最优分数位 minwise 哈希算法的研究

袁鑫攀 龙 军 张祖平 罗跃逸 桂卫华

(中南大学信息科学与工程学院 长沙 410083)

摘 要 在信息检索中,minwise 哈希算法用于估值集合的相似度;b 位 minwise 哈希算法则通过存储哈希值的 b 位来估算相似度,从而节省了存储空间和计算时间。分数位 minwise 哈希算法对各种精度和存储空间需求有着更加广泛的可选择性。对于给定的分数位 f,构建 f 的方式有很多。分析了有限的分数位组合方式,给出最优化分数位的理论分析。大量的实验验证了此方法的有效性。

关键词 相似度估值,哈希,最优分数位

中图法分类号 TP301.6 **文献标识码** A

Research on Optimal Fractional Bit Minwise Hashing

YUAN Xin-pan LONG Jun ZHANG Zu-ping LUO Yue-yi GUI Wei-hua

(School of Information Science and Engineering, Central South University, Changsha 410083, China)

Abstract In information retrieval, minwise hashing algorithm is often used to estimate similarities among documents, and b-bit minwise hashing is capable of gaining substantial advantages in terms of computational efficiency and storage space by only storing the lowest b bits of each(minwise) hashed value(e.g., $b=1$ or 2). Fractional bit minwise hashing has a wider range of selectivity for accuracy and storage space requirements. For the fixed fraction f, there are so many combinations of f. We theoretically analyzed limited combinations of fractional bit. The optimal fractional bit was found. Experimental results demonstrate the effectiveness of this method.

Keywords Similarity estimation, Hasing, Optimal fractional bit

1 引言

随着 Web 信息爆炸性增长,海量网页中存在超大量的相似信息,这些相似性文档一方面消耗了高额的检索资源,另一方面影响了用户的使用。与此同时,电子学术资源获取的便利及电子资源本身简单的"复制"、"粘贴"功能,为学术资源的抄袭、非法扩散等不道德行为提供了方便。

本文的研究背景是有大量项目的某基金对相似性检测[1]的需求,即避免一个项目多人申报、一个项目多年重复申报和窃取他人项目成果等不良现象发生。随着每年申报项目数量的剧增,项目文本集合大小将达数十万。随着每年申报项目的数量的剧增,项目文本集合大小将达数百万。对于这种大规模文本集合,Broder 等[2]的 minwise 哈希算法是一种相对成熟、性能相对稳定的文档相似性度量技术。minwise 哈希算法将集合的求交集问题转换为某一个事件发生的概率问题,即通过大量的实验次数 k(样本大小)来估计事件的发生概率,从而估计文档的相似率。作为估值集合相似度的常用方法,minwise 算法后来被大多数的文本相似性度量技术所借鉴,被广泛用于网页去重[3]、无线传感器网络[4]、网络社区分类[5]、文本重用[6]、连接图压缩[7]等领域。minwise 哈希算法也有了相当多的理论和实验方法的创新和发展[8,9]。

2010 年,Li 等[10]的 b 位 minwise 在 minwise 哈希算法的基础上将 $b=64$ 缩小到 1 位,降低了存储空间和计算时间。b 位 minwise 哈希在 3 者相似性检测、大型机器学习、基于最大似然估计(MLE)改进的估计算法上有了新的理创新和应用发展[11]。分数位 minwise 哈希算法[12]对各种精度和存储空间需求有着更加广泛的可选择性。本文找到了构建分数位的最小方差组合,建立了最优分数位 minwise 哈希,完善了 minwise 的理论体系。

1.1 minwise 哈希算法

通过 shingling 每一个文档得到相关集合 S_d。给定两个集合 S_1, S_2。S_1, S_2 包含于 $\{0,1,2,\cdots,D-1\}$。相似度 $R(1,2)$ 表示为:

$$R=\frac{|S_1\cap S_2|}{|S_1\cup S_2|}=\frac{a}{f_1+f_2-a},\ f_1=|S_1|,\ f_2=|S_2|,$$

$$a=|S_1\cap S_2|$$

计算两篇文档的相似度就是计算两个 shingles 集合的交集 a,假定一个在 Ω 上的随机置换群 π。$\pi:\Omega\to\Omega,\Omega=\{0,1,$

