(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110264311 B
(45)授权公告日 2023.04.18

(21)申请号 201910461767.7

(22)申请日 2019.05.30

(65)同一申请的已公布的文献号
申请公布号 CN 110264311 A

(43)申请公布日 2019.09.20

(73)专利权人 佛山科学技术学院
地址 528000 广东省佛山市南海区狮山镇广云路33号

(72)发明人 苏俊健 王东 麦志领 何佳奋 纪淇纯 叶新华

(74)专利代理机构 广州嘉权专利商标事务所有限公司 44205
专利代理师 谢泳祥

(51)Int.Cl.
G06Q 30/0601(2023.01)
G06F 16/35(2019.01)
G06F 16/9536(2019.01)
G06F 40/284(2020.01)

(56)对比文件
CN 105869024 A,2016.08.17
CN 108427670 A,2018.08.21
CN 107291693 A,2017.10.24
CN 109635204 A,2019.04.16
李爱国、库向阳.数据预处理.《数据挖掘原理、算法及应用》.2012,第25-27页.

审查员 卢济敏

权利要求书4页 说明书9页 附图1页

(54)发明名称

一种基于深度学习的商业推广信息精准推荐方法及系统

(57)摘要

本发明公开了一种基于深度学习的商业推广信息精准推荐方法及系统,通过训练样本数据集训练LSTM神经网络和通过测试样本数据集测试LSTM神经网络准确度得到训练好的LSTM神经网络作为推荐系统,得到功能完善的推荐系统的分类器,缩短发展客户所需要的时间,提高发现客户的精准度。本公开的方法及系统是一套成型的、高效的、面向区域行业的推荐系统;在各个区域,有很多区域性的企业集中在同一片区域,线下发展客户已接近饱和。而线下挖掘潜在客户需要一定的人力物力成本,并且成功率不高。这类企业迫切需要一套成型的系统来指引,节省发展客户所需要的成本。

CN 110264311 B

1.一种基于深度学习的商业推广信息精准推荐方法,其特征在于,所述方法包括以下步骤:

步骤1,采集商业推广信息数据;

步骤2,对所采集到的商业推广信息数据进行预处理和清洗得到商业推广信息数据集;

步骤3,对商业推广信息数据集进行降维和特征选择;

步骤4,将特征选择得到的商业推广信息数据集划分为训练样本数据集和测试样本数据集;

步骤5,训练样本数据集和测试样本数据集通过word2vec模型得到用于训练的词向量;

在步骤5中,训练样本数据集和测试样本数据集通过word2vec模型得到用于训练的词向量的方法包括以下步骤:

步骤5.1,分词:由于中文特殊性,通过分词库对商业推广信息中语句进行分词得到词库,分词库包括Jieba词库、IK词库、mmseg词库、word词库;

步骤5.2,统计词频:遍历步骤5.1中分词后形成的词库,统计出现过的词语的频率并且对其进行编号;

步骤5.3,构造树形结果:依据出现步骤5.2中各个词的出现概率,构造Huffman树;

步骤5.4,生成节点所在的二进制码:将各个词的出现概率转换为二进制编码来表示步骤5.3中Huffman树中的各个节点;

步骤5.5,初始化各非叶子节点的中间向量和叶子节点中的词向量:所述Huffman树中的各个节点,都存储有一个长为m的向量,但叶子节点和非叶结点中的向量的含义不同,叶子节点中存储的是各词的词向量,是作为神经网络的输入的;而非叶结点中存储的是中间向量,对应于神经网络中隐含层的参数,与输入一起决定分类结果;

步骤5.6,训练中间向量和词向量:使用CBOW模型或Skip-Gram模型训练中间向量和词向量,最后得到商业推广信息的对应词向量;

步骤6,通过训练样本数据集训练LSTM神经网络和通过测试样本数据集测试LSTM神经网络准确度得到训练好的LSTM神经网络作为推荐系统;

在步骤3中,对商业推广信息数据集进行降维和特征选择的方法包括主成分分析方法、独立成分分析、线性判别分析、局部线性嵌入、拉普拉斯特征映射、多维缩放、等度量映射中任意一种降维方法;而特征选择使用基于单独最优的特征选择法的改良算法,所述特征为商业推广信息数据集中的公司的所在地、经营范围、注册信息经过文本向量化后的向量;单独最优的特征选择算法计算出每个特征单独使用时的可分性判据值,然后根据可分性判据值从大到小进行排序,取前30个可分性判据值较大的特征作为特征组合;所述基于单独最优的特征选择法的改良算法为以下公式:$J(\boldsymbol{X})=\sum_{i=1}^{n}\left(J(\mathrm{x(i)})+N(\mathrm{x(i)})/\mathrm{M}\right)$其中,$x(i)=(x(1),x(2),x(3),\cdots,x(n))$,x(i)代表第$i$个特征,$n$为特征个数,$J(X)$表示该特征集合的可分性判据,$N(x(i))$表示第$i$个特征的不是缺失的数据量个数,$M$表示数据量的总量,$N(x(i))/M$表示了第$i$个特征在数据中的缺失程度;

在步骤4中,将特征选择得到的商业推广信息数据集划分为训练样本数据集和测试样本数据集的方法包括:留出法、交叉验证法、自助法任意一种方法;

