(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 108009847 B
(45)授权公告日 2021.06.15

(21)申请号 201711230471.1
(22)申请日 2017.11.30
(65)同一申请的已公布的文献号
申请公布号 CN 108009847 A
(43)申请公布日 2018.05.08
(73)专利权人 西安电子科技大学
地址 710071 陕西省西安市太白南路2号西安电子科技大学
(72)发明人 赵纪伟 杨清海 鲁焕 秦猛
(74)专利代理机构 西安长和专利代理有限公司 61227
代理人 黄伟洪 李霞
(51)Int.Cl.
G06Q 30/02(2012.01)
G06K 9/62(2006.01)
(56)对比文件
CN 106959966 A,2017.07.18
CN 106920147 A,2017.07.04
CN 107169801 A,2017.09.15
CN 102004979 A,2011.04.06
CN 103886090 A,2014.06.25
CN 104834686 A,2015.08.12
WO 2017057921 A1,2017.04.06
曹军 等.外卖用户差评影响因素研究——基于文本评论和Word2vec.《现代商贸工业》.2017,(第2期),第55-56页.
董文.基于LDA和Word2Vec的推荐算法研究.《中国优秀硕士学位论文全文数据库信息科技辑》.2015,(第08期),摘要,第6-60页.
唐明 等.基于Word2vec的一种文档向量表示.《计算机科学》.2016,第43卷(第6期),第214-217页.
Eissa M.Alshari et al.Improvement of Sentiment Analysis Based on Clustering of Word2Vec Features.《2017 28th International Workshop on Database and Expert Systems Applications》.2017,第123-126页.
Goldberg,Yoav et al.word2vec Explained: deriving Mikolov et al.″s negative-sampling word-embedding method.《arXiv e-prints》.2014,第1-5页.
审查员 任丽霞

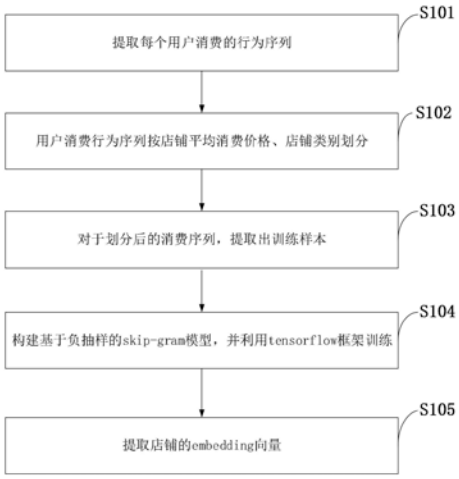
权利要求书1页 说明书4页 附图4页

(54)发明名称

外卖场景下店铺embedding特征提取的方法

(57)摘要

本发明属于计算机应用技术领域,公开了一种外卖场景下店铺embedding特征提取的方法、计算机、计算机程序。提取每个用户消费的行为序列;用户消费行为序列按店铺平均消费价格、店铺类别划分;对于划分后的消费序列,提取出训练样本;构建基于负抽样的skip-gram模型,并利用tensorflow框架训练;提取店铺的embedding特征向量。本发明对店铺进行embedding特征提取,获取了店铺某些方面的特征信息,将数据从高维的one-hot特征空间转换为指定维度的embedding特征空间;结合商户的embedding特征与线上模型,带来线上下单模型的整体性能提升。



CN 108009847 B

1.一种外卖场景下店铺embedding特征提取的方法,其特征在于,所述方法包括以下步骤:

步骤一:提取每个用户消费的行为序列;

步骤二:用户消费行为序列按店铺平均消费价格、店铺类别划分;

步骤三:对于划分后的消费行为序列,提取出训练样本;

步骤四:构建基于负抽样的skip-gram模型,并利用tensorflow框架训练;

步骤五:提取店铺的embedding特征向量;

其中,所述步骤一包括:应用hive提取用户一段时间内的用户所有下单店铺及其ID ,并将user_id—shop_id数据写入表格user_shop,利用 hive构建店铺字典shop_dict,其组织形式为{shop_id:index},其中index为店铺在字典中的顺序;

所述步骤二包括:将原始的user_shop表中的shop_id用字典中shop_idx索引代替;按照店铺的平均客单价、店铺一级类别进行分组去重,获得构建训练样本所需要的doc,其数据组织形式为:{user_id,array <shop_id_1,shop_id_2,...,shop_id_m>};对context内只有一家店铺的数据进行过滤,最终获得的数据组织形式为:array<shop_id_1,shop_id_2,...,shop_id_m>, 并将其存入HIVE表,用于后续训练样本的构建;

