

(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

(43) 国际公布日
2025 年 2 月 6 日 (06.02.2025)



(10) 国际公布号
WO 2025/026068 A1

(51) 国际专利分类号:
G06F 16/35 (2019.01)

(21) 国际申请号: PCT/CN2024/105819

(22) 国际申请日: 2024 年 7 月 16 日 (16.07.2024)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202310955488.2 2023年7月31日 (31.07.2023) CN

(71) 申请人: 阿里巴巴(中国)有限公司(**ALIBABA (CHINA) CO., LTD.**) [CN/CN]; 中国浙江省杭州市滨江区长河街道网商路 699 号 4 号楼 5 楼 508 室, Zhejiang 310052 (CN)。

(72) 发明人: 龙定坤(**LONG, Dingkun**); 中国浙江省杭州市西湖区三墩镇灯彩街 1008 号云谷园区, Zhejiang 310030 (CN)。张延钊(**ZHANG, Yanzhao**); 中国浙江省杭州市西湖区三墩镇灯彩街 1008 号云谷园区, Zhejiang 310030 (CN)。刘楚(**LIU, Chu**); 中国浙江省杭州市西湖区三墩镇灯彩街 1008 号云谷园区, Zhejiang 310030 (CN)。谢朋峻(**XIE, Pengjun**); 中国浙江省杭州市西湖区三墩镇灯彩街 1008 号云谷园区, Zhejiang 310030 (CN)。黄非(**HUANG, Fei**); 美国加州桑尼维尔阿尔曼街 525 号 4 楼, Sunnyvale 94085 (US)。

(74) 代理人: 北京同钧律师事务所(**BEIJING TONGJUN LAW FIRM**); 中国北京市海淀区西直门北大街 32 号院 1 号楼 7 层 808 室, Beijing 100082 (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,

(54) **Title:** TEXT DATA PROCESSING METHOD, TEXT REPRESENTATION METHOD, TEXT CLUSTERING METHOD, AND DEVICE

(54) 发明名称: 文本数据处理、文本表示、文本聚类的方法及设备

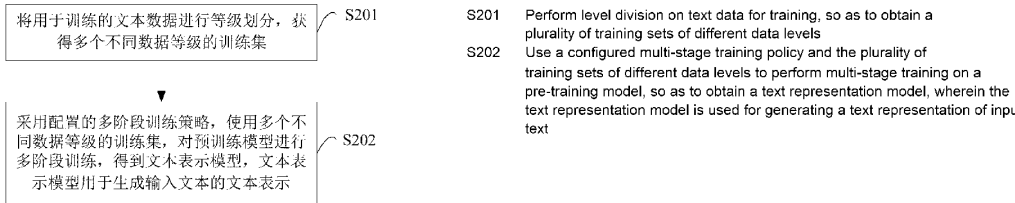


图 2

(57) **Abstract:** Provided in the present application are a text data processing method, a text representation method, a text clustering method, and a device. The text data processing method in the present application comprises: performing level division on text data for training, so as to obtain a plurality of training sets of different data levels; and using a configured multi-stage training policy and the plurality of training sets of different data levels to perform multi-stage training on a pre-training model, so as to obtain a text representation model. Compared with one instance (one stage) of training on a pre-training model on the basis of all training data, performing multi-stage training on the pre-training model in stages on the basis of a plurality of training sets of different data levels can better improve the representation capability of the pre-training model, thereby improving the text representation capability of a text representation model obtained by means of training, improving the text representation quality, and further improving the effects of downstream text processing tasks.

(57) 摘要: 本申请提供一种文本数据处理、文本表示、文本聚类的方法及设备。本申请的方法, 通过将用于训练的文本数据进行等级划分, 获得多个不同数据等级的训练集, 采用配置的多阶段训练策略, 使用多个不同数据等级的训练集, 对预训练模型进行多阶段训练, 得到文本表示模型, 相较于基于全部训练数据对预训练模型进行一次性(一个阶段)的训练, 基于多个不同数据等级的训练集分阶段地对预训练模型进行多阶段训练, 更能提升预训练模型的表示能力, 从而提升训练得到的文本表示模型的文本表示能力, 提升文本表示质量, 进而可提升下游文本处理任务的效果。

[续页]

WO 2025/026068 A1

BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, 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, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, 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, ME, 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))。