(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 12 月 30 日 (30.12.2020)

WIPO | PCT

(10) 国际公布号
WO 2020/258481 A1

(51) 国际专利分类号:
G06F 16/335 (2019.01)

(21) 国际申请号: PCT/CN2019/102201

(22) 国际申请日: 2019 年 8 月 23 日 (23.08.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910582849.7 2019年6月28日 (28.06.2019) CN

(71) 申请人: 平安科技（深圳）有限公司 (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN]; 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。

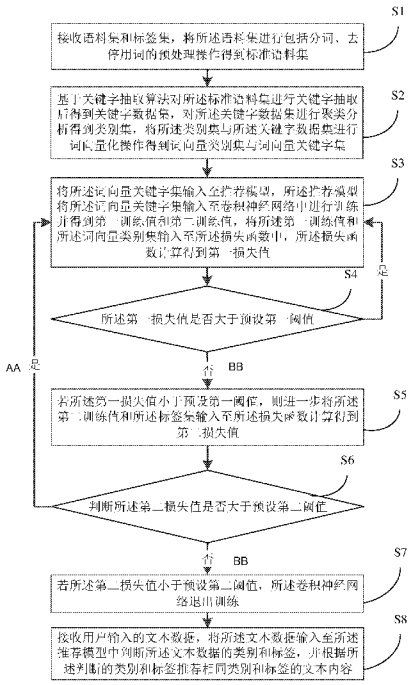
(72) 发明人: 金戈 (JIN, Ge); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。 徐亮 (XU, Liang); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。

(74) 代理人: 深圳市沃德知识产权代理事务所（普通合伙） (SHENZHEN WORLD INTELLECTUAL PROPERTY AGENCY (GENERAL PARTNERSHIP)); 中国广东省深圳市福田区园岭街道八卦四路10号中浩大厦1528-1530室于志光, Guangdong 518000 (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,

(54) **Title:** METHOD AND APPARATUS FOR INTELLIGENTLY RECOMMENDING PERSONALIZED TEXT, AND COMPUTER-READABLE STORAGE MEDIUM

(54) 发明名称: 个性化文本智能推荐方法、装置及计算机可读存储介质



S1 Receive a corpus and a tag set, and perform a preprocessing operation comprising word segmentation and stop word removal on the corpus to obtain a standard corpus
S2 Perform keyword extraction on the standard corpus on the basis of a keyword extraction algorithm to obtain a keyword data set, perform cluster analysis on the keyword data set to obtain a category set, and perform a word vectorization operation on the category set and the keyword data set to obtain a word vector category set and a word vector keyword set
S3 Input the word vector keyword set into a recommendation model; the recommendation model inputs the word vector keyword set into a convolutional neural network for training and obtains a first training value and a second training value, and inputs the first training value and the word vector category set into a loss function; and the loss function calculates a first loss value
S4 Is the first loss value greater than a preset first threshold
S5 If the first loss value is smaller than the preset first threshold, then further input the second training value and the tag set into the loss function to calculate a second loss value
S6 Determine whether the second loss value is greater than a preset second threshold
S7 If the second loss value is smaller than the preset second threshold, then the convolutional neural network exits training
S8 Receive text data inputted by a user, input the text data into the recommendation model to determine the category and tag of the text data, and recommend text content having the same category and tag according to the determined category and tag
AA Yes
BB No

图 1

(57) **Abstract:** A method and apparatus for intelligently recommending personalized text, and a computer-readable storage medium, which relate to artificial intelligence technology and achieve the accurate recommendation of personalized text. The method comprises: receiving a corpus and a tag set; preprocessing the corpus to obtain a standard corpus; performing keyword extraction on the standard corpus to obtain a keyword data set; performing cluster analysis on the keyword data set to obtain a category set; performing a word vectorization operation on the category set and the keyword data set to obtain a word vector category set and a word vector keyword set;

[见续页]

WO 2020/258481 A1

GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, 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))。

inputting the word vector category set and the word vector keyword set into a recommendation model for training until the recommendation model exits training; receiving text data inputted by a user; determining the category and tag of the text data; and recommending text content having the same category and tag from a database according to the determined category and tag.

(57) 摘要：一种个性化文本智能推荐方法、装置以及计算机可读存储介质，涉及人工智能技术，实现精准的个性化文本推荐。所述方法包括：接收语料集和标签集，将所述语料集进行预处理得到标准语料集，对所述标准语料集进行关键字抽取后得到关键字数据集，对所述关键字数据集进行聚类分析得到类别集，将所述类别集与所述关键字数据集进行词向量化操作得到词向量类别集与词向量关键字集，将所述词向量类别集与所述词向量关键字集输入至推荐模型训练，直至所述推荐模型退出训练，接收用户输入的文本数据，判断所述文本数据的类别和标签，并根据判断的所述类别和标签从数据库中推荐相同类别和标签的文本内容。

个性化文本智能推荐方法、装置及计算机可读存储介质

本申请要求于 2019 年 6 月 28 日提交中国专利局，申请号为201910582849.7、发明名称为“ 个性化文本智能推荐方法、装置及计算机可读存储介质” 的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及人工智能技术领域，尤其涉及一种个性化文本推荐方法、装置及计算机可读存储介质。

背景技术

随着互联网信息的快速增长,用户每天会浏览大量的文本数据，如果能从用户所浏览的文本数据中提取有用的关键字信息，从而进行个性化推荐，则能更高效的利用计算资源，并节约用户时间。目前国内外学者对推荐算法进行了相关研究，研究发现，其中数据稀疏性问题、冷启动问题以及用户兴趣获取问题都是影响推荐效果的重要因素。因此基于所述研究结果，现有的一些网站例如：电影、音乐、小说等使用神经规则引擎方法进行个性化推荐，所述神经规则引擎方法虽然精确，但是其方法僵硬脆弱，推荐的内容往往与用户实际所需的内容大相径庭，因此，个性化推荐的准确率有待进一步加强。

发明内容

本申请提供一种个性化文本智能推荐方法、装置及计算机可读存储介质，其主要目的在于当用户输入文本数据时，给用户精准的推荐与所述文本数据内容相近的文本数据。

为实现上述目的，本申请提供的一种个性化文本智能推荐方法，包括：接收包括基础文本数据集和场景文本数据集的语料集和标签集，将所述语料集进行包括分词、去停用词的预处理操作得到标准语料集；基于关键字抽取算法对所述标准语料集进行关键字抽取后得到关键字数据集，对所述关键字数据集进行聚类分析得到类别集，将所述类别集与所述关键字数据集进行词

向量化操作得到词向量类别集与词向量关键字集；将所述词向量关键字集输入至推荐模型，所述推荐模型将所述词向量关键字集输入至卷积神经网络中进行训练并得到第一训练值和第二训练值，将所述第一训练值和所述词向量类别集输入至损失函数中，所述损失函数计算得到第一损失值，判断所述第一损失值与预设第一阈值的大小，若所述第一损失值大于预设第一阈值，则所述卷积神经网络继续训练，若所述第一损失值小于预设第一阈值，则将所述第二训练值和所述标签集输入至所述损失函数计算得到第二损失值，判断所述第二损失值与预设第二阈值的大小，若所述第二损失值大于预设第二阈值，所述卷积神经网络继续训练，若所述第二损失值小于预设第二阈值，所述卷积神经网络退出训练；接收用户输入的文本数据，将所述文本数据输入至所述推荐模型中判断所述文本数据的类别和标签，并根据判断的所述类别和标签从数据库中推荐相同类别和标签的文本内容。

此外，为实现上述目的，本申请还提供一种个性化文本智能推荐装置，该装置包括存储器和处理器，所述存储器中存储有可在所述处理器上运行的个性化文本智能推荐程序，所述个性化文本智能推荐程序被所述处理器执行时实现如下步骤：接收包括基础文本数据集和场景文本数据集的语料集和标签集，将所述语料集进行包括分词、去停用词的预处理操作得到标准语料集；基于关键字抽取算法对所述标准语料集进行关键字抽取后得到关键字数据集，对所述关键字数据集进行聚类分析得到类别集，将所述类别集与所述关键字数据集进行词向量化操作得到词向量类别集与词向量关键字集；将所述词向量关键字集输入至推荐模型，所述推荐模型将所述词向量关键字集输入至卷积神经网络中进行训练并得到第一训练值和第二训练值，将所述第一训练值和所述词向量类别集输入至损失函数中，所述损失函数计算得到第一损失值，判断所述第一损失值与预设第一阈值的大小，若所述第一损失值大于预设第一阈值，则所述卷积神经网络继续训练，若所述第一损失值小于预设第一阈值，则将所述第二训练值和所述标签集输入至所述损失函数计算得到第二损失值，判断所述第二损失值与预设第二阈值的大小，若所述第二损失值大于预设第二阈值，所述卷积神经网络继续训练，若所述第二损失值小于预设第二阈值，所述卷积神经网络退出训练；接收用户输入的文本数据，将所述文本数据输入至所述推荐模型中判断所述文本数据的类别和标签，并根据判断

的所述类别和标签从数据库中推荐相同类别和标签的文本内容。

此外，为实现上述目的，本申请还提供一种计算机可读存储介质，所述计算机可读存储介质上存储有个性化文本智能推荐程序，所述个性化文本智能推荐程序可被一个或者多个处理器执行，以实现如上所述的个性化文本智能推荐方法的步骤。