在步骤6中,通过训练样本数据集训练LSTM神经网络和通过测试样本数据集测试LSTM神经网络准确度得到训练好的LSTM神经网络作为推荐系统的方法包括以下步骤:

步骤6.1,使用词向量训练LSTM神经网络:将训练样本数据集通过LSTM神经网络的遗忘门,开始LSTM神经网络的信息丢弃动作,信息丢弃动作由遗忘门中的sigmoid层实现,将查看sigmoid层前一个输出和当前词向量的输入,决定上一个状态学习的信息是否保留,所述LSTM神经网络包括输入门、遗忘门和输出门;

步骤6.2,将信息丢弃后的训练样本数据集通过LSTM神经网络的输入门,开始LSTM神经网络的信息更新动作,信息更新动作由输入门中的sigmoid层实现,然后tanh层将会改变LSTM神经网络的各个细胞状态,学习出新的知识;

步骤6.3,将信息更新后的训练样本数据集通过LSTM神经网络的输出门,输出一个向量,这个向量取决于步骤6.2中的细胞状态;首先,运行sigmoid层得到向量确定细胞状态的输出部分,把细胞状态通过tanh层进行处理,并将它和sigmoid门的输出相乘,得到LSTM网络的输出信息;

步骤6.4,使用测试数据输入步骤6.3中得到的LSTM网络,得到输出结果,与测试数据中的标签比对,验证网络的准确性,若准确度达到要求,则完成训练得到训练好的LSTM神经网络,将训练好的LSTM神经网络作为推荐系统。

2.根据权利要求1所述的一种基于深度学习的商业推广信息精准推荐方法,其特征在于,在步骤2中,对所采集到的商业推广信息数据进行预处理和清洗得到商业推广信息数据集的方法为,进行预处理,预处理为通过对所采集到的商业推广信息数据进行分类或分组前所做的审核、筛选、排序的处理,即数据审核完整性和准确性、数据筛选、数据排序,即数据清理、数据集成、数据变换、数据归约;对所采集到的商业推广信息数据进行清洗是指数据清洗,发现并纠正数据文件中可识别的错误的最后一道程序,包括检查数据一致性,处理无效值和缺失值;商业推广信息数据通过预处理和清洗,利用数理统计、数据挖掘或预定义的清理规则将脏数据转化为满足数据质量要求的数据,即清洗后的商业推广信息数据集。

3.一种基于深度学习的商业推广信息精准推荐系统,其特征在于,所述系统包括:存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机程序,所述处理器执行所述计算机程序运行在以下系统的单元中:

数据集采集单元,用于采集商业推广信息数据;

数据预处理单元,用于对所采集到的商业推广信息数据进行预处理和清洗得到商业推广信息数据集;

特征选择单元,用于对商业推广信息数据集进行降维和特征选择;

训练样本划分单元,用于将特征选择得到的商业推广信息数据集划分为训练样本数据集和测试样本数据集;

向量化单元,用于训练样本数据集和测试样本数据集通过word2vec模型得到用于训练的词向量;

训练样本数据集和测试样本数据集通过word2vec模型得到用于训练的词向量的方法包括以下步骤:

步骤5.1,分词:由于中文特殊性,通过分词库对商业推广信息中语句进行分词得到词库,分词库包括Jieba词库、IK词库、mmseg词库、word词库;

步骤5.2,统计词频:遍历步骤5.1中分词后形成的词库,统计出现过的词语的频率并且对其进行编号;

步骤5.3,构造树形结果:依据出现步骤5.2中各个词的出现概率,构造Huffman树;

步骤5.4,生成节点所在的二进制码:将各个词的出现概率转换为二进制编码来表示步骤5.3中Huffman树中的各个节点;

步骤5.5,初始化各非叶子节点的中间向量和叶子节点中的词向量:所述Huffman树中的各个节点,都存储有一个长为*m*的向量,但叶子节点和非叶结点中的向量的含义不同,叶子节点中存储的是各词的词向量,是作为神经网络的输入的;而非叶结点中存储的是中间向量,对应于神经网络中隐含层的参数,与输入一起决定分类结果;

步骤5.6,训练中间向量和词向量:使用CBOW模型或Skip-Gram模型训练中间向量和词向量,最后得到商业推广信息的对应词向量;

推荐系统获得单元,用于通过训练样本数据集训练LSTM神经网络和通过测试样本数据集测试LSTM神经网络准确度得到训练好的LSTM神经网络作为推荐系统;

对商业推广信息数据集进行降维和特征选择的方法包括主成分分析方法、独立成分分析、线性判别分析、局部线性嵌入、拉普拉斯特征映射、多维缩放、等度量映射中任意一种降维方法;而特征选择使用基于单独最优的特征选择法的改良算法,所述特征为商业推广信息数据集中的公司的所在地、经营范围、注册信息等信息经过文本向量化后的向量;单独最优的特征选择算法计算出每个特征单独使用时的可分性判据值,然后根据可分性判据值从大到小进行排序,取前30个可分性判据值较大的特征作为特征组合;但结合实际情况,在采集到的商业推广信息数据集中,存在信息缺失的现象,所以在选择特征时,需要考虑单个特征信息缺失度,所述基于单独最优的特征选择法的改良算法为以下公式:

$J(X)=\sum_{i=1}^{n}\left(J(\mathrm{x(i)})+N(\mathrm{x(i)})/\mathrm{M}\right)$ 其中,$x(i)=(x(1),x(2),x(3),\cdots,x(n))$,x(*i*)代表第*i*个特征,*n*为特征个数,*J*(*X*)表示该特征集合的可分性判据,*N*(*x*(*i*))表示第*i*个特征的不是缺失的数据量个数,*M*表示数据量的总量,*N*(*x*(*i*))/*M*表示了第*i*个特征在数据中的缺失程度;

将特征选择得到的商业推广信息数据集划分为训练样本数据集和测试样本数据集的方法包括:留出法、交叉验证法、自助法任意一种方法;

通过训练样本数据集训练LSTM神经网络和通过测试样本数据集测试LSTM神经网络准确度得到训练好的LSTM神经网络作为推荐系统的方法包括以下步骤:

步骤6.1,使用词向量训练LSTM神经网络:将训练样本数据集通过LSTM神经网络的遗忘门,开始LSTM神经网络的信息丢弃动作,信息丢弃动作由遗忘门中的sigmoid层实现,将查看sigmoid层前一个输出和当前词向量的输入,决定上一个状态学习的信息是否保留,所述LSTM神经网络包括输入门、遗忘门和输出门;

步骤6.2,将信息丢弃后的训练样本数据集通过LSTM神经网络的输入门,开始LSTM神经网络的信息更新动作,信息更新动作由输入门中的sigmoid层实现,然后tanh层将会改变LSTM神经网络的各个细胞状态,学习出新的知识;

步骤6.3,将信息更新后的训练样本数据集通过LSTM神经网络的输出门,输出一个向量,这个向量取决于步骤6.2中的细胞状态;首先,运行sigmoid层得到向量确定细胞状态的

输出部分，把细胞状态通过tanh层进行处理，并将它和sigmoid门的输出相乘，得到LSTM网络的输出信息；

步骤6.4，使用测试数据输入步骤6.3中得到的LSTM网络，得到输出结果，与测试数据中的标签比对，验证网络的准确性，若准确度达到要求，则完成训练得到训练好的LSTM神经网络，将训练好的LSTM神经网络作为推荐系统。

一种基于深度学习的商业推广信息精准推荐方法及系统

技术领域

[0001] 本公开涉及机器学习推荐算法和深度学习技术领域,具体涉及一种基于深度学习的商业推广信息精准推荐方法及系统。

背景技术

[0002] 传统推荐系统一般为商品与顾客之间的推荐,为不同的顾客推荐不同的商品。而传统的推荐算法一般有:基于内容的推荐(Content Based,CB),协同过滤(Collaborative Filtering CF),混合推荐方法等。而本专利抛开传统推荐模式,实现上游企业与下游企业之间的推荐,使用基于深度学习的推荐算法来解决上下游企业之间复杂的关系。

[0003] 目前的传统推荐系统存在以下两方面问题:

[0004] 1)国内现阶段推荐算法的研究领域主要集中在对商品的精准推荐,对于客户推荐的研究还比较少。而国内研究推荐系统的算法一般采用机器学习算法,但对于现在日益复杂的企业关系和多元化的数据关系,机器学习所具有的学习效率渐渐不能满足需求。

[0005] 2)脱离传统数据挖掘的框架,数据由网页自动抓取和筛选,降低收集数据的成本或解决缺乏数据的困境

发明内容

[0006] 为解决上述问题,本公开提供一种基于深度学习的商业推广信息精准推荐方法及系统的技术方案,通过训练样本数据集训练LSTM神经网络和通过测试样本数据集测试LSTM神经网络准确度得到训练好的LSTM神经网络作为推荐系统,得到功能完善的推荐系统的分类器,基于深度学习和网络技术,减少企业发展客户时所需的成本,缩短发展客户所需要的时间,提高发现客户的精准度。

[0007] 为了实现上述目的,根据本公开的一方面,提供一种基于深度学习的商业推广信息精准推荐方法,所述方法包括以下步骤:

[0008] 步骤1,采集商业推广信息数据;

[0009] 步骤2,对所采集到的商业推广信息数据进行预处理和清洗得到商业推广信息数据集;

[0010] 步骤3,对商业推广信息数据集进行降维和特征选择;

[0011] 步骤4,将特征选择得到的商业推广信息数据集划分为训练样本数据集和测试样本数据集;

[0012] 步骤5,训练样本数据集和测试样本数据集通过word2vec模型得到用于训练的词向量;

[0013] 步骤6,通过训练样本数据集训练LSTM(Long Short-Term Memory)神经网络和通过测试样本数据集测试LSTM神经网络准确度得到训练好的LSTM神经网络作为推荐系统。

[0014] 进一步地,在步骤1中,采集商业推广信息数据的方法包括但不限于:采集开源的数据集网站如kaggle数据集作为商业推广信息数据;通过二次开发后的网络爬虫技术对类

