

(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 2025/194986 A1

(43) 国际公布日
2025 年 9 月 25 日 (25.09.2025)

WIPO | PCT

(51) 国际专利分类号:
G06F 16/33 (2025.01)

(21) 国际申请号: PCT/CN2025/071604

(22) 国际申请日: 2025 年 1 月 9 日 (09.01.2025)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202410332295.6 2024 年 3 月 22 日 (22.03.2024) CN

(71) 申请人: 上海得物信息集团有限公司 (SHANGHAI DEWU INFORMATION GROUP CO., LTD.) [CN/CN]; 中国上海市虹口区同心路 723 号 13 幢 6416 室 200080 (CN)。

(72) 发明人: 张章伟 (ZHANG, Zhangwei); 中国上海市虹口区同心路 723 号 13 幢 6416 室 200080 (CN)。安鹏 (AN, Peng); 中国上海市虹口区同心路 723 号 13 幢 6416 室 200080 (CN)。沙爱晖 (SHA, Aihui); 中国上海市虹口区同心路 723 号 13 幢 6416 室 200080 (CN)。周斌 (ZHOU, Bin); 中国上海市虹口区同心路 723 号 13 幢 6416 室 200080 (CN)。

(74) 代理人: 北京超凡宏宇知识产权代理有限公司 (CHOFN INTELLECTUAL PROPERTY); 中国北京市海淀区北四环西路 68 号左岸工社 1215-1218 室 100080 (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: SAMPLE CLEANING METHOD AND APPARATUS, AND ELECTRONIC DEVICE AND STORAGE MEDIUM

(54) 发明名称: 一种样本的清洗方法、装置、电子设备及存储介质

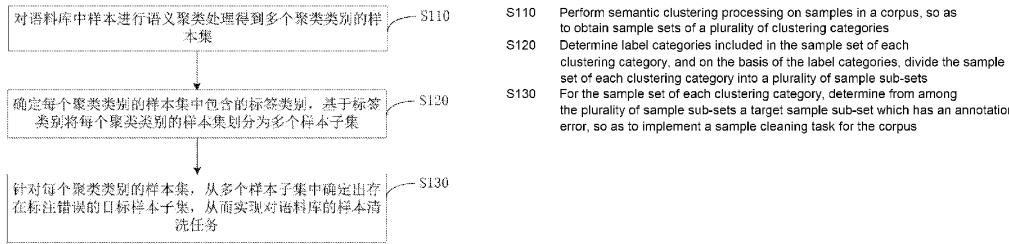


图 1

(57) Abstract: The present application relates to the technical field of artificial intelligence. Disclosed are a sample cleaning method and apparatus, and an electronic device and a storage medium. The method comprises: performing semantic clustering processing on samples in a corpus, so as to obtain sample sets of a plurality of clustering categories; determining label categories included in the sample set of each clustering category, and on the basis of the label categories, dividing the sample set of each clustering category into a plurality of sample sub-sets; and for the sample set of each clustering category, determining from among the plurality of sample sub-sets a target sample sub-set which has an annotation error, so as to implement a sample cleaning task for the corpus. By means of the technical solution provided in the present application, a sample that has an annotation error in a corpus can be relatively quickly located, and the accuracy of the determination of the sample which has the annotation error is relatively high, thereby effectively reducing the cost investments of time and labor.

(57) 摘要: 本申请公开了一种样本的清洗方法、装置、电子设备及存储介质, 涉及人工智能技术领域。其中, 该方法包括: 对语料库中样本进行语义聚类处理得到多个聚类类别的样本集; 确定每个聚类类别的样本集中包含的标签类别, 基于标签类别将每个聚类类别的样本集划分为多个样本子集; 针对每个聚类类别的样本集, 从多个样本子集中确定出存在标注错误的目标样本子集, 从而实现对语料库的样本清洗任务。本申请提供的技术方案, 可以较快地定位到语义库中标注错误的样本, 所确定的标注错误样本准确率较高, 有效降低了时间和人力的成本投入。

[见续页]

WO 2025/194986 A1

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, UY, 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))。