所述步骤三包括:训练样本的构建,利用python进行数据转换,并利用HIVE进行处理;对于每一个context,分别应用 get_train_samples函数构建训练样本,返回trains的RDD;最终获取的RDD写入到HIVE表中,用于接下来训练样本的导出;

所述步骤四包括:构建用于训练的tensorflow图,是整体的训练函数,包括训练数据的获取以及batch的生成;所述构建基于负抽样的skip-gram模型用于使用context的中间店铺来预测其他店铺;

所述步骤五包括:将数据利用t-sne进行可视化,获取店铺的embedding向量的直观感知,将embedding向量映射到2维,对各店铺的距离进行图形化展示。

2.如权利要求1所述的外卖场景下店铺embedding特征提取的方法,其特征在于,所述提取每个用户消费的行为序列具体包括:从数据库中,提取用户一定时间内的用户所有下单店铺及其ID,按照用户ID将数据分组,构建每个用户ID的下单店铺集合,存储到数据库中。

3.如权利要求1所述的外卖场景下店铺embedding特征提取的方法,其特征在于,所述用户消费行为序列按店铺平均消费价格、店铺类别划分具体包括:将提取出的用户下单店铺序列按店铺的平均消费价格划分为多个等级,同时将不同价格区间的店铺划分为子集合。

4.如权利要求1所述的外卖场景下店铺embedding特征提取的方法,其特征在于,所述提取店铺的embedding特征向量具体包括:用tensorflow训练好的向量保存到本地,并在实际操作中提取需要的店铺embedding特征。

外卖场景下店铺embedding特征提取的方法

技术领域

[0001] 本发明属于计算机应用技术领域，尤其涉及一种外卖场景下店铺embedding特征提取的方法。

背景技术

[0002] Google在2013年开源了词向量计算的工具——word2vec，引起了工业界和学术界的关注。word2vec可以在百万数量级的词典和上亿的数据集上进行高效地训练；得到的训练结果——词向量(word embedding)，很好地度量词与词之间的相似性。在外卖领域，针对线上业务实时性的需求，原始的dense特征与one-hot特征并不能满足及时毫秒的预测时延要求，需要对店铺的特征进行整体的抽象；从特征工程的角度来看，现有的特征工程工作主要是从单维度进行的，很难从一个整体的角度来反映店铺的特征；传统的FM算法在样本空间很大的时候很难保证效率。

[0003] 综上所述，现有技术存在的问题是：由于现有的特征工程工作主要从单维度，而不是从整体的角度来反映店铺特征，这就会导致特征空间趋于扁平；现有特征量线下为1000万维左右，线上为300维左右，大的特征量导致算法复杂度较高，很难满足线上实时计算的需求。若能解决这几个核心问题，可以有效降低线上业务的时延，将推荐过程控制在40ms以内，更好地符合推荐业务的需求，方便用户从大量店铺中寻找到自己感兴趣的店铺。

发明内容

[0004] 针对现有技术存在的问题，本发明提供了一种外卖场景下店铺embedding特征提取的方法、计算机、计算机程序。

[0005] 本发明是这样实现的，一种外卖场景下店铺embedding特征提取的方法，所述外卖场景下店铺embedding特征提取的方法包括：提取每个用户消费的行为序列；用户消费行为序列按店铺平均消费价格、店铺类别划分；对于划分后的消费序列，提取出训练样本；构建基于负抽样的skip-gram模型，并利用tensorflow框架训练；提取店铺的embedding特征向量。

[0006] 进一步，所述提取每个用户消费的行为序列具体包括：从数据库中，提取用户一定时间内的用户所有下单店铺及其ID，按照用户ID将数据分组，构建每个用户ID的下单店铺集合，存储到数据库中。

[0007] 进一步，所述用户消费行为序列按店铺平均消费价格、店铺类别划分具体包括：将提取出的用户下单店铺序列按店铺的平均消费价格划分为多个等级，同时将不同价格区间的店铺划分为子集合。

[0008] 进一步，所述提取店铺的embedding特征向量具体包括：用tensorflow训练好的向量保存到本地，并在实际操作中提取需要的店铺embedding特征。

