一种基于 Simhash 算法的重复域名数据去重方法

侯开茂，韩庆敏，吴云峰，黄　兵，张久发，柴处处
（中国电子信息产业集团有限公司第六研究所，北京 100083）

摘　要： 随着数字科学技术的发展，各领域需要传输和存储的数据量急剧上升。然而传输和存储的数据中重复数量占据了很大的比例，这不仅会增加使用数据的成本，也会影响处理数据的效率。域名是一种存储量大而且对处理速率有极高要求的数据，为了节约域名解析系统的存储成本，提高传输效率，本文在原有数据去重技术的基础上，引入了 Simhash 算法，结合域名数据的结构特征，改进数据分词和指纹值计算方式，提出了一种基于 Simhash 算法的重复域名数据去重方法。实验结果表明，相比于传统的数据去重技术，该方法对删除重复域名数据效率更高，具有较好的实际应用价值。

关键词： 数据去重；域名；Simhash；数据分块
中图分类号： TP391　　**文献标识码：** A　　**DOI：** 10.19358/j.issn.2096-5133.2022.04.011
引用格式： 侯开茂，韩庆敏，吴云峰，等. 一种基于 Simhash 算法的重复域名数据去重方法[J].信息技术与网络安全，2022，41(4)：71-76.

Method for deleting duplicate domain name data based on Simhash algorithm

Hou Kaimao，Han Qingmin，Wu Yunfeng，Huang Bing，Zhang Jiufa，Chai Chuchu
（The 6th Research Institute of China Electronics Corporation，Beijing 100083，China）

Abstract： With the development of digital science and technology，the amount of data that needs to be transmitted and stored in various fields has risen sharply. However，the number of repetitions in these data occupies a large proportion. This not only increases the cost of using data，but also reduces the efficiency of data processing. Domain name is a kind of data with large storage capacity and extremely high requirements for processing speed. In order to save storage cost and improve transmission efficiency，this paper proposes a method for deleting duplicate domain name data based on Simhash algorithm. Compared with the traditional data deduplication technology，this method combines the structural characteristics of the domain name data，and introduces the Simhash algorithm to design a deduplication method for the domain name data. The experimental results show that compared with the traditional data deduplication technology，this method is more efficient in deleting duplicate domain name data and has better practical application value.

Key words： data deduplication；domain name；Simhash；data block

0 引言

随着电子信息技术的发展，各行各业都产生了大量的数据信息，根据国际数据公司（International Data Corporation，IDC）的最新预测：到 2023 年，中国的数据量将达到 40 ZB，并且随着 5G 技术的普及，数据量增长将会迎来又一个新的高潮[1]。有研究发现，这些数据中超过 60%都是重复冗余数据[2]，传输和存储这些冗余数据不仅造成了存储资源和网络资源的严重浪费，也降低了使用数据的效率。并且随着时间推移，这些数据带来的冗余问题会越来越严重。域名[3]（Domain Name）作为互联网中频繁使用的数据类型之一，是一种特殊的数据形式，其对字符的变化敏感度极高，一个字符的变化往往会对使用结果产生严重的影响。因此，处理重复域名数据需要采用精确而且高效的去重技术。

已有重复数据处理技术中，完全文件检测（Whole File Detection，WFD）技术[4]无法对内容进行查重处理，固定分块（Fixed-Sized Partition，FSP）检测技术、可变分块检测技术和滑动块检测技术都是针对数据共有特征的粗粒度去重，直接用于重复域名的处

理效果并不理想。因此,本文在已有重复数据检测技术的基础上,引入 Simhash 算法,结合域名数据的结构特征,改进计算文本特征值的方式,提出了一种基于 Simhash 算法的重复域名数据去重方法。经过实验对比看出,该方法对于处理重复域名数据效果更好,同时在时间开销上也和原有技术差别不大,对于处理重复域名数据具有比传统去重技术更好的实用价值。

