第 43 卷 第 6A 期
2016 年 6 月

计 算 机 科 学
Computer Science

Vol. 43 No. 6A
June 2016

连接位极大似然动态过滤算法

曹 阳[1] 袁鑫攀[2] 龙 军[1]

(中南大学信息科学与工程学院 长沙 410083)[1] (湖南工业大学计算机与通信学院 株洲 412000)[2]

摘 要 Minwise Hash 极大似然估计子 R_{Mle} 综合考虑所有事件的发生概率,可以提高估计精度,但降低了估计的效率。连接位 Minwise Hash 估计子 $R_{Minwise,c}$ 可以成倍减少比对次数,动态阈值过滤器能够进一步提高 Minwise Hash 算法和其变种算法的效率。结合连接位极大似然估计子和动态阈值过滤器,提出了连接位极大似然动态过滤算法 $R(T_{Mle,c})$。实验表明,$R(T_{Mle,c})$ 具有精度和效率兼顾的特性,计算时间最少,并且在 $k>300$ 的条件下,其准确度与 R_{Mle} 的近乎相等。

关键词 极大似然,相似性检测,哈希,连接位,动态阈值

中图法分类号 TP301.6 **文献标识码** A

Dynamic Filtering Algorithm of Connected Bit Maximum Likelihood Minwise Hash

CAO Yang[1] YUAN Xin-pan[2] LONG Jun[1]

(School of Information Science and Engineering, Central South University, Changsha 410083, China)[1]

(School of Computer and Communication, Hunan University of Technology, Zhuzhou 412000, China)[2]

Abstract Maximum likelihood estimator (R_{Mle}) can improve the average accuracy. Connected bit Minwise Hash ($R_{Minwise,c}$) can exponentially improve efficiency of similarity estimation. Dynamic threshold filter makes further improvement on efficiency. Combining $R_{Minwise,c}$ and dynamic threshold filter, the maximum likelihood dynamic filtering algorithm of Connected bit Minwise Hash was proposed. Experimental results demonstrate that $R(T_{Mle,c})$ can get second-best precision and efficiency, and is the most cost effective business in the estimator options (R_{Mle}, $R_{Mle,c}$, $R_{Minwise,c}$, $R(T_{Mle,c})$.

Keywords Maximum likelihood estimator, Similarity measure, Hash, Connected bit, Dynamic threshold

1 引言

Minwise Hash 算法[1]作为估值集合相似度的常用统计学方法,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[2]、无线传感器网络[3]、网络社区分类[4]、文本重用[5]、连接图压缩[6]。Minwise Hash 算法也有了相当多的理论和实验方法的创新和发展[6-10]。b 位 Minwise Hash[8] 将 $b=64$ 位缩小到 1 位,降低了存储空间和计算时间。b 位 Minwise Hash 在大型机器学习、基于最大似然估计(MLE)而改进的估计算法中有了新的理论创新和应用发展[9]。连接位 Minwise Hash 算法[10]能成倍地减少比对的次数,提升算法的性能

本文综合 Minwise Hash 的连接位估计子 $R_{Minwise,c}$ 和极大似然估计子 R_{Mle} 得到连接位极大似然估计子 $R_{Mle,c}$,最后辅以动态阈值过滤器,提出了连接位极大似然动态过滤算法。

2 连接位极大似然估计

2.1 Minwise Hash

设全集 $\Omega=\{0,1,\cdots,D-1\}$,通过 shingling 文档 d 得到相关 shingles 集合 S_d。文档 S_x 和 S_y 的相似度 $R(x,y)$ 定义为:

$R(x,y)=\dfrac{|S_x\cap S_y|}{|S_x\cup S_y|}=\dfrac{a}{f_x+f_y-a}$,其中 $f_x=|S_x|$, $f_y=|S_y|$, $a=|S_x\cap S_y|$。假定一个在 Ω 上的随机独立置换群:$\pi.\ \Omega\rightarrow\Omega$, $\Omega=\{0,1,\cdots,D-1\}$,通过 k 个独立随机的置换群 π_1, π_2, $\cdots$, π_k 群,就把任意一个文档 d 的 shingles 集合转换为:

$$\overline{S_d}=(\min\{\pi_1(S_d)\},\min\{\pi_2(S_d)\},\cdots,\min\{\pi_k(S_d)\})$$

