基于数据摘要奇偶性的集合相似性近似算法

贾建伟 陈 峻
(扬州大学信息工程学院 扬州 225002)

摘 要 在应用 b 位哈希函数近似计算两个集合的 Jaccard 相似性时,如果有多个元素与输入元素的 Jaccard 相似性都很高(接近于 1),那么 b 位哈希函数不能对这些元素进行很好的区分。为了提高数据摘要函数的准确性并提高基于相似性的应用的性能,提出了一种基于数据摘要奇偶性的集合相似性近似算法。在应用 minwise 哈希函数得到两个变异集合后,用两个 n 位指示向量来表示变异集合中的元素在指示向量中出现的奇偶性,并基于这两个奇偶性向量来估计原集合间的 Jaccard 相似性。通过马尔科夫链和泊松分布两种模型对奇偶性数据摘要进行了推导,并证明了这两种方法的等价性。Enron 数据集上的实验表明,提出的奇偶性数据摘要算法与传统的 b 位哈希函数相比具有更高的准确性,并且在重复文档检测和关联规则挖掘两种应用中具有更高的性能。

关键词 数据摘要,集合相似性,奇偶性,近似算法
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.6.050

Set Similarity Approximation Algorithm Based on Parity of Data Sketch

JIA Jian-wei CHEN Ling
(College of Information Engineering, Yangzhou University, Yangzhou 225002, China)

Abstract Jaccard similarity is one of the most important methods in set similarity computation. When approximately computing the Jaccard similarity of two sets using the b-bits hash function, if there are multiple elements being similar to the input element with similarity up to 1, the b-bits hash function can't differentiate these elements very well. In order to improve the accuracy of data sketch and the application performance based on set similarity, this paper proposed a set similarity approximation algorithm based on parity of data sketch. After getting the two permutation sets with minwise hash function, we used two n-bits indicator vectors to represent the parity of elements in the permutation set appearing in indicator vectors, and estimated the Jaccard similarity of original sets based on these two parity vectors. We inferred the parity sketch based on both Markov chain and Poisson distribution models, and verified their equivalence. Experiments on Enron dataset show that the proposed parity sketch is more accurate than the b-bits hash function, and performs much better in both applications of duplicate document detection and associate rule mining.

Keywords Data sketch, Set similarity, Parity, Approximation algorithm

集合的相似性计算[1]是数据库、机器学习和信息检索等研究领域的基础研究内容之一。在 Web 文档中,每个文档可以看作一个集合,集合中的元素为该文档所包含的单词或者词组。在推荐系统中,每一个用户或者项都可以看作一个集合,而该用户或者项对应的所有数据即为集合中的元素。集合的相似性估计技术可以广泛应用于 Web 页面的重复检测[2,3]、协同过滤技术[4,5]以及关联规则挖掘[6,7]等。

Jaccard 系数是集合的相似性计算所采用最广泛的方法之一。给定两个已知的集合 $S_1, S_2 \in \Omega=\{0,1,\cdots,D-1\}$,那么它们的 Jaccard 系数为 $J(S_1,S_2)=|S_1 \cap S_2|/|S_1 \cup S_2|$。然而,随着数据集规模的不断增大,基于集合相似性的搜索将花费大量的计算资源,并且需要很长的运行时间。为了减小集合相似性计算的计算量,研究人员[8,9]通常将集合 S 用一种摘要的数据结构 $D(S)$ 来表示,并在集合的摘要上进行相似性的近似计算。

如果摘要函数为 $D(\cdot)$,那么 S_1 和 S_2 的 Jaccard 相似性可近似表示为 $J(S_1,S_2) \approx J(D(S_1),D(S_2))$。在 minwise 哈希函数 $\min(\cdot)$[10] 下,如果 $\pi:\Omega \to \Omega$ 为随机变异函数[11],那么 S_1 和 S_2 的 Jaccard 相似性为:

$$J(S_1,S_2)=\Pr[\min(\pi(S_1))=\min(\pi(S_2))] \quad (1)$$

经过 k 次随机变异后,可以得到如下的两个集合:

$$S_j=\{(i,\min(\pi_i(S_j)))\mid i=1,\cdots,k;j=1,2\} \quad (2)$$

通过计算相等变异的比例 $\hat{J}=|S_1 \cap S_2|/k$ 便可以得到 J 的估计值。如果采用 32 位或者 64 位大小来保存每个 minwise 哈希值,那么每个集合所需的存储空间为 32k 或者 64k。为了进一步提高空间的利用效率,Li 和 König[12] 提出了一种 b 位 minwise 哈希函数。b 位 minwise 哈希函数的基本思想是:如果两个元素的哈希值相同,那么它们最低的 b 位也是相