到稿日期:2011-10-09 返修日期:2012-01-20 本文受国家自然科学基金项目(M0921005,60873081,60970095,61003033),湖南省杰出青年基金(11JJ1012),教育部新世纪优秀人才支持计划(NCET-10-0787)资助。

袁鑫攀(1982—),博士生,主要研究方向为信息检索、数据挖掘,E-mail:xpyuanfly@163.com;龙 军(1972—),男,博士,副教授,主要研究方向为网格计算、分布式系统(通信作者);张祖平(1966—),男,博士,教授,博士生导师,主要研究方向为信息融合与信息系统、参数计算与生物计算、网络容错路由算法及协议;罗跃逸(1986—),男,硕士生,主要研究方向为数据挖掘;桂卫华(1950—),男,博士,教授,博士生导师,主要研究方向为复杂过程的建模与控制、大系统控制理论与应用。

…,$D-1$},定义相似度 R：

$$R=\frac{|S_1\cap S_2|}{|S_1\cup S_2|}=\Pr(\min\{\pi(S_1\}=\min\{\pi(S_2)\} \tag{1}$$

通过 k 个置换群，得到 R 的无偏估计：

$$\overset{\vee}{R}_M=\frac{1}{k}\sum_{j=1}^{k}1\{\min(\pi_j(S_1))=\pi_j(S_2))\} \tag{2}$$

$$Var(\overset{\vee}{R}_M)=\frac{1}{k}R(1-R) \tag{3}$$

式中，k 表示样本大小(或者实验次数)。

1.2 b 位 minwise 哈希算法

b 位 minwise 哈希算法[2]在 minwise 哈希算法的基础上进行了改进。通过使用更少的位数 b，来估算相似度 R。定义 z_1,z_2 是一个随机置换群 π 作用在集合 S_1 和 S_2 后的最小值：

$$z_1=\min\{\pi(S_1)\},z_2=\min\{\pi(S_2)\}$$

定义 $e_{1,i}$ 是 z_1 的最低 i 位，$e_{2,i}$ 是 z_2 的最低 i 位。则通过推导[2]可以得出：

$$E_b=\Pr(\prod_{i=1}^{b}1\{e_{1,i}=e_{2,i}\}=1)=C_{1,b}+(1-C_{2,b})R \tag{4}$$

式中

$$r_1=\frac{f_1}{D},r_2=\frac{f_2}{D},s=\frac{a}{D} \tag{5}$$

$$C_{1,b}=A_{1,b}\frac{r_2}{r_1+r_2}+A_{2,b}\frac{r_1}{r_1+r_2} \tag{6}$$

$$C_{2,b}=A_{1,b}\frac{r_1}{r_1+r_2}+A_{2,b}\frac{r_2}{r_1+r_2} \tag{7}$$

$$A_{1,b}=\frac{r_1[1-r_1]^{2^b-1}}{1-[1-r_1]^{2^b}} \tag{8}$$

$$A_{2,b}=\frac{r_2[1-r_2]^{2^b-1}}{1-[1-r_2]^{2^b}} \tag{9}$$

R 的无偏估计：

$$\overset{\vee}{R}_b=\frac{\hat{E}_b-C_{1,b}}{1-C_{2,b}} \tag{10}$$

$$\hat{E}_b=\frac{1}{k}\sum_{j=1}^{k}(\prod_{i=1}^{b}1\{e_{1,i,\pi_j}=e_{2,i,\pi_j}\}=1) \tag{11}$$

根据二项分布的属性

$$Var(\overset{\vee}{R}_b)=\frac{Var(\hat{E}_b)}{[1-C_{2,b}]^2}=\frac{1}{k}\frac{E_b(1-E_b)}{[1-C_{2,b}]^2}=\frac{1}{k}\frac{[C_{1,b}+(1-C_{2,b})R][1-C_{1,b}-(1-C_{2,b})R]}{[1-C_{2,b}]^2} \tag{12}$$

对于很大的 $b(A_{1,b},A_{2,b}\to 0,C_{1,b},C_{2,b}\to 0)$，$Var(\overset{\vee}{R}_b)$ 近似为 $Var(\overset{\vee}{R}_M)$。

