基于关键词重提取的密文文本相似性度量方法研究

李志华 陈超群 李 村 胡振宇 张华伟
(江南大学物联网工程学院计算机科学系 无锡 214122)

摘 要 针对密文的相似性度量问题,提出了一种新的密文文本相似性度量方法。该方法通过定义关键词的有效作用域、相对作用域、分散域的概念,有效克服了现有的关键词权重量化方法不能对篇幅不同、结构不同的文档进行相对公平的关键词权重量化的不足,同时显著减少了文本度量时所依赖的关键词数量。进一步对文档的关键词进行重提取,并建立文档的关键词密文索引条目,通过密文的索引条目来度量密文的相似性。将该方法在真实文档上进行实验,并同其它算法进行比较,结果表明所提出的方法在准确率和召回率两方面优于其它参与比较的算法,并能在准确率和召回率之间取得比较好的平衡。

关键词 关键词重提取,相似性度量,密文文本,作用域
中图法分类号 TP309.5 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.8.020

Similarity Measure Algorithm of Cipher-text Based on Re-extracted Keywords

LI Zhi-hua CHEN Chao-qun LI Cun HU Zhen-yu ZHANG Hua-wei
(Department of Computer Science, School of IOT Engineering, Jiangnan University, Wuxi 214122, China)

Abstract To solve the similarity of dissimilarity measurement between the cipher texts, a new similarity measure algorithm of cipher-text based on re-extracted keywords called SMCTBRK was proposed. Through defining the new concepts of effective scope, relative scope, distributed scope of the keywords, and re-extracting the keywords in documents, the SMCTBRK constructs the encryption index item for the compared documents depending on the less amounts of re-extracted keywords. Here, the encryption index item is organized as the feature vector. Further, the SMCTBRK computes the similarity between the different cipher texts by the encryption index item instead of the separated keywords. Experiments on real documents were conducted. And the results show that the SMCTBRK is more promised than the Shingling algorithm and the Simhash algorithm on accuracy and recall ratio.

Keywords Re-extracted keywords, Similarity measure, Cipher texts, Effective scope

1 引言

文本通常根据其固有的特征来度量其相似性。文本的特征包括文本的内容特征和非内容特征两种。非内容特征包括文档的大小、类型、位置、拥有者等特征,通常比较具体。文本的内容特征包括文本的结构特征、关键词词频统计特征和关键词的语义特征等。文本的相似性度量在信息检索[1,2]、文本分类[3,4]、文本查重[5,6]等领域得到了广泛应用。当前,随着云存储技术的发展,长期存储海量信息成为可能,但是云端信息的隐私保护、数据信息防泄漏等问题显得更加突出。首先,用户将文档托管到云存储系统中,用户失去了对文档的控制,为了保证文档的安全性,在用户上传文档之前对文档进行加密、以密文的形式对文档进行云端存储不失为一种良好的解决办法,但是又将会为云端的文件检索造成困难;其次,文档的关键词特征不应该泄露文档的内容信息。总之,这一云端密文存储的应用模式具有比较好的应用价值,但同样为学术界带来了密文检索、密文相似性度量等新学术问题的挑战。

本文以密文的相似性度量为研究内容,要求用户在文档加密之前重提取文档的关键词特征,该关键词特征同样要求以密文形式存在。通过对文本关键词的重提取,在关键词的权重量化、密文索引的建立、密文的相似性度量等方面进行研究,提出了一种新的密文文本相似性度量方法。

2 相关工作

当前文本相似度度量的研究工作主要局限于明文文档,最著名的是 Shingling 算法、Simhash 算法以及它们的后续改进算法[4,5]。Shingling 算法主要适用于英语文本,算法首先用一个滑动窗口对文档进行 Shingle 划分,提取文本的特征并用 Shingle 表示,通过比较两个文档相同的 Shingle 数目来计算文档的相似度[4,5]。由于英文文本是一些由字母组成的单词集合,并且单词与单词之间有空格隔开,因此处理对象比较简单。当对 Shingling 算法进行改进并将其应用于中文文本相似性度量时,由于表示中文文档特征的基本单元是"词语",因此要对中文文档首先进行文本格式分析、有效文本提取和分词处理等预处理工作,然后再进行文本特征表示、文本特征抽取,最后进行文本相似度计算,其中分词处理是中文文本处