淘宝店家网页或者同城交易信息网通过爬虫进行抓取数据集，获得商业推广信息数据；利用百度快照中保留的txt格式的网页备份，从中获取所需要的信息作为商业推广信息数据。

[0015] 进一步地，在步骤2中，对所采集到的商业推广信息数据进行预处理和清洗得到商业推广信息数据集的方法为，由于获取到的商业推广信息数据极其庞大、混杂乃至无用，所以需要进行预处理，预处理为通过对所采集到的商业推广信息数据进行分类或分组前所做的审核、筛选、排序等必要的处理，即数据审核完整性和准确性、数据筛选、数据排序，即数据清理、数据集成、数据变换、数据归约；对所采集到的商业推广信息数据进行清洗是指数据清洗，发现并纠正数据文件中可识别的错误的最后一道程序，包括检查数据一致性，处理无效值和缺失值等；商业推广信息数据通过预处理和清洗，利用数理统计、数据挖掘或预定义的清理规则将脏数据转化为满足数据质量要求的数据，即清洗后的商业推广信息数据集。

[0016] 进一步地，在步骤3中，对商业推广信息数据集进行降维和特征选择的方法包括但不限于主成分分析方法(Principal Component Analysis，PCA)、独立成分分析(Independent Component Analysis，ICA)、线性判别分析(Linear Discriminant Analysis，LDA)、局部线性嵌入(Locally Linear Embedding，LLE)、拉普拉斯特征映射(Laplacian Eigenmaps)、多维缩放(MultiDimensional scaling，MDS)、等度量映射(Equal Metric Mapping)中任意一种降维方法；而特征选择使用基于单独最优的特征选择法的改良算法，所述特征为商业推广信息数据集中的公司的所在地、经营范围、注册信息等信息经过文本向量化后的向量；单独最优的特征选择算法计算出每个特征单独使用时的可分性判据值，然后根据可分性判据值从大到小进行排序，取前30个可分性判据值较大的特征作为特征组合；但结合实际情况，在采集到的商业推广信息数据集中，存在信息缺失的现象，所以在选择特征时，需要考虑单个特征信息缺失度，所述基于单独最优的特征选择法的改良算法为以下公式：$J(\boldsymbol{X})=\sum_{i=1}^{n}\left(J(\mathrm{x(i)})+N(\mathrm{x(i)})/\mathrm{M}\right)$，其中，x(i)＝(x(1)，x(2)，x(3)，…，x(n))，x(i)代表第i个特征，n为特征个数，J(X)表示该特征集合的可分性判据，N(x(i))表示第i个特征的不是缺失的数据量个数，M表示数据量的总量，N(x(i))/M表示了第i个特征在数据中的缺失程度，改良了信息缺失的现象。

[0017] 进一步地，在步骤4中，将特征选择得到的商业推广信息数据集划分为训练样本数据集和测试样本数据集的方法包括：留出法、交叉验证法、自助法任意一种方法。

[0018] 留出法为直接将商业推广信息数据集划分为两个互斥的集合，其中一个集合作为训练样本数据集，留下的集合作为测试样本数据集。

[0019] 交叉验证法为将商业推广信息数据集划分为个大小相等的互斥子集，即每个子集都尽可能保持数据分布的一致性，即从中通过分层采样得到，然后，每次用个子集的并集作为训练样本数据集，剩下的那个子集作为测试样本数据集。

[0020] 自助法为对进行商业推广信息数据集进行采样产生：每次随机从中商业推广信息数据集挑选一个样本，将其拷贝一份放入训练样本数据集中，保持不变的作为测试样本数据集，重复以上过程次。其中，有部分样本会多次出现在商业推广信息数据集中的数据集作为训练样本数据集，而另一部分不会出现在商业推广信息数据集的数据集作为测试样本数据集。

[0021] 由于需要输出一类带有类似属性的有购买该特定企业服务或产品的客户。所以考虑到可能会使数据集有不同个数的属性，所以用LSTM作为核心网络来训练模型。LSTM是一种RNN(Recurrent Neural Network)特殊的类型，可以学习长期依赖信息。LSTM通过“门”(gate)来控制丢弃没用的信息或者提高有益信息的比重，同时模型里面增添了记忆细胞(cell)，收集记忆之前相关的信息，实现遗忘或记忆的功能，这种携带记忆的特性使该网络对长期使用该产品有很大优势，因为网络在使用中也不断的学习提高记忆能力，提高产品的使用周期。

[0022] 进一步地，在步骤5中，训练样本数据集和测试样本数据集通过word2vec模型得到用于训练的词向量的方法包括以下步骤：

[0023] 步骤5.1，分词：由于中文特殊性，通过分词库对商业推广信息中语句进行分词得到词库，分词库包括但不限于Jieba词库、IK词库、mmseg词库、word词库；

[0024] 步骤5.2，统计词频：遍历步骤5.1中分词后形成的词库，统计出现过的词语的频率并且对其进行编号；

[0025] 步骤5.3，构造树形结果：依据出现步骤5.2中各个词的出现概率，构造Huffman树；

[0026] 步骤5.4，生成节点所在的二进制码：将各个词的出现概率转换为二进制编码来表示步骤5.3中Huffman树中的各个节点；

[0027] 步骤5.5，初始化各非叶子节点的中间向量和叶子节点中的词向量：所述Huffman树中的各个节点，都存储有一个长为m的向量，但叶子节点和非叶结点中的向量的含义不同，叶子节点中存储的是各词的词向量，是作为神经网络的输入的；而非叶结点中存储的是中间向量，对应于神经网络中隐含层的参数，与输入一起决定分类结果；