2 最优分数位 minwise 哈希算法

2.1 分数位的定义

设 b_1,b_2 为整数位，例如 $b_1=1,b_2=2$，若样本总数 k 为 1000，$b_1=1$ 的样本数 k_1 为 500，$b_2=2$ 的样本数 k_2 为 500。定义 $b=b_1$ 所占权重 w_1 及 $b=b_2$ 所占权重 $w_2(w_1=\frac{k_1}{k},w_2=\frac{k_2}{k})$。定义分数位 f 为：

$$f=w_1b_1+w_2b_2,w_1+w_2=1,b_1\neq b_2 \tag{13}$$

设 $y_1=e_{1,1}e_{1,2}\cdots,e_{1,b_x},y_2=e_{2,1}e_{2,2}\cdots e_{2,b_x}(b_x=b_1 \text{ or } b_2)$，则可以推导出：

$$\begin{aligned}E_f&=\Pr(y_1=y_2)=\Pr(\prod_{i=1}^{b_x}1\{e_{1,i}=e_{2,i}\}=1),b_x=b_1 \text{ or } b_2\\&=\Pr(b_x=b_1)\Pr(\prod_{i=1}^{b_1}1\{e_{1,i}=e_{2,i}\}=1)+\Pr(b_x=b_2)\Pr(\prod_{i=1}^{b_2}1\{e_{1,i}=e_{2,i}\}=1)\\&=w_1(C_{1,b_1}+(1-C_{2,b_1})R)+w_2(C_{1,b_2}+(1-C_{2,b_2})R)\end{aligned} \tag{14}$$

可得出分数位的无偏估计：

$$\overset{\vee}{R}_f=\frac{\hat{E}_f-(w_1C_{1,b_1}+w_2C_{1,b_2})}{1-(w_1C_{2,b_1}+w_2C_{2,b_2})} \tag{15}$$

$$\hat{E}_f=\frac{1}{k}\sum_{j=1}^{k}\{\prod_{i=1}^{b_x}1\{e_{1,i,\pi_j}=e_{2,i,\pi_j}\}=1\} \tag{16}$$

式中，$e_{1,i,\pi_j}(e_{2,i,\pi_j})$ 表示在 π_j 作用下 $z_1(z_2)$ 的最低 i 位。当 $w_1=1,w_2=0(w_1=0,w_2=1)$ 时，分数位估值式(15)简化为整数位估值式(10)，表明整数位估值公式只是分数位估值公式的一个特例。通过推导可得分数位的方差为：

$$\begin{aligned}&Var(\overset{\vee}{R}_f)\\&=\frac{Var(\hat{E}_f)}{[1-(w_1C_{2,b_1}+w_2C_{2,b_2})]^2}\\&=\frac{Var(w_1\hat{E}_{b_1}+w_2\hat{E}_{b_1})}{[1-(w_1C_{2,b_1}+w_2C_{2,b_2})]^2}\\&=\frac{1}{k}\frac{w_1^2Var(\hat{E}_{b_1})+w_2^2Var(\hat{E}_{b_2})+2w_1w_2Cov(\hat{E}_{b_1},\hat{E}_{b_2})}{[1-(w_1C_{2,b_1}+w_2C_{2,b_2})]^2}\\&=\frac{1}{k}\frac{w_1^2E_{b_1}(1-E_{b_1})+w_2^2E_{b_2}(1-E_{b_2})}{[1-(w_1C_{2,b_1}+w_2C_{2,b_2})]^2}\end{aligned} \tag{17}$$

式中

$$\begin{aligned}E_{b_1}&=C_{1,b_1}+(1-C_{2,b_1})R\\E_{b_2}&=C_{1,b_2}+(1-C_{2,b_2})R\end{aligned} \tag{18}$$