1 重复数据检测技术

现有的重复数据检测技术都是通过检测出文件中重复数据并进行删除,只保留唯一的数据对象,然后使用指向此唯一数据对象的指针代替其他重复数据,以达到所有数据都只存储一次的效果。目前主要有完全文件检测技术[4]、固定分块检测技术、可变分块检测技术和滑动块检测技术[5]。

1.1 完全文件检测技术

完全文件检测技术是将目标文件作为检测单位,以文件为粒度进行相同数据查找的方法。该算法首先对整个文件进行 hash 计算,得到一个文件级的 hash 值,然后将计算得到的 hash 值与已经存储的所有 hash 值进行比较,如果与存储 hash 值相同,则可以判断文件数据重复,只需用一个指向已经存储数据的指针替换该文件即可,不必对数据进行实际的存储操作。如果没有相同的值,则将该文件存储为新文件[6]。

1.2 固定分块检测技术

固定分块[7-8]检测技术是将文件划分为同等大小的文件块,并以块为单位进行 hash 计算的检测方法。该算法检测重复数据的主要过程为:(1)定义分块策略,指定一个独立于文件内容的固定文件块大小值,以此值为标准将整个文件分为若干块;(2)对每个文件块进行 hash 运算,得到一个可以唯一标识该块的指纹值;(3)将该值与系统中存储的指纹值进行比较,如果存在相同的值,则可以判断该文件块数据重复,用一个指向已经存储文件块的指针替换该文件块即可,不对该文件块进行实际的存储操作。如果没有相同的值,则新分配一块存储空间存储该文件块。由于该技术对数据变化敏感度很高,文件块中一两个字符的不同都会对检测结果产生极大影响,因此在实际应用中该技术多用于图片、视频等变化比较少的数据查重。

1.3 可变分块检测技术

可变分块检测技术基于内容对文件进行分块,所以分出的文件块大小也不相等。对内容计算最常用的是 CDC(Content-Defined-Chunking)算法,该算法基于 Rabin 指纹[9]对文件内容进行计算,然后根据计算结果进行分块。其过程为:(1)从文件头开始,采用固定大小的滑动窗口对文件内容进行覆盖;(2)在窗口的每个边界采用 Rabin 指纹函数计算出该窗口边界下文件块的指纹值;(3)当指纹值满足某个预定条件(例如某个特殊数字的倍数)时,就将该窗口边界作为文件块的边界;(4)重复以上过程,直到所有文件被划分为不同大小的块。与 FSP 技术不同的是该方法对数据变化不敏感,数据发生变化时也只会影响邻近数据块的计算值,但是分块大小的设定会影响该方法的去重效果。

1.4 滑动块检测技术

滑动块检测技术[10]采用两级处理的方式对文件进行查重处理。该技术首先对文件基于文件粒度进行处理,如果文件级比较值相同,则采用更细粒度的算法进行进一步处理。该方法结合了文件级检测和内容级检测的优点,查重的正确率更高。但是该技术过程繁琐,效率也比较低,处理相同数据量的数据能耗比其他方法高,因此不适合用来处理大规模的数据。

1.5 小结

本节介绍了 4 种常见的重复数据检测技术,对处理过程和算法特征进行了深入分析。在几种算法中,完全文件检测算法以文件为粒度对数据进行检验,检测过程简单,检测效率高,但是不能用于对文件内部数据的重复内容进行检测。固定分块检测技术对数据变化敏感度很高,文件块中一两个字符的不同都会对检测结果产生极大影响,因此不适合处理需要频繁更新的数据。而可变分块技术中的 CDC 算法检测技术弥补了前二者在文件内容处理上的不足,对于相差几个字节的数据可以有很好的检测效果。滑动块检测技术结合了文件级检测和内容级检测的优点,查重的正确率更高,但是,该方法存在计算过程繁琐、效率低下、能耗高等缺点,不适合实时性要求高、成本预算低的文件处理。基于以上对算法适用场景的分析以及域名数据对符号变化敏感的特点,基于 CDC 算法的检测技术最适合与 Simhash 算法结合进行重复域名数据检测,因

