Computer Science and Application 计算机科学与应用, 2020, 10(5), 971-977
Published Online May 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.105100

Hans汉斯

Similar Fragment Queries Based on Substitution Errors

Fan Zhang, Yuqi Xie, Chen Rao, Mingchun Wang

College of Information and Intelligent Science and Technology, Hunan Agricultural University, Changsha Hunan
Email: fanfanzizizi@163.com

Received: May 1st, 2020; accepted: May 13th, 2020; published: May 20th, 2020

Abstract

The key to deciphering an unknown language is to look for similar sequences of letter fragments. In this paper, a new algorithm for finding similar fragments is developed. First, the index structure is built and the fragments are divided at intervals. Then the similarity formula and the similarity matrix are established based on the hamming distance to represent the similarity between the two fragments. In combination with practice, the similarity threshold formula is established on the basis of substitution errors in a large number of text records, and the formula is used to judge whether it is the similar fragment to be searched. Finally, the similar fragments of multiple text and their corresponding positions are obtained. In addition, the average accuracy evaluation algorithm is used, and the analysis and experiments show that the algorithm has good accuracy and search efficiency.

Keywords

Similar Pieces, Hamming Distance, Threshold Value, Locating

基于替换错误的相似片段查找

张　帆，谢宇奇，饶　晨，王明春

湖南农业大学信息与智能科学技术学院，湖南　长沙
Email: fanfanzizizi@163.com

收稿日期：2020年5月1日；录用日期：2020年5月13日；发布日期：2020年5月20日

摘　要

破译未知语言的关键是寻找相似的字母片段序列。本文针对相似片段的查找，编写了一种新的算法。首

文章引用: 张帆, 谢宇奇, 饶晨, 王明春. 基于替换错误的相似片段查找[J]. 计算机科学与应用, 2020, 10(5): 971-977.
DOI: 10.12677/csa.2020.105100

先建立索引结构，多次间隔划分得到片段。然后基于海明距离建立相似公式和相似矩阵用于表示两个片段之间的相似度。结合实际，在大量文本记录时发生替换错误的基础下建立相似阈值公式，并通过该公式判断是否为要求查找的相似片段。最后获得了多段文本的相似片段以及其对应的位置。此外使用平均准确率评价算法，经分析和实验表明，该算法有较高的准确率和查找效率。

关键词

相似片段，海明距离，阈值，查找定位

1. 引言

破译某种未知语言的关键是寻找一段相似的字母序列片段，因为这些片段很可能具备某种固定含义，类似词汇或词根。对照查找相似片段序列的算法，通常采用的方法是先引入特定的索引结构，如后缀树、后继数组等[1] [2]。基于这些索引结构的算法不仅计算量大，而且要求索引结构模式的首字母必须相同，效率有待提高且局限性较大。因此本文提出一种新的索引结构，优先考虑计算的复杂度，计算海明距离并根据阈值划分，在此基础上编写一种新的算法。该算法适用于查找首字母不同的未知语言字母序列的相似片段，经分析和实验表明，该算法可以得到较好的查找结果且有较优的查找效率。

2. 数据来源和模型假设

下载英文版《双城记》，将里面的字母装换成大写，删去标点符号以及字母 U 到 Z，造出研究未知语言的基础文本数据。为了简化，作以下假设：1) 未知语言由 A 到 T 这 20 个大写字母组成；2) 文本在获取过程中，有一些位置发生了替换错误，即某个字母被篡改成了其他字母；3) 文本在获取过程中只将替换错误纳入考虑范围，不考虑丢失了某个字母或增加了原本不存在的字母；4) 文本中各种字母出现的可能性符合一般规律。

3. 相似片段的获得以及位置的记录

3.1. 建立索引结构

对于某段长度为 m 的文本，用 $Z_i = z_1, z_2, \cdots, z_d \ (1 \le i \le n)$ 的字符串表示其中的相似片段，Z_i 称为一个模式。根据模式长度 d 进行 d 次划分，每次划分从文本的第 i 个位置间隔 d 取字符，组成的字符串即为 Z_i。按照这样的规则划分得到片段，构造出元胞划分矩阵和划分行向量。其中元胞划分矩阵 T 的维度为 $d \times (m-i+1)/d$，T 的每一个元素为一个长度是 d 的模式片段；划分行向量 H 的维度为 $1 \times n$，T 的非空元素按行依次存放到 H 中。