当 k 非常大时，$\hat{E}_{b_1}\approx E_{b_1},\hat{E}_{b_2}\approx E_{b_2},\hat{E}_{b_1},\hat{E}_{b_2}$ 的协方差为：

$$\begin{aligned}Cov(\hat{E}_{b_1},\hat{E}_{b_2})&=E(\hat{E}_{b_1}\cdot\hat{E}_{b_2})-E(\hat{E}_{b_1})\cdot E(\hat{E}_{b_2})\\&\approx E\{[C_{1,b_1}+(1-C_{2,b_1})R]\cdot[C_{1,b_2}+(1-C_{2,b_2})R]\}-E(C_{1,b_1}+(1-C_{2,b_1})R)\cdot E(C_{1,b_2}+(1-C_{2,b_2})R)\\&=[C_{1,b_1}+(1-C_{2,b_1})R]\cdot[C_{1,b_2}+(1-C_{2,b_2})R]-[C_{1,b_1}+(1-C_{2,b_1})R]\cdot[C_{1,b_2}+(1-C_{2,b_2})R]\\&=0\end{aligned} \tag{19}$$

2.2 最优分数位

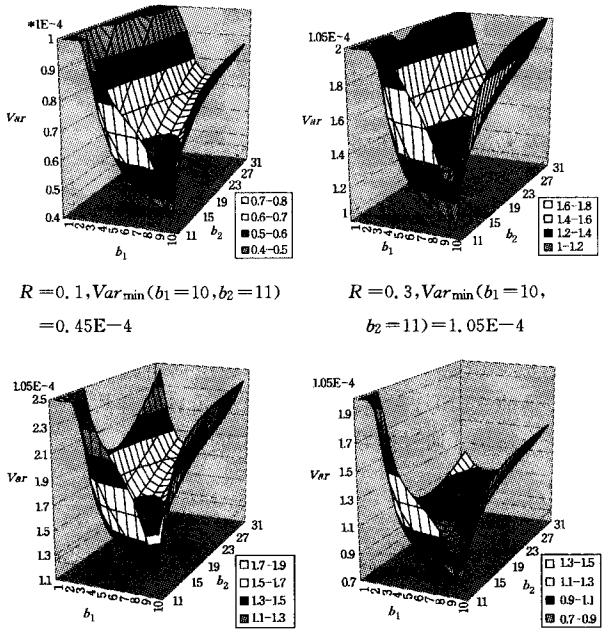
对于给定的分数位 f，组合 f 的方式千差万别。例如当 $f=1.5$ 时，可以用 $w_1=0.5,w_2=0.5,b_1=1,b_2=2$ 组合 $f=1.5$，同样也可以用 $w_1=0.75,w_2=0.25,b_1=1,b_2=3$ 组合 $f=1.5$。另外，还存在其他很多的分数位 f 的组合。由 $f=w_1b_1+w_2b_2,w_1+w_2=1,b_1\neq b_2$ 可得：

$$w_1=\frac{b_2-f}{b_2-b_1},w_2=\frac{f-b_1}{b_2-b_1} \tag{20}$$

$\forall b_1,b_2,f=f_0,1\leqslant b_1<f_0<b_2\leqslant 32,Var(\overset{\vee}{R}_f)=v(b_1,b_2,f)$，$Var(\overset{\vee}{R}_f)$ 变换为：

$$\begin{aligned}Var(\overset{\vee}{R}_f)&=Var(b_1,b_2,f)\\&=\frac{1}{k}\frac{(\frac{b_2-f}{b_2-b_1})^2E_{b_1}(1-E_{b_1})+(\frac{f-b_1}{b_2-b_1})^2E_{b_2}(1-E_{b_2})}{[1-((\frac{b_2-f}{b_2-b_1})C_{2,b_1}+(\frac{f-b_1}{b_2-b_1})C_{2,b_2})]^2}\end{aligned} \tag{21}$$

因为 b_1，b_2 的取值是离散并且有限的，故组合 f 的种类数是有限的。当给定一个 $f=10.4$，$k=1000$ 时，对式(21)中 b_1，b_2 进行有限量的取值，计算得出方差数据，如图 1 所示。