到稿日期:2015-07-20 返修日期:2015-11-16 本文受江苏省科技厅产学研前瞻项目(BY2013015-23)资助。
李志华(1969-),男,博士,教授,主要研究领域为网络技术、信息安全等;陈超群(1990-),男,硕士生,主要研究领域为云计算与云安全;李 村(1991-),男,硕士生,主要研究领域为云计算与云安全;胡振宇(1991-),男,硕士生,主要研究领域为信息安全;张华伟(1988-),男,硕士生,主要研究领域为云计算、分布式计算。

理中必不可少的预处理工作之一[6]。但是，Shingling 算法存在以下不足：对文档的每个特征 Shingle 同等对待[4-7]，没有权衡各特征在特定文本中的重要程度，这显然是不合理的；另外，通过设置一个权重门限值来抽取权重大的 Shingle 参加文本的相似度度量[4]，在此，该门限值仅仅是一个经验值[8]；没有考虑 Shingle 的作用范围，显然不适合结构化特征非常鲜明的中文文本之相似性度量。

Simhash 算法以文档中出现的单词作为文档的特征[4,5]，同时，根据单词出现的频率为每个特征设置一个权重，并生成单词的权重序列，选取权重超过一定阈值的特征组成特征向量，并将高维的特征向量映射成一个唯一的 n 位指纹值。此指纹值除了能提供原始内容是否相等等信息以外，还能额外提供不相等内容的差异程度信息。所以可以将计算文档之间的相似度问题转化为计算两个文档的指纹值之间的汉明距离，用汉明距离来表示文档之间的相似度[5-7]。显然，Simhash 算法是基于单词和单词的权重序列的[8,9]，但是在 Simhash 算法中单词的权重仅仅考虑了词频信息，没有考虑文档的结构，如大小标题中所蕴含的关键词特征信息和关键词特征的作用范围等因素，并不能完整地覆盖一篇文档的全部信息。

显然，两个算法都需要提取大量的文档特征，在云存储环境下，网络传输代价会比较高。另外，对于加密后的密文文本度量问题，首先，通过统计关键词在文档中的词频特性来确定关键词的权重的策略显然行不通，因为在密文文档中难以统计关键词和关键词出现的频率；其次，只要关键词出现的位置相同，而不论其所在文档的篇幅差异有多大，最后关键词量化的权重都相同，这对于结构显明的中文文本之相似性度量显然是不公平的，也不适合密文文本度量；最后，为了不泄漏密文信息，对于密文度量而言，关键词越少越好。针对以上不足，本文从以下几方面进行研究改进，首先，通过对文档关键词进行重提取，根据所提取文档中反映文档主题的关键信息之关键词建立密文索引，有效避免了关键词所组成的文本特征向量过于稀疏的不足；其次，考虑到从密文中提取特征关键词的困难，本文通过度量密文索引的相似性来衡量密文之间的相似度；再次，本文通过提出有效作用域和分散域两个概念来描述关键词的范围属性，从而起到减少关键词数量的效果；最后，在上述研究基础上，本文提出了基于关键词重提取的密文文本相似性度量（Similarity measurement of cipher text based on re-extracted keywords，SMCTBRK）方法。

3 SMCTBRK 方法

SMCTBRK 文本相似性度量方法包含 3 方面内容：关键词提取与权重量化、密文索引建立、密文文本相似度的计算。

3.1 关键词提取与权重量化

在中文文本中，词频和语义是选择关键词的两个主要依据，在二者之间根据语义选择关键词的好处是：可通过选择数量较少的关键词得到的度量效果与通过词频所选择的大量关键词进行度量的效果相当，但是在传统根据语义选择关键词进行文本度量的方法中，没有考虑关键词的作用范围这一特征，显然在考虑了关键词的作用范围的情况下，文本度量时所依赖的关键词数量将会变得更少。在此，以结构化鲜明的中文文本为例来说明此问题，如图 1 所示。