本文获得了 30 段长度在 5000~8000 个字母范围的文本，片段的模式长度为 15。元胞划分矩阵如表 1 所示。

3.2. 基于海明距离建立相似矩阵

建立相似公式和相似矩阵：

Table 1. The cellular partition matrix of a text
表 1. 一段文本的元胞划分矩阵

	第 1 列	第 2 列	第 3 列	……	第 495 列	第 496 列
第 1 行	BOOKTHEFIRSTREC	ALLEDTOLIFECHAP	TERITHEPERIODEI	……	ALITTLEAHEADINT	OTHEMISTANDDARK
第 2 行	OOKTHEFIRSTRECA	LLEDTOLIFECHAPT	ERITHEPERIODEIT	……	LITTLEAHEADINTO	THEMISTANDDARKN
第 3 行	OKTHEFIRSTRECAL	LEDTOLIFECHAPTE	RITHEPERIODEITA	……	ITTLEAHEADINTOT	HEMISTANDDARKNE
第 4 行	KTHEFIRSTRECALL	EDTOLIFECHAPTER	ITHEPERIODEITAS	……	TTLEAHEADINTOTH	EMISTANDDARKNES
第 5 行	THEFIRSTRECALLE	DTOLIFECHAPTERI	THEPERIODEITAST	……	TLEAHEADINTOTHE	MISTANDDARKNESS
第 6 行	HEFIRSTRECALLED	TOLIFECHAPTERIT	HEPERIODEITASTH	……	LEAHEADINTOTHEM	[]
……	……	……	……	……	……	……
第 15 行	CALLEDTOLIFECHA	PTERITHEPERIODE	ITASTHEBESTOFTI	……	TOTHEMISTANDDAR	[]

海明距离是两个等长字母片段之间对应位置的不同字符个数。对于长度 d 的两个片段 x 和 y，它们的海明距离为 l，那么相似公式[3]为

$$\gamma_{xy} = \frac{d-l}{d} \tag{1}$$

相似公式可以求出任意两个片段的相似度，将结果用相似矩阵表示。任意一个片段与它本身的相似度为 1，在这里的目的是找到达到相似阈值的片段，则相似矩阵 R 是主对角线为 0 的对称矩阵。相似矩阵构造过程示意图如图 1 所示。

$$R = \begin{pmatrix} 0 & \cdots & \gamma_{1n} \\ \vdots & \ddots & \vdots \\ \gamma_{1n} & \cdots & 0 \end{pmatrix} \tag{2}$$

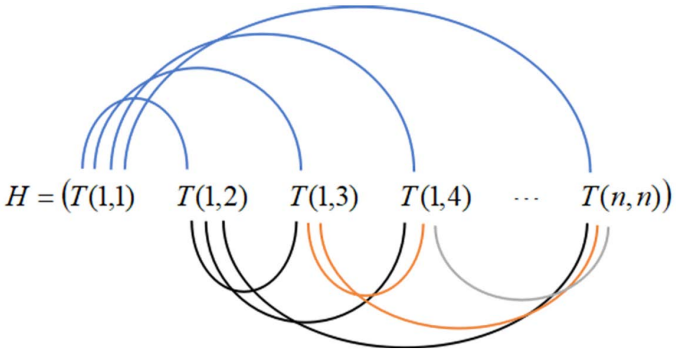

Figure 1. Schematic diagram of similar matrix construction process
图 1. 相似矩阵构造过程示意图

本文获得了 30 段文本的相似矩阵，如表 2 所示。

Table 2. A matrix of similarity for a piece of text
表 2. 一段文本的相似矩阵

	第 1 列	第 2 列	……	第 n − 1 列	第 n 列
第 1 行	0	0.8667	……	0	0
第 2 行	0.8667	0	……	0	0
……	……	……	……	……	……
第 n − 1 行	0	0	……	0	0.8000
第 n 行	0	0	……	0.8000	0

3.3. 基于替换错误建立相似阈值公式

对于任意一个片段，最多出现的替换错误的字母个数为 e，那么相似阈值公式为

$$\gamma_{阈} = \frac{d-e}{d} \tag{3}$$

基于相似阈值对相似矩阵进行逻辑判断

$$\begin{cases} \gamma_{xy} \geq \gamma_{阈}, 赋值为1 \\ \gamma_{xy} < \gamma_{阈}, 赋值为0 \end{cases} \tag{4}$$

从而得到对应的 0-1 矩阵，按列相加该矩阵中的元素得到一个行向量，该行向量的第 x 个元素表示在最多只会出现 e 个字母的替换错误的情况下，与片段 x 相似的片段个数。由此可以得到一段文本中的相似片段以及在划分行向量对应的位置。算法设计流程图如图 2 所示。