到稿日期:2015-06-04 返修日期:2015-07-24 本文受国家自然科学基金面上项目(61070240)资助。
贾建伟(1974—),男,硕士,讲师,主要研究方向为数据挖掘、计算机技术应用及计算物理;陈 峻(1951—),男,教授,博士生导师,主要研究方向为数据挖掘和人工智能。

同的；如果两个元素的哈希值不同，那么它们的最低 b 位不同的概率为 $1-1/2^b$。

在采用 Jaccard 相似性的搜索应用中，如果有多个元素与输入元素的 Jaccard 系数都很高（接近于 1），那么 b 位 minwise 哈希函数不能对这些元素进行很好的区分。为此，本文基于一种固定长度的奇偶性摘要来计算两个集合的 Jaccard 相似性。应用 minwise 哈希函数得到两个变异集合后，用两个 n 位指示向量来表示变异集合中的元素在指示向量中出现的奇偶性，并基于这两个奇偶性向量来估计原集合间的 Jaccard 相似性。

1 集合相似性近似计算

1.1 奇偶性数据摘要构建

本文提出一种奇偶性数据摘要构建方法，该方法的示意图如图 1 所示。当计算$\mathbb{S}_1$和$\mathbb{S}_2$的 Jaccard 相似性时，首先根据式(1)得到相应的集合 S_1 和 S_2。接下来，应用函数 $odd(\cdot)$分别将 S_1 和 S_2 映射为两个 n 位数据摘要 $odd(S_1)$和 $odd(S_2)$。在摘要映射之前，令 $odd(S_1)$和 $odd(S_2)$的所有位为 0。在摘要的映射过程中，对于 S_1 和 S_2 的共有部分 x，将 $odd(S_1)$和 $odd(S_2)$的对应位置为 1；对于 S_1 和 S_2 的独有部分 x'和 x''，分别置对应的位为 1。

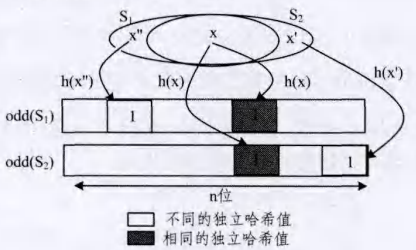

图 1 odd 摘要构建示意图

对于集合 S，$odd(S)=s$ 是 S 的一个长度为 n 位的线性指示向量。令 $h:\Omega\to[n]$为完全随机的哈希函数，那么在 $odd(\cdot)$函数下，s 的第 i 位为

$$s_i=\bigoplus_{x\in S}1_{h(x)=i} \tag{3}$$

其中，$\oplus$为异或操作符，$0\leqslant i<n$。$odd(\cdot)$摘要的算法如算法 1 所示。首先，令 n 位向量 s 的所有位都为 0；接下来，选取一个随机哈希函数 h；最后，对于 S 中的每个元素 x，根据函数 $h(x)$对 $s_{h(x)}$的奇偶性进行翻转。

算法 1 $odd(S,n)$

输入：集合 S，哈希值的长度 n

1. 初始化 s←$[0]^n$；
2. 随机选取哈希函数 h：$\Omega\to[n]$；
3. $\forall x\in S, s_{h(x)}=s_{h(x)}\oplus 1$；
4. 返回 s。

由于 $odd(S,n)$记录了 S 中的元素哈希到 s 中某个具体位的元素个数的奇偶性，因此 S_1 和 S_2 的对称差奇偶性摘要等于 S_1 和 S_2 的奇偶性摘要的异或，即

$$odd(S_1\Delta S_2)=odd(S_1)\oplus odd(S_2) \tag{4}$$

其中，Δ为对称差操作符。

1.2 原始数据量估计

用 m 表示集合 S 含有的元素个数，用 n 表示奇偶性摘要 $s=odd(S)$的位数。由于哈希函数 h 是完全随机的，因此 $odd(S)$的构建过程等价于将 m 个球独立地投入到 n 个篮子中，并且用 s_i 记录第 i 个篮子中包含的球的奇偶个数。下面分析如何从观察到的 $odd(S)$的奇偶性数据中分析出 m 的估计值$\hat{m}$。