通常二级标题、三级标题中的关键词的语义是服从一级标题的，所以一级标题中的关键词的作用范围完全可以覆盖二级标题、三级标题中的关键词，甚至覆盖文本正文中的所有关键词。基于此，完全可以从一级标题中提取语义丰富的关键词进行文本度量，这样在度量时可明显减少所依赖关键词的数量。为此，首先给出如下新定义。

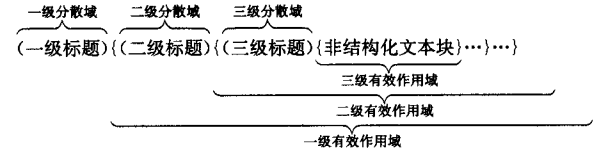

图 1 结构化文本的基本结构

定义 1 关键词 key 作用的范围称为关键词的有效作用域，关键词作用范围用有效关键词的长度表示，符号表示成 L_u。

如在结构化文本中的一级标题中的关键词作用范围为整篇文档，则它的有效作用域可以表示为该文档除了一级标题之外的有效关键词的长度。另外，关键词有效作用域范围中内容的描述是分散到多个关键词中的，例如文档的标题是对整篇文档的概括，标题中包含多个关键词，那么这些关键词的有效作用域就是整篇文档，标题中包含的关键词越多，次标题中每个关键词包含的文档信息就会越少。

定义 2 若有效关键词 $key_1, key_2, \cdots, key_n$ 具有相同的有效作用域 L_u，则 $\frac{L\_u}{n}$ 称为关键词 $key_1, key_2, \cdots, key_n$ 的相对作用域，用符号 L_r 表示，关键词 $key_1, key_2, \cdots, key_n$ 称为有效作用域 L_u 的分散域。

根据定义 1、定义 2，可以得到关键词 key_i 在文本 doc_j 中第 k 次出现的权重，按式(1)计算：

$$W\_Loc(key_i, doc_j, k) = a * \frac{L\_u_{i,j,k}}{n_{i,j,k}} \tag{1}$$

其中，a 为常数，$L\_u_{i,j,k}$ 表示关键词 key_i 在文本 doc_j 中第 k 次出现所处的作用域，$n_{i,j,k}$ 表示关键词 key_i 在文本 doc_j 中第 k 次出现所处的分散域。

通过关键词权重合并可以得到文本 doc_j 中关键词 key_i 的权重，如式(2)所示：

$$w(key_i, doc_j) = \frac{\sum_{k=1}^{k \le count(key_i, doc_j)} W\_Loc(key_i, doc_j, k)}{Stat(key_i)} \tag{2}$$

其中，$Stat(key_i)$ 为关键词 key_i 的词频统计，$count(key_i, doc_j)$ 为关键词 key_i 在文本 doc_j 中出现的次数。

3.2 密文索引的建立

将关键词按照其权重从大到小排序，选择前 s 个关键词作为文本的关键信息关键词。根据所提取的关键信息关键词，构造文档 doc_j 的关键词索引条目，如图 2 所示，其中 $E(key_i)$ 为关键词 key_i 的密文。该关键词索引同文档 doc_j 的密文拼接在一起充当 doc_j 在密文状态下的索引摘要，可供密文检索、密文调度、密文相似性度量之用。

doc_j	$(E(key_1), w(doc_j, key_1))$	…	$(E(key_s), w(doc_j, key_s))$

图 2 关键词密文索引

可见，在关键词的密文索引条目中，关键词与其密文同样是一一对应的，因此在文本的关键词密文索引中关键词并没有丧失其应有的统计特性。

3.3 密文文本的相似性度量

假设有两个待度量文本 doc_i, doc_j，并且它们均需要加密后以密文的形式存储。为了度量它们的相似性，首先借助分词算法进行关键词提取[4,9,10]，再根据其关键词密文索引构造