[0009] 本发明对于店铺的embedding提取，获取了店铺某些方面的特征信息，将数据从高维的one-hot转变为指定维度向量的特征提取方法；结合商户的embedding与在线下单模

型,带来整体模型的性能提升。而推荐系统的传统CF算法都是利用item2item关系计算商品间相似性。但在实际应用中,用户和物品数量都非常大,这种情况下,评分矩阵会极度稀疏,对算法的效率产生消极影响;同时由于这个问题的存在,两个用户之间的相似度很有可能为零,产生“邻居传递损失”现象;不同的物品名称可能对应相似的物品,基于相似度计算的推荐系统不能发现这样的潜在关系,而是把它们当不同的物品对待。而本发明在一定程度上克服了这些缺点,应用了浅层的神经网络,解决了one-hot维度过高的问题,获取了上下文信息,相比于skip-gram概率模型或者基于神经网络的embedding模型,降低了计算的时间复杂度和空间复杂度。本发明在NLP领域的成功应用,也证明了这是一种获取高维one-hot数据的低维嵌入表示的行而有效的方法。

附图说明

[0010] 图1是本发明实施例提供的外卖场景下店铺embedding特征提取的方法流程图。

[0011] 图2是本发明实施例提供的skip-gram模型示意图。

[0012] 图3是本发明实施例提供的第一幅结果展示图。

[0013] 图4是本发明实施例提供的第二幅结果展示图。

[0014] 图5是本发明实施例提供的embedding特征时延与原始特征的时延比较示意图。

具体实施方式

[0015] 为了使本发明的目的、技术方案及优点更加清楚明白,以下结合实施例,对本发明进行进一步详细说明。应当理解,此处所描述的具体实施例仅仅用以解释本发明,并不用于限定本发明。

[0016] 本发明则解决了one-hot维度过高的问题,同时embedding也表征了上下文信息,相比于skip-gram概率模型或者基于神经网络的embedding模型,无论是计算的时间复杂度,还是空间复杂度,都带来了很大程度上的提升。本发明的计算机配置:Spark、hadoop计算集群,其中Spark必须配置HIVE数据库;Python开发环境;显卡GeForce GTX TITAN X。本发明的存储的配置信息:128G运行内存;硬盘500G以上。

[0017] 下面结合附图对本发明的应用原理作详细的描述。

[0018] 如图1所示,本发明实施例提供的外卖场景下店铺映射embedding向量的方法包括以下步骤:

[0019] S101:提取每个用户消费的行为序列;

[0020] S102:用户消费行为序列按店铺平均消费价格、店铺类别划分;

[0021] S103:对于划分后的消费序列,提取出训练样本;

[0022] S104:构建基于负抽样的skip-gram模型,并利用tensorflow框架训练;

[0023] S105:提取店铺的embedding特征向量。

[0024] 下面结合附图对本发明的应用原理作进一步的描述。

[0025] 本发明实施例提供的外卖场景下店铺映射embedding向量的方法具体包括以下步骤:

[0026] 步骤一,应用hive提取user_id—shop_id数据写入表格user_shop。利用hive构建了店铺字典shop_dict,其组织形式为{shop_id:index},其中index为店铺在字典中的顺

序。然后将原始的user_shop表中的shop_id用字典中shop_idx索引代替；

[0027] 步骤二，将原始的user_shop表中的shop_id用字典中shop_id索引代替。用户六个月以来的消费店铺序列，按照店铺的平均客单价、店铺一级类别进行分组去重，获得构建训练样本所需要的doc，其数据组织形式为：{user_id,array<shop_id_1,shop_id_2,...,shop_id_m>}。对context内只有一家店铺的数据进行过滤。最终获得的数据组织形式为：array<shop_id_1,shop_id_2,...,shop_id_m>，并将其存入HIVE表，便于后续训练样本的构建。

[0028] 步骤三，训练样本的构建，利用python进行数据转换的，在第三版模型时，利用了HIVE进行处理，大大提高了效率；对于每一个context，分别应用get_train_samples函数构建训练样本，返回trains的RDD；最终获取的RDD写入到HIVE表中，用于接下来训练样本的导出。

[0029] 步骤四，构建用于训练的tensorflow图，是整体的训练函数，包括训练数据的获取以及batch的生成。由于总的训练样本大约30亿，每一个batch有1024，因此，在此处设置每10000步计算一次loss值；每100000步显示一次loss值，也就是大约七分钟左右显示一次；每1000000步评估一次并存储embedding到本地。