首先，应用马尔科夫链模型基于奇偶性摘要数据对原始的数据量进行估计。假设 n 个篮子中任意一个包含的球数的奇偶性为一个具有两个状态的马尔科夫链模型，这两个状态分别为奇数和偶数，那么状态转换的概率为 $1/n$。当随机投出 i 个球后，每个篮子含有奇数个球的概率为 p_i，在马尔科夫链模型下通过归纳可得：

$$p_i=\frac{1-(1-2/n)^i}{2} \tag{5}$$

在投出 m 个球后，令 $X_i\in\{0,1\}$为第 i 个篮子包含的球的个数的奇偶性变量，那么 $X=\sum_{0\leqslant i<n}X_i$ 的数学期望为

$$E[X]=n\frac{1-(1-2/n)^m}{2} \tag{6}$$

经过进一步推导后可以得到 m 的估计值：

$$\hat{m}=\frac{\ln(1-2E[X]/n)}{\ln(1-2/n)} \tag{7}$$

接下来，应用泊松分布模型对原始的数据量进行估计。将 m 个球独立地投入到 n 个篮子中可以近似看作每个篮子都含有均值为 $\mu=m/n$ 的服从泊松分布的球数，并且这些篮子之间是相互独立的。引理 1 给出了均值为 μ 的泊松分布与奇偶性分布之间的关系。

引理 1[13] 令 Q 为服从均值为 μ 的泊松分布的随机变量，那么 Q 为奇数的概率为 $p=(1-e^{-2\mu})/2$。

证明：Q 为奇数的概率为

$$p=\sum_{i\text{为奇数}}\frac{e^{-\mu}\mu^i}{i!}=e^{-\mu}\sum_{i\text{为奇数}}\frac{\mu^i}{i!}$$

$$=e^{-\mu}\frac{e^{\mu}-e^{-\mu}}{2}=\frac{1-e^{-2\mu}}{2}$$

在均值为 $\mu=m/n$ 的泊松分布下，第 i 个篮子中含有的球数的奇偶性为 Y_i，并且 Y_i 与 Y_j 是相互独立的$(i\neq j)$，那么含有奇数个球的篮子的总数为 $Y=\sum_{0\leqslant i<n}Y_i$，并且 Y 的数学期望为：

$$E[Y]=np=n\frac{1-e^{-2m/n}}{2} \tag{8}$$

经过进一步推导后可以得到 m 的估计值：

$$\hat{m}=-\frac{n}{2}\ln(1-2E[Y]/n) \tag{9}$$

当函数 $odd(\cdot)$的位数 n 足够大时，根据自然对数的定义$\lim_{n\to\infty}(1+1/n)^n=e$，可得 $\ln(1-2/n)=-2/n$。在式(7)和式(9)中，当 $E[X]$和 $E[Y]$取相同的观测值 z 时，式(7)和式(9)等价为

$$\hat{m}=-\frac{n}{2}\ln(1-2z/n) \tag{10}$$

1.3 Jaccard 相似性估计

上文介绍了如何根据集合$\mathbb{S}_1$和$\mathbb{S}_2$推导相应的 minwise 哈希值 S_1 和 S_2，以及如何利用 S_1 和 S_2 构建奇偶性哈希值 $odd(S_1)$和 $odd(S_2)$。下面介绍如果根据 $odd(S_1)$和$odd(S_2)$估计$\mathbb{S}_1$和$\mathbb{S}_2$的 Jaccard 相似性。

在构建集合 S_1 和 S_2 时，如果 J 为$\mathbb{S}_1$和$\mathbb{S}_2$的 Jaccard 相似性，那么 S_1 和 S_2 的对称差含有的元素个数的数学期望为

$$E[|S_1\Delta S_2|]=2k(1-J) \tag{11}$$

其中，k 为式(2)所示的随机变异的次数。在根据 S_1 和 S_2 得到 $odd(S_1)$ 和 $odd(S_2)$ 后，可以根据 $odd(S_1)$ 和 $odd(S_2)$ 对 $|S_1\Delta S_2|$ 的值进行估计。如果用 $|S_1\hat{\Delta}S_2|$ 表示 $|S_1\Delta S_2|$ 的估计值，那么有

$$|S_1\hat{\Delta}S_2|=-\frac{n}{2}\ln(1-2\frac{|odd(S_1)\Delta odd(S_2)|}{n}) \tag{12}$$