[0028] 步骤5.6，训练中间向量和词向量：使用CBOW(Continuous Bag-Of-Words Model)模型或Skip-Gram模型，将词A附近的n-1个词的词向量相加作为系统的输入，并且按照词A在步骤5.4中生成的二进制码，依次进行分类并按照分类结果训练中间向量和词向量，最后得到商业推广信息的对应词向量；词A为单词，训练的过程主要有输入层(input)，映射层(projection)和输出层(output)三个阶段；输入层即为某个单词A(词A)周围的n-1个单词的词向量。如果n取5，则词A(可记为w(t))，前两个和后两个的单词为w(t-2)，w(t-1)，w(t+1)，w(t+2)。相对应的，那4个单词的词向量记为v(w(t-2))，v(w(t-1))，v(w(t+1))，v(w(t+2))。从输入层到映射层比较简单，将那n-1个词向量相加即可。

[0029] 进一步地，在步骤6中，通过训练样本数据集训练LSTM神经网络和通过测试样本数据集测试LSTM神经网络准确度得到训练好的LSTM神经网络作为推荐系统的方法包括以下步骤：

[0030] 步骤6.1，使用词向量训练LSTM神经网络：将训练样本数据集通过LSTM神经网络的遗忘门，开始LSTM神经网络的信息丢弃动作，信息丢弃动作由遗忘门中的sigmoid层实现，将查看sigmoid层前一个输出和当前词向量的输入，决定上一个状态学习的信息是否保留，所述LSTM神经网络包括输入门、遗忘门和输出门；

[0031] 步骤6.2，将信息丢弃后的训练样本数据集通过LSTM神经网络的输入门，开始LSTM神经网络的信息更新动作，信息更新动作由输入门中的sigmoid层实现，然后tanh层将会改变LSTM神经网络的各个细胞状态，学习出新的知识；

[0032] 步骤6.3，将信息更新后的训练样本数据集通过LSTM神经网络的输出门，输出一个

向量,这个向量取决于步骤6.2中的细胞状态;首先,运行sigmoid层得到向量确定细胞状态的输出部分,把细胞状态通过tanh层进行处理,并将它和sigmoid门的输出(向量)相乘,得到LSTM网络的输出信息;

[0033] 步骤6.4,使用测试数据输入步骤6.3中得到的LSTM网络,得到输出结果,与测试数据中的标签比对,验证网络的准确性,若准确度达到要求,则完成训练得到训练好的LSTM神经网络,将训练好的LSTM神经网络作为推荐系统。

[0034] 本发明还提供了一种基于深度学习的商业推广信息精准推荐系统,所述系统包括:存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机程序,所述处理器执行所述计算机程序运行在以下系统的单元中:

[0035] 数据集采集单元,用于采集商业推广信息数据;

[0036] 数据预处理单元,用于对所采集到的商业推广信息数据进行预处理和清洗得到商业推广信息数据集;

[0037] 特征选择单元,用于对商业推广信息数据集进行降维和特征选择;

[0038] 训练样本划分单元,用于将特征选择得到的商业推广信息数据集划分为训练样本数据集和测试样本数据集;

[0039] 向量化单元,用于训练样本数据集和测试样本数据集通过word2vec模型得到用于训练的词向量;

[0040] 推荐系统获得单元,用于通过训练样本数据集训练LSTM神经网络和通过测试样本数据集测试LSTM神经网络准确度得到训练好的LSTM神经网络作为推荐系统。

[0041] 本公开的有益效果

[0042] 本发明提供一种基于深度学习的商业推广信息精准推荐方法及系统:

[0043] 1)针对潜在客户的挖掘方面没有成型的线上服务系统,线下的基于社交网络的客户推荐基本走到极致;

[0044] 2)虽然有企业针对上述情况进行研究并开发系统,但据了解,这种简单的系统不能适应现实中各种复杂的情况,工作成效不大。据调查,国内基本没有专门针对公司或企业进行信息推荐的公司,甚至没有一套成型的、高效的、面向区域行业的推荐系统;

[0045] 3)在各个区域,有很多区域性的企业集中在同一片区域,线下发展客户已接近饱和。而线下挖掘潜在客户需要一定的人力物力成本,并且成功率不高。这类企业迫切需要一套成型的系统来指引,节省发展客户所需要的成本;

[0046] 4)通过应用LSTM在自然语言处理方面取得了很好的效果,应用在语言翻译器中,具有很高的准确率和自适应能力,能够完美的适应现实中各种复杂的条件,适合作为本产品的核心网络。

附图说明

[0047] 通过对结合附图所示出的实施方式进行详细说明,本公开的上述以及其他特征将更加明显,本公开附图中相同的参考标号表示相同或相似的元素,显而易见地,下面描述中的附图仅仅是本公开的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据这些附图获得其他的附图,在附图中:

[0048] 图1所示为一种基于深度学习的商业推广信息精准推荐方法的流程图;