如下形式的“关键词密文-权重”的键值对集合。

$U(doc_i)=\{(E(key_{i1}),w(doc_i,key_{i1})),(E(key_{i2}),w(doc_i,key_{i2})),\cdots,(E(key_{is}),w(doc_i,key_{is}))$

$U(doc_j)=\{(E(key_{j1}),w(doc_j,key_{j1})),(E(key_{j2}),w(doc_j,key_{j2})),\cdots,(E(key_{js}),w(doc_j,key_{js}))$

在此，$U(doc_i)$键的集合为$U_key(doc_i)=\{E(key_{i1}),E(key_{i2}),\cdots,E(key_{is})\}$，$U(doc_j)$键的集合为$U_key(doc_j)=\{E(key_{j1}),E(key_{j2}),\cdots,E(key_{js})\}$。

假设$U_key(doc_i)$与$U_key(doc_j)$的并集为：

$UU(doc_i,doc_j)=\{(key_1,w_{i1}),(key_2,w_{i2}),\cdots,(key_m,w_{im})\}$

并且，$U(doc_i)$对应于$U(doc_j)$的并集计算成：

$UU(doc_i,doc_j)=\{(key_1,w_{i1}),(key_2,w_{i2}),\cdots,(key_m,w_{im})\}$

当$key_k\in U_key(doc_i)$时，$w_{ik}=w(doc_i,key_k)$，否则$w_{ik}=0$。

同理可得$U(doc_i)$对应于$U(doc_j)$的并集计算成：

$UU(doc_j,doc_i)=\{(key_1,w_{j1}),(key_2,w_{j2}),\cdots,(key_m,w_{jm})\}$

当$key_k\in U_key(doc_j)$时，$w_{jk}=w(doc_j,key_k)$，否则$w_{jk}=0$。

假设UU的权重值为：

$UU\_value(doc_i,doc_j)=(w_{i1},w_{i2},\cdots,w_{im})$

$UU\_value(doc_j,doc_i)=(w_{j1},w_{j2},\cdots,w_{jm})$

最后，文档doc_i与文档doc_j的相似性计算成两个向量的余弦夹角，如式(3)所示：

$$re(doc_i,doc_j)=\frac{UU\_value(doc_i,doc_j)UU\_value(doc_j,doc_i)^{\mathrm{T}}}{|UU\_value(doc_i,doc_j)||UU\_value(doc_j,doc_i)|} \tag{3}$$

基于上述研究，本文提出基于关键词重提取的密文文本相似性度量SMCTBRK方法，描述如下。

Step1 重提取doc_i，doc_j的分词关键词集合；

Step2 根据式(1)、式(2)对关键词进行权重量化，得到关键词-权重集合；

Step3 将关键词-权重集合中的元素按权重从大到小进行筛选，选取一定数目的关键词-权重对；

Step4 将筛选出的关键词-权重对的关键词部分加密，得到关键词密文-权重对集合，将关键词密文-权重对按照图2所示方式构造文本doc_i，doc_j的关键词密文索引；

Step5 根据文本doc_i，doc_j的关键词密文索引重新构造文本doc_i，doc_j的关键词密文-权重的键值对集合$U(doc_i)$，$U(doc_j)$；

Step6 根据式(3)计算文本doc_i，doc_j的相似性$re(doc_i,doc_j)$。

该方法不需要存储整个文档的词语集合，并且鉴于中文文档的结构特征，算法充分考虑了各级标题中关键词的有效作用域，着重针对各级标题中的关键词进行了提取，因此仅仅需要存储所提取之数量不算多的关键词集合，这样为大规模文本数据节省了大量的物理存储空间。另外，由于文本的度量需要比较两两子项目之间的相似度，需要计算C_n^2次，即时间复杂度为$O(n^2)$。该方法适合在本地加密并在创建好其关键词密文索引后上传到需要进行云端存储的应用场合，有利于保护数据的隐私并防止数据信息泄漏。