D_x, D_y 相似度 R 的估计子为:

$$\begin{aligned}\hat{R}_{Minwise}(x,y)&=\Pr(\min\{\pi(S_x)\}=\min\{\pi(S_y)\}\\&=\frac{1}{k}\sum_{i=1}^{k}1\{\min(\pi_i(S_x))=\min(\pi_i(S_y))\}\end{aligned}\quad(1)$$

式中,k 表示实验次数(或者样本大小)。

2.2 极大似然估计子

由文献[9]可知,定义极大似然函数:

$l(a)=\log(\Pr_{=}^{k_=}\cdot\Pr_{<}^{k_<}\cdot\Pr_{>}^{k_>})=k_=\log\Pr_=+k_<\log\Pr_<+k_>\log\Pr_>$。求解 $l(a)$ 的最大值,则有 $l'(a)=k_=\dfrac{(\Pr_=)'}{\Pr_=}+k_<\dfrac{(\Pr_<)'}{\Pr_<}+k_>\dfrac{(\Pr_>)'}{\Pr_>}=0$。

$$k_=\frac{f_1+f_2}{a}-k_<\frac{f_2}{f_1-a}-k_>\frac{f_1}{f_2-a}=0\quad(2)$$

本文受国家自然科学基金项目(61472450,61379110),湖南省自然科学青年基金项目(2015JJ3058),湖南工业大学自然科学基金(2014HZX17),湖南省教育厅科技研究项目(14C0325),国家级大学生创新创业训练计划项目(201511535006),湖南省研究生科研创新项目(CX2015B567)资助。

曹 阳(1992—),女,硕士生,主要研究方向为大数据处理,E-mail:caoyangx2015@163.com;袁鑫攀(1982—),男,博士,讲师,主要研究方向为信息检索、数据挖掘(通信作者);龙 军(1972—),男,博士,教授,主要研究方向为网构化软件。

式(2)的解为 Minwise 极大似然估计子 R_{Mle}。

2.3 连接位极大似然估计子

由文献[10]可知，连接位 Minwise Hash 相似度 R 的估计子为：

$$\hat{R}_{xy,b,n}=\frac{\hat{G}^{\frac{1}{n}}_{xy,b,n}-C_{x,b}}{1-C_{y,b}} \tag{3}$$

令连接位的极大似然概率为：$\Pr_{MLE\text{-}C}=\Pr^{k_{c,=}}_{c,=}\cdot\Pr^{k_{c,<}}_{c,<}\cdot\Pr^{k_{c,>}}_{c,>}$，同理需 $k_{c,=}\frac{(\Pr_{c,=})'}{\Pr_{c,=}}+k_{c,<}\frac{(\Pr_{c,<})'}{\Pr_{c,<}}+k_{c,>}\frac{(\Pr_{c,>})'}{\Pr_{c,>}}=0$ 成立。故需解出 $\Pr_{c,=},\Pr_{c,<},\Pr_{c,>}$ 的表达式。由文献[18]可知，$\Pr_{b,=},\Pr_{b,<},\Pr_{b,>}$ 的表达式为：

$$\Pr_{b,=}=P_b=\frac{a}{f_1+f_2-a}(1-C_{2,b})+C_{1,b} \tag{4}$$

$$\Pr_{b,<}=\frac{1}{1-[1-r_2]^{2^b}}\frac{(r_1-s)}{1-[1-(r_1+r_2-s)]^{2^b}}A+\frac{[1-r_1]^{2^b}-1}{1-[1-r_1]^{2^b}}\frac{(r_2-s)}{1-[1-(r_1+r_2-s)]^{2^b}}B \tag{5}$$