[0049] 图2所示为一种基于深度学习的商业推广信息精准推荐系统图。

具体实施方式

[0050] 以下将结合实施例和附图对本公开的构思、具体结构及产生的技术效果进行清楚、完整的描述，以充分地理解本公开的目的、方案和效果。需要说明的是，在不冲突的情况下，本申请中的实施例及实施例中的特征可以相互组合。

[0051] 如图1所示为根据本公开的一种基于深度学习的商业推广信息精准推荐方法的流程图，下面结合图1来阐述根据本公开的实施方式的一种基于深度学习的商业推广信息精准推荐方法。

[0052] 本公开提出一种基于深度学习的商业推广信息精准推荐方法，具体包括以下步骤：

[0053] 步骤1，采集商业推广信息数据；

[0054] 步骤2，对所采集到的商业推广信息数据进行预处理和清洗得到商业推广信息数据集；

[0055] 步骤3，对商业推广信息数据集进行降维和特征选择；

[0056] 步骤4，将特征选择得到的商业推广信息数据集划分为训练样本数据集和测试样本数据集；

[0057] 步骤5，训练样本数据集和测试样本数据集通过word2vec模型得到用于训练的词向量；

[0058] 步骤6，通过训练样本数据集训练LSTM(Long Short-Term Memory)神经网络和通过测试样本数据集测试LSTM神经网络准确度得到训练好的LSTM神经网络作为推荐系统。

[0059] 进一步地，在步骤1中，采集商业推广信息数据的方法包括但不限于：采集开源的数据集网站如kaggle数据集作为商业推广信息数据；通过二次开发后的网络爬虫技术对类淘宝店家网页或者同城交易信息网通过爬虫进行抓取数据集，获得商业推广信息数据；利用百度快照中保留的txt格式的网页备份，从中获取所需要的信息作为商业推广信息数据。

[0060] 进一步地，在步骤2中，对所采集到的商业推广信息数据进行预处理和清洗得到商业推广信息数据集的方法为，由于获取到的商业推广信息数据极其庞大、混杂乃至无用，所以需要进行预处理，预处理为通过对所采集到的商业推广信息数据进行分类或分组前所做的审核、筛选、排序等必要的处理，即数据审核完整性和准确性、数据筛选、数据排序，即数据清理、数据集成、数据变换、数据归约；对所采集到的商业推广信息数据进行清洗是指数据清洗，发现并纠正数据文件中可识别的错误的最后一道程序，包括检查数据一致性，处理无效值和缺失值等；商业推广信息数据通过预处理和清洗，利用数理统计、数据挖掘或预定义的清理规则将脏数据转化为满足数据质量要求的数据，即清洗后的商业推广信息数据集。

[0061] 进一步地，在步骤3中，对商业推广信息数据集进行降维和特征选择的方法包括但不限于主成分分析方法(PCA)、独立成分分析(ICA)、线性判别分析(LDA)、局部线性嵌入(LLE)、拉普拉斯特征映射、多维缩放(MDS)、等度量映射中任意一种降维方法；而特征选择使用基于单独最优的特征选择法的改良算法，所述特征为商业推广信息数据集中的公司的所在地、经营范围、注册信息等信息经过文本向量化后的向量；单独最优的特征选择算法计

算出每个特征单独使用时的可分性判据值，然后根据可分性判据值从大到小进行排序，取前30个可分性判据值较大的特征作为特征组合；但结合实际情况，在采集到的商业推广信息数据集中，存在信息缺失的现象，所以在选择特征时，需要考虑单个特征信息缺失度，所述基于单独最优的特征选择法的改良算法为以下公式：$J(\boldsymbol{X})=\sum_{i=1}^{n}\left(J(\mathrm{x(i)})+N(\mathrm{x(i)})/\mathrm{M}\right)$，其中，x(i)＝(x(1)，x(2)，x(3)，…，x(n))，x(i)代表第i个特征，n为特征个数，J(X)表示该特征集合的可分性判据，N(x(i))表示第i个特征的不是缺失的数据量个数，M表示数据量的总量，N(x(i))/M表示了第i个特征在数据中的缺失程度。

[0062] 进一步地，在步骤4中，将特征选择得到的商业推广信息数据集划分为训练样本数据集和测试样本数据集的方法包括：留出法、交叉验证法、自助法任意一种方法。

[0063] 留出法为直接将商业推广信息数据集划分为两个互斥的集合，其中一个集合作为训练样本数据集，留下的集合作为测试样本数据集。

[0064] 交叉验证法为将商业推广信息数据集划分为个大小相等的互斥子集，即每个子集都尽可能保持数据分布的一致性，即从中通过分层采样得到，然后，每次用个子集的并集作为训练样本数据集，剩下的那个子集作为测试样本数据集。

[0065] 自助法为对进行商业推广信息数据集进行采样产生：每次随机从中商业推广信息数据集挑选一个样本，将其拷贝一份放入训练样本数据集中，保持不变的作为测试样本数据集，重复以上过程次。其中，有部分样本会多次出现在商业推广信息数据集中的数据集作为训练样本数据集，而另一部分不会出现在商业推广信息数据集的数据集作为测试样本数据集。

