山东农业大学学报(自然科学版),2017,48(3):467-471 VOL.48 NO.3 2017
Journal of Shandong Agricultural University (Natural Science Edition) doi:10.3969/j.issn.1000-2324.2017.03.029

基于音位的网络盗版文本查重方法

金哲凡,俞定国,林生佑,周忠成
浙江传媒学院, 浙江 杭州 310018

摘 要: 传统的文本查重算法是对文本作分词以构建关键词向量，而对于某些特殊应用的网络盗版检测，分词的开销则未必合理和必要。因此，本文提出一种基于汉语音位信息的文本查重方法。文本被表达为声、韵、调三个空间向量，以余弦距离作相似性度量。提出两种相似性判断公式，一种假定三向量独立分布；一种取其线性组合，系数可由音位元素的信息熵算出，通过大文本统计得出信息熵的估计值，以传统的关键词向量/SimHash 方法做参照产生语料，对其作统计得到模型参数。实验结果表明该方法有一定的精确率和很好的召回率，计算开销低于传统的方法，适合需要过滤大量 TN 类型文本的场合。

关键词: 音位; 盗版文本; 查重
中图法分类号: TP391 **文献标识码**: A **文章编号**: 1000-2324(2017)03-0467-05

Method for Checking Duplicate Text of Network Piracy Based on Phoneme

JIN Zhe-fan, YU Ding-guo, LIN Sheng-you, ZHOU Zhong-cheng
Zhejiang University of Media and Communications, Hangzhou 310018, *China*

Abstract: The traditional method checking repetition takes a text as a participle to establish some key vectors, however the piratical cost may not be reasonable or necessary for the discovery of the online copyright violation in some special APP. Therefore this paper proposed a method checking repetition with Chinese phonology. A text was represented by three vectors in spaces of Chinese initial, final and tone and cosine distance was used as a measurement of similarity. Two decision models were proposed. One assumed the three vectors were independent each other, while the other took a linear combination of the three, which needed to calculate the factors using information entropies that could be evaluated by large-corpus counting. Training corpus was generated with the old term-vector/SimHash method being used as a standard and threshold values were calculated. Test results showed the proposed method had a good precision and a very good recall ratio, and computational cost was lowed comparing to traditional methods based on term vectors to be suitable for filtering out a large amount of TN documents.

Keywords: Phonology; piratical text; checking repetition

文本查重是根据一定相似度模型从数据流中发现相重文本的过程。它在搜索引擎构建、抄袭检测、新闻分类等领域有广泛的应用。文本查重是一种特殊的文本过滤[1]，过滤条件是目标文本与源文本相似度大于阈值。文本查重方法基于两种基础技术：文本向量空间模型和文本指纹，前者解决相似性度量的问题，后者优化检索。

向量空间模型的作用是将无结构的文本表示成计算机易处理的特征向量，文本间的相似性问题随之转变成向量间距离的问题。特征提取算法包括 TF-IDF、词频方法、互信息方法、信息增益方法等。其中 TF-IDF 用关键词的权重做特征，权重计算兼顾了关键词在全局的重要性和在局部的频率这两种信息，使用广泛，是经典方法[2]。有些场合如文献[3]修改 TF-IDF 的权重公式以优化排序。针对中文，文献[4]在特征选取中考虑了词频，也考虑了标点符号，并且将文本的位置因素加入在内；文献[5]的提出"动词中心词"的概念，将文本中的部分动词组成动词序列作为一种特征；文献[6]用以中文句号为基础的特征实现了大规模的新闻网页查重。

特征向量确定后，文本间的相似性可用某种空间距离来表示，如余弦距离、数量积、相关系数、指数相似系数、几何平均最小、算数平均最小等[7,8]。特征向量与距离公式配合，就可以进行文本查重的计算。以网络盗版检测为例，网络文本的特征向量如果与源文本特征向量的余弦值大于阈值，就可以认为它有盗版嫌疑。现实中某些应用，如 Google 的搜索引擎对存储空间和计算时间特别敏感，需要使用文本指纹技术。它将文本的特征向量通过 Hash 函数映射为一定字长比如 64 bit 的二进制数，称为指纹，文本的比较通过指纹进行。长度固定的指纹适合构造指纹库，可进行快速检索。