$$\Pr_{b,>}=1-\Pr_{b,<}-\Pr_{b,=} \tag{6}$$

其中，$s=\frac{a}{D}$。

若连接 2 个 b 位，显然有 $\Pr_{c,<}=\Pr^2_{b,<}+2\Pr_{b,<}\cdot\Pr_{b,=}+\Pr_{b,<}\cdot\Pr_{b,>}$；$\Pr_{c,>}=\Pr^2_{b,>}+2\Pr_{b,>}\cdot\Pr_{b,=}+\Pr_{b,>}\cdot\Pr_{b,<}$；$\Pr_{c,=}=\Pr^2_{b,=}$。若连接 3 个 b 位，则有 $\Pr_{c,<}=\Pr^3_{b,<}+3\Pr^2_{b,<}\cdot\Pr_{b,=}+3\Pr_{b,<}\cdot\Pr^2_{b,=}+\Pr_{b,<}\cdot\Pr^2_{b,>}+4\Pr_{b,<}\cdot\Pr_{b,=}\cdot\Pr_{b,>}$；$\Pr_{c,>}=\Pr^3_{b,>}+3\Pr^2_{b,>}\cdot\Pr_{b,=}+3\Pr_{b,>}\cdot\Pr^2_{b,=}+\Pr_{b,>}\cdot\Pr^2_{b,<}+4\Pr_{b,>}\cdot\Pr_{b,=}\cdot\Pr_{b,<}$；$\Pr_{c,=}=\Pr^3_{b,=}$。要让极大似然函数 $l(a)=\log(\Pr^{k_{c,=}}_{c,=}\cdot\Pr^{k_{c,<}}_{c,<}\cdot\Pr^{k_{c,>}}_{c,>})$ 最大，则

$$k_{c,=}\frac{(\Pr_{c,=})'}{\Pr_{c,=}}+k_{c,<}\frac{(\Pr_{c,<})'}{\Pr_{c,<}}+k_{c,>}\frac{(\Pr_{c,>})'}{\Pr_{c,>}}=0 \tag{7}$$

式(7)的解即为连接位极大似然的估计子 $R_{Mle,c}$。

3 连接位极大似然动态阈值过滤算法

3.1 动态阈值过滤器

定义 k 为总比对次数，T 为预设阈值，文档对为 $\{(S_1,S_2),(S_3,S_4),\cdots,(S_{2n-1},S_{2n})\}$。相似度度量的目的是在 k 次比对后输出相似度估计值大于 T 的集合：$\{(S_{2i-1},S_{2i})\mid R(S_{2i-1},S_{2i})>T_0,1\leqslant i\leqslant n\}$。

动态阈值过滤器设定了数个比对点 $(k_1,k_2,k_3,\cdots,k_n)$，将比对的总过程划分为若干个子过程，在每个比对点 $k_i(1<i<n)$ 上设定上界阈值 $T_U(k_i)$，提前过滤相似度小于下界阈值的文档对。定义 k 为当前比对次数，随机变量 X 是文档对 S_{2i-1},S_{2i} 的 Minwise 哈希值相等的次数，即 $\{X=\sum_{j=1}^{k}1\{\min(\pi_j(S_{2i-1}))=\min(\pi_j(S_{2i}))\}\}$。假设某对文档经过 k_i 次对比后的相似度 $\hat{R}_M(k_i)\geqslant T$，给定参数 $m(0<m\leqslant k)$，由二项分布概率公式，事件 $\{X\leqslant m\}$ 的概率：

$$\Pr(X\leqslant m)\leqslant\sum_{i=0}^{m}\binom{k}{i}T_0^i(1-T_0)^{k-i} \tag{8}$$

定理 1

$$\left.\begin{array}{l}\text{select } m \text{ making } \Pr(X\leqslant m)\\ \text{is small probability}\\ \hat{R}_M(k_i)\leqslant T_L(k_i)=m/k_i\end{array}\right\}\Rightarrow\hat{R}_M(k_i)\leqslant T \tag{9}$$