4 性能评价与分析

4.1 实验环境

实验的硬件环境为因特尔双核E5800处理器，4GB内存，操作系统为Windows 7，32bit，MyEclipse 10，JDK1.6，Xpdf，加密算法为国密SMS4算法[8]。

4.2 实验数据简介

在中国知网上随机下载51篇中文文档，将其分成3组，分别标记为G1，G2，G3，每组包含17篇文档。为了上下文表达和实验的方便，对G1，G2，G3进行统一编排，其中每组17篇文档中包含“云计算调度”主题的文章11篇，分成A，B两组，A组5篇，编号为A_1－A_5；B组6篇，编号为B_1－B_6；干扰文章6篇，编号为C_1－C_6。然后分别对G1，G2，G3组进行实验。在实验之前首先对这些文章进行英文字符排除处理，然后把A组文档分别与B，C组中的文档逐一进行相似性比较，并对这3组文档的实验结果按“A与B、A与C”的相似性计算结果分别求平均值。

4.3 评价标准

本文采用传统的准确率[4]、召回率[4]以及新定义的平衡系数这3个关键指标来对SMCTBRK方法进行性能评价。

(1)准确率和召回率

假设在进行文本相似性度量时的实验结果如表1所列，则准确率定义成被检测相似文本中实际相似文本所占的比例，如式(4)所示：

$$precision=\frac{DS}{DS+DN} \tag{4}$$

召回率定义成实际相似文本中被准确检测出的比例，如式(5)所示：

$$recall=\frac{DS}{DS+US} \tag{5}$$

表1 文本相似性度量的实验结果

算法	实际相似的文本数量	实际不相似的文本数量
检测结果为相似的文本数量	DS	DN
检测结果为不相似的文本数量	US	UN

(2)平衡系数

通常，准确率会随着召回率的增长而增长，但是有时候准确率和召回率会出现不匹配的情况，即准确率达到用户需求而召回率没有达到用户需求，或者召回率达到用户需求而准确率没有达到用户需求，亦即准确率和召回率存在失衡情况。可见，只依靠准确率和召回率不能准确评价文本相似性度量方法的性能优劣。为了衡量准确率和召回率的平衡性，本文提出了平衡系数的概念。准确率和召回率的平衡系数计算如式(6)所示。

$$F_\beta\_balance=(1+\beta^2)\times\frac{precision\times recall}{\beta^2\times precision+recall} \tag{6}$$

其中$\beta=\frac{weight(recall)}{weight(precision)}$，$weight(recall)$和$weight(precision)$分别为召回率与准确率的权重。显然，$F_\beta\_balance\to 1$，准确率与召回率的平衡性越好；$F_\beta\_balance\to 0$，准确率与召回率的平衡性越差。可见，平衡系数是关于算法性能的一个评价指标，其越趋近于“1”，说明算法的运行结果越理想；反之，亦然。

4.4 实验结果及分析

对4.2节所提供的样本首先用SMS4加密算法进行加密，对加密后的密文分别用Shingling，Simhash，SMCTBRK 3个算法进行实验，并对实验结果进行分析和比较。其中，Shingling，Simhash算法对关键词的预处理采用与SMCTBRK相同的“关键词提取与权重量化”方法。

4.4.1 实验结果

实验 1 Shingling 算法

Shingling 算法在 4.2 节样本上的实验结果如表 2 所列，表中列出的是 G1，G2，G3 中各自的 A 组样本与 B，C 组样本之间相似性的平均值。

表 2 Shingling 算法的测试结果

编号 \ 相似系数	A_1	A_2	A_3	A_4	A_5
B_1	0.492	0.449	0.345	0.363	0.554
B_2	0.562	0.546	0.472	0.457	0.600
B_3	0.434	0.613	0.492	0.423	0.521
B_4	0.382	0.563	0.352	0.337	0.451
B_5	0.628	0.617	0.518	0.473	0.633
B_6	0.997	0.467	0.366	0.415	0.572
C_1	0.287	0.309	0.262	0.298	0.349
C_2	0.308	0.433	0.351	0.346	0.424
C_3	0.239	0.323	0.205	0.206	0.286
C_4	0.290	0.334	0.219	0.208	0.295
C_5	0.516	0.506	0.430	0.437	0.514
C_6	0.321	0.483	0.329	0.290	0.378