本文获得了最多出现 4 个字母的替换错误的 30 段文本的相似片段及其对应位置，如表 3 所示。

Table 3. Similar segments and their locations
表 3. 相似片段以及对应位置

第 1 段文本		……	第 14 段文本	
相似片段	对应 H 的位置		相似片段	对应 H 的位置
OFTHEPASSENGERS	4230		ETHEPASSENGERBO	527
EDTHEPASSENGERS	445		……	……
LETHEPASSENGERS	5254	……	DTHEPASSENGERAS	4933
OFITSPASSENGERS	6351		TTHEPASSENGERSI	6638
相似片段总个数：4			相似片段总个数：9	
第 15 段文本		……	第 30 段文本	
相似片段	对应 H 的位置		相似片段	对应 H 的位置
ETHEPASSENGERBO	5625		THEPASSENGERASM	161
……	……		……	……
STHEPASSENGEROP	5313	……	SOFTHEPASSENGER	4415
FTHEPASSENGERSB	6164		AIDTHEPASSENGER	5684
相似片段总个数：21			相似片段总个数：15	

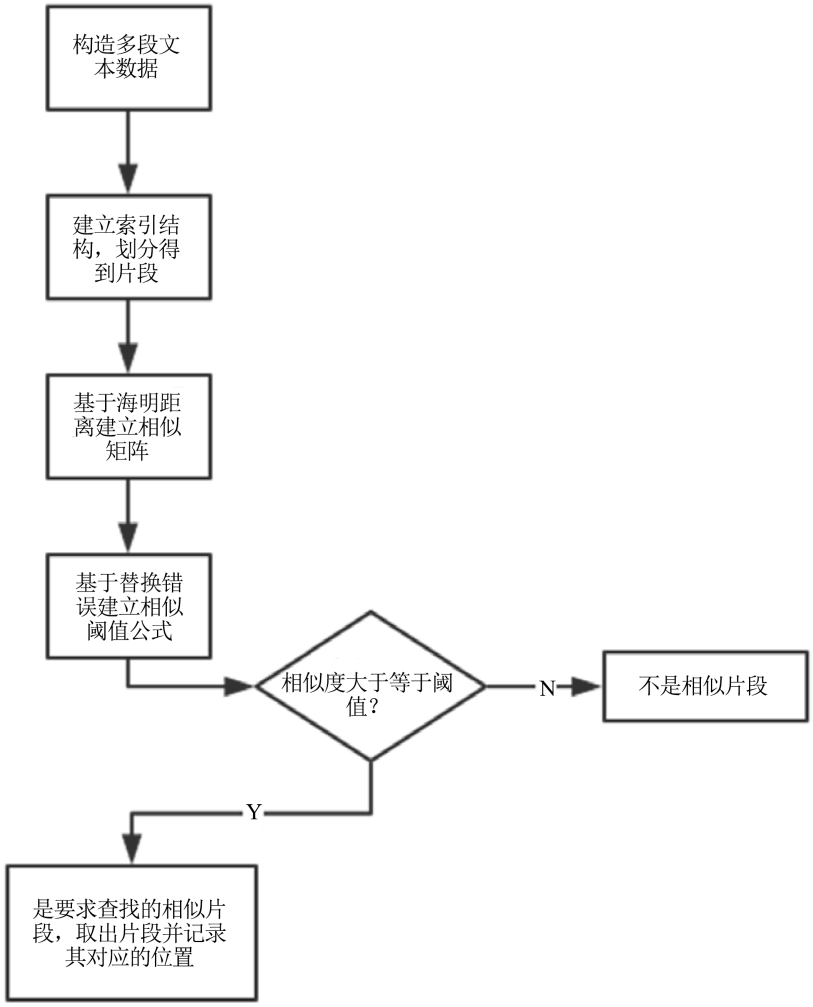

Figure 2. Flow chart of algorithm design
图 2. 算法设计流程图

4. 用平均准确率评价算法

利用 MATLAB 软件随机生成相同模式长度的片段，作为原始片段。对原始片段随机位置进行字母替换，得到的新片段插入一段空文本中，并记录插入的片段与原始片段相似的片段个数。

重复上述操作直到文本长度达到预先给定的范围，用这样的方法生成多段文本并将它们作为测试文本。

选定一个原始片段和一个测试文本，使用本文算法查找测试文本中与原始片段相似的片段个数，与之前记录的数目对比，计算出此次对比的相似片段查找准确率。准确率的计算公式为

$$\rho = \frac{M}{N} \tag{5}$$

重新选择原始片段和测试文本，计算准确率，直到所有文本中的纪录数目对应的准确率全部计算出来。平均准确率的计算公式为

$$\rho_{avg} = \frac{\sum \rho}{P} \tag{6}$$

根据上文，在这里同样取 P 为 30，如表 4 所示获得了 30 段文本的准确率。则求得算法的相似查找片段平均准确率为 0.9757。因此该算法具有较高的准确率，是查找相似片段的有效算法[4]。其中算法分析流程图如图 3 所示。