证明：将小概率事件 $\Pr(X\leqslant m)$ 看作不可能发生的事件。假设 $\hat{R}_M(k_0)\geqslant T_0$，经过 k 次比对后，如果不可能发生的事件 $\{X\leqslant m\}$ 发生了，则假设 $\hat{R}_M(k_0)\geqslant T_0$ 不成立，即 $\hat{R}_M(k_0)<T_0$。而事件 $\{X\leqslant m\}$ 的充分必要条件为 $\hat{R}_M(k)\leqslant T_L(k_i=m/k$，定理 1 得证。假设 $k=1000$，动态阈值过滤过程如图 1 所示。

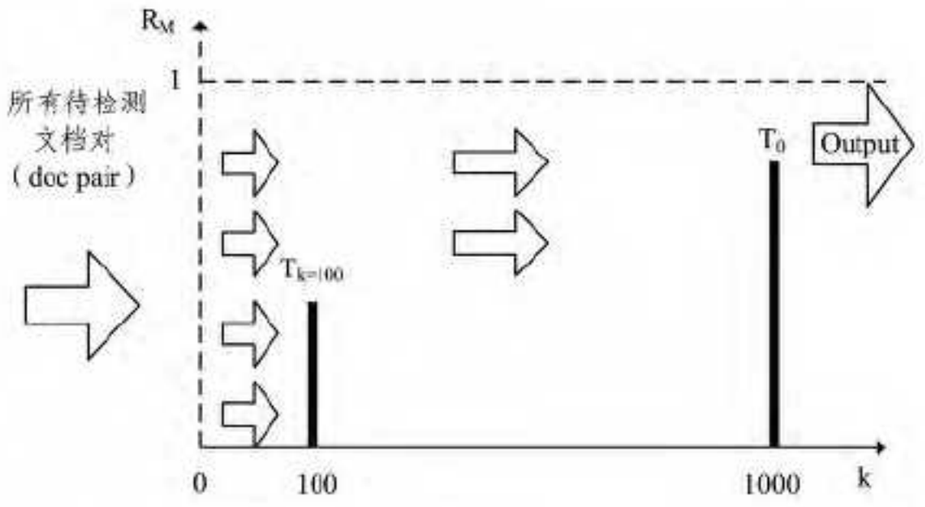


图 1 动态阈值过滤器效果图

3.2 连接位极大似然动态阈值过滤算法

动态阈值过滤器是一种基于二项分布和小概率事件构建的 Minwise Hash 估计过滤器，理论上能够应用于 Minwise Hash 及其变种算法，乃至所有符合二项分布的估计子。式(7)得到的连接位极大似然的估计子 $R_{Mle,c}$ 符合二项分布，能够添加动态双重阈值过滤器。本文将连接位极大似然的估计子 $R_{Mle,c}$ 和动态阈值过滤器相结合，提出连接位极大似然动态阈值过滤算法。

给定相似度阈值 T，总比对次数 k。把连接位极大似然的估计过程划分为若干过程，根据动态双重阈值过滤器的定义，在每个比对点上设置 $R_{Mle,c}$ 的阈值 T_{Lmc}。在 $t(t<k)$ 次实验中，令随机变量 X 为连接位极大似然指纹相等的次数，记事件：$\{X\leqslant m\}$，则事件发生的概率为：

$$\Pr(X\leqslant m)=\sum_{i=0}^{m}\binom{t}{i}T^i(1-T)^{t-i}=\sum_{i=0}^{m}\binom{t}{i}(R_{Mle},c)^i(1-R_{Mle},c)^{t-i} \tag{10}$$

根据式(10)可知，找到一个 m 使 $\{X\leqslant m\}$ 为小概率事件，则 $T_{Lmc}=m/k_t$。$R_{Mle,c}$ 通过动态阈值过滤器，在每一个比对点提前输出相似度大于上界阈值的文档对，提前过滤相似度低于下界阈值的文档对。