此，在后面的内容中，本文将采用基于 CDC 算法的检测技术对数据进行分块去重。

2 基于 Simhash 算法的域名数据去重

传统去重算法检测出的数据往往包含了一些相似而不相同的数据，造成检测结果不理想的情况。这是因为不同的数据文件具有不同的结构特征，采用传统的重复数据检测技术，只能通过统一的方法针对具有共性的部分作出检测，而不能针对具体的数据对症下药。因此，本文在分析域名数据结构特征的基础上，提出了基于 Simhash[11]算法的域名数据去重方法。

2.1 域名数据文件特征分析

数据的语义特征对于查重技术和算法的选择至关重要，不同的查重技术和算法对于不同格式的文件的去重效果有很大的影响。因此，在对数据处理之前，有必要对域名数据的特征进行分析。

域名是互联网上用于解决计算机名称和 IP 地址之间映射关系的一种方法。一个完整的域名由一串由“.”分隔符分隔的字符串组成。域名右边第一个被“.”隔开的字符串代表的是顶级域名(TLD，也称为一级域名)，依次向左分别是二级域名，三级域名……以中华人民共和国教育部官方网站域名(www.moe.gov.cn)为例，“cn”“gov”“moe”依次代表了一级域名、二级域名和三级域名[12]。

每一个域名都唯一对应一个 IP 地址，这种对应关系以资源记录(Resource Records，RR)的形式存储在域名解析文件中。一个资源记录对应一个域名和 IP 地址的映射关系及其他相关信息。图 1 是常用的资源记录格式，其各字段含义如下：

Name：主机域名；

TTL：该记录有效时间；

IN：表示一个标准的 DNS Internet 类；

RR-Type：记录类型，如 A、AAAA、SOA、NS、MX、CNAME 等；

Value：IP 地址。

Name	TTL	IN	RR-Type	Value

图 1 资源记录格式

例如：(www.node3.com IN A 1.1.1.1)代表了一条域名为“www.node3.com”，IP 地址为 1.1.1.1 的 IPv4 资源记录。

2.2 数据分块

通过特征分析发现，域名数据可以以文件为粒度进行重复数据检测，也可以先以资源记录为粒度进行分块，进一步以字段进行分词检测。本文采用先分块再分词的方法进行查重。首先基于 CDC 算法对文件进行分块，再对分块结果按字段分词处理。本小节主要介绍分块的过程，分词的过程将在下一部分介绍。数据分块的流程如图 2 所示，具体过程为：