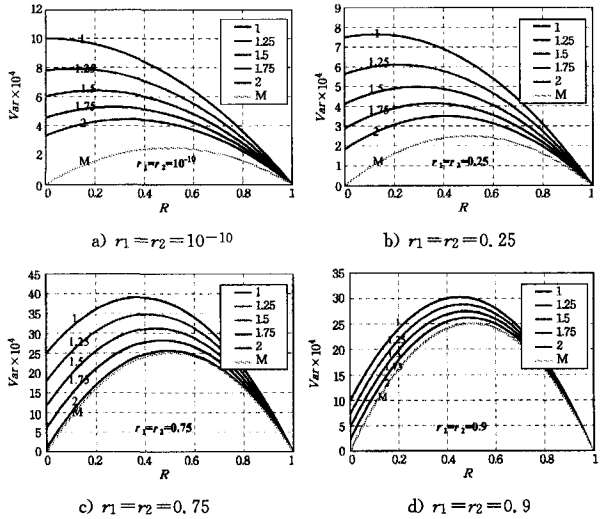
$R=0.1, Var_{\min}(b_1=10, b_2=11)=0.45\text{E}-4$　　$R=0.3, Var_{\min}(b_1=10, b_2=11)=1.05\text{E}-4$

$R=0.5, Var_{\min}(b_1=10, b_2=11)=1.25\text{E}-4$　　$R=0.8, Var_{\min}(b_1=10, b_2=11)=0.801E-4$

图 1　不同 b_1 和 b_2 组合下的分数位 Var 图

图 1 分析了在给定 $k=1000$，$r_1=r_2=10^{-10}$，分数位 $f=10.4$，$1\leqslant b_1\leqslant 10$，$11\leqslant b_2\leqslant 32$ 时，只有当 $b_1=\lfloor f\rfloor$，$b_2=\lceil f\rceil$ 时，$Var(\overset{\smile}{R_f})$ 才最小，即 $Var(\lfloor f\rfloor, \lceil f\rceil, f)$ 是在分数位 f 下最优。

图 2 表明在 $k=1000$，选定的 $r_1=r_2$（从 10^{-10} 到 0.9），$b=1$、$b=2$、$f=1.25(w_1=0.75, w_2=0.25, b_1=1, b_2=2)$，$1.5(w_1=0.5, w_2=0.5, b_1=1, b_2=2)$，$1.75(w_1=0.25, w_2=0.75, b_1=1, b_2=2)$时，相似度($R$)-方差($Var$)的关系。方差随着位数增大而变小，定义 $b_1=\lfloor f\rfloor$，$b_2=\lceil f\rceil$ 组合最优分数位 f 来计算方差曲线，从而满足精度和存储空间更加广泛的可选择性需求。



a) $r_1=r_2=10^{-10}$　　b) $r_1=r_2=0.25$

c) $r_1=r_2=0.75$　　d) $r_1=r_2=0.9$

图 2　最优分数位 R-Var 关系图

3　实验结果及分析

3.1　实验方差

以某基金的申报项目为数据来源，对 135,653 个项目文本对进行了最优分数位 minwise 哈希算法的相似度度量实验。实际数据方差的测量方法是：

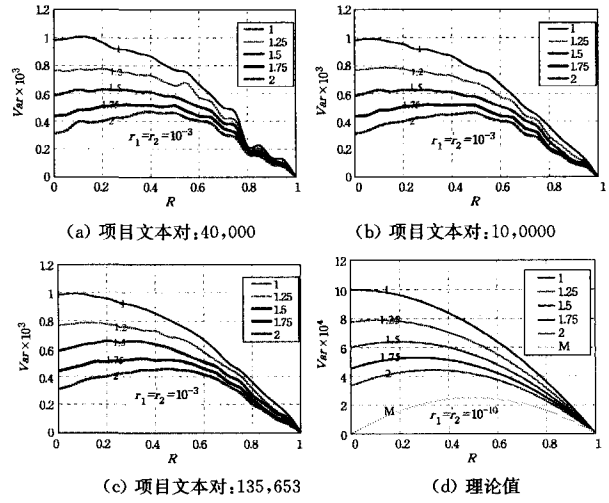
$$Var(R=R_0)=Avg\{[R(\text{真实值})-R(\text{估计值})]^2\}$$