4 实验结果及分析

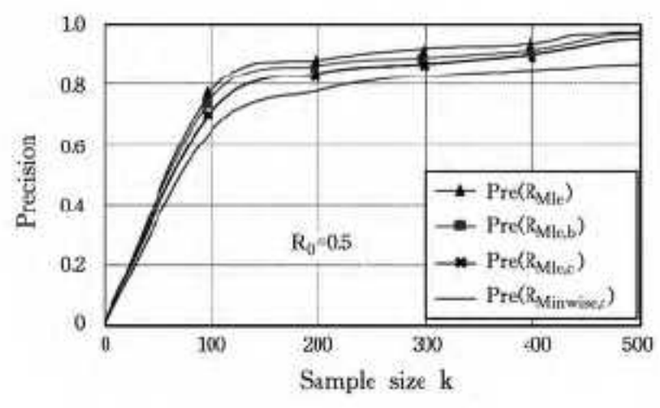
4.1 准确率和召回率

以某基金的申报项目为数据来源，对 10 万份文档进行了相似性度量实验。定义 R_0 为相似度阈值，准确率和召回率的定义如式(11)、式(12)所示。比较极大似然估计子 (R_{Mle})、b 位 Minwise Hash 极大似然估计子 $(R_{Mle,b})$，连接位极大似然估计子 $(R_{Mle,c})$ 和连接位 Minwise Hash 估计子 $(R_{Minwise,c})$ 的准确率和召回率。V_a 为其实相似度，V_e 为测试相似度。

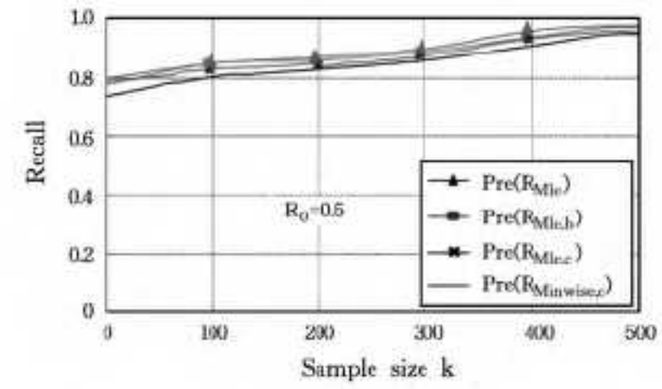
$$precision=\frac{|V_a>R_0|\cap|V_e>R_0|}{|V_e>R_0|} \tag{11}$$

$$Recall=\frac{|V_a>R_0|\cap|V_e>R_0|}{|V_a>R_0|} \tag{12}$$

图 2 显示在相同的阈值 R_0 和比对次数 k 的情况下，4 种估计子的召回曲线没有明显的差别，准确率曲线有一定的变化。



(a)Precision of $R_0=0.5$



(b)Recall of $R_0=0.5$

图 2 R_{Mle},$R_{Mle,b}$, $R_{Mle,c}$, $R_{Minwise,c}$的准确率和召回率

(1)随着比对次数 k 的增加,估计精度会提升,表明 k 越大,极大似然估计子的估计方差越小,估计精度就越接近真实相似度。

(2)在相同的阈值 T 和比对次数 k 的条件下,准确率排序为:$Pre(R_{Mle})>Pre(R_{Mle,b})>Pre(R_{Mle,c})>Pre(R_{Minwise,c})$。例如,当 $T=0.5$、$k=300$ 时,准确率排序为:$Pre(R_{Mle})=93\%>Pre(R_{Mle,b})=89\%>Pre(R_{Mle,c})=87\%>Pre(R_{Minwise,c})=65\%$。

4.2 时间性能

图 3 显示了 3 种估计子(R_{Mle},$R_{Mle,c}$, $R_{Minwise,c}$)和 $R_{Mle,c}$ 增加过滤器 $R(T_{Mle,c})$ 的时间性能,当计算规模相同时,计算所需的时间排序为:$\mathrm{Time}(R(T_{Mle,c}))>\mathrm{Time}(R_{Minwise,c})>\mathrm{Time}(R_{Mle,c})>\mathrm{Time}(R_{Mle})$。