本申请提出的个性化文本智能推荐方法、装置及计算机可读存储介质。本申请将文本数据分为基础文本和场景文本，提高了初期对文本数据内容的精确划分；同时构建了概率分词模型并最大化概率分词模型，提高对所述文本数据的特征提取，高效并最大化的利用到已有特征；另外基于深度学习的卷积神经网络可有效的利用所述特征进行学习，提高对文本数据的推荐能力。因此，本申请可为用户实现精确的个性化文本推荐。

附图说明

图1为本申请一实施例提供的个性化文本智能推荐方法的流程示意图；

图2为本申请一实施例提供的个性化文本智能推荐装置的内部结构示意图；

图3为本申请一实施例提供的个性化文本智能推荐装置中个性化文本智能推荐程序的模块示意图。

本申请目的的实现、功能特点及优点将结合实施例，参照附图做进一步说明。

具体实施方式

应当理解，此处所描述的具体实施例仅仅用以解释本申请，并不用于限定本申请。

本申请提供一种个性化文本智能推荐方法。参照图 1 所示，为本申请一实施例提供的个性化文本智能推荐方法的流程示意图。该方法可以由一个装置执行，该装置可以由软件和/或硬件实现。

在本实施例中，个性化文本智能推荐方法包括：

S1、接收语料集和标签集，将所述语料集进行包括分词、去停用词的预处理操作得到标准语料集。

本申请较佳实施例所述语料集包括文本数据，所述语料集可分为基础文本数据集和场景文本数据集。

进一步地所述基础文本数据集包括微博评论集、影电观后感集、音乐评论集等。所述微博评论集、所述影电观后感集、所述乐评论集都包括若干条数据。所述场景文本数据集包括股票评论集、政府工作报告评论集、公司季度年度财务报表评论集、大学生就业情况评论集。

优选地，所述标签集注明所述语料集内各文本数据所属领域。如文本数据：“威金病毒主要通过网络共享传播，病毒会感染电脑中所有的.EXE 可执行文件，传播速度十分迅速。威金病毒运行后，修改注册表自启动项，以使自己随系统一起运行，向系统文件目录下生成以下病毒文件”，在标签集中注明为“ 计算机领域”。

本申请较佳实施例中，所述分词包括根据所述语料集建立概率分词模型P(S)和最大化所述概率分词模型P(S) ,并利用最大化的所述概率分词模型P(S)对所述语料集执行分词操作。

其中，所述概率分词模型P(S)为：

$$\mathrm{P(S)} = \mathrm{P}(W_1, W_2, \dots, W_m) = \prod_{i=1}^{m} p(W_i | W_{i-1})$$

其中，$W_1, W_2, \dots, W_m$为所述语料集包括的词，m为所述语料集的数量，$p(W_i|W_{i-1})$表示在词W_{i-1}出现的情况下词W_i出现的概率；

所述最大化的所述概率分词模型P(S)：

$$argmax\mathrm{P(S)} = argmax \prod_{i=1}^{m} \frac{count(W_{i-1}, W_i)}{count(W_{i-1})}$$

其中，$count(W_{i-1}, W_i)$表示词W_{i-1}和词W_i同时出现在所述语料集内同一篇文本的文本数量，$count(W_{i-1})$表示词W_{i-1}出现在所述语料集内的文本数量，$argmax$表示最大化操作。

进一步所述停用词是文本数据中没有什么实际意义的词，且对文本的情感分析没有什么影响，但出现频率高的词，所述停用词包括常用的代词、介词等。如所述影电观后感集中用户 A 的影评为：其实大话西游中的至尊宝就像现实中的我们，都曾以为自己会成为盖世英雄，以为自己是这个世界的唯一，可是渐渐发现，自己与别人并没有什么不同。蓦然回首，原来最怀念的，是当初的我们。有一天你走着，别人会指着你的背影说：“他好像条狗啊！”，所以这部电影真的很耐人寻味发人深省。在去除停用词后则变为“ 其实大话

西游至尊宝像现实我们，都曾以为自己成为盖世英雄，以为自己这个世界唯一，渐渐发现，自己别人没有什么不同。蓦然回首，原来最怀念，是当初我们。有一天你走着，别人指着你背影说："他好像条狗！"，所以这部电影真耐人寻味发人深省"

本申请较佳实施例，所述去停用词的方法为停用词表过滤法，基于已构建好的停用词表和所述语料集的词进行一一匹配，若匹配成功，则该词为停用词，且将所述该词从所述语料集中删除。

S2、基于关键字抽取算法对所述标准语料集进行关键字抽取后得到关键字数据集，对所述关键字数据集进行聚类分析得到类别集，将所述类别集与所述关键字数据集进行词向量化操作得到词向量类别集与词向量关键字集。

较佳实施例所述关键字抽取算法包括：计算所述标准语料集中任意两词W_i, W_j之间的依存关联度$\mathrm{Dep}(W_i, W_j)$：

$$\mathrm{Dep}(W_i, W_j) = \frac{1}{b^{len(W_i, W_j)}}$$

其中，$len(W_i, W_j)$表示词语W_i和W_j之间的依存路径长度，b是超参数；

计算所述标准语料集中任意两词W_i, W_j之间的引力值$f_{grav}(W_i, W_j)$：

$$f_{grav}(W_i, W_j) = \frac{tfidf(W_i) * tfidf(W_j)}{d^2}$$

其中，$tfidf(W_i)$、$tfidf(W_j)$表示词W_i, W_j的词频-逆文本频率指数，d表示词W_i和W_j的词向量之间的欧式距离；

根据所述依存关联度$\mathrm{Dep}(W_i, W_j)$和所述引力值$f_{grav}(W_i, W_j)$判断所述标准语料集中任意两词W_i, W_j之间的权重系数$\mathrm{weight}(W_i, W_j)$：

$$\mathrm{weight}(W_i, W_j) = \mathrm{Dep}(W_i, W_j) * f_{grav}(W_i, W_j)$$

按照所述权重系数大小选择权重系数$\mathrm{weight}(W_i, W_j)$大的词，完成所述关键字抽取，得到关键字数据集。

本申请较佳实施例所述聚类分析包括随机化类别中心位置和最优化类别中心位置。

其中，所述随机化类别中心位置包括确定类别中心个数和随机生成所述类别中心的坐标位置，所述类别中心个数为所述基础文本数据集的种类和场景文本数据集的种类的总和。

所述最优化类别中心位置为：

$$\mathrm{dist}(x_i, x_j) = \sqrt{\sum_{d=1}^{D}(x_{i,d} - x_{j,d})^2}$$

其中，x_i, x_j为所述标准语料集的数据，$\mathrm{dist}(x_i, x_j)$为所述标准语料集数据之间的位置距离，D为所述类别中心个数。

较优地所述类别集是通过所述聚类分析后，得到的具有相似文本的文本集。如所述标准语料集中有文本数据 A：真生气，我不介意情怀买单，但我不想毫无诚意情怀买单，台词、表演、剧情、人物情感变化拿捏都不在水准，转场剪辑像支离破碎拼凑。出场噱头原来不过片场花絮，到处植入硬广摆明圈钱标语。希望你以后停止消费自己，如果要加一段期限，我希望直到宇宙毁灭。文本数据 B：垃圾，垃圾，垃圾，就会卖情怀，现在连情怀都不好好卖，一点内容都没有，只有出场噱头、植入广告，剪辑像是看 PPT 一样破碎不堪，没有剧情没有表演，台词对白僵硬。由于所述文本数据 A 与所述文本数据 B 在所述聚类分析中被判别有很多相同用词，属于相同类别，因此被划分为同一类别集中。

较佳实施例所述词向量化操作采用 Word2Vec 算法，所述 Word2Vec 算法包括输入层、投影层和输出层，所述输入层接收所述关键字数据集，所述输出层输出得到所述词向量集，所述投影层$\zeta(\omega, j)$为：

$$\zeta(\omega, j) = \left(1 - d_j^{\omega}\right) log\left[\sigma\left(X_{\omega}\theta_{j-1}^{\omega}\right)\right] + d_j^{\omega}\, log\left[1 - \sigma\left(X_{\omega}\theta_{j-1}^{\omega}\right)\right]$$

其中，$d_j^{\omega}\epsilon\{d_2^{\omega}, d_3^{\omega}, \ldots, d_{l^{\omega}}^{\omega}\}$，表示在路径$\omega$内，第$j$个结点对应的霍夫曼编码，$\theta$为所述 Word2Vec 模型的迭代因子，σ表示 sigmoid 函数，X_{ω}为所述关键字数据集。

本申请较佳实施例所述霍夫曼编码是根据数据通信知识使用 0,1 码的不同排列来表示所述关键字数据集。

S3、将所述词向量关键字集输入至推荐模型，所述推荐模型将所述词向量关键字集输入至卷积神经网络中进行训练并得到第一训练值和第二训练值，将所述第一训练值和所述词向量类别集输入至所述损失函数中，所述损失函数计算得到第一损失值。

