

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局



(10) 国际公布号
WO 2022/166676 A1

(43) 国际公布日
2022 年 8 月 11 日 (11.08.2022)

(51) 国际专利分类号:
G06F 40/284 (2020.01) *G06F 16/35* (2019.01)
G06F 21/60 (2013.01)

(21) 国际申请号: PCT/CN2022/073677

(22) 国际申请日: 2022 年 1 月 25 日 (25.01.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202110161186.9 2021年2月5日 (05.02.2021) CN

(71) 申请人: 支付宝(杭州)信息技术有限公司 (ALIPAY (HANGZHOU) INFORMATION TECHNOLOGY CO., LTD.) [CN/CN]; 中国浙江省杭州市西湖区西溪路 556 号 8 层 B 段 801-11, Zhejiang 310000 (CN)。

(72) 发明人: 吴若凡 (WU, Ruofan); 中国浙江省杭州市西湖区西溪路 556 号 8 层 B 段 801-11, Zhejiang 310000 (CN)。石磊磊 (SHI, Leilei); 中国浙江省杭州市西湖区西溪路 556 号 8 层 B 段 801-11, Zhejiang 310000 (CN)。陈永环 (CHEN, Yonghuan); 中国浙江省杭州市西湖区西溪路 556 号 8 层 B 段 801-11, Zhejiang 310000 (CN)。朱耀伟 (ZHU, Yaowei); 中国浙江省杭州市西湖区西溪路 556 号 8 层 B 段 801-11, Zhejiang 310000 (CN)。

(74) 代理人: 北京博思佳知识产权代理有限公司 (BEIJING BESTIPR INTELLECTUAL PROPERTY LAW CORPORATION); 中国北京市海淀区上地三街 9 号嘉华大厦 B 座 409, Beijing 100085 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU,

(54) Title: METHOD AND APPARATUS FOR ESTIMATING SEGMENTED WORD FREQUENCY IN DIFFERENTIAL PRIVACY PROTECTION DATA

(54) 发明名称: 估计差分隐私保护数据中分词频度的方法及装置

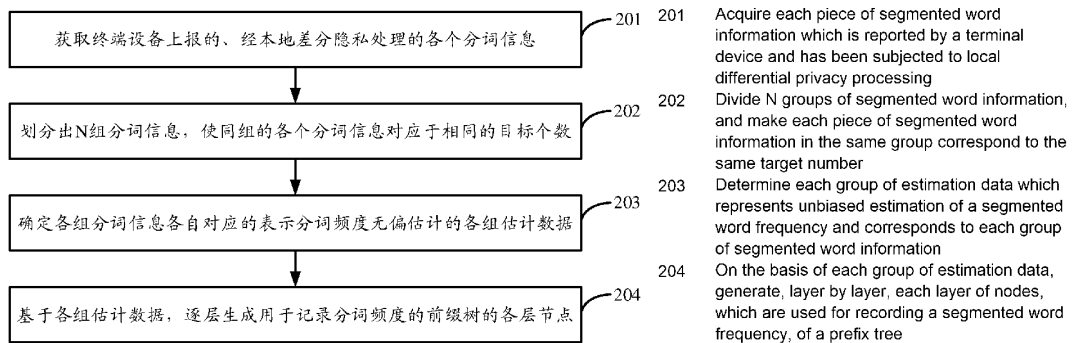


图 2

(57) Abstract: Provided are a method and apparatus for estimating a segmented word frequency in differential privacy protection data, and an electronic device. The method comprises: acquiring each piece of segmented word information which is reported by a terminal device and has been subjected to local differential privacy processing; dividing N groups of segmented word information, and making each piece of segmented word information in the same group correspond to the same target number; determining each group of estimation data which represents unbiased estimation of a segmented word frequency and corresponds to each group of segmented word information; and on the basis of each group of estimation data, generating, layer by layer, each layer of nodes, which are used for recording a segmented word frequency, of a prefix tree, wherein during the process of generating the nth layer of nodes of the prefix tree, some alternative n-element segmented words are selected as n-element segmented words represented by the nth layer of nodes, and there is no need to traverse all n-element segmented words composed of preset word units. Therefore, the calculation amount is greatly reduced and the calculation efficiency is improved, and the n-element segmented words which are represented by the nth layer of nodes and are screened out on the basis of frequency significance distribution information of the segmented words are more rational.

[见续页]

WO 2022/166676 A1

CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国(除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

(57) 摘要: 本说明书提供一种估计差分隐私保护数据中分词频度方法、装置及电子设备, 根据该方法, 获取终端设备上报的、经本地差分隐私处理的各个分词信息; 划分出N组分词信息, 使同组的各个分词信息对应于相同的目标个数; 确定各组分词信息各自对应的表示分词频度无偏估计的各组估计数据; 基于所述各组估计数据, 逐层生成用于记录分词频度的前缀树的各层节点。其中, 在生成前缀树的第n层节点的过程中, 选择部分备选n元分词作为第n层节点表示的n元分词, 无需遍历由预设的词语单元构成的所有n元分词, 不仅大大降低了计算量, 提高了计算效率, 而且基于分词的频度显著性分布信息而筛选出来的第n层节点表示的n元分词更具合理性。