其中，$R(\text{真实值})=\dfrac{|S_1\cap S_2|}{|S_1\cup S_2|}$。例如，若 3 个文档对(文档 1，文档 2)，(文档 3，文档 4)，(文档 5，文档 6)的 R(真实值)=0.2，其在 $b=1$ 时的估计值为 $R(b=1)=0.18, 0.19, 0.22$，则：

$$Var(R=0.2, b=1)=\frac{(0.18-0.2)^2+(0.19-0.2)^2+(0.22-0.2)^2}{3}=0.0003$$

对所有的项目进行了 $b=1,2$，$f=1.25(w_1=0.75, w_2=0.25, b_1=1, b_2=2)$，$1.5(w_1=0.5, w_2=0.5, b_1=1, b_2=2)$，$1.75(w_1=0.25, w_2=0.75, b_1=1, b_2=2)$的相似度方差度量实验。实验中，样本数 $k=1000$，shingle 总数 $D\approx$1e6，而一个文档的 shingles 数大约为 1000，故 $r_1=r_2=$1e−3。

图 3 显示了项目文本对在 4 万、10 万、13.5653 万时的实验测量的方差和理论方差图。当数据量变大，实验的曲线逐渐与理论的曲线吻合，符合 2.2 节中方差的理论推导，证明了最优分数位的方差在整数位之间，对于存储空间和精度的需求有着更加广泛的可选择性。



(a) 项目文本对:40,000　　(b) 项目文本对:10,0000

(c) 项目文本对:135,653　　(d) 理论值

图 3　实验数据的方差

3.2　准确率和召回率(precision and recall)

设 R_0 为相似度阈值，定义相似度度量的准确率和召回率的测量方法：

$$precision=\frac{\text{正确检测到相似度}>R_0\text{ 的文本对}}{\text{所有检测到相似度}>R_0\text{ 的文本对}}$$

$$recall=\frac{\text{正确检测到相似度}>R_0\text{ 的文本对}}{\text{实际存在的相似度}>R_0\text{ 的文本对}}$$

在 k 变化下，进行另一组实验，测量 $b=1,2$，$f=1.25(w_1=0.75, w_2=0.25, b_1=1, b_2=2)$，$1.5(w_1=0.5, w_2=0.5, b_1=1, b_2=2)$，$1.75(w_1=0.25, w_2=0.75, b_1=1, b_2=2)$时相似度度量的准确率和召回率，如图 4 所示。图 4 显示了在相似度 $\geqslant R_0$ 时，分数位哈希算法的准确率和召回率的实验结果。可

以看到，最优分数位 minwise 哈希的准确率和召回率在整数位之间，验证了此算法的有效性。

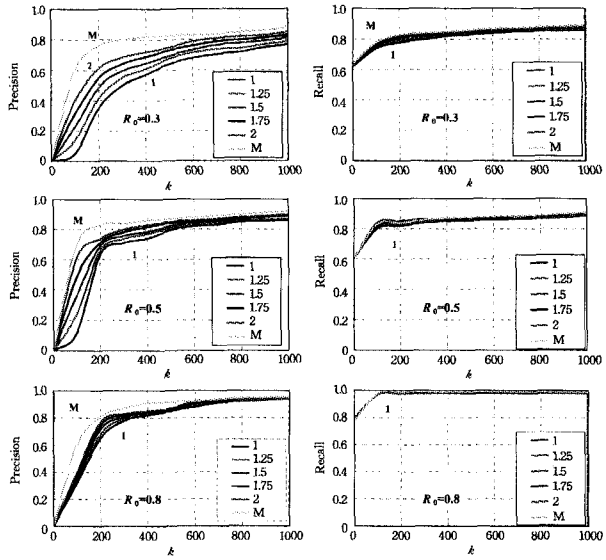


图 4　估值的准确率和召回率

结束语　min-wise 哈希算法广泛地应用于海量数据下的信息检索。b 位哈希算法将 $b=32$ 位缩小到 1 或者 2 位，降低了存储空间和计算时间。分数位 minwise 哈希算法对各种精度和存储空间需求有着更加广泛的可选择性。本文找到了构建分数位的最小方差组合，通过大量的基金项目数据对分数位估值的准确率和召回率进行实验分析，证明了最优分数位 minwise 哈希的准确率和召回率在整数位之间，验证了最优分数位 minwise 哈希算法的有效性。