优选地所述卷积神经网络包括卷积层、池化层、第一全连接层和第二全连接层。所述卷积层接收所述词向量关键字集并对所述词向量关键字集进行卷积操作得到卷积集。

进一步地所述卷积操作为：

$$\omega' = \frac{\omega + 2p - k}{s} + 1$$

其中ω'为所述卷积集，ω为所述词向量关键字集，k为卷积核的大小，s为所述卷积操作的步幅，p为数据补零矩阵。

本申请较佳实施例将所述卷积集输入至所述池化层，所述池化层寻找所述卷积集中各词向量数值最大的词向量并组成池化集。

进一步地将所述池化集同时输入至所述第一全连接层和所述第二全连接层，所述第一全连接层和所述第二全连接层根据激活函数输出所述训练值。所述激活函数为：

$$y = \frac{1}{1 + e^{-\omega}}$$

其中y为所述第一训练值或第二训练值，e为无限不循环小数。

较佳地所述第一损失值$E1$为：

$$E1 = \sum_{j=1}^{m} \sum_{x \in m} |x - \mu_j|^2$$

其中，x为所述第一训练值，μ_j为所述词向量类别集，m为所述类别集的数量。

S4、判断所述第一损失值与预设第一阈值的大小。

本申请较佳实施例所述预设第一阈值一般设定为 0.5。

若所述第一损失值大于预设第一阈值，则返回 S3，所述卷积神经网络继续训练。

当所述第一损失值大于所述预设第一阈值时，表明所述卷积神经网络对所述关键字数据集内各关键字的类别分类与所述聚类分析得到所述类别集误差较大，证明所述卷积神经网络识别类别能力较差，需继续训练。

S5、若所述第一损失值小于预设第一阈值，则进一步将所述第二训练值和所述标签集输入至所述损失函数计算得到第二损失值。

本申请较佳实施例所述第二损失值$E2$为：

$$E2 = \sum_{j=1}^{m} \sum_{x \in m} |x - \mu_j|^2$$

其中，x为所述第二训练值，μ_j为所述标签集，m为所述标签集的数量。

S6、判断所述第二损失值与预设第二阈值的大小。

较佳地所述预设第二阈值一般设置为 0.01。

若所述第二损失值大于预设第二阈值，返回 S3，所述卷积神经网络继续训练。

若所述第二损失值大于预设第二阈值，表明所述卷积神经网络对所述关键字数据集内各关键字的领域分类与所述标签集误差较大。如所述卷积神经网络接受到“操作系统”关键字，所述“操作系统”关键字在所述标签集中注明为“计算机”领域，但所述卷积神经网络可能会将所述“操作系统”关键字识别为“艺术”领域，表明所述卷积神经网络领域识别能力较差，需继续训练。

S7、若所述第二损失值小于预设第二阈值，所述卷积神经网络退出训练。

S8、接收用户输入的文本数据，将所述文本数据输入至所述推荐模型中判断所述文本数据的类别和标签，并根据所述判断的类别和标签推荐相同类别和标签的文本内容。

较佳地，如用户输入文本数据 X 为：历经四年，终于要划上句号了，今年的骑士，堪称是史上最烂的总决赛参赛球队，能拼到今天已经是难能可贵，老套的说一句吧，是“虽败犹荣”。至于 NBA 的两大命题，东部詹姆斯厉害和联盟抗勇无敌则仍然是未解之谜，只能在下个赛季拭目以待了。所述推荐模型根据所述聚类分析后得出与所述文本数据 X 有较多相同用词的文本数据，同时分析出所述文本数据 X 的标签输入 NBA 体育类，因此会智能化的推荐出相同类别和相同标签的文本供用户阅读。

发明还提供一种个性化文本智能推荐装置。参照图 2 所示，为本申请一实施例提供的个性化文本智能推荐装置的内部结构示意图。

在本实施例中，所述个性化文本智能推荐装置 1 可以是 PC（Personal Computer，个人电脑），或者是智能手机、平板电脑、便携计算机等终端设备，也可以是一种服务器等。该个性化文本智能推荐装置 1 至少包括存储器 11、处理器 12，通信总线 13，以及网络接口 14。

其中，存储器 11 至少包括一种类型的可读存储介质，所述可读存储介质包括闪存、硬盘、多媒体卡、卡型存储器（例如，SD 或 DX 存储器等）、磁性存储器、磁盘、光盘等。存储器 11 在一些实施例中可以是个性化文本智能推荐装置 1 的内部存储单元，例如该个性化文本智能推荐装置 1 的硬盘。存储器 11 在另一些实施例中也可以是个性化文本智能推荐装置 1 的外部存储设

备，例如个性化文本智能推荐装置 1 上配备的插接式硬盘，智能存储卡（Smart Media Card，SMC），安全数字（Secure Digital，SD）卡，闪存卡（Flash Card）等。进一步地，存储器 11 还可以既包括个性化文本智能推荐装置 1 的内部存储单元也包括外部存储设备。存储器 11 不仅可以用于存储安装于个性化文本智能推荐装置 1 的应用软件及各类数据，例如个性化文本智能推荐程序 01 的代码等，还可以用于暂时地存储已经输出或者将要输出的数据。

处理器 12 在一些实施例中可以是一中央处理器（Central Processing Unit，CPU）、控制器、微控制器、微处理器或其他数据处理芯片，用于运行存储器 11 中存储的程序代码或处理数据，例如执行个性化文本智能推荐程序 01 等。

通信总线 13 用于实现这些组件之间的连接通信。

网络接口 14 可选的可以包括标准的有线接口、无线接口（如 WI-FI 接口），通常用于在该装置 1 与其他电子设备之间建立通信连接。

可选地，该装置 1 还可以包括用户接口，用户接口可以包括显示器（Display）、输入单元比如键盘（Keyboard），可选的用户接口还可以包括标准的有线接口、无线接口。可选地，在一些实施例中，显示器可以是 LED 显示器、液晶显示器、触控式液晶显示器以及 OLED（Organic Light-Emitting Diode，有机发光二极管）触摸器等。其中，显示器也可以适当的称为显示屏或显示单元，用于显示在个性化文本智能推荐装置 1 中处理的信息以及用于显示可视化的用户界面。

图 2 仅示出了具有组件 11-14 以及个性化文本智能推荐程序 01 的个性化文本智能推荐装置 1，本领域技术人员可以理解的是，图 1 示出的结构并不构成对个性化文本智能推荐装置 1 的限定，可以包括比图示更少或者更多的部件，或者组合某些部件，或者不同的部件布置。

在图 2 所示的装置 1 实施例中，存储器 11 中存储有个性化文本智能推荐程序 01；处理器 12 执行存储器 11 中存储的个性化文本智能推荐程序 01 时实现如下步骤：

步骤一、接收语料集和标签集，将所述语料集进行包括分词、去停用词的预处理操作得到标准语料集。

本申请较佳实施例所述语料集包括文本数据，所述语料集可分为基础文本数据集和场景文本数据集。

进一步地所述基础文本数据集包括微博评论集、影电观后感集、音乐评论集等。所述微博评论集、所述影电观后感集、所述乐评论集都包括若干条数据。所述场景文本数据集包括股票评论集、政府工作报告评论集、公司季度年度财务报表评论集、大学生就业情况评论集。

优选地，所述标签集注明所述语料集内各文本数据所属领域。如文本数据：“威金病毒主要通过网络共享传播，病毒会感染电脑中所有的.EXE 可执行文件，传播速度十分迅速。威金病毒运行后，修改注册表自启动项，以使自己随系统一起运行，向系统文件目录下生成以下病毒文件”，在标签集中注明为“ 计算机领域”。

本申请较佳实施例中，所述分词包括根据所述语料集建立概率分词模型P(S)和最大化所述概率分词模型P(S) ,并利用最大化的所述概率分词模型P(S)对所述语料集执行分词操作。

其中，所述概率分词模型P(S)为：

$$\mathrm{P(S)} = \mathrm{P}(W_1, W_2, \ldots, W_m) = \prod_{i=1}^{m} p(W_i | W_{i-1})$$

其中，$W_1, W_2, \ldots, W_m$为所述语料集包括的词，m为所述语料集的数量，$p(W_i|W_{i-1})$表示在词W_{i-1}出现的情况下词W_i出现的概率；

所述最大化的所述概率分词模型P(S)：

$$argmax\mathrm{P(S)} = argmax \prod_{i=1}^{m} \frac{count(W_{i-1}, W_i)}{count(W_{i-1})}$$

其中，$count(W_{i-1}, W_i)$表示词W_{i-1}和词W_i同时出现在所述语料集内同一篇文本的文本数量，$count(W_{i-1})$表示词W_{i-1}出现在所述语料集内的文本数量，$argmax$表示最大化操作。

进一步所述停用词是文本数据中没有什么实际意义的词，且对文本的情感分析没有什么影响，但出现频率高的词，所述停用词包括常用的代词、介词等。如所述影电观后感集中用户 A 的影评为：其实大话西游中的至尊宝就像现实中的我们，都曾以为自己会成为盖世英雄，以为自己是这个世界的唯一，可是渐渐发现，自己与别人并没有什么不同。蓦然回首，原来最怀念的，是当初的我们。有一天你走着，别人会指着你的背影说：“他好像条狗啊！”，所以这部电影真的很耐人寻味发人深省。在去除停用词后则变为“ 其实大话西游至尊宝像现实我们，都曾以为自己成为盖世英雄，以为自己这个世界唯一，渐渐发现，自己别人没有什么不同。蓦然回首，原来最怀念，是当初我

们。有一天你走着，别人指着你背影说：“他好像条狗！”，所以这部电影真耐人寻味发人深省”

本申请较佳实施例，所述去停用词的方法为停用词表过滤法，基于已构建好的停用词表和所述语料集的词进行一一匹配，若匹配成功，则该词为停用词，且将所述该词从所述语料集中删除。

步骤二、基于关键字抽取算法对所述标准语料集进行关键字抽取后得到关键字数据集，对所述关键字数据集进行聚类分析得到类别集，将所述类别集与所述关键字数据集进行词向量化操作得到词向量类别集与词向量关键字集。