收稿日期: 2016-08 03 **修回日期**: 2016-08-23
基金项目: 浙江省公益技术应用研究项目(2016C33196);浙江省公益性技术应用研究项目(2017C33105)
作者简介: 金哲凡(1974-),男,副教授,博士.主要研究方向为并行计算，信息处理，图形学. E-mail:jinzf@zjicm.edu.cn

数字优先出版:2017-05-02 http://www.cnki.net

从原始文本到特征向量、再到文本指纹是一个单向不可逆的信息减少过程。64 bit 的指纹实际只保留了 64 维向量空间的方向信息。在各种指纹算法之中，Google 的 SimHash 保留了较多向量间的相似性，可根据指纹间的海明距离反映文档间的差异程度，因此优于 MD5 等 Hash 算法，是主要使用的指纹技术。根据 Google 的经验，64 位 SimHash 值的海明距离在 3~5 之间可认为是同一文本。

中文文本的分词是提取关键词向量的前置步骤。分词算法已非常成熟，基于统计的方法是其主流[9]；与人工智能新技术结合的、基于大规模神经网络学习的方法是当前的热点[10,11]。分词算法至少是 $O(n^2)$ 的复杂度。定性地看，关键词向量可以看做文本“含义”的一种统计表达，大部分文本处理应用如摘要生成、倒排索引、机器翻译等的后续计算需要对文本的含义作一定程度的理解，因此分词的计算开销是完全必要的。而在一些特殊应用盗版检测中，文本查重、判断是否相同文本是唯一重要的计算，“含义”并不是必须的。如能用其他特征、如音位代替关键词，或作为关键词向量方法的前置和补充，可以避免大量分词计算，提高速度。

中文是极其独特的语言，“字”是独一无二的“音/意”载体，是其他语言没有的构造单位[12]。字的音位构成规整一致，音节占用时长和书写占位大体平均。从字的二进制表示得到其拼音只需一次内存访问的开销，远低于最好的分词算法。以字的音位统计信息作为特征进行文本查重，是一个符合汉字规律的方法。

本文的查重方法对文本提取声母、韵母、声调三个特征向量，以余弦距离为度量，实验了两种文本相似性模型。首先通过大样本的训练，获得模型参数的估计。之后以二种模型对大量文本进行测试，实验结果证明可以达到区分和过滤的目的。与传统的基于关键词向量方法比较，本文方法避免了分词，计算开销较低。

1　音位特征选取

国家汉语拼音标准规定了 23 个声母、24 个韵母和 16 个整体认读音节。一些语言学的统计工作[12]把 *w*、*y* 排除在声母之外，而韵母包括了三拼。本文如下处理：

(1)声母为标准的 23 个加上零声母，共 24 个。

(2)韵母为标准的 24 个加 10 个三拼韵母：ia, ua, uo, uai,iao,ian, iang, uan, uang, iong，共 34 个。

(3)声调为“阴、阳、上、去、轻”5 种不变。

(4)继承“ü”去两点的规则，除了 nü、lü、nüe、lüe 四个音节之外，都作“u”计。

如此覆盖汉语拼音标准下包括整体认读音节的所有情况，使从汉字到音位的映射可做到 1 字 1 声 1 韵 1 调。设文本 d 是字 z_k 的序列，如忽略标点、数字等非汉字元素，字长 n 的文本为：$d=(z_1 z_2 z_3 \ldots z_k \ldots z_n)$

其中 $z_k \in Z$，Z 为汉字集。汉字 z 的音位由声母 a、韵母 b 和声调 c 组成。若对多音字取其第一种发音，则 $z_k=(a_k, b_k, c_k)$，其中 $a_k \in S$, $b_k \in Y$, $c_k \in T$。$S=\{s_1, s_2, s_3, \ldots s_i \ldots s_{24}\}$,是声母集合；$Y=\{y_1, y_2, y_3, \ldots y_i \ldots y_{34}\}$,是韵母集合；$T=\{t_1, t_2, t_3, \ldots t_5\}$,是声调集合。