其中，$|odd(S_1)\Delta odd(S_2)|$ 表示 $odd(S_1)$ 和 $odd(S_2)$ 的对称差中位值为 1 的个数。如果用 $|S_1\hat{\Delta}S_2|$ 来代替 $E[|S_1\Delta S_2|]$，那么 S_1 和 S_2 的 Jaccard 相似性为

$$\hat{J}^{odd}=1-\frac{|S_1\hat{\Delta}S_2|}{2k}$$

$$=1+\frac{n}{4k}\ln(1-\frac{2|odd(S_1)\Delta odd(S_2)|}{n}) \tag{13}$$

2 实验结果与分析

实验采用的数据集为 Enron 电子邮件数据集，该数据集的详细信息参见文献[14]。在算法的性能对比中，将本文提出的奇偶性哈希函数记为 Odd，并与经典的 b 位 minwise 哈希函数[12]进行了对比。在评价指标中，分别选取了 MSE (Mean Square Error)、准确率和召回率。

2.1 参数估计

集合 S_1 和 S_2 的相似性为 J，令 $\alpha=2k(1-J)/n$ 为 S_1 和 S_2 中不相同的元素所占的比例。当参数 α 改变时，摘要的空间利用率也随之发生变化，通过调整 α 的值来观察 Odd 算法的平均 MSE。首先，固定 $J=0.9$ 并调整 n 在(500,1000)中的取值，观察 Odd 算法的平均 MSE 的变化情况，结果如图 2 所示。接下来，固定 $n=800$ 并调整 J 在(0.75,0.95)中的取值，实验结果如图 3 所示。从图 2 和图 3 中可以看出，当 α 在 0.5 附近取值时，Odd 算法的平均 MSE 达到了最小值。

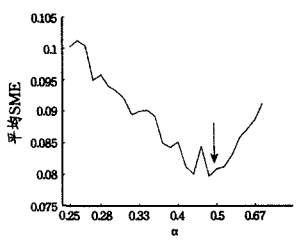

图 2 平均 MSE 随着 α 的变化 ($J=0.9, n\in(500,1000)$)

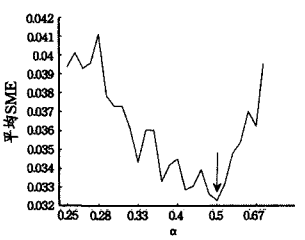

图 3 平均 MSE 随着 α 的变化 ($J\in(0.75,0.95), n=800$)

2.2 准确性评估

在评估算法的准确性时，将 Odd 算法与 b 位 minwise 哈希函数进行了对比。在 b 位 minwise 哈希函数中，分别选取了 $b=1,2$ 和 4；在 Odd 算法中，选取了 $n=512$ 和 1024 两种情况。图 4 和 5 分别为在 $n=512$ 和 1024 两种情况下算法的误差随着 Jaccard 相似性的变化情况。在此处，误差的衡量采用了 $-\log(MSE)$。从图 4 和图 5 中可以看出：在 b 位 minwise 哈希函数中，随着 b 的取值的减小，算法的误差逐渐减小；当 Jaccard 相似性超过 0.75 后，Odd 算法的 $-\log(MSE)$ 值始终处于其它方法的上方，这表明 Odd 算法在 $J>0.75$ 时具有最小的误差。

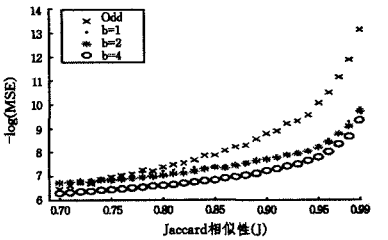

图 4 算法的误差对比($n=512$)

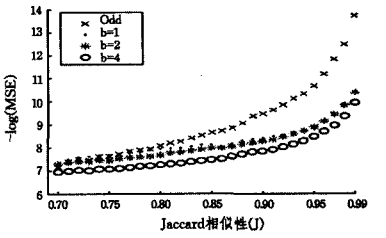

图 5 算法的误差对比($n=1024$)

2.3 应用性能评估

通过重复文档检测和关联规则挖掘两个具体的应用场景对比了 Odd 算法和 b 位 minwise 哈希函数的性能。

在重复文档检测中，令 b 位 minwise 哈希函数中的 b 取 $\frac{1}{2}$ 和 1 两种情况，分别对比了算法的准确率和召回率，实验结果分别如图 6－图 9 所示。从这 4 幅图中可以看出，不论 b 的取值为 $\frac{1}{2}$ 还是 1，Odd 算法的准确率和召回率都明显高于 b 位 minwise 哈希函数，这表明 Odd 算法与 b 位 minwise 哈希函数相比具有更好的重复文档检测性能。