[0066] 由于需要输出一类带有类似属性的有购买该特定企业服务或产品的客户。所以考虑到可能会使数据集有不同个数的属性，所以用LSTM网络作为核心网络来训练模型。LSTM是一种RNN特殊的类型，可以学习长期依赖信息。LSTM通过“门”来控制丢弃没用的信息或者提高有益信息的比重，同时模型里面增添了记忆细胞，收集记忆之前相关的信息，实现遗忘或记忆的功能，这种携带记忆的特性使该网络对长期使用该产品有很大优势，因为网络在使用中也不断的学习提高记忆能力，提高产品的使用周期。

[0067] 进一步地，在步骤5中，训练样本数据集和测试样本数据集通过word2vec模型得到用于训练的词向量的方法包括以下步骤：

[0068] 步骤5.1，分词：由于中文特殊性，通过分词库对商业推广信息中语句进行分词得到词库，分词库包括但不限于Jieba词库、IK词库、mmseg词库、word词库；

[0069] 步骤5.2，统计词频：遍历步骤5.1中分词后形成的词库，统计出现过的词语的频率并且对其进行编号；

[0070] 步骤5.3，构造树形结果：依据出现步骤5.2中各个词的出现概率，构造Huffman树；

[0071] 步骤5.4，生成节点所在的二进制码：将各个词的出现概率转换为二进制编码来表示步骤5.3中Huffman树中的各个节点；

[0072] 步骤5.5，初始化各非叶子节点的中间向量和叶子节点中的词向量：所述Huffman树中的各个节点，都存储有一个长为m的向量，但叶子节点和非叶结点中的向量的含义不同，叶子节点中存储的是各词的词向量，是作为神经网络的输入的；而非叶结点中存储的是中间向量，对应于神经网络中隐含层的参数，与输入一起决定分类结果；

[0073] 步骤5.6,训练中间向量和词向量:使用CBOW(Continuous Bag-Of-Words Model)模型或Skip-Gram模型,将词A附近的n-1个词的词向量相加作为系统的输入,并且按照词A在步骤5.4中生成的二进制码,依次进行分类并按照分类结果训练中间向量和词向量,最后得到商业推广信息的对应词向量;词A为单词,训练的过程主要有输入层(input),映射层(projection)和输出层(output)三个阶段;输入层即为某个单词A(词A)周围的n-1个单词的词向量。如果n取5,则词A(可记为w(t)),前两个和后两个的单词为w(t-2),w(t-1),w(t+1),w(t+2)。相对应的,那4个单词的词向量记为v(w(t-2)),v(w(t-1)),v(w(t+1)),v(w(t+2))。从输入层到映射层比较简单,将那n-1个词向量相加即可。

[0074] 进一步地,在步骤6中,通过训练样本数据集训练LSTM神经网络和通过测试样本数据集测试LSTM神经网络准确度得到训练好的LSTM神经网络作为推荐系统的方法包括以下步骤:

[0075] 步骤6.1,使用词向量训练LSTM神经网络:将训练样本数据集通过LSTM神经网络的遗忘门,开始LSTM神经网络的信息丢弃动作,信息丢弃动作由遗忘门中的sigmoid层实现,将查看sigmoid层前一个输出和当前词向量的输入,决定上一个状态学习的信息是否保留,所述LSTM神经网络包括输入门、遗忘门和输出门;

[0076] 步骤6.2,将信息丢弃后的训练样本数据集通过LSTM神经网络的输入门,开始LSTM神经网络的信息更新动作,信息更新动作由输入门中的sigmoid层实现,然后tanh层将会改变LSTM神经网络的各个细胞状态,学习出新的知识;

[0077] 步骤6.3,将信息更新后的训练样本数据集通过LSTM神经网络的输出门,输出一个向量,这个向量取决于步骤6.2中的细胞状态;首先,运行sigmoid层得到向量确定细胞状态的输出部分,把细胞状态通过tanh层进行处理,并将它和sigmoid门的输出(向量)相乘,得到LSTM网络的输出信息;

[0078] 步骤6.4,使用测试数据输入步骤6.3中得到的LSTM网络,得到输出结果,与测试数据中的标签比对,验证网络的准确性,若准确度达到要求,则完成训练得到训练好的LSTM神经网络,将训练好的LSTM神经网络作为推荐系统。

[0079] 通过推荐系统可以阐述该方法如何实现将客户精准推荐给企业,帮助企业更有效率地发展大量的、潜在的和精准的客户。利用“互联网+金融”的思维,基于深度学习和网络技术,减少企业发展客户时所需的成本,缩短发展客户所需要的时间,提高发现客户的精准度。

[0080] 本公开的实施例采用一种基于深度学习的商业推广信息精准推荐系统的挖掘出的数据的准确率(0.12)远高于传统的基于内容的方法的准确率(0.067),留存率也有明显改善;更容易发觉热点商业推广信息。将LSTM深度神经网络应用在商业推广信息挖掘上,遗忘不相关的噪声信息,并加深强关联信息的记忆,在算法上优胜劣汰,择优选择特征。采用Huffman树进行相关商业推广信息分词,极大的提升的分词速度,分词计算时间大概是一般的穷举的1/20。

[0081] 本公开的实施例提供的一种基于深度学习的商业推广信息精准推荐系统,如图2所示为本公开的一种基于深度学习的商业推广信息精准推荐系统图,该实施例的一种基于深度学习的商业推广信息精准推荐系统包括:处理器、存储器以及存储在所述存储器中并可在所述处理器上运行的计算机程序,所述处理器执行所述计算机程序时实现上述一种基