令 $f(s_i,d)$、$f(y_i,d)$、$f(t_i,d)$是声母 s_i、韵母 y_i、声调 t_i 在文档 d 中的频率，即

$$f(s_i,d)=\frac{\sum_{k=1}^{n} I(a_k=s_i)}{n}, f(y_i,d)=\frac{\sum_{k=1}^{n} I(b_k=y_i)}{n}, \quad f(t_i,d)=\frac{\sum_{k=1}^{n} I(c_k=t_i)}{n}$$

其中 I 为指示函数，参数表达式成立时为 1，否则为 0。则文档 d 可表示为三个特征向量的组合，$d=(\overline{s_d}, \overline{y_d}, \overline{t_d})$，其中 $\overline{s_d}=\langle f(s_1,d), f(s_2,d), f(s_3,d), \ldots f(s_{24},d)\rangle$，$\overline{y_d}=\langle f(y_1,d), f(y_2,d), f(y_3,d), \ldots f(y_{34},d)\rangle$，$\overline{t_d}=\langle f(t_1,d), f(t_2,d), f(t_3,d), f(t_4,d), f(t_5,d)\rangle$ 设有两个文档 d_1, d_2，可在 $\overline{s_d}, \overline{y_d}, \overline{t_d}$ 空间各定义余弦距离如下：

$$\cos\_s(d_1,d_2)=\frac{\overline{s_{d1}}\bullet\overline{s_{d2}}}{\|\overline{s_{d1}}\|*\|\overline{s_{d2}}\|}, \cos\_y(d_1,d_2)=\frac{\overline{y_{d1}}\bullet\overline{y_{d2}}}{\|\overline{y_{d1}}\|*\|\overline{y_{d2}}\|}, \cos\_t(d_1,d_2)=\frac{\overline{t_{d1}}\bullet\overline{t_{d2}}}{\|\overline{t_{d1}}\|*\|\overline{t_{d2}}\|}$$ 余弦距离 $\cos\_s(d_1,d_2)$、$\cos_y(d_1,d_2)$和 $\cos_t(d_1,d_2)$可对 d_1、d_2 间的相似度作出基于音位的度量。

2　相似性模型

实验了两种文本相似性模型：

(1)假定声、韵、调独立分布，当 d_1、d_2 在 $\overline{s_d}$, $\overline{y_d}$, $\overline{t_d}$ 三个空间上的余弦都超过阈值时，它们被认为相似（相重）。即文本 d_1，d_2 相似的条件为：

$$I\left[\cos\_s\left(d_1,d_2\right)\geq g_s\right]I\left[\cos\_y\left(d_1,d_2\right)\geq g_y\right]I\left[\cos\_t\left(d_1,d_2\right)\geq g_t\right]=1$$

其中 I 为指示函数，g_s，g_y，g_t 为三个空间上的阈值。

假定声、韵、调在文本相似性上的贡献度可通过权值表达，将它们线性组合来定义相似度指标 Similarity：$Similarity=\alpha cos\_s(d_1,d_2)+\beta cos\_y(d_1,d_2)+\theta cos\_t(d_1,d_2)$，其中$\alpha+\beta+\theta=1$。

文本 d_1，d_2 相似的条件为：$Similarity > g_{similairty}$，$g_{similairty}$ 为相似度阈值。

对模型 2 的权重系数α、β、θ，作如下考虑：某一特征向量对相似度的贡献应和它包含的信息量相关，而信息量可用信息熵 H 度量。令声母、韵母、声调向量的信息熵为 H_s、H_y、H_t，可定义：

$$\alpha=\frac{H_s}{H_s+H_y+H_t},\beta=\frac{H_y}{H_s+H_y+H_t},\theta=\frac{H_t}{H_s+H_y+H_t}$$

其中，$H_s=-\sum_{i=1}^{24}p\left(s_i\right)log_2p\left(s_i\right)$，$H_y=-\sum_{i=1}^{34}p\left(y_i\right)log_2p\left(y_i\right)$，$H_t=-\sum_{i=1}^{5}p\left(t_i\right)log_2p\left(t_i\right)$

其中 $p(s_i)$、$p(y_i)$、$p(t_i)$是声母 s_i、韵母 y_i、声调 t_i在文本中出现的概率。

$p(s_i)$、$p(y_i)$、$p(t_i)$可通过对大语料统计的频率值来近似。对 1,41 1,996 篇、共 481,065,247 字的现代汉语语料作统计的结果如表 1~3 所示(none 表示零声母)。