(1)将整个文件读入临时缓冲区；

(2)创建一个临时列表(List)；

(3)从文件头部开始将缓冲区文件读入临时列表，从 Name 字段开始读取，到 Value 字段结束，停止读入；

(4)将临时列表中的数据写入到列表中，并将临时列表清空；

(5)重复第(3)、(4)两步，直到将整个文件都写入列表。

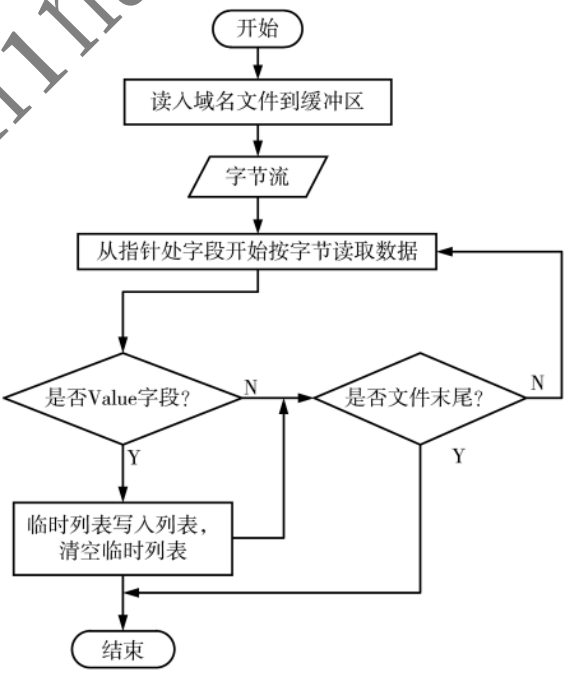

图 2 数据分块流程图

经过以上处理，得到一个列表，列表的每个元素就是一个资源记录，这样就可以以资源记录为粒度对数据进行比较。

2.3 计算 Simhash 值

由于资源记录具有明显的结构特征，通过结构属性进行结构分词比基于语义特征分词效率更高，

因此本文通过结构分词之后再结合 Simhash 算法，从列表中按数据块即资源记录计算指纹值。Simhash 算法以分词结果中关键词为特征值[13]，以关键词出现的频率为权重，各关键词的权重集合作为一个特征向量 N，然后采用 MD5 算法产生一个 m 位的签名 G，再对特征向量加权，对于最终向量的每一位如果大于 0 则为 1，否则为 0，这样就能得到最终的 Simhash 的指纹签名。

在本文中，计算权重采用经典的 TF－IDF(Term Frequency－Inverse Document Frequency)算法，TF 词频指的是关键词在文件中出现的频率，关键词 w_i 的词频 $f_{w_i,j}$ 表示为：

$$f_{w_i,j}=\frac{x_{i,j}}{\sum_k x_{k,j}} \tag{1}$$

其中，$x_{i,j}$ 表示关键词 w_i 在文档 y_j 中出现的次数，$\sum_k x_{k,j}$ 表示所有关键词的词频总和。

计算出权重即词频，指纹值计算的具体过程可以描述为：

(1)确定目标指纹值的维度 m。

(2)初始化向量 $V=(v_1,v_2,\cdots,v_m)$，初始值为 0。

(3)对列表的每一个元素(资源记录)根据属性分词，得到 Name、TTL、IN、RR－Type、Value 五个分词。

(4)采用 MD5[14]算法，针对每一个分词计算出一个数字签名。对于签名的每一位，如果是 1，则用权重和码值相乘；如果是 0，则用权重和码值负相乘，得到每个关键词的特征向量。

(5)将所有特征向量进行如图 3 所示列合并，得到和向量 S，对 S 进行降维处理，对于 S 的每一位，如果大于 0 则为 1，否则为 0，得到的结果就是目标资源记录的 Simhash 指纹值。

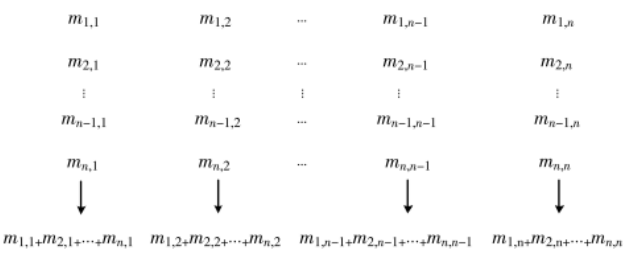

图 3 n 维向量列相加

采用 Simhash 算法计算一条资源记录指纹值的算法过程如图 4 所示。

2.4 数据去重

经过数据分块和指纹值计算，文件里的每一条资源记录都有且只有一个指纹值 P 对其进行了标识。因此，判断数据的重复性问题被简化成了判断指纹值重复的问题。对指纹值 P，将其与已存储的指纹值 P_i 进行异或运算：

$$f(p_i)=P\oplus P_i \tag{2}$$

如果 P 与其中一个存储值异或结果是 0，则说明两数相等，其对应的资源记录与之前存储的资源记录重复，放弃对该条记录的存储。如果 P 与所有存储值异或结果为 1，则说明该指纹值与所有存储值不相等，即该资源记录为新数据，将该数据和指纹值存入系统。重复以上步骤直到列表为空，就完成了对文件的所有数据进行去重处理。