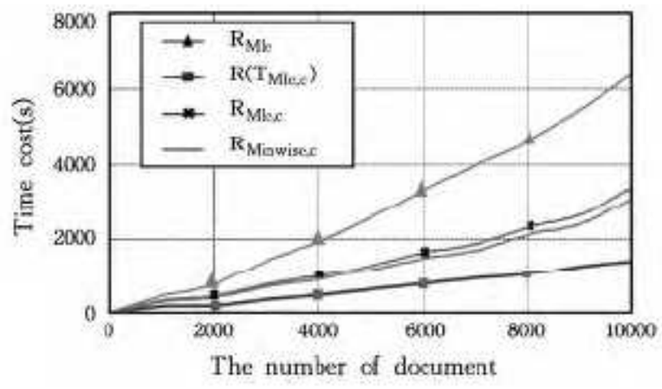


图 3 R_{Mle},$R_{Mle,c}$,$R_{Minwise,c}$和$R(T_{Mle,c})$的时间性能

综合图 2 和图 3 可知,根据不同的需求可以选择最适合自己的估计子:

(1) 如果将精度需求排在第一位,可以选择 R_{Mle},但是计算耗费的时间是最多的。

(2)如果想要兼顾精度和效率,可以选择 $R(T_{Mle,c})$,它的计算时间最少,并且在 $k>300$ 的条件下,$R_{Mle,c}$ 与 R_{Mle} 相比,准确度近乎相等。

结束语 本文提出了连接位极大似然动态过滤算法,通过实验对 $R(T_{Mle,c})$ 的可行性进行了分析。测试了 4 种估计子 R_{Mle},$R_{Mle,b}$,$R_{Mle,c}$, $R_{Minwise,c}$ 的准确率和召回率,其中 R_{Mle} 精度最高,$R_{Mle,c}$ 次之。测试了 3 种估计子(R_{Mle}, $R_{Mle,c}$, $R_{Minwise,c}$)和 $R_{Mle,c}$ 增加过滤器 $R(T_{Mle,c})$ 的时间性能,综合实验的结果表明 $R(T_{Mle,c})$ 是一种精度和效率兼顾的相似度估计算法,它的计算时间最少,并且在 $k>300$ 的条件下,$R_{Mle,c}$ 与 R_{Mle} 相比,准确度近乎相等。

参 考 文 献

[1] 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

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

[3] 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

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

[5] Buehrer G, Chellapilla K. A scalable pattern mining approach to web graph compression with communities [C]// Proceedings of the International Conference on Web Search and Web Data Mining (WSDM'08). New York, USA: ACM, 2008: 95-106

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

[7] Charikar M S. Similarity estimation techniques from rounding algorithms [C]// Proceedings of the Thiry-fourth Annual ACM Symposium on Theory of Computing (STOC'02). New York, USA: ACM, 2002: 380-388

[8] Li P, König A C. b-bit minwise hashing [C]// Proceedings of the 19th International Conference on World Wide Web (WWW'10). New York, USA: ACM, 2010: 671-680

[9] Li P, König A C. Theory and applications of b-bit minwise hashing [J]. Communications of the ACM, 2011, 54(8): 101-109

[10] Yuan Xin-pan, Long Jun, Zhang Zu-ping, et al. Connected bit minwise hashing [J]. Journal of Computer Research and Development, 2013, 50(4): 883-890

(上接第 399 页)

[23] Danon L, Diaz-Guilera A, Duch J, et al. Comparing community structure identification [J]. Journal of Statistical Mechanics: Theory and Experiment, 2005, 2005(9): 09008

[24] 马英红,李慧嘉,张晓东. 赋权网络中的弱化免疫研究[J]. 管理科学学报, 2010, 13(10): 32-39

[25] Girvan M, Newman M E J. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2002, 99(12): 7821-7826

[26] Hoffman J M, Wiggins C H. Bayesian Approach to Network Modularity[J]. Physical Review Letters, 2008, 100(25): 258701

[27] Zachary W. An information flow model for conflict and fission in small groups[J]. Journal of Anthropological Research, 1977, 28: 452-47

[28] Ronhovde P, Nussinov Z. Local resolution-limit-free Potts model for community detection[J]. Physical Review E, 2010, 81(4): 046114