表 1 声母频率统计

Table 1 Frequencies of Chinese initials

b	p	m	f	d	t	n	l	g	k
4.314%	1.723%	2.773%	2.940%	9.419%	3.202%	1.986%	5.022%	5.062%	1.985%
h	j	q	x	zh	ch	sh	r	z	c
4.365%	8.121%	3.382%	6.185%	6.337%	3.461%	7.218%	2.331%	3.497%	1.639%
s	w	y	none						
1.668%	3.264%	9.005%	1.099%						

表 2 韵母频率统计

Table 2 Frequencies of Chinese finals

iang	uang	iong	ang	eng	ing	ong	uai	iao	ian
1.877%	0.631%	0.033%	3.686%	3.202%	4.040%	4.456%	0.136%	1.741%	4.304%
uan	ai	ei	ui	ao	ou	iu	ie	ue	er
2.808%	3.892%	3.287%	2.077%	3.516%	3.496%	0.923%	1.349%	0.974%	0.427%
an	en	in	un	vn	ia	ua	uo	a	o
4.137%	3.088%	2.702%	1.285%	0.000%	1.199%	0.578%	3.056%	2.970%	0.611%
e	i	u	v						
8.386%	15.997%	6.567%	2.569%						

表 3 声调频率统计

Table 3 Frequencies of Chinese tones

yin	yang	shang	qu	qing
21.775%	21.200%	17.134%	35.816%	4.075%

文献[12]列出了对 7754 个现代汉字的音位的静态统计结果，表 1-表 3 是对动态文本的统计且所取统计项目不同，故有较大差异。但两者还是有一些相似点，比如“*d*”是使用频率最高的声母，“*i*”是使用频率最高的韵母，去声是出现频率最高的声调。统计程序通过 Java 语言实现。用声、韵、调频率的统计值作为概率值的估计，得到 H_s、H_y、H_t 的估计值：H_s’=4.3644；H_y’=4.5300；H_t’=2.1081。进而得到模型 2 系数α、β、θ的估计值：α’=0.3967；β’=0.4117；θ’=0.1916。

3 实验和测试结果

实验分为参数计算和过滤测试两部分。参数计算目的是获得前述 2 个相似性模型的参数的估计值：模型 1 参数为 g_s，g_y，g_t；模型 2 参数为 $g_{similairty}$。

计算过程用传统的“关键词向量+SimHash 指纹”办法作为参照，以行业的经验值、指纹海明距离 3 作为文档相重的阈值。用随机字替换的办法给源文本掺入噪声，直至海明距离为阈值 3 为止。训练选用包含 925 个文本共 534,924 汉字的现代汉语语料，命名为 D，首先对其掺入噪声获得语料 D’，掺噪声的流程如下：

(1) 预先准备噪声模板 NoiseTemplate.txt，这是一个包含 7000 余字的现代汉语文本。

(2) 对 D 中文本 d，获取关键词向量及其 SimHash 指纹 u_1。

(3) 从噪声模板中随机取一个字 z，选择 d 文中一随机位置，用 z 替换原文字。

(4) 获取 d 的新指纹 u_2。

(5) 计算 u_1 和 u_2 的海明距离 H_dist。若 H_dist<3，跳转 3，循环。若 H_dist==3，转 6，出循环。若 H_dist>3，比较本次掺噪声前的文本和掺噪声后的文本的指纹哪个更接近 3，取接近者为输出文本，转 6。

(6) 若最终 H_dist==3，d 的处理结束。否则，若累积尝试次数小于上限，转 2，文本 d 的处理重新开始；否则若尝试次数大于上限，结束。

有时掺入一个字的噪声会导致海明距离跃迁，比如从 2 跳到 6，此时回到原状、重新尝试，直至语料中所有文本 d 都得到了对应的含噪声为海明距离 3 的相似文本 d'。