较佳实施例所述关键字抽取算法包括：计算所述标准语料集中任意两词W_i, W_j之间的依存关联度$\mathrm{Dep}(W_i, W_j)$：

$$\mathrm{Dep}(W_i, W_j) = \frac{1}{b^{len(W_i, W_j)}}$$

其中，$len(W_i, W_j)$表示词语W_i和W_j之间的依存路径长度，b是超参数；

计算所述标准语料集中任意两词W_i, W_j之间的引力值$f_{grav}(W_i, W_j)$：

$$f_{grav}(W_i, W_j) = \frac{tfidf(W_i) * tfidf(W_j)}{d^2}$$

其中，$tfidf(W_i)$、$tfidf(W_j)$表示词W_i, W_j的词频-逆文本频率指数，d表示词W_i和W_j的词向量之间的欧式距离；

根据所述依存关联度$\mathrm{Dep}(W_i, W_j)$和所述引力值$f_{grav}(W_i, W_j)$判断所述标准语料集中任意两词W_i, W_j之间的权重系数$\mathrm{weight}(W_i, W_j)$：

$$\mathrm{weight}(W_i, W_j) = \mathrm{Dep}(W_i, W_j) * f_{grav}(W_i, W_j)$$

按照所述权重系数大小选择权重系数$\mathrm{weight}(W_i, W_j)$最大的词，完成所述关键字抽取，得到关键字数据集。

本申请较佳实施例所述聚类分析包括随机化类别中心位置和最优化类别中心位置。

其中，所述随机化类别中心位置包括确定类别中心个数和随机生成所述类别中心的坐标位置，所述类别中心个数为所述基础文本数据集的种类和场景文本数据集的种类的总和。

所述最优化类别中心位置为：

$$\mathrm{dist}(x_i, x_j) = \sqrt{\sum_{d=1}^{D}(x_{i,d} - x_{j,d})^2}$$

其中，x_i, x_j为所述标准语料集的数据，$\mathrm{dist}(x_i, x_j)$为所述标准语料集数据之间的位置距离，D为所述类别中心个数。

较优地所述类别集是通过所述聚类分析后，得到的具有相似文本的文本集。如所述标准语料集中有文本数据 A：真生气，我不介意情怀买单，但我不想毫无诚意情怀买单，台词、表演、剧情、人物情感变化拿捏都不在水准，转场剪辑像支离破碎拼凑。出场嘘头原来不过片场花絮，到处植入硬广摆明圈钱标语。希望你以后停止消费自己，如果要加一段期限，我希望直到宇宙毁灭。文本数据 B：垃圾，垃圾，垃圾，就会卖情怀，现在连情怀都不好好卖，一点内容都没有，只有出场嘘头、植入广告，剪辑像是看 PPT 一样破碎不堪，没有剧情没有表演，台词对白僵硬。由于所述文本数据 A 与所述文本数据 B 在所述聚类分析中被判别有很多相同用词，属于相同类别，因此被划分为同一类别集中。

较佳实施例所述词向量化操作采用 Word2Vec 算法，所述 Word2Vec 算法包括输入层、投影层和输出层，所述输入层接收所述关键字数据集，所述输出层输出得到所述词向量集，所述投影层$\zeta(\omega, j)$为：

$$\zeta(\omega, j) = \left(1 - d_j^{\omega}\right) log\left[\sigma\left(X_{\omega}\theta_{j-1}^{\omega}\right)\right] + d_j^{\omega}\, log\left[1 - \sigma\left(X_{\omega}\theta_{j-1}^{\omega}\right)\right]$$

其中，$d_j^{\omega} \epsilon \{d_2^{\omega}, d_3^{\omega}, \ldots, d_l^{\omega}\}$，表示在路径$\omega$内，第$j$个结点对应的霍夫曼编码，$\theta$为所述 Word2Vec 模型的迭代因子，σ表示 sigmoid 函数，X_{ω}为所述关键字数据集。

本申请较佳实施例所述霍夫曼编码是根据数据通信知识使用 0,1 码的不同排列来表示所述关键字数据集。

步骤三、将所述词向量关键字集输入至推荐模型，所述推荐模型将所述词向量关键字集输入至卷积神经网络中进行训练并得到第一训练值和第二训练值，将所述第一训练值和所述词向量类别集输入至所述损失函数中，所述损失函数计算得到第一损失值。

优选地所述卷积神经网络包括卷积层、池化层、第一全连接层和第二全连接层。所述卷积层接收所述词向量关键字集并对所述词向量关键字集进行卷积操作得到卷积集。

进一步地所述卷积操作为：

$$\omega' = \frac{\omega + 2p - k}{s} + 1$$

其中ω'为所述卷积集，ω为所述词向量关键字集，k为卷积核的大小，s为所述卷积操作的步幅，p为数据补零矩阵。

本申请较佳实施例将所述卷积集输入至所述池化层，所述池化层寻找所述卷积集中各词向量数值最大的词向量并组成池化集。

进一步地将所述池化集同时输入至所述第一全连接层和所述第二全连接层，所述第一全连接层和所述第二全连接层根据激活函数输出所述训练值。所述激活函数为：

$$y = \frac{1}{1 + e^{-\omega}}$$

其中y为所述第一训练值或第二训练值，e为无限不循环小数。

较佳地所述第一损失值$E1$为：

$$E1 = \sum_{j=1}^{m} \sum_{x \in m} |x - \mu_j|^2$$

其中，x为所述第一训练值，μ_j为所述词向量类别集，m为所述类别集的数量。

步骤四、判断所述第一损失值与预设第一阈值的大小。

本申请较佳实施例所述预设第一阈值一般设定为 0.5。

若所述第一损失值大于预设第一阈值，则返回步骤三，所述卷积神经网络继续训练。

当所述第一损失值大于所述预设第一阈值时，表明所述卷积神经网络对所述关键字数据集内各关键字的类别分类与所述聚类分析得到所述类别集误差较大，证明所述卷积神经网络识别类别能力较差，需继续训练。

步骤五、若所述第一损失值小于预设第一阈值，则进一步将所述第二训练值和所述标签集输入至所述损失函数计算得到第二损失值。

本申请较佳实施例所述第二损失值$E2$为：

$$E2 = \sum_{j=1}^{m} \sum_{x \in m} |x - \mu_j|^2$$

其中，x为所述第二训练值，μ_j为所述标签集，m为所述标签集的数量。

步骤六、判断所述第二损失值与预设第二阈值的大小。

较佳地所述预设第二阈值一般设置为 0.01。

若所述第二损失值大于预设第二阈值，返回步骤三，所述卷积神经网络继续训练。

若所述第二损失值大于预设第二阈值，表明所述卷积神经网络对所述关键字数据集内各关键字的领域分类与所述标签集误差较大。如所述卷积神经网络接受到“操作系统”关键字，所述“操作系统”关键字在所述标签集中注明为“计算机”领域，但所述卷积神经网络可能会将所述“操作系统”关键字识别为“艺术”领域，表明所述卷积神经网络领域识别能力较差，需继续训练。

步骤七、若所述第二损失值小于预设第二阈值，所述卷积神经网络退出训练。

步骤八、接收用户输入的文本数据，将所述文本数据输入至所述推荐模型中判断所述文本数据的类别和标签，并根据所述判断的类别和标签推荐相同类别和标签的文本内容。

较佳地，如用户输入文本数据 X 为：历经四年，终于要划上句号了，今年的骑士，堪称是史上最烂的总决赛参赛球队，能拼到今天已经是难能可贵，老套的说一句吧，是“虽败犹荣”。至于 NBA 的两大命题，东部詹姆斯厉害和联盟抗勇无敌则仍然是未解之谜，只能在下个赛季拭目以待了。所述推荐模型根据所述聚类分析后得出与所述文本数据 X 有较多相同用词的文本数据，同时分析出所述文本数据 X 的标签输入 NBA 体育类，因此会智能化的推荐出相同类别和相同标签的文本供用户阅读。

可选地，在其他实施例中，个性化文本智能推荐程序还可以被分割为一个或者多个模块，一个或者多个模块被存储于存储器 11 中，并由一个或多个处理器（本实施例为处理器 12）所执行以完成本申请，本申请所称的模块是指能够完成特定功能的一系列计算机程序指令段，用于描述个性化文本智能推荐程序在个性化文本智能推荐装置中的执行过程。

例如，参照图 3 所示，为本申请个性化文本智能推荐装置一实施例中的个性化文本智能推荐程序的程序模块示意图，该实施例中，所述个性化文本智能推荐程序可以被分割为源数据接收模块 10、特征提取模块 20、特征分析模块 30 以及个性化文本输出模块 40，示例性地：

所述源数据接收模块 10 用于：接收包括基础文本数据集和场景文本数据

集的语料集和标签集，将所述语料集进行包括分词、去停用词的预处理操作得到标准语料集。

所述特征提取模块 20 用于：基于关键字抽取算法对所述标准语料集进行关键字抽取后得到关键字数据集，对所述关键字数据集进行聚类分析得到类别集，将所述类别集与所述关键字数据集进行词向量化操作得到词向量类别集与词向量关键字集。

所述特征分析模块 30 用于：将所述词向量关键字集输入至推荐模型，所述推荐模型将所述词向量关键字集输入至卷积神经网络中进行训练并得到第一训练值和第二训练值，将所述第一训练值和所述词向量类别集输入至损失函数中，所述损失函数计算得到第一损失值，判断所述第一损失值与预设第一阈值的大小，若所述第一损失值大于预设第一阈值，则所述卷积神经网络继续训练，若所述第一损失值小于预设第一阈值，则将所述第二训练值和所述标签集输入至所述损失函数计算得到第二损失值，判断所述第二损失值与预设第二阈值的大小，若所述第二损失值大于预设第二阈值，所述卷积神经网络继续训练，若所述第二损失值小于预设第二阈值，所述卷积神经网络退出训练。