为了使实验结果更直观，将表 2 中 A 组与 B 组文档的相似系数按从大到小的顺序排列，A 组与 C 组文档的相似系数按从小到大的顺序排列，其分布图如图 3 所示。

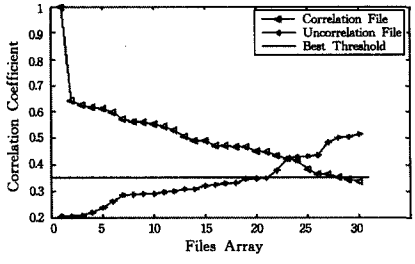

图 3 Shingling 算法的相似系数

当相似性阈值取 0.3520 时，检测结果如表 3 所列，表中是 Shingling 算法在 G1，G2，G3 组上实验结果的平均值。

表 3 Shingling 算法的实验结果

算法	实际相似的文本数量	实际不相似的文本数量
检测结果为相似的文本数量	28	9
检测结果为不相似的文本数量	2	21

实验 2 Simhash 算法

Simhash 算法在 4.2 节样本上的实验结果如表 4 所列，表中列出的是 G1，G2，G3 中各自的 A 组样本与 B，C 组样本之间相似性的平均值。

表 4 Simhash 算法的测试结果

编号 \ 相似系数	A_1	A_2	A_3	A_4	A_5
B_1	0.228	0.010	0.239	0.001	0.172
B_2	0.067	0.012	0.142	0.006	0.093
B_3	0.103	0.085	0.141	0.079	0.114
B_4	0.089	0.005	0.259	0.002	0.012
B_5	0.055	0.105	0.009	0.065	0.141
B_6	1.000	0.030	0.094	0.018	0.424
C_1	0.023	0.030	0.020	0.001	0.019
C_2	0.011	0.008	0.024	0.007	0.003
C_3	0.018	0.008	0.008	0.010	0.005
C_4	0.042	0.045	0.025	0.009	0.098
C_5	0.009	0.015	0.012	0.001	0.023
C—6	0.018	0.126	0.114	0.004	0.024

将表 4 中 A 组与 B 组文档的相似系数按照从大到小的顺序排序，A 组与 C 组文档的相似系数按照从小到大的顺序排序，排序后的相似系数分布如图 4 所示。

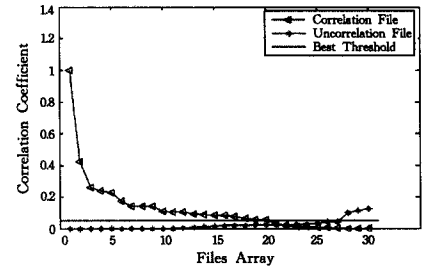

图 4 Simhash 算法的相似系数

当相似性阈值取 0.0498 时，检测结果如表 5 所列，表中是 Simhash 算法在 G1，G2，G3 组上实验结果的平均值。

表 5 Simhash 算法的实验结果

算法	实际相似的文本数量	实际不相似的文本数量
检测结果为相似的文本数量	20	3
检测结果为不相似的文本数量	10	27

实验 3 SMCTBRK 方法

SMCTBRK 方法在 4.2 节样本上的实验结果如表 6 所列，表中列出的是 G1，G2，G3 中各自的 A 组样本与 B，C 组样本之间相似性的平均值。

表 6 SMCTBRK 方法的测试结果

编号 \ 相似系数	A_1	A_2	A_3	A_4	A_5
B_1	0.822	0.502	0.607	0.610	0.687
B_2	0.844	0.535	0.729	0.670	0.712
B_3	0.825	0.526	0.714	0.671	0.686
B_4	0.897	0.555	0.625	0.677	0.755
B_5	0.073	0.006	0.148	0.005	0.033
B_6	0.995	0.565	0.682	0.686	0.797
C_1	0.001	0.014	0.097	0.052	0.015
C_2	0.034	0.042	0.241	0.068	0.004
C_3	0.001	0.011	0.007	0.02	0.280
C_4	0.003	0.014	0.088	0.034	0.001
C_5	0.013	0.001	0.130	0.025	0.024
C_6	0.001	0.035	0.221	0.078	0.001