3 实验分析

3.1 实验环境和实验数据

实验仿真环境采用 ASUS 电脑，处理器为 Intel COREi7 8th Gen，8 个核心处理器，8 GB 内存，512 GB 固态硬盘，64 位 Windows 10 系统。以域名数据文件为实验对象，将本文的实验结果与传统数据去重技

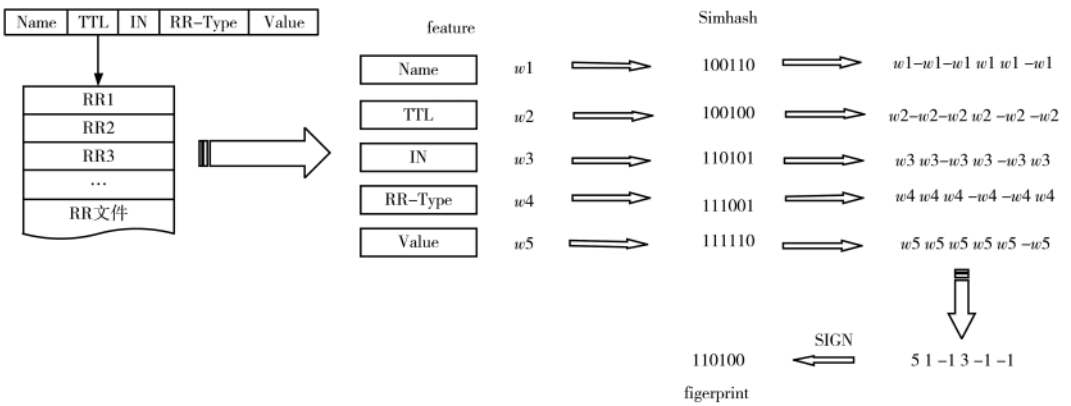

图 4 计算 Simhash 指纹值

术结果进行对比。数据集是从 3 000 万条 CN 顶级域下的域名数据中筛选出的 200 万条无重复域名的数据，将该数据集平均分成 5 组，依次命名为 A 组、B 组、C 组、D 组、E 组。每组包含 40 万条互不重复的域名数据。再将这 200 万条域名数据复制五份，分别命名为 1 组、2 组、3 组、4 组、5 组，然后分别将 A 组数据和 1 组数据组合成一组，命名为 A1 组；将 B 组数据和 2 组数据组合成一组，命名为 B2 组；依次类推，组合出 C3 组、D4 组、E5 组。实验数据准备过程如图 5 所示。经过以上处理，每组数据包含了 240 万条数据，其中有 40 万对是重复数据。实验的目标就是通过设计程序，将 40 万对重复数据中的 40 万条重复数据进行查找和删除。

3.2 实验参数选取

重复数据删除率和执行算法所耗时长是评价数据去重性能的两个重要指标。为了评价本文介绍的方法在实际应用中的性能，本文设计了一系列实验，从重复数据删除率和算法执行时间两个维度与传统的数据去重算法进行对比实验。