所述个性化文本输出模块 40 用于：接收用户输入的文本数据，将所述文本数据输入至所述推荐模型中判断所述文本数据的类别和标签，并根据判断的所述类别和标签从数据库中推荐相同类别和标签的文本内容。

上述源数据接收模块 10、特征提取模块 20、特征分析模块 30 以及个性化文本输出模块 40 等程序模块被执行时所实现的功能或操作步骤与上述实施例大体相同，在此不再赘述。

此外，本申请实施例还提出一种计算机可读存储介质，所述计算机可读存储介质上存储有个性化文本智能推荐程序，所述个性化文本智能推荐程序可被一个或多个处理器执行，以实现如下操作：

接收包括基础文本数据集和场景文本数据集的语料集和标签集，将所述语料集进行包括分词、去停用词的预处理操作得到标准语料集；

基于关键字抽取算法对所述标准语料集进行关键字抽取后得到关键字数据集，对所述关键字数据集进行聚类分析得到类别集，将所述类别集与所述

关键字数据集进行词向量化操作得到词向量类别集与词向量关键字集；

将所述词向量关键字集输入至推荐模型，所述推荐模型将所述词向量关键字集输入至卷积神经网络中进行训练并得到第一训练值和第二训练值，将所述第一训练值和所述词向量类别集输入至损失函数中，所述损失函数计算得到第一损失值，判断所述第一损失值与预设第一阈值的大小，若所述第一损失值大于预设第一阈值，则所述卷积神经网络继续训练，若所述第一损失值小于预设第一阈值，则将所述第二训练值和所述标签集输入至所述损失函数计算得到第二损失值，判断所述第二损失值与预设第二阈值的大小，若所述第二损失值大于预设第二阈值，所述卷积神经网络继续训练，若所述第二损失值小于预设第二阈值，所述卷积神经网络退出训练；

接收用户输入的文本数据，将所述文本数据输入至所述推荐模型中判断所述文本数据的类别和标签，并根据判断的所述类别和标签从数据库中推荐相同类别和标签的文本内容。

本申请计算机可读存储介质具体实施方式与上述个性化文本智能推荐装置和方法各实施例基本相同，在此不作累述。

需要说明的是，上述本申请实施例序号仅仅为了描述，不代表实施例的优劣。并且本文中的术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、装置、物品或者方法不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、装置、物品或者方法所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括该要素的过程、装置、物品或者方法中还存在另外的相同要素。

通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件，但很多情况下前者是更佳的实施方式。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在如上所述的一个存储介质(如 ROM/RAM、磁碟、光盘)中，包括若干指令用以使得一台终端设备(可以是手机，计算机，服务器，或者网络设备等)执行本申请各个实施例所述的方法。

以上仅为本申请的优选实施例，并非因此限制本申请的专利范围，凡是利用本申请说明书及附图内容所作的等效结构或等效流程变换，或直接或间接运用在其他相关的技术领域，均同理包括在本申请的专利保护范围内。

权 利 要 求 书

1、一种个性化文本智能推荐方法，其特征在于，所述方法包括：

接收包括基础文本数据集和场景文本数据集的语料集和标签集，将所述语料集进行包括分词、去停用词的预处理操作得到标准语料集；

基于关键字抽取算法对所述标准语料集进行关键字抽取后得到关键字数据集，对所述关键字数据集进行聚类分析得到类别集，将所述类别集与所述关键字数据集进行词向量化操作得到词向量类别集与词向量关键字集；

将所述词向量关键字集输入至推荐模型，所述推荐模型将所述词向量关键字集输入至卷积神经网络中进行训练并得到第一训练值和第二训练值，将所述第一训练值和所述词向量类别集输入至损失函数中，所述损失函数计算得到第一损失值，判断所述第一损失值与预设第一阈值的大小，若所述第一损失值大于预设第一阈值，则所述卷积神经网络继续训练，若所述第一损失值小于预设第一阈值，则将所述第二训练值和所述标签集输入至所述损失函数计算得到第二损失值，判断所述第二损失值与预设第二阈值的大小，若所述第二损失值大于预设第二阈值，所述卷积神经网络继续训练，若所述第二损失值小于预设第二阈值，所述卷积神经网络退出训练；

接收用户输入的文本数据，将所述文本数据输入至所述推荐模型中判断所述文本数据的类别和标签，并根据判断的所述类别和标签从数据库中推荐相同类别和标签的文本内容。

2、如权利要求 1 所述的个性化文本智能推荐方法，其特征在于，所述基础文本数据集包括微博评论集、影电观后感集、音乐评论集；

所述场景文本数据集包括股票评论集、政府工作报告评论集、公司季度年度财务报表评论集、大学生就业情况评论集。

3、如权利要求 1 所述的个性化文本智能推荐方法，其特征在于，所述分词包括：

根据所述语料集建立概率分词模型P(S)和最大化所述概率分词模型P(S)，并利用最大化的所述概率分词模型P(S)对所述语料集执行分词操作；

其中，所述概率分词模型P(S)为：

$$\mathrm{P(S)} = \mathrm{P}(W_1, W_2, \ldots, W_m) = \prod_{i=1}^{m} p(W_i \mid W_{i-1})$$

其中，$W_1, W_2, \ldots, W_m$为所述语料集包括的词，m为所述语料集的数量，

$p(W_i|W_{i-1})$表示在词W_{i-1}出现的情况下词W_i出现的概率；

所述最大化的所述概率分词模型P(S)：

$$argmax\mathrm{P(S)} = argmax\prod_{i=1}^{m}\frac{count(W_{i-1},W_i)}{count(W_{i-1})}$$

其中，$count(W_{i-1},W_i)$表示词W_{i-1}和词W_i同时出现在所述语料集内同一篇文本的文本数量，$count(W_{i-1})$表示词W_{i-1}出现在所述语料集内的文本数量，$argmax$表示最大化操作。

4、如权利要求 1 所述的个性化文本智能推荐方法，其特征在于，基于关键字抽取算法对所述标准语料集进行关键字抽取后得到关键字数据集，包括：

计算所述标准语料集中任意两词W_i,W_j之间的依存关联度$\mathrm{Dep}(W_i,W_j)$：

$$\mathrm{Dep}(W_i,W_j) = \frac{1}{b^{len(W_i,W_j)}}$$

其中，$len(W_i,W_j)$表示词语W_i和W_j之间的依存路径长度，b是超参数；

计算所述标准语料集中任意两词W_i,W_j之间的引力值$f_{grav}(W_i,W_j)$：

$$f_{grav}(W_i,W_j) = \frac{tfidf(W_i)*tfidf(W_j)}{d^2}$$

其中，$tfidf(W_i)$、$tfidf(W_j)$表示词W_i,W_j的词频-逆文本频率指数，d表示词W_i和W_j的词向量之间的欧式距离；

根据所述依存关联度$\mathrm{Dep}(W_i,W_j)$和所述引力值$f_{grav}(W_i,W_j)$判断所述标准语料集中任意两词W_i,W_j之间的权重系数$\mathrm{weight}(W_i,W_j)$：

$$\mathrm{weight}(W_i,W_j) = \mathrm{Dep}(W_i,W_j)*f_{grav}(W_i,W_j)$$

按照所述权重系数大小选择权重系数$\mathrm{weight}(W_i,W_j)$最大的词，完成所述关键字抽取，得到关键字数据集。

5、如权利要求 2 所述的个性化文本智能推荐方法，其特征在于，基于关键字抽取算法对所述标准语料集进行关键字抽取后得到关键字数据集，包括：

计算所述标准语料集中任意两词W_i,W_j之间的依存关联度$\mathrm{Dep}(W_i,W_j)$：

$$\mathrm{Dep}(W_i,W_j) = \frac{1}{b^{len(W_i,W_j)}}$$

其中，$len(W_i,W_j)$表示词语W_i和W_j之间的依存路径长度，b是超参数；

计算所述标准语料集中任意两词W_i,W_j之间的引力值$f_{grav}(W_i,W_j)$：

$$f_{grav}(W_i,W_j) = \frac{tfidf(W_i)*tfidf(W_j)}{d^2}$$

其中，$tfidf(W_i)$、$tfidf(W_j)$表示词W_i,W_j的词频-逆文本频率指数，d表示词W_i和W_j的词向量之间的欧式距离；

根据所述依存关联度$\mathrm{Dep}(W_i,W_j)$和所述引力值$f_{grav}(W_i,W_j)$判断所述标准

语料集中任意两词W_i, W_j之间的权重系数$\text{weight}(W_i, W_j)$：

$$\text{weight}(W_i, W_j) = \text{Dep}(W_i, W_j) * f_{grav}(W_i, W_j)$$

按照所述权重系数大小选择权重系数$\text{weight}(W_i, W_j)$最大的词，完成所述关键字抽取，得到关键字数据集。

6、如权利要求 3 所述的个性化文本智能推荐方法，其特征在于，基于关键字抽取算法对所述标准语料集进行关键字抽取后得到关键字数据集，包括：

计算所述标准语料集中任意两词W_i, W_j之间的依存关联度$\text{Dep}(W_i, W_j)$：

$$\text{Dep}(W_i, W_j) = \frac{1}{b^{len(W_i, W_j)}}$$

其中，$len(W_i, W_j)$表示词语W_i和W_j之间的依存路径长度，b是超参数；

计算所述标准语料集中任意两词W_i, W_j之间的引力值$f_{grav}(W_i, W_j)$：

$$f_{grav}(W_i, W_j) = \frac{tfidf(W_i) * tfidf(W_j)}{d^2}$$

其中，$tfidf(W_i)$、$tfidf(W_j)$表示词W_i, W_j的词频-逆文本频率指数，d表示词W_i和W_j的词向量之间的欧式距离；

根据所述依存关联度$\text{Dep}(W_i, W_j)$和所述引力值$f_{grav}(W_i, W_j)$判断所述标准语料集中任意两词W_i, W_j之间的权重系数$\text{weight}(W_i, W_j)$：