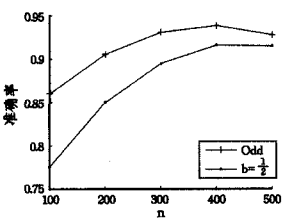

图 6 算法在重复文档检测中的准确率对比($b=\frac{1}{2}$)

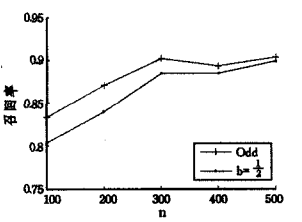

图 7 算法在重复文档检测中的召回率对比($b=\frac{1}{2}$)

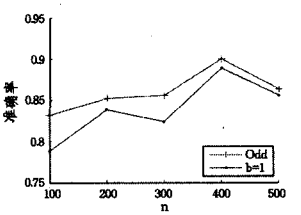

图 8 算法在重复文档检测中的准确率对比($b=1$)

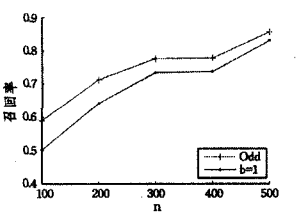

图 9 算法在重复文档检测中的召回率对比($b=1$)

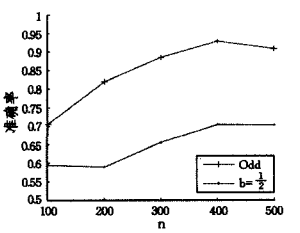

图 10 算法在关联规则挖掘中的准确率对比

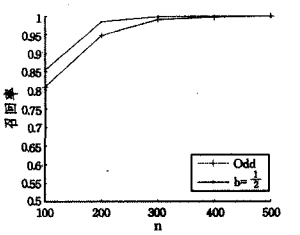

图 11 算法在关联规则挖掘中的召回率对比

(下转第 311 页)

参 考 文 献

[1] Haritaoglu I, Harwood D, Davis LS, et al. Real-time Surveillance of People and Their Activities[J]. Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(1): 809-830

[2] Doll'ar P, Wojek C, Schiele B, et al. Pedestrian detection: an evaluation of the state of the art[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2012, 34(4): 743-761

[3] Jia Hui-xing, Zhang Yu-jin. Summary of driver assistance systems based on computer vision pedestrian detection[J]. Automation Journal, 2007, 33(1): 84-90(in Chinese)
贾慧星，章毓晋. 车辆辅助驾驶系统中基于计算机视觉的行人检测研究综述[J]. 自动化学报，2007，33(1)：84-90

[4] Chen Yi-ru. Research and implementation of pedestrian detection algorithm based on vision[D]. Hangzhou: Zhejiang University, 2014(in Chinese)
陈益如. 基于视觉的行人检测算法的研究与实现[D]. 杭州：浙江大学，2014

[5] Wang Ning-bo. Pedestrian detection based on RGB-D[D]. Hangzhou: Zhejiang University, 2013(in Chinese)
王宁波. 基于 RGB-D 的行人检测[D]. 杭州：浙江大学，2013

[6] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]// Proc. IEEE CVPR. 2005: 886-893

[7] Dollar P, Wojek C, Schiele B, et al. Pedestrian detection: A benchmark[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2009: 304-311

[8] Benenson R, Omran M, Hosang J, et al. Ten Years of Pedestrian Detection, What Have We Learned? [M]// Computer Vision-ECCV 2014 Workshops. 2015: 613-627

[9] Wang Cheng-liang, Zhou Jia, Huang Sheng. Fast moving human detection based on Gauss mixture model and PCA-HOG[J]. Computer Application Research, 2012, 29(6): 2156-2160(in Chinese)
汪成亮，周佳，黄晟. 基于高斯混合模型与 PCA-HOG 的快速运动人体检测[J]. 计算机应用研究，2012，29(6)：2156-2160

[10] Wang X, Han T X, Yan S. An HOG-LBP human detector with partial occlusion handling[C]// 2009 IEEE 12th International Conference on Computer Vision. IEEE, 2009: 32-39

[11] Maji S, Berg A, Malik J. Classification using inter section kernel SVMs is efficient[C]// Proc. IEEE CVPR. 2008: 1-8

