用了模板更新的方法，大大提高了聚类结果的准确率。

其中模板更新的代价则是牺牲算法的运算时间，因为每聚类完成一次后，以往的数据(如模板峰值权数向量、行间距约束等)都需要重新计算，所以改进后聚类算法的时间复杂度是成倍的增加。本文所提出的 1-范数最小相似模型算法在处理信息量较少的碎片时，容易出现误差。这也是本文产生人工误差较多的地方。因此，在以后的工作中，我们会在本文的模型和算法的基础上加入对字母的模式识别技术，希望可以进一步减少人工干预，提高拼接的准确率。

参考文献 (References)

[1] Prandtstetter, M. and Raidl, G.R. (2009) Meta-Heuristics for Reconstructing cross Cut Shredded Text Documents. Institute of Computer Graphics and Algorithms Vienna University of Technology, GECCO'09, 349-356. http://dx.doi.org/10.1145/1569901.1569950

[2] Butler, P., Chakraborty, P. and Ramakrishan, N. (2012) The Deshredder: A Visual Analytic Approach to Reconstructing Shredded Documents. *IEEE Symposium on Visual Analytics Science and Technology*, Seattle, 14-19 October 2012, 14-19. http://dx.doi.org/10.1109/vast.2012.6400560

[3] 鲁嘉琪. 基于文字信息的碎纸片拼接复原算法[J]. 现代电子技术, 2014, 37(4): 28-31.

[4] 尹玉萍, 刘万军, 张冲, 刘永超. 基于动态聚类的文档碎纸片自动拼接算法[J]. 计算机工程与应用, 2014, 50(18): 162-170.

[5] Sleit, A., Massad, Y. and Musaddaq, M. (2013) An Alternative Clustering Approach for Reconstructing cross Cut Shredded Text Documents. *Telecommunication Systems*, **52**, 1491-1501. http://dx.doi.org/10.1007/s11235-011-9626-x

[6] 张宇, 刘雨东, 计钊. 向量相似度测量方法[J]. 声学技术, 2008, 28(4): 532-535.