Table 4. Accuracy of algorithm
表 4. 算法的准确率

文本	准确率	文本	准确率	文本	准确率	文本	准确率	文本	准确率
第 1 段	1.0000	第 7 段	0.9375	第 13 段	0.9787	第 19 段	0.9535	第 25 段	1.0000
第 2 段	0.9250	第 8 段	1.0000	第 14 段	0.9804	第 20 段	1.0000	第 26 段	1.0000
第 3 段	0.9783	第 9 段	0.9474	第 15 段	1.0000	第 21 段	0.9800	第 27 段	0.9787
第 4 段	0.9524	第 10 段	0.9565	第 16 段	1.0000	第 22 段	1.0000	第 28 段	0.9796
第 5 段	1.0000	第 11 段	0.9783	第 17 段	0.9706	第 23 段	0.9630	第 29 段	0.9811
第 6 段	0.9444	第 12 段	0.9839	第 18 段	0.9697	第 24 段	0.9615	第 30 段	0.9710

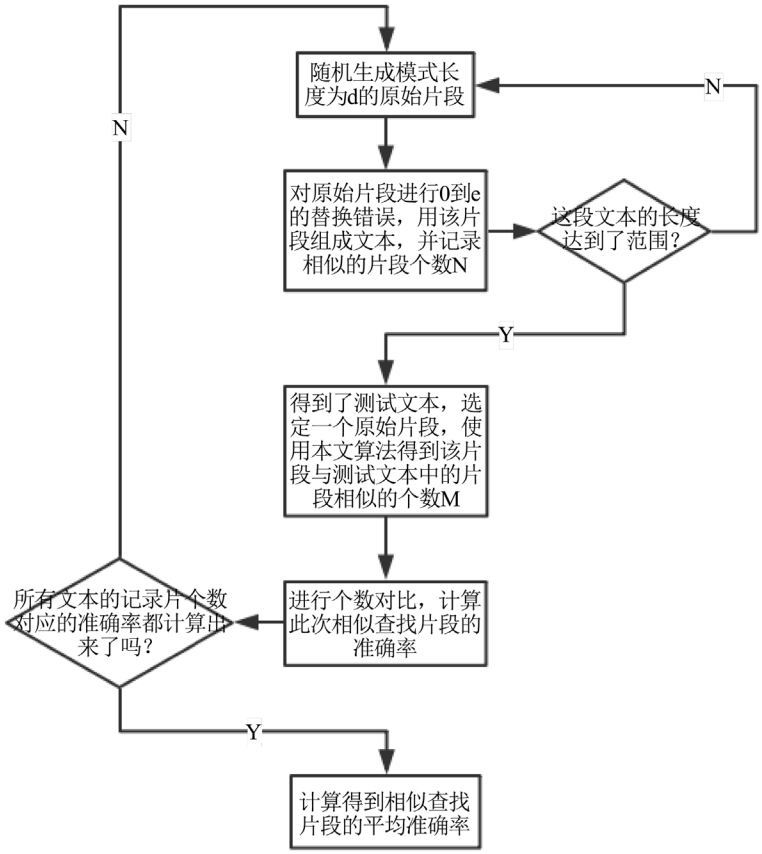

Figure 3. Flow chart of algorithm analysis
图 3. 算法分析流程图

5. 结语

本文编写了一种针对相似片段查找的算法，该算法以海明矩阵为基础，构建了相似度表达公式，迭代运算直到满足相似度阈值的片段，同时考虑了语言文本记录时的替换错误，为破译未知语言提供了很

好的出发点。

参考文献

[1] 郭顺, 管河山, 姜青山. 一种新的 DNA 序列重复片段的查找算法[C]//中国计算机学会. 第二十五届中国数据库学术会议(NDBC2008)论文集, 2008: 414-418.

[2] 王镝, 赵毅, 陈白尘, 等. DNA序列中基于后继数组索引的SATR查找算法[J]. 东北大学学报(自然科学版), 2007, 28(2): 184-188.

[3] 赵毅. 基于海明距离的 DNA 序列中相似性重复片段查找技术研究[D]: [硕士学位论文]. 沈阳: 东北大学, 2007.

[4] 朱扬勇, 熊赟. DNA 序列数据挖掘技术[J]. 软件学报, 2007, 18(11): 2766-2781.