语料 D={d_i|i=1..925}掺噪声后得语料 D'={d_i'|i=1..925}，对每对文本 d_i 与 d_i'，提取文字音位的声、韵、调成分，计算各成分频率，获得向量($\overrightarrow{s_d}$, $\overrightarrow{y_d}$, $\overrightarrow{t_d}$)和($\overrightarrow{s_d}$', $\overrightarrow{y_d}$', $\overrightarrow{t_d}$')，之后计算它们在 S、Y、T 空间的夹角 cos_s(d_i,d_i')，cos_y(d_i,d_i')和 cos_t(d_i,d_i')。 针对模型 2，则按如下公式得一组 Similarity 计算值：Similarity$_i$=α' cos_s(d_i,d_i')+ β' cos_y(d_i,d_i')+ θ' cos_t(d_i,d_i')。最终结果如表 4。

表 4 模型参数训练结果

Table 4 Training results for parameters of models

	均值 Average	最大值 Max	最小值 Min	均方差 Standard deviation
cos_s	0.978	0.992	0.953	0.00174
cos_y	0.979	0.995	0.932	0.00168
cos_t	0.989	0.992	0.964	0.00044
Similarity	0.981	0.989	0.962	0.00140

两个相似性模型需要的参数为阈值，且表 4 中均方差较小，故不妨取最小值做模型参数。得模型 1 的参数为：g_s=0.953，g_y=0.932，g_t=0.964；模型 2 的参数 $g_{similairty}$=0.962。过滤测试包括两部分工作：测试语料生成和测试。先对源语料 E 随机掺入噪声，流程如下：

(1) 预先准备噪声模板 NoiseTemplate.txt，这是一个包含 7000 余字的现代汉语文本。

(2) 对 E 中文本 e，若其长度为 l，在[0,[kl]]区间内产生随机数 r。系数 k 是一个经验值，0<k<1。

(3) 循环 r 次，每次取噪声模板中一个随机汉字替换 e 中一个随机位置的汉字。最终得到掺入噪声后的文档 e'。

(4) 求得 e 和 e'文本指纹的海明距离 Hamming(e,e')。

取与训练语料无交集、共 455,464 字的 1000 个文本作为源语料 E，用上述流程掺入噪声后得 E'，E'与 E 的差距如表 5。

表 5 测试语料生成结果

Table 5 Results of generating testing corpus

Hamming dist in [0,3]	Hamming dist in (3-10]	Hamming dist > 10
592 docs	278 docs	130 docs

根据 Google 的经验值，若 Hamming(e,e')≤3，则 e 和 e'是相重文本，若用基于音位的方法测得正，则为 TP；若测得负，则为 FN。若 Hamming(e,e')>3，则 e 和 e'是不同文本，若用基于音位的方法测得正，则为 FP；若测得负，则为 TN。

将训练所得参数代入模型 1 和模型 2，对 E 和 E'作查重计算，统计各文本分类结果，并计算精确率 P、召回率 R、调和指标 F_1 和运行时间 t，结果如表 6：

表 6 测试结果

Table 6 Test results

Model	TP	FN	FP	TN	P	R	F_1	t(s)
Model-1	565	27	58	350	90.69%	95.44%	93.00%	6.773
Model-2	572	20	49	359	92.11%	96.62%	94.31%	6.952

试将模型使用的参数（阈值）都调低 10%，再运行测试，结果如表 7。

表 7 调低阈值后测试结果

Table 7 Test results after lowing threshold value

Model	TP	FN	FP	TN	P	R	F_1	t(s)
Model-1	581	11	68	340	89.52%	98.14%	93.63%	6.760
Model-2	585	7	57	351	91.12%	98.82%	94.81%	6.945

可以看出两个模型都可以过滤大部分相似文本，模型 2 略好于模型 1，且都有很好的召回率。

降低阈值后，精确率 P 下降，召回率 R 升高。作为对比，对同数据集进行分词、关键词提取和算关键词向量余弦操作，执行时间 t'=28.355 s。运行时间的计取都剔除了文件读取等无关操作。对表 3 和表 4 中运行时间取均值，模型 1 均值 t_1，模型 2 均值 t_2，t_1/t' =23.86%，t_2/t'=24.51%。文本基于音位的方法在计算时间上明显优于关键词向量方法。