(上接第 256 页)

在关联规则挖掘中，令 b 位 minwise 哈希函数中的 b 为 $\frac{1}{2}$，分别对比了算法的准确率和召回率，实验结果如图 10 和图 11 所示。从这两幅图中可以看出，当 $b=\frac{1}{2}$ 时，Odd 算法在关联规则挖掘时的准确率和召回率都高于 b 位 minwise 哈希函数。

结束语 集合的相似性计算是数据库、机器学习和信息检索等研究领域的基础研究内容之一。为了提高数据摘要函数的准确性及基于相似性的应用性能，本文提出了一种基于数据摘要奇偶性的集合相似性近似算法。首先，应用 minwise 哈希函数得到两个变异集合；然后，用两个 n 位指示向量来表示变异集合中的元素在指示向量中出现的奇偶性，并基于这两个奇偶性向量来估计原集合间的 Jaccard 相似性；最后，通过马尔科夫链和泊松分布两种模型对奇偶性数据摘要进行了推导，并证明了这两种方法的等价性。Enron 数据集上的实验表明，本文提出的奇偶性数据摘要算法与传统的 b 位哈希函数相比具有更高的准确性，并且在重复文档检测和关联规则挖掘两种应用下具有更高的性能。

参 考 文 献

[1] Arasu A, Ganti V, Kaushik R. Efficient exact set-similarity joins [C]// Proceedings of the 32nd International Conference on Very Large Data Bases. VLDB Endowment, 2006: 918-929

[2] Xiao C, Wang W, Lin X, et al. Efficient similarity joins for near-duplicate detection[J]. ACM Transactions on Database Systems (TODS), 2011, 36(3): 15-20

[3] Manku G S, Jain A, Das Sarma A. Detecting near-duplicates for Web crawling[C]// Proceedings of the 16th International Conference on World Wide Web. ACM, 2007: 141-150

[4] Zhao Qin-qin, Lu Kai, Wang Bin. SPCF: A Memory Based Collaborative Filtering Algorithm via Propagation[J]. Chinese Journal of Computers, 2013, 36(3): 671-676(in Chinese)
赵琴琴，鲁凯，王斌. SPCF：一种基于内存的传播式协同过滤推荐算法[J]. 计算机学报，2013，36(3)：671-676

[5] Koren Y, Bell R. Advances in collaborative filtering[M]. Recommender systems handbook. Springer US, 2011: 145-186

[6] Li Shu-kui. Research on Time series similarity problem [D]. Wuhan: Huazhong University of Science and Technology, 2008 (in Chinese)
李俊奎. 时间序列相似性问题研究 [D]. 武汉：华中科技大学，2008

[7] Feng Yu-cai, Feng Jian-lin. Incremental Updating Algorithms for Mining Association Rules[J]. Journal of Software, 1998(4): 301-306(in Chinese)
冯玉才，冯剑琳. 关联规则的增量式更新算法[J]. 软件学报，1998，9(4)：301-306

[8] Stein B. Principles of hash-based text retrieval[C]// Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2007: 527-534

[9] 凌康. 基于位置敏感哈希的相似性搜索技术研究[D]. 南京：南京大学，2012

[10] Indyk P. A small approximately min-wise independent family of hash functions[J]. Journal of Algorithms, 2011, 38(1): 84-90

[11] Broder A Z, Charikar M, Frieze A M, et al. Min-wise independent permutations [J]. Journal of Computer and System Sciences, 2010, 60(3): 630-659

[12] Li P, König C. b-Bit minwise hashing[C]// Proceedings of the 19th International Conference on World Wide Web. ACM, 2010: 671-680

[13] Schuster E F, Philippou A N. The odds in some odd-even games [J]. American Mathematical Monthly, 1975, 82(6): 646-648

[14] Shetty J, Adibi J. The Enron email dataset database schema and brief statistical report[J]. Information Sciences Institute Technical Report, University of Southern California, 2004, 4(4): 210-215

[15] Yu Xiao-sheng, Hu Sun-zhi. Research on Eliminating Duplicate Records Based on SNM Improved Algorithm [J]. Journal of Chongqing University of Technology(Natural Science), 2016, 30(4): 91-96(in Chinese)
余肖生，胡孙枝. 基于 SNM 改进算法的相似重复记录消除[J]. 重庆理工大学学报(自然科学)，2016，30(4)：91-96