于深度学习的商业推广信息精准推荐系统实施例中的步骤。

[0082] 所述系统包括:存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机程序,所述处理器执行所述计算机程序运行在以下系统的单元中:

[0083] 数据集采集单元,用于采集商业推广信息数据;

[0084] 数据预处理单元,用于对所采集到的商业推广信息数据进行预处理和清洗得到商业推广信息数据集;

[0085] 特征选择单元,用于对商业推广信息数据集进行降维和特征选择;

[0086] 训练样本划分单元,用于将特征选择得到的商业推广信息数据集划分为训练样本数据集和测试样本数据集;

[0087] 向量化单元,用于训练样本数据集和测试样本数据集通过word2vec模型得到用于训练的词向量;

[0088] 推荐系统获得单元,用于通过训练样本数据集训练LSTM神经网络和通过测试样本数据集测试LSTM神经网络准确度得到训练好的LSTM神经网络作为推荐系统。

[0089] 所述一种基于深度学习的商业推广信息精准推荐系统可以运行于桌上型计算机、笔记本、掌上电脑及云端服务器等计算设备中。所述一种基于深度学习的商业推广信息精准推荐系统,可运行的系统可包括,但不仅限于,处理器、存储器。本领域技术人员可以理解,所述例子仅仅是一种基于深度学习的商业推广信息精准推荐系统的示例,并不构成对一种基于深度学习的商业推广信息精准推荐系统的限定,可以包括比例子更多或更少的部件,或者组合某些部件,或者不同的部件,例如所述一种基于深度学习的商业推广信息精准推荐系统还可以包括输入输出设备、网络接入设备、总线等。

[0090] 所称处理器可以是中央处理单元(Central Processing Unit,CPU),还可以是其他通用处理器、数字信号处理器(Digital Signal Processor,DSP)、专用集成电路(Application Specific Integrated Circuit,ASIC)、现成可编程门阵列(Field-Programmable Gate Array,FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等,所述处理器是所述一种基于深度学习的商业推广信息精准推荐系统运行系统的控制中心,利用各种接口和线路连接整个一种基于深度学习的商业推广信息精准推荐系统可运行系统的各个部分。

[0091] 所述存储器可用于存储所述计算机程序和/或模块,所述处理器通过运行或执行存储在所述存储器内的计算机程序和/或模块,以及调用存储在存储器内的数据,实现所述一种基于深度学习的商业推广信息精准推荐系统的各种功能。所述存储器可主要包括存储程序区和存储数据区,其中,存储程序区可存储操作系统、至少一个功能所需的应用程序(比如声音播放功能、图像播放功能等)等;存储数据区可存储根据手机的使用所创建的数据(比如音频数据、电话本等)等。此外,存储器可以包括高速随机存取存储器,还可以包括非易失性存储器,例如硬盘、内存、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)、至少一个磁盘存储器件、闪存器件、或其它易失性固态存储器件。

[0092] 尽管本公开的描述已经相当详尽且特别对几个所述实施例进行了描述,但其并非旨在局限于任何这些细节或实施例或任何特殊实施例,而是应当将其视作是通过参考所附

权利要求考虑到现有技术为这些权利要求提供广义的可能性解释，从而有效地涵盖本公开的预定范围。此外，上文以发明人可预见的实施例对本公开进行描述，其目的是为了提供有用的描述，而那些目前尚未预见的对本公开的非实质性改动仍可代表本公开的等效改动。


采集商业推广信息数据
对所采集到的商业推广信息数据进行预处理和清洗得到商业推广信息数据集
对商业推广信息数据集进行降维和特征选择
将特征选择得到的商业推广信息数据集划分为训练样本数据集和测试样本数据集
训练样本数据集和测试样本数据集通过word2vec模型得到用于训练的词向量
通过训练样本数据集训练LSTM神经网络和通过测试样本数据集测试LSTM神经网络准确度得到训练好的LSTM神经网络作为推荐系统

图1

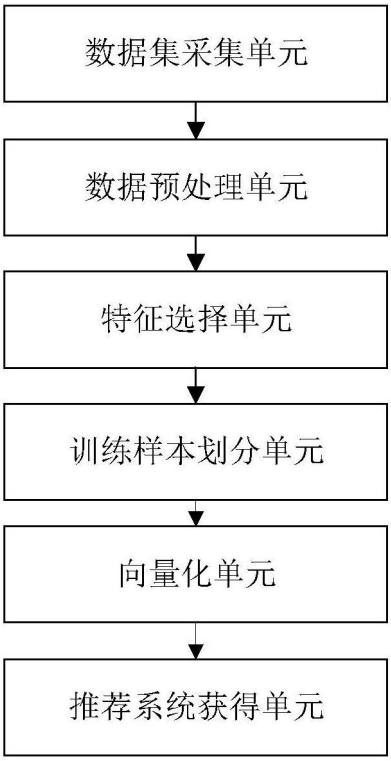
数据集采集单元
数据预处理单元
特征选择单元
训练样本划分单元
向量化单元
推荐系统获得单元

图2