将表 6 中 A 组与 B 组文档的相似系数按照从大到小的顺序排列，A 组与 C 组文档的相似系数按照从小到大的顺序排序，排序后的相似系数的分布如图 5 所示。

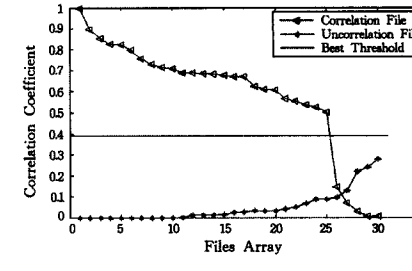

图 5 SMCTBRK 方法的相似系数

当相似性阈值取 0.3906 时，检测结果如表 7 所列，表中是 SMCTBRK 方法在 G1，G2，G3 组上实验结果的平均值。

表 7　SMCTBRK 文本相似性度量结果

算法	实际相似的文本数量	实际不相似的文本数量
检测结果为相似的文本数量	25	0
检测结果为不相似的文本数量	5	30

4.4.2　实验结果及分析

根据 4.3 节的相关评价标准，分别计算 3 种算法的 3 个主要性能指标，结果如表 8 所列。

表 8　3 种算法的性能比较

算法	准确率(%)	召回率(%)	平衡系数
Shingling	75.68	93.33	0.84
Simhash	86.96	66.67	0.75
SMCTBRK	100	83.33	0.91

表 8 比较了 3 种文本相似性度量方法的准确率、召回率以及平衡系数，可以看出，SMCTBRK 在召回率方面略低于 Shingling 算法，但在准确率和平衡系数方面高于 Shingling 算法和 Simhash 算法；而且从图 3－图 5 可以看出，SMCTBRK 方法的相关文本之相似系数与无关文本的相似系数之可区分性高于其它两种参与对比的算法。另外，虽然 3 种算法的时间复杂度相同，均为 $O(n^2)$，但是由于 SMCTBRK 方法在计算时所依赖的关键词数量比 Shingling 算法、Simhash 算法要少，因此在应用时，SMCTBRK 方法实际所消耗的时间要比另外两种算法少。所以，从这个角度而言，SMCTBRK 方法的效率要高于其它两种算法的。

结束语　本文提出了 SMCTBRK 方法，主要工作来自 3 个方面：1)优化了原有的关键词权重度量方法，在权重度量过程中考虑了关键词的范围属性，使关键词的度量在篇幅和文章结构差异性较大的文本中更加公平，并能显著减少在文本度量过程中所需要依赖的关键词数量，同时不影响最后的度量效果；2)通过更合理地计算关键词的权重，通过抽取更能代表文本主要内容的关键词，并根据它们构建文档的密文索引，有效地防止了数据信息泄漏；3)通过计算文档之密文索引向量之间的余弦夹角而不是具体的关键词来度量文档之间的相似性，起到了保护信息隐私的作用。在真实文档上进行了实验，结果表明 SMCTBRK 方法在准确率和平衡系数方面高于 Shingling 算法和 Simhash 算法，但在召回率方面略低于 Shingling 算法，需要进一步改进。

参 考 文 献

[1] Wang C, Cao N, Li J, et al. Secure ranked keyword search over encrypted cloud data[C]// Proceedings of ICDCS. Genova, Italy, 2010: 253-262

[2] Sebastiani F. Machine learning in automated text categorization, acmcs[J]. ACM Computing Surveys, 2002, 34(1): 1-47