$$\text{weight}(W_i, W_j) = \text{Dep}(W_i, W_j) * f_{grav}(W_i, W_j)$$

按照所述权重系数大小选择权重系数$\text{weight}(W_i, W_j)$最大的词，完成所述关键字抽取，得到关键字数据集。

7、如权利要求 1 所述的个性化文本智能推荐方法，其特征在于，所述聚类分析包括随机化类别中心位置和最优化类别中心位置；

其中，所述随机化类别中心位置包括确定类别中心个数和随机生成所述类别中心的坐标位置，所述类别中心个数为所述基础文本数据集的种类和场景文本数据集的种类的总和；

所述最优化类别中心位置为：

$$\text{dist}(x_i, x_j) = \sqrt{\sum_{d=1}^{D}(x_{i,d} - x_{j,d})^2}$$

其中，x_i, x_j为所述标准语料集的数据，$\text{dist}(x_i, x_j)$为所述标准语料集数据之间的位置距离，D为所述类别中心个数。

8、一种个性化文本智能推荐装置，其特征在于，所述装置包括存储器和处理器，所述存储器上存储有可在所述处理器上运行的个性化文本智能推荐

程序，所述个性化文本智能推荐程序被所述处理器执行时实现如下步骤：

接收包括基础文本数据集和场景文本数据集的语料集和标签集，将所述语料集进行包括分词、去停用词的预处理操作得到标准语料集；

基于关键字抽取算法对所述标准语料集进行关键字抽取后得到关键字数据集，对所述关键字数据集进行聚类分析得到类别集，将所述类别集与所述关键字数据集进行词向量化操作得到词向量类别集与词向量关键字集；

将所述词向量关键字集输入至推荐模型，所述推荐模型将所述词向量关键字集输入至卷积神经网络中进行训练并得到第一训练值和第二训练值，将所述第一训练值和所述词向量类别集输入至损失函数中，所述损失函数计算得到第一损失值，判断所述第一损失值与预设第一阈值的大小，若所述第一损失值大于预设第一阈值，则所述卷积神经网络继续训练，若所述第一损失值小于预设第一阈值，则将所述第二训练值和所述标签集输入至所述损失函数计算得到第二损失值，判断所述第二损失值与预设第二阈值的大小，若所述第二损失值大于预设第二阈值，所述卷积神经网络继续训练，若所述第二损失值小于预设第二阈值，所述卷积神经网络退出训练；

接收用户输入的文本数据，将所述文本数据输入至所述推荐模型中判断所述文本数据的类别和标签，并根据判断的所述类别和标签从数据库中推荐相同类别和标签的文本内容。

9、如权利要求 8 所述的个性化文本智能推荐装置，其特征在于，所述基础文本数据集包括微博评论集、影电观后感集、音乐评论集；

所述场景文本数据集包括股票评论集、政府工作报告评论集、公司季度年度财务报表评论集、大学生就业情况评论集。

10、如权利要求 8 所述的个性化文本智能推荐装置，其特征在于，所述分词包括：

根据所述语料集建立概率分词模型P(S)和最大化所述概率分词模型P(S)，并利用最大化的所述概率分词模型P(S)对所述语料集执行分词操作；

其中，所述概率分词模型P(S)为：

$$\mathrm{P(S)} = \mathrm{P}(W_1, W_2, \ldots, W_m) = \prod_{i=1}^{m} p(W_i \mid W_{i-1})$$

其中，$W_1, W_2, \ldots, W_m$为所述语料集包括的词，m为所述语料集的数量，$p(W_i|W_{i-1})$表示在词W_{i-1}出现的情况下词W_i出现的概率；

所述最大化的所述概率分词模型P(S)：

$$argmax\mathrm{P(S)} = argmax\prod\prod_{i=1}^{m}\frac{count(W_{i-1},W_i)}{count(W_{i-1})}$$

其中，$count(W_{i-1},W_i)$表示词W_{i-1}和词W_i同时出现在所述语料集内同一篇文本的文本数量，$count(W_{i-1})$表示词W_{i-1}出现在所述语料集内的文本数量，$argmax$表示最大化操作。

11、如权利要求 8 所述的个性化文本智能推荐装置，其特征在于，基于关键字抽取算法对所述标准语料集进行关键字抽取后得到关键字数据集，包括：

计算所述标准语料集中任意两词W_i,W_j之间的依存关联度$\mathrm{Dep}(W_i,W_j)$：

$$\mathrm{Dep}(W_i,W_j) = \frac{1}{b^{len(W_i,W_j)}}$$

其中，$len(W_i,W_j)$表示词语W_i和W_j之间的依存路径长度，b是超参数；

计算所述标准语料集中任意两词W_i,W_j之间的引力值$f_{grav}(W_i,W_j)$：

$$f_{grav}(W_i,W_j) = \frac{tfidf(W_i)*tfidf(W_j)}{d^2}$$

其中，$tfidf(W_i)$、$tfidf(W_j)$表示词W_i,W_j的词频-逆文本频率指数，d表示词W_i和W_j的词向量之间的欧式距离；

根据所述依存关联度$\mathrm{Dep}(W_i,W_j)$和所述引力值$f_{grav}(W_i,W_j)$判断所述标准语料集中任意两词W_i,W_j之间的权重系数$\mathrm{weight}(W_i,W_j)$：

$$\mathrm{weight}(W_i,W_j) = \mathrm{Dep}(W_i,W_j)*f_{grav}(W_i,W_j)$$

按照所述权重系数大小选择权重系数$\mathrm{weight}(W_i,W_j)$最大的词，完成所述关键字抽取，得到关键字数据集。

12、如权利要求 9 所述的个性化文本智能推荐装置，其特征在于，基于关键字抽取算法对所述标准语料集进行关键字抽取后得到关键字数据集，包括：

计算所述标准语料集中任意两词W_i,W_j之间的依存关联度$\mathrm{Dep}(W_i,W_j)$：

$$\mathrm{Dep}(W_i,W_j) = \frac{1}{b^{len(W_i,W_j)}}$$

其中，$len(W_i,W_j)$表示词语W_i和W_j之间的依存路径长度，b是超参数；

计算所述标准语料集中任意两词W_i,W_j之间的引力值$f_{grav}(W_i,W_j)$：

$$f_{grav}(W_i,W_j) = \frac{tfidf(W_i)*tfidf(W_j)}{d^2}$$

其中，$tfidf(W_i)$、$tfidf(W_j)$表示词W_i,W_j的词频-逆文本频率指数，d表

示词W_i和W_j的词向量之间的欧式距离；

根据所述依存关联度$\mathrm{Dep}(W_i,W_j)$和所述引力值$f_{grav}(W_i,W_j)$判断所述标准语料集中任意两词W_i,W_j之间的权重系数$\mathrm{weight}(W_i,W_j)$：

$$\mathrm{weight}(W_i,W_j)=\mathrm{Dep}(W_i,W_j)*f_{grav}(W_i,W_j)$$

按照所述权重系数大小选择权重系数$\mathrm{weight}(W_i,W_j)$最大的词，完成所述关键字抽取，得到关键字数据集。

13、如权利要求 10 所述的个性化文本智能推荐装置，其特征在于，基于关键字抽取算法对所述标准语料集进行关键字抽取后得到关键字数据集，包括：

计算所述标准语料集中任意两词W_i,W_j之间的依存关联度$\mathrm{Dep}(W_i,W_j)$：

$$\mathrm{Dep}(W_i,W_j)=\frac{1}{b^{len(W_i,W_j)}}$$

其中，$len(W_i,W_j)$表示词语W_i和W_j之间的依存路径长度，b是超参数；

计算所述标准语料集中任意两词W_i,W_j之间的引力值$f_{grav}(W_i,W_j)$：

$$f_{grav}(W_i,W_j)=\frac{tfidf(W_i)*tfidf(W_j)}{d^2}$$

其中，$tfidf(W_i)$、$tfidf(W_j)$表示词W_i,W_j的词频-逆文本频率指数，d表示词W_i和W_j的词向量之间的欧式距离；

根据所述依存关联度$\mathrm{Dep}(W_i,W_j)$和所述引力值$f_{grav}(W_i,W_j)$判断所述标准语料集中任意两词W_i,W_j之间的权重系数$\mathrm{weight}(W_i,W_j)$：

$$\mathrm{weight}(W_i,W_j)=\mathrm{Dep}(W_i,W_j)*f_{grav}(W_i,W_j)$$

按照所述权重系数大小选择权重系数$\mathrm{weight}(W_i,W_j)$最大的词，完成所述关键字抽取，得到关键字数据集。

14、如权利要求 8 所述的个性化文本智能推荐装置，其特征在于，所述聚类分析包括随机化类别中心位置和最优化类别中心位置；

其中，所述随机化类别中心位置包括确定类别中心个数和随机生成所述类别中心的坐标位置，所述类别中心个数为所述基础文本数据集的种类和场景文本数据集的种类的总和；

所述最优化类别中心位置为：

$$\mathrm{dist}(x_i,x_j)=\sqrt{\sum_{d=1}^{D}(x_{i,d}-x_{j,d})^2}$$

其中，x_i,x_j为所述标准语料集的数据，$\mathrm{dist}(x_i,x_j)$为所述标准语料集数据

之间的位置距离，D为所述类别中心个数。

15、一种计算机可读存储介质，其特征在于，所述计算机可读存储介质上存储有个性化文本智能推荐程序，所述个性化文本智能推荐程序可被一个或者多个处理器执行，以实现如下步骤：