3.2.1 重复数据删除率

重复数据删除率[15] R_i 定义为：

$$R_i=\frac{\sum_{1}^{n} D(i)-E(i)}{\sum_{1}^{n} D(i)} \tag{3}$$

其中，$D(i)$表示文件中重复的数据量，$E(i)$表示实验过后还剩的重复数据量。

通过图 6 的实验结果可以看出，本文采用的算法在域名数据去重效率上相比于其他传统的去重算法效果更好。

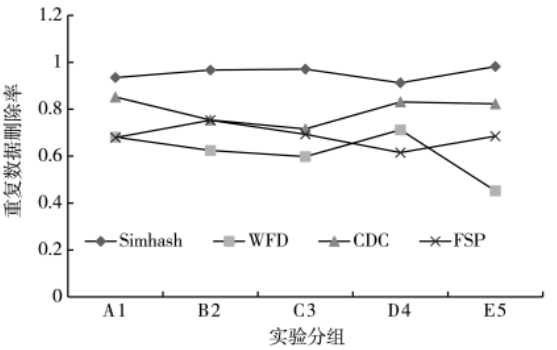

图 6 算法重复数据删除率对比

3.2.2 算法执行时间

本文对各算法执行时间做了对比，结果如图 7 所示。

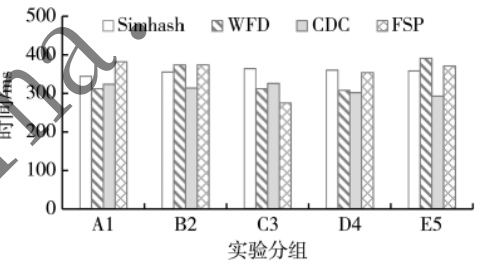

图 7 算法执行时间对比

通过对比可以看出，本文所研究的算法执行时间稍微多于其他算法，这可能与本文在处理域名数据时，处理粒度较细有关。但是本文介绍的方法时间上比其他几种算法更稳定，这说明该算法

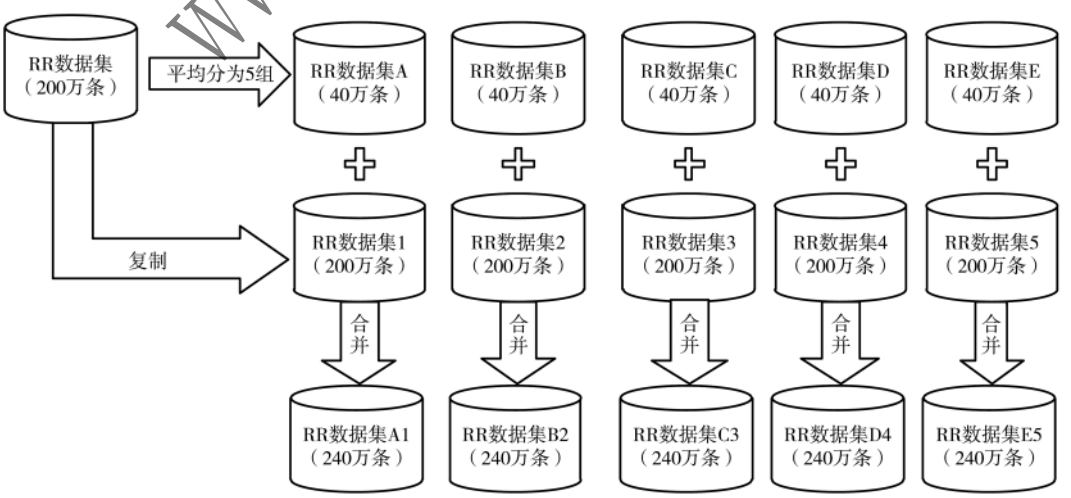

图 5 数据分组

对数据变化敏感度不大，更适合处理需要频繁更新的数据。

4 结论

本文在传统数据去重技术的基础上，引入 Simhash 算法，结合特定数据文件类型——域名数据的特征，对数据分块的过程以及特征值指纹计算方法进行了改进。一方面，针对域名数据的结构化特点，采用了基于 CDC 算法可变分块的数据分块方法对文件进行分块；另一方面，基于域名数据资源记录的字段特征，引入了基于 Simhash 算法的指纹值计算方法。相比于传统去重技术笼统的去重方式，该方法专门针对域名数据的去重问题进行研究，提高了检测重复域名数据的效率。在未来的工作中，一方面继续优化算法，减少算法的执行时间；另一方面，可以将这种有的放矢的方式应用到其他数据领域，使得数据去重技术体现出更好的实用价值。

参考文献

[1] 栗翘楚.分布式存储打开千亿级市场 深入推动行业数字化转型[EB/OL].(2021-03-25).http://m.people.cn/n4/2021/0325/c125-14907855.html.