参 考 文 献

[1] 鲍军鹏，沈钧毅，刘晓东，等. 自然语言文档复制检测研究综述[J]. 软件学报，2003，14(10)：1753-1760

[2] Broder A Z, Charikar M, Frieze A M, et al. Min-wise independent permutations[J]. Journal of Computer Systems and Sciences, 2000, 60(3): 630-659

[3] Broder A Z. On the resemblance and containment of documents [C] // Proceedings of Compression and Complexity of Sequences. Washington, DC, USA: IEEE Computer Society, 1997: 21-29

[4] Kalpakis K, Tang S. Collaborative data gathering in wireless sensor networks using measurement co-occurrence [J]. Computer Communications, 2008, 31(10): 1979-1992

[5] Dourisboure Y, Geraci F, Pellegrini M. Extraction and classification of dense implicit communities in the Web graph [J]. ACM Transactions on the Web(TWEB), 2009, 3(2): 1-36

[6] Bendersky M, Croft W B. Finding text reuse on the Web [C]// WSDM'09 Proceedings of the Second ACM International Conference on Web Search and Data Mining. New York, USA: ACM, 2009: 262-271

[7] Buehrer G, Chellapilla K. A scalable pattern mining approach to Web graph compression with communities [C] // WSDM '08 Proceedings of the international conference on Web search and Web data mining. New York, USA: ACM, 2008: 95-106

[8] Indyk P. A small approximately min-wise independent family of hash functions [J]. Journal of Algorithm, 2001, 38(1): 84-90

[9] Charikar M S. Similarity estimation techniques from rounding algorithms [C]// STOC'02 Proceedings of the thiry-fourth annual ACM symposium on Theory of computing. New York, USA: ACM, 2002: 380-388

[10] Li P, König A C. b-Bit minwise hashing [C]// WWW'10 Proceedings of the 19th international conference on World Wide Web. New York, USA: ACM, 2010: 671-680

[11] Li P, König A C. Theory and Applications of b-Bit Minwise Hashing [J]. Communications of the ACM, 2011, 54(8): 101-109

[12] Yuan X P, Long J, Zhang Z P, et al. f-Fractional Bit Minwise Hashing[J]. Journal of Software, 2012, 7(1): 228-236

(上接第 157 页)

维度的策略无法识别的攻击。另外，提出了策略实施的类型系统并进行了可靠性证明。

参 考 文 献

[1] Denning D E. A lattice model of secure information flow [J]. Communications of the ACM, 1976, 19(5): 236-243

[2] Goguen J A, Meseguer J. Security policies and security models [C]// IEEE Symposium on Security and Privacy. 1982: 11-20

[3] Sabelfeld A, Sands D. Declassification: dimensions and principles [J]. Journal of Computer Security, 2009, 17(5): 517-548

[4] Sabelfeld A, Myers A C. A model for delimited information release[J]. Software Security Theories and Systems, 2004, 3233: 174-191

[5] Askarov A, Sabelfeld A. Gradual Release: unifying declassification, encryption and key release policies[C]// IEEE Symposium on Security and Privacy. 2007: 207-221

[6] Lux A, Mantel H. Declassification with explicit reference points [C]// 14th European Symposium on Research in Computer Security. 2009: 69-85

[7] Lux A, Mantel H. Who can declassify? Formal Aspects in Security and Trust[J]. Lecture Notes in Computer Science, 2009, 5491: 35-49

[8] Askarov A, Hunt S, Sabelfeld A, et al. Termination insensitive noninterference leaks more than just a bit[C]// Computer Security-ESORICS. 2008: 333- 348

[9] Sabelfeld A, Myers A C. Language-based information flow security[J]. Selected Areas in Communications, 2003, 21(1): 5-19

[10] Askarov A, Myers A C. A semantic framework for declassification and endorsement [J]. Programming Languages and Systems, Lecture Notes in Computer Science, 2010, 6012: 64-84

[11] 唐和平，黄曙光，张亮. 动态信息流分析的漏洞利用检测系统[J]. 计算机科学，2010，37(7)：148-151

[12] 李伟楠，李翰超，石文昌. 基于信息流源的访问控制研究[J]. 计算机科学，2011，38(3)：34-39