[3] Hemalatha S, Raja K, Arasu T. Duplicate Detection of Query Results from Multiple Web Databases [J]. IJCA Special Issue on Computational Science—New Dimension & Perspectives, 2011(2): 71-75

[4] Zhang Zu-ping, Xu Xin, Long Jun, et al. Parameters Correlation and optimization in Text Similarity Measurement[J]. Journal of Chinese Computer Systems, 2011, 32(5): 983-989(in Chinese)
张祖平，徐昕，龙军，等. 文本相似性度量中参数相关性与优化配置研究[J]. 小型微型计算机系统，2011，32(5)：983-989

[5] Song Qin-bao, Yang Xiang-rong, Shen Jun-yi, et al. A Detection Algorithm for the Illegal Coping and Distributing of Digital Goods[J]. Chinese Journal of Computers, 2002, 25(11): 1207-1213(in Chinese)
宋擒豹，杨向荣，沈钧毅，等. 数字商品非法复制的检测算法[J]. 计算机学报，2002，25(11)：1207-1213

[6] Li Ya-zhou. The research and improvement of an automatic construction system of text classification corpus[D]. Wuhan: Wuhan University of Technology, 2011(in Chinese)
李亚洲. 文本分类语料库自动构建系统的研究与改进[D]. 武汉：武汉理工大学，2011

[7] Ye Shao-zhi, Wen Ji-rong, Ma Wei-ying. A systematic study on parameter correlation in large scale duplicate document detection [J]. Knowledge and Information Systems, 2008, 14(2): 217-232

[8] Li Rui-lin, Sun Bing, Li Chao, et al. Differential Fault Analysis on SMS4 using a single fault[J]. Information Processing Letters, 2011, 111(4): 156-163

[9] Shi Kan-sheng, Liu Hai-tao, Song Wen-tao. A Text Clustering Method Based on Speech to Text and Improved Center Selection [J]. Pattern Recognition and Artificial Intelligence, 2012, 25(6): 996-1001(in Chinese)
施侃晟，刘海涛，宋文涛. 基于词性和中心点改进的文本聚类方法[J]. 模式识别与人工智能，2012，25(6)：996-1001

[10] Xu Ge, Wang Hou-feng. The Development of Topic Models in Natural Language Processing[J]. Chinese Journal of Computers, 2011, 34(8): 1423-1436(in Chinese)
徐戈，王厚峰. 自然语言处理中主题模型的发展[J]. 计算机学报，2011，34(8)：1423-1436

(上接第 91 页)

[7] Biham E, Biryukov A, Shamir A. Cryptanalysis of Skipjack reduced to 31 rounds using impossible differentials [M]// Advances in Cryptology—Eurocrypt'99. Springer Berlin Heidelberg, 1999: 12-23

[8] Biham E, Shamir A. Differential cryptanalysis of DES-like cryptosystems[C]// Proc of CRYPTO'90. 1991: 2-21

[9] Liu Xuan, Liu Feng, Meng Shuai. Impossible differential cryptanalysis of lightweight block cipher ESF[J]. Computer Engineering & Science, 2013, 35(9): 89-95(in Chinese)
刘宣，刘枫，孟帅. 轻量级分组密码算法 ESF 的不可能差分分析[J]. 计算机工程与科学，2013，35(9)：89-95

[10] Chen Jie, Hu Yu-pu, Zhang Yue-yu. Impossible differential attack on the 17-round block cipher SMS4[J]. Journal of Xidian University, 2008, 35(3): 455-458(in Chinese)
陈杰，胡予濮，张跃宇. 用不可能差分法分析 17 轮 SMS4 算法[J]. 西安电子科技大学学报，2008，35(3)：455-458

[11] Liu Qing, Wei Hong-ru. New Related-key Rectangle Attack on Full ARIRANG Encryption Mode[J]. Computer Science, 2013, 40(8): 109-114(in Chinese)
刘青，卫宏儒. 对完整轮数 ARIRANG 加密模式的新的相关密钥矩形攻击[J]. 计算机科学，2013，40(8)：109-114