接收包括基础文本数据集和场景文本数据集的语料集和标签集，将所述语料集进行包括分词、去停用词的预处理操作得到标准语料集；

基于关键字抽取算法对所述标准语料集进行关键字抽取后得到关键字数据集，对所述关键字数据集进行聚类分析得到类别集，将所述类别集与所述关键字数据集进行词向量化操作得到词向量类别集与词向量关键字集；

将所述词向量关键字集输入至推荐模型，所述推荐模型将所述词向量关键字集输入至卷积神经网络中进行训练并得到第一训练值和第二训练值，将所述第一训练值和所述词向量类别集输入至损失函数中，所述损失函数计算得到第一损失值，判断所述第一损失值与预设第一阈值的大小，若所述第一损失值大于预设第一阈值，则所述卷积神经网络继续训练，若所述第一损失值小于预设第一阈值，则将所述第二训练值和所述标签集输入至所述损失函数计算得到第二损失值，判断所述第二损失值与预设第二阈值的大小，若所述第二损失值大于预设第二阈值，所述卷积神经网络继续训练，若所述第二损失值小于预设第二阈值，所述卷积神经网络退出训练；

接收用户输入的文本数据，将所述文本数据输入至所述推荐模型中判断所述文本数据的类别和标签，并根据判断的所述类别和标签从数据库中推荐相同类别和标签的文本内容。

16、如权利要求 15 所述的计算机可读存储介质，其特征在于，所述基础文本数据集包括微博评论集、影电观后感集、音乐评论集；

所述场景文本数据集包括股票评论集、政府工作报告评论集、公司季度年度财务报表评论集、大学生就业情况评论集。

17、如权利要求 15 所述的个性化文本智能推荐装置，其特征在于，所述分词包括：

根据所述语料集建立概率分词模型P(S)和最大化所述概率分词模型P(S)，并利用最大化的所述概率分词模型P(S)对所述语料集执行分词操作；

其中，所述概率分词模型P(S)为：

$$\mathrm{P(S)} = \mathrm{P}(W_1, W_2, \ldots, W_m) = \prod_{i=1}^{m} p(W_i \mid W_{i-1})$$

其中，$W_1, W_2, \ldots, W_m$为所述语料集包括的词，m为所述语料集的数量，$p(W_i|W_{i-1})$表示在词W_{i-1}出现的情况下词W_i出现的概率；

所述最大化的所述概率分词模型$\mathrm{P(S)}$：

$$argmax\mathrm{P(S)} = argmax \prod_{i=1}^{m} \frac{count(W_{i-1}, W_i)}{count(W_{i-1})}$$

其中，$count(W_{i-1}, W_i)$表示词W_{i-1}和词W_i同时出现在所述语料集内同一篇文本的文本数量，$count(W_{i-1})$表示词W_{i-1}出现在所述语料集内的文本数量，$argmax$表示最大化操作。

18、如权利要求 15 所述的计算机可读存储介质，其特征在于，基于关键字抽取算法对所述标准语料集进行关键字抽取后得到关键字数据集，包括：

计算所述标准语料集中任意两词W_i, W_j之间的依存关联度$\mathrm{Dep}(W_i, W_j)$：

$$\mathrm{Dep}(W_i, W_j) = \frac{1}{b^{len(W_i, W_j)}}$$

其中，$len(W_i, W_j)$表示词语W_i和W_j之间的依存路径长度，b是超参数；

计算所述标准语料集中任意两词W_i, W_j之间的引力值$f_{grav}(W_i, W_j)$：

$$f_{grav}(W_i, W_j) = \frac{tfidf(W_i) * tfidf(W_j)}{d^2}$$

其中，$tfidf(W_i)$、$tfidf(W_j)$表示词W_i, W_j的词频-逆文本频率指数，d表示词W_i和W_j的词向量之间的欧式距离；

根据所述依存关联度$\mathrm{Dep}(W_i, W_j)$和所述引力值$f_{grav}(W_i, W_j)$判断所述标准语料集中任意两词W_i, W_j之间的权重系数$\mathrm{weight}(W_i, W_j)$：

$$\mathrm{weight}(W_i, W_j) = \mathrm{Dep}(W_i, W_j) * f_{grav}(W_i, W_j)$$

按照所述权重系数大小选择权重系数$\mathrm{weight}(W_i, W_j)$最大的词，完成所述关键字抽取，得到关键字数据集。

19、如权利要求 16 或 17 所述的计算机可读存储介质，其特征在于，基于关键字抽取算法对所述标准语料集进行关键字抽取后得到关键字数据集，包括：

计算所述标准语料集中任意两词W_i, W_j之间的依存关联度$\mathrm{Dep}(W_i, W_j)$：

$$\mathrm{Dep}(W_i, W_j) = \frac{1}{b^{len(W_i, W_j)}}$$

其中，$len(W_i, W_j)$表示词语W_i和W_j之间的依存路径长度，b是超参数；

计算所述标准语料集中任意两词W_i, W_j之间的引力值$f_{grav}(W_i, W_j)$：

$$f_{grav}(W_i, W_j) = \frac{tfidf(W_i) * tfidf(W_j)}{d^2}$$

其中，$tfidf(W_i)$、$tfidf(W_j)$表示词W_i, W_j的词频-逆文本频率指数，d表示词W_i和W_j的词向量之间的欧式距离；

根据所述依存关联度$\mathrm{Dep}(W_i, W_j)$和所述引力值$f_{grav}(W_i, W_j)$判断所述标准语料集中任意两词W_i, W_j之间的权重系数$\mathrm{weight}(W_i, W_j)$：