实际应用是一个用于互联网盗版发现的系统。对出版社等拥有的原作库提取音位特征，网络爬虫连续获取网络文本，对其内容逐个提取音位信息，用本文方法进行前置过滤，之后再作同一性（查重）检测。之后进行违法性检测，找到真正的侵权项目。网络盗版行为猖獗，但在海量的文本流中涉嫌盗版的毕竟是少数，绝大部分是无关的。由于内容库文本数量巨大，系统效率很大程度取决于能否将这 99%以上的无关文本快速排除，因此在精确率和速度之间，系统更关注速度；在精确率和召回率之间，系统更关注召回率。本文方法有很好速度和召回率，非常适合做前置过滤。精确率随阈值的下调而下降是个问题，对于 FP 类型的文本，可以在后续步骤用其他方法滤去，比如用传统的关键词向量+SimHash 的办法作交叉验证，或人工验证。由于这部分文本数量已极少，因此不影响系统整体性能。

实验程序的分词、向量提取和 SimHash 计算使用了软件 simHash，汉字的拼音转换使用了软件包 pinyin4j，语料使用了搜狐实验室的全网新闻语料资源。音位相关的程序用 Java 实现，Simhash 相关的程序用 gcc 实现，用 Java 本地进程调用机制处理二者的协同。

4　结 论

文本查重方法基于两种基础技术：以 TF/IDF 为代表的向量表示和以 SimHash 为代表的文本指纹。研究者提出了各种改进，针对特殊的领域进行优化。本文基于汉语“字”音位均匀的特点，提出基于音位的查重办法。文本被表示为声、韵、调三个空间的向量，相似性以余弦距离度量。提出两种距离模型，一种假定三向量独立分布；一种假定三向量可线性组合，其系数由音位元素的信息熵算出。测试结果表明两种模型都可以实现过滤，且有召回率优于精确率的特点。这个特点非常适合于类似网络盗版文本发现的应用：输入文档的数量巨大，要求快速处理，但其中 99%以上是不涉嫌盗版（True Negative 型）的。将本文方法用于其前置过滤，由于音位频率的计算只需一次内存访问的开销，不需进行分词，因此效率高于基于关键词向量的方法。

语言是含义和发音的综合物。关键词向量是对含义的统计表达而不顾及其发音；本文方法利用了汉语的发音而不顾及其含义。定性地考虑，前者相当于人通过默读区分文档，后者相当于不识字的人通过辨音区分文档。两者都是可行的，但必定有各有特点。基于音位的方法优点是不需分词，可以以较快速度实现一定精确率的过滤。它可以单独使用，也可与其他方法联合。在必要的场合，音位向量也可通过 SimHash 产生指纹以加快检索。未来还需要深入的研究以拓展其应用。

参考文献

[1] Korde V, Mahender CN. Text Classification and Classifiers: A Survey[J]. International Journal of Artificial Intelligence & Applications, 2012,3(2):45-52

[2] Wang XH, Cao J, Liu Y, *et al*. Text Clustering Based on the Improved TF-IDF by the Iterative Algorithm[J]. IEEE Symposium on Electrical & Electronics Engineering, 2012,34(7):140-143

[3] 李镇君,周竹荣.基于Document Triage的TF-IDF算法的改进[J].计算机应用,2015,35(12):3506-3510

[4] 宋 兰,孙茂松.中文文本全文查重的实验研究[C]//宋兰.全国第八届计算语言学联合学术会议(JSCL-2005),2005:147-157

[5] 刘小军,赵 栋,姚卫东.一种用于中文文本查重的双因子相似度算法[J].计算机仿真,2007,24(12):312-314

[6] 韦永壮,袁春风,黄宜华.CCDet:一种高效的大规模中文重复网页检测方法[J].计算机研究与发展,2013,50(S2):140-152

[7] 华秀丽,朱巧明,李培峰.语义分析与词频统计相结合的中文文本相似度量方法研究[J].计算机应用研究,2012,29(3):833-836

[8] 张佩云,陈传明,黄 波.基于子树匹配的文本相似度算法[J].模式识别与人工智能,2014,27(3):226-234

[9] 黄昌宁,赵 海.中文分词十年回顾[J].中文信息学报,2007,21(3):8-19

[10] 韩冬煦,常宝宝.中文分词模型的领域适应性方法[J].计算机学报,2015,38(2):272-281

[11] 来斯惟,徐立恒,陈玉博,等.基于表示学习的中文分词算法探索[J].中文信息学报,2013,27(5):8-14

[12] 冯志伟.语言与数学[M].北京:世界图书出版社,2011:360