

(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
2020 年 6 月 4 日 (04.06.2020)



(10) 国际公布号
WO 2020/108063 A1

- (51) 国际专利分类号:
G06F 17/27 (2006.01) *G06F 16/23* (2019.01)
- (21) 国际申请号: PCT/CN2019/108024
- (22) 国际申请日: 2019 年 9 月 26 日 (26.09.2019)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201811416994.X 2018 年 11 月 26 日 (26.11.2018) CN
- (71) 申请人: 阿里巴巴集团控股有限公司 (ALIBABA GROUP HOLDING LIMITED) [—/CN]; 开曼群岛大开曼资本大厦一座四层 847 号邮箱, Grand Cayman (KY)。
- (72) 发明人: 李怀松 (LI, Huaisong); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集

团法务部, Zhejiang 311121 (CN)。潘健民 (PAN, Jianmin); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。周绪刚 (ZHOU, Xugang); 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (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, 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,

(54) Title: FEATURE WORD DETERMINING METHOD, APPARATUS, AND SERVER

(54) 发明名称: 特征词的确定方法、装置和服务器

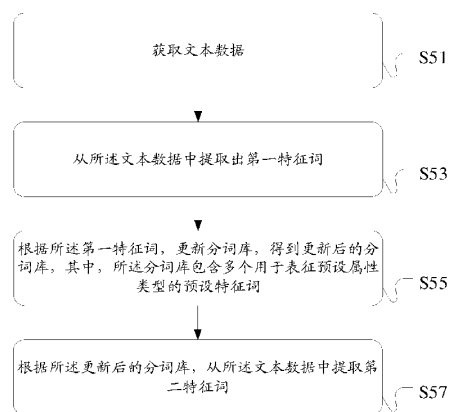


图 5

(57) Abstract: A feature word determining method, apparatus, and server. The method comprises: acquiring text data (S51); extracting a first feature word from the text data (S53); on the basis of the first feature word, updating a word segment base to obtain an updated word segment base, the word segment base comprising a plurality of preset feature words used for representing preset attribute types (S55); and, on the basis of the updated word segment base and the preset feature words, extracting a second feature word from the text data (S57). The present method first performs new word extraction on the text data to obtain the first feature word, then uses the first feature word to update the word segment base, and can then use the updated word segment base and the preset feature words to extract a new feature word from the text data as a second feature word, and thereby avoids the inaccurate feature word extraction caused by incorrect word segmentation during the feature word extraction process, and achieves the technical effect of precisely mining new feature words conforming to requirements from text data.

[见续页]

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, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, 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: 一种特征词的确定方法、装置和服务器。其中, 该方法包括: 获取文本数据 (S51); 从所述文本数据中提取出第一特征词 (S53); 根据所述第一特征词, 更新分词库, 得到更新后的分词库, 其中, 分词库包含多个用于表征预设属性类型的预设特征词 (S55); 根据更新后的分词库和预设特征词, 从文本数据中提取第二特征词 (S57)。该方法通过先对文本数据进行新词提取得到第一特征词; 再利用第一特征词对分词库进行更新; 进而可以利用更新后的分词库和预设特征词从文本数据提取出新的特征词作为第二特征词, 从而避免了提取特征词的过程中, 由于分词错误导致的特征词提取不准确, 达到能精确地从文本数据中挖掘出符合要求的新的特征词的技术效果。