$$\mathrm{weight}(W_i, W_j) = \mathrm{Dep}(W_i, W_j) * f_{grav}(W_i, W_j)$$

按照所述权重系数大小选择权重系数$\mathrm{weight}(W_i, W_j)$最大的词，完成所述关键字抽取，得到关键字数据集。

20、如权利要求 15 所述的计算机可读存储介质，其特征在于，所述聚类分析包括随机化类别中心位置和最优化类别中心位置；

其中，所述随机化类别中心位置包括确定类别中心个数和随机生成所述类别中心的坐标位置，所述类别中心个数为所述基础文本数据集的种类和场景文本数据集的种类的总和；

所述最优化类别中心位置为：

$$\mathrm{dist}(x_i, x_j) = \sqrt{\sum_{d=1}^{D}(x_{i,d} - x_{j,d})^2}$$

其中，x_i, x_j为所述标准语料集的数据，$\mathrm{dist}(x_i, x_j)$为所述标准语料集数据之间的位置距离，D为所述类别中心个数。

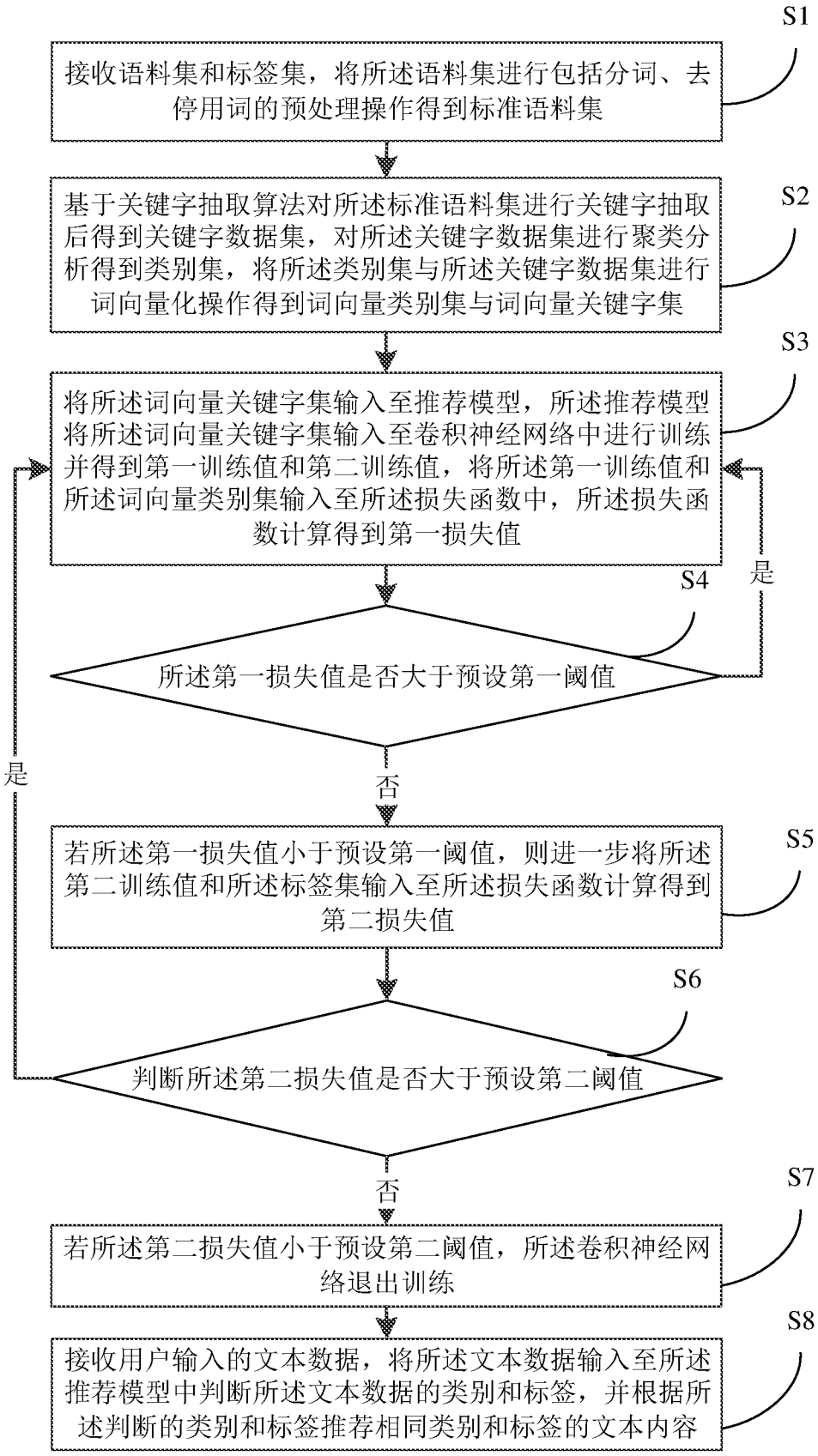

S1
接收语料集和标签集，将所述语料集进行包括分词、去停用词的预处理操作得到标准语料集
S2
基于关键字抽取算法对所述标准语料集进行关键字抽取后得到关键字数据集，对所述关键字数据集进行聚类分析得到类别集，将所述类别集与所述关键字数据集进行词向量化操作得到词向量类别集与词向量关键字集
S3
将所述词向量关键字集输入至推荐模型，所述推荐模型将所述词向量关键字集输入至卷积神经网络中进行训练并得到第一训练值和第二训练值，将所述第一训练值和所述词向量类别集输入至所述损失函数中，所述损失函数计算得到第一损失值
S4
所述第一损失值是否大于预设第一阈值
是
否
S5
若所述第一损失值小于预设第一阈值，则进一步将所述第二训练值和所述标签集输入至所述损失函数计算得到第二损失值
S6
判断所述第二损失值是否大于预设第二阈值
是
否
S7
若所述第二损失值小于预设第二阈值，所述卷积神经网络退出训练
S8
接收用户输入的文本数据，将所述文本数据输入至所述推荐模型中判断所述文本数据的类别和标签，并根据所述判断的类别和标签推荐相同类别和标签的文本内容

图 1

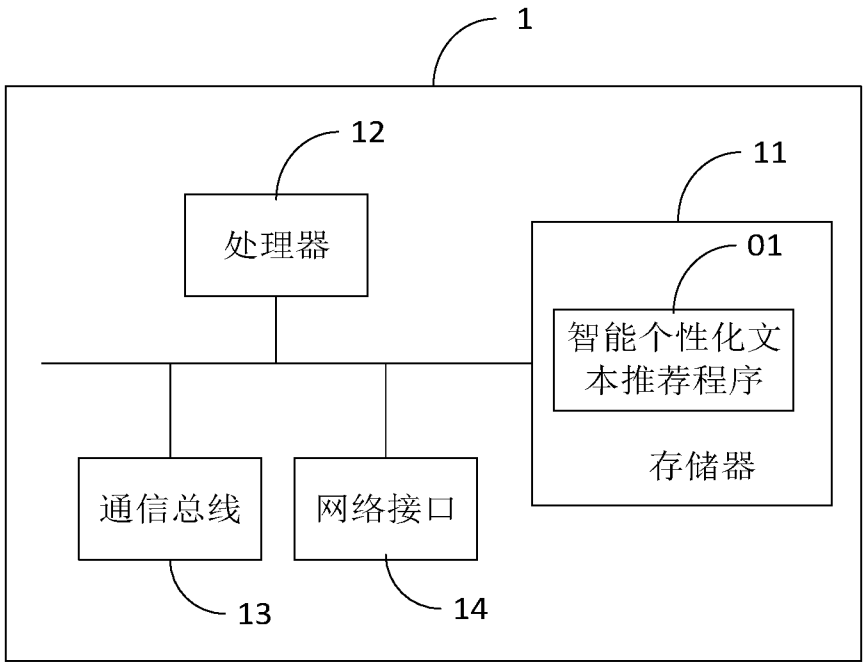

1
12
处理器
11
01
智能个性化文本推荐程序
存储器
通信总线
网络接口
13
14

图 2

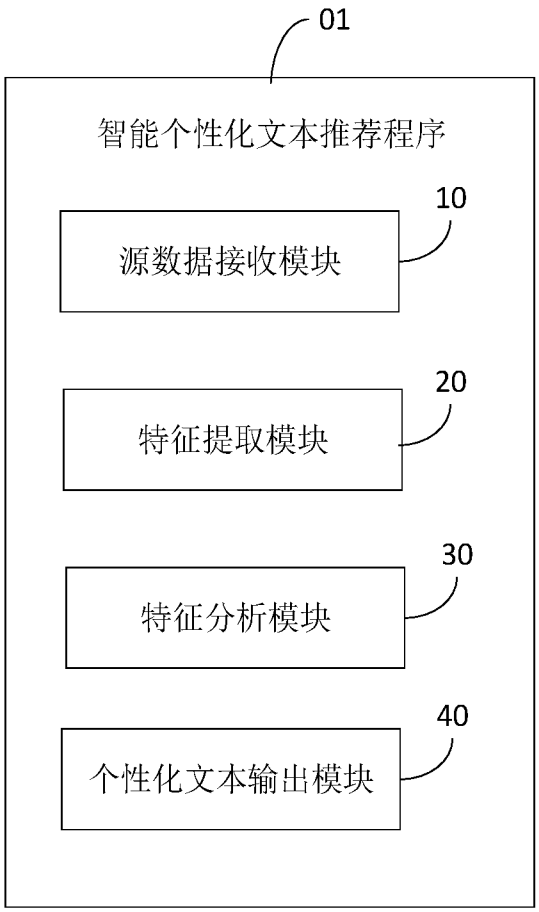

01
智能个性化文本推荐程序
10
源数据接收模块
20
特征提取模块
30
特征分析模块
40
个性化文本输出模块

图 3

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2019/102201

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/335(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

WPI, EPODOC, CNPAT, CNKI, IEEE, GOOGLE: 文本, 场景, 语料, 标签, 分词, 类别, 集, 关键字, 聚类, 向量, 神经网络, 损失, text, scene, corpus, label, word, classification, set, keyword, cluster, vector, neural, lost

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 106776881 A (INSTITUTE OF SOFTWARE, CHINESE ACADEMY OF SCIENCES) 31 May 2017 (2017-05-31) description, paragraphs [0014]-[0062]	1-20
A	CN 107315797 A (JIANGXI HONGDU AVIATION INDUSTRY GROUP CO., LTD.) 03 November 2017 (2017-11-03) entire document	1-20
A	CN 104298732 A (INSTITUTE OF COMPUTING TECHNOLOGY CHINESE ACADEMY OF SCIENCES) 21 January 2015 (2015-01-21) entire document	1-20
A	US 2015227589 A1 (MICROSOFT CORPORATION) 13 August 2015 (2015-08-13) entire document	1-20

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
14 March 2020	**26 March 2020**
Name and mailing address of the ISA/CN **China National Intellectual Property Administration (ISA/CN)** **No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088** **China** Facsimile No. **(86-10)62019451**	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2019/102201

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 106776881 A	31 May 2017	None	
CN 107315797 A	03 November 2017	None	
CN 104298732 A	21 January 2015	None	
US 2015227589 A1	13 August 2015	None	

国际检索报告

国际申请号
PCT/CN2019/102201

A. 主题的分类

G06F 16/335(2019.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

WPI, EPODOC, CNPAT, CNKI, IEEE, GOOGLE:文本，场景，语料，标签，分词，类别，集，关键字，聚类，向量，神经网络，损失，text, scene, corpus, label, word, classification, set, keyword, cluster, vector, neural, lost

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
A	CN 106776881 A (中国科学院软件研究所) 2017年 5月 31日 (2017-05-31) 说明书第[0014]-[0062]段	1-20
A	CN 107315797 A (江西洪都航空工业集团有限责任公司) 2017年 11月 3日 (2017-11-03) 全文	1-20
A	CN 104298732 A (中国科学院计算技术研究所) 2015年 1月 21日 (2015-01-21) 全文	1-20
A	US 2015227589 A1 (MICROSOFT CORPORATION) 2015年 8月 13日 (2015-08-13) 全文	1-20

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2020年 3月 14日	2020年 3月 26日
ISA/CN的名称和邮寄地址 中国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451	受权官员 胡百乐 电话号码 86-(10)-53961519

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2019/102201

检索报告引用的专利文件	公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN 106776881 A	2017年 5月 31日	无	
CN 107315797 A	2017年 11月 3日	无	
CN 104298732 A	2015年 1月 21日	无	
US 2015227589 A1	2015年 8月 13日	无	