[2] MCKNIGHT J, ASARO T, BABINEAU B. Digital archiving: end user survey and market forecast 2006-2010[EB/OL].[2021-12-22].http://www.enterprisestrategygroup.com/ESGPublications/ReportDetail.asp?ReportID=591.

[3] LIU C, ALBITZ P.DNS 与 BIND(5th ed)[M].房向明，译.北京：人民邮电出版社，2014.

[4] BOLOSKY B, CORBIN S, GOEBEL D, et al. Single instance storage in Windows 2000[C]//Proceedings of the 4th USENIX Windows System Symposium.Berkeley: USENIX Association, 2000: 13-24.

[5] POLICRONIADES C, PRATT I.Alternatives for detecting redundancy in storage systems data[C]//Proceedings of USENIX Technical Conference. Berkeley, CA, USA: USENIX Association, 2004.

[6] 敖莉，舒继武，李明强.重复数据删除技术[J].软件学报，2010，21(5)：916-929.

[7] 彭双和，图尔贡·麦提萨比尔，周巧凤.基于 Simhash 的中文文本去重技术研究[J].计算机技术与发展，2017，27(11)：137-140，145.

[8] BOBBARJUNG D R, JAGANNATHAN S, DUBNICKI C. Improving duplicate elimination in storage systems[J]. ACM Transactions on Storage, 2006, 2(4): 424-448.

[9] BHAGWAT D, POLLACK K, LONG D D E, et al. Providing high reliability in a minimum redundancy archival storage system[C]//Proc. of the 14th Int'l Symp. on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems(MASCOTS 2006). Washington: IEEE Computer Society Press, 2006: 413-421.

[10] DENEHY T E, HSU W W. Duplicate management for reference data[R]. IBM Research Division, 2003.

[11] FU Z J, SUN X M, LIU Q, et al. Achieving effective cloud search services: multi-keyword ranked search over encrypted cloud data supporting synonym query[J]. IEEE Transactions on Consumer Electronics, 2014, 60(1): 164-172.

[12] 周汝佳.基于语义指纹和 Simhash 的文本去重方法研究[D].南昌：江西财经大学，2021.

[13] 陈春玲，陈琳，熊晶，等.基于 Simhash 算法的重复数据删除技术的研究与改进[J].南京邮电大学学报(自然科学版)，2016，36(3)：85-91.

[14] ZHANG Z P, XU X, LONG J, et al. Parameters correlation and optimization in text similarity measuremen[J]. Journal of Chinese Computer Systems, 2011, 32(5): 983-988.

[15] 陈丹伟，唐平，周书桃.基于沙盒技术的恶意程序检测模型[J].计算机科学，2012，39(S1)：12-14.

(收稿日期：2022-01-20)

作者简介：

侯开茂(1997-)，男，硕士研究生，主要研究方向：计算机网络、DNS 域名系统等。

韩庆敏(1979-)，女，硕士，正高级工程师，主要研究方向：工业软件、自动化控制系统、智能制造、现场总线、网络标识分析技术、工业互联网安全等。

吴云峰(1977-)，男，硕士，正高级工程师，主要研究方向：安全可信编程编译技术、工控安全威胁监测与溯源技术、网络标识分析技术、互联网域名安全等。

版权声明

经作者授权，本论文版权和信息网络传播权归属于《信息技术与网络安全》杂志，凡未经本刊书面同意任何机构、组织和个人不得擅自复印、汇编、翻译和进行信息网络传播。未经本刊书面同意，禁止一切互联网论文资源平台非法上传、收录本论文。

截至目前，本论文已经授权被中国期刊全文数据库（CNKI）、万方数据知识服务平台、中文科技期刊数据库（维普网）、JST 日本科技技术振兴机构数据库等数据库全文收录。

对于违反上述禁止行为并违法使用本论文的机构、组织和个人，本刊将采取一切必要法律行动来维护正当权益。

特此声明！

《信息技术与网络安全》编辑部

中国电子信息产业集团有限公司第六研究所