[0030] (a)构建基于负抽样的skip-gram模型

[0031] Skip-gran模型是使用中间店铺来预测其他店铺(context)。如图2所示，输入向量为one-hot向量x(one-hot，分类方法，通常需要把数据的各个属性转化为一个向量表示，这样每条数据的特征就是一个向量，向量上的每个维度就表示了一个特征属性)。在输出端，变成了多路的输出：

[0032] $$p(w_{c,j}=w_{o,c}\mid w_I)=y_{c,j}=\frac{\exp(u_{c,j})}{\sum_{j'=1}^{V}\exp(u_{j'})};$$

[0033] 其中，$w_{c,j}$是预测在context出现的第c个店铺，其索引位于店铺第j个，而$w_{o,j}$是真实存在的context中第c个店铺。不过由于输出层共享权重矩阵W'则有：

[0034] $u_{c,j}=u_j=V_w'^{T}_j\cdot h$，对于c=1,2,...,C；

[0035] 损失函数变为：

[0036] $$E=-\log(w_{o,1},w_{o,2},...,w_{o,c}\mid w_I)=-\log\prod_{c=1}^{C}\frac{\exp(u_{c,j_c^*})}{\sum_{j'=1}^{V}\exp(u_{j'})}=-\sum_{c=1}^{C}u_{c,j_c^*}+C\cdot\log\sum_{j'=1}^{V}\exp(u_{j'})\ ;$$

[0037] 对context的第c个店铺的输出层的第j个神经元score求偏导：

[0038] $$\frac{\partial E}{\partial u_{c,j}}=y_{c,j}-t_{c,j}=e_{c,j};$$

[0039] 出于表述上的简洁性，定义一组V维向量El={El_1,El_2,...,El_V}作为输出层的预测误差在所有context单位的累加和：

[0040] $$EI_j=\sum_{c=1}^{C}e_{c,j};$$

[0041] 接下来，对W'求偏导：

[0042] $$\frac{\partial E}{\partial w_{ij}^{'}}=\sum_{c=1}^{C}\frac{\partial E}{\partial u_{c,j}}\cdot\frac{\partial u_{c,j}}{\partial w_{ij}^{'}}=EI_j\cdot h_i;$$

[0043] 更新W’：

[0044] $$V_{wjv}^{'(new)}=V_{wj}^{'(old)}-\eta\cdot EI_j\cdot h,$$对于j＝1,2,...,V

[0045] $$V_{wI}^{(new)}=V_{wI}^{(old)}-\eta\cdot EH,$$其中$$EH_i=\sum_{j=1}^{V}EI_j\cdot w_{ij}^{'}$$；

[0046] (b)利用tensorflow框架训练

[0047] Tensorflow数据处理部分，利用tensorflow提供的tf.teain.AdamOptimizer优化器来处理，控制学习速度。通过使用动量(参数的移动平均数)来改善传统梯度下降，促进超参数动态调整。

[0048] 步骤五，将数据利用t-sne进行可视化，获取embedding向量的直观感知，将embedding向量映射到2维，对各店铺的距离进行图形化展示。如图4所示，可以看出，在选择的1000家店铺，有些店铺是能聚类到一起的，而有些店铺是分散到平面的。

[0049] 图3是本发明实施例提供的第一幅结果展示图。从店铺集合中随机选取一家店铺，然后再获取与该店铺在embedding特征空间中最近的10家店铺，展示结果如图3所示。我们可以看到，与选取的店铺最近的十家店铺，其类别大致相同；价格接近，都属于相同的价格区间；距离较近，大部分在2公里以内，属于同一个商圈。

[0050] 图4是本发明实施例提供的第二幅结果展示图。将店铺的embedding向量通过t-sne算法映射到二维平面中，然后显示出来。通过图4可以发现，店铺在embedding特征空间中具有明显的聚类效果，有很多独立的簇是聚到一起的；同时，通过比较同一簇的店铺，可以发现它们大多是属于同一商圈的。

[0051] 图5是本发明实施例提供的embedding特征时延与原始特征的时延比较示意图。可以发现，在应用embedding特征之后，在保证相近的AUC值条件下，融合embedding特征后线上模型的时延明显低于使用原始特征模型的时延。

[0052] 以上所述仅为本发明的较佳实施例而已，并不用以限制本发明，凡在本发明的精神和原则之内所作的任何修改、等同替换和改进等，均应包含在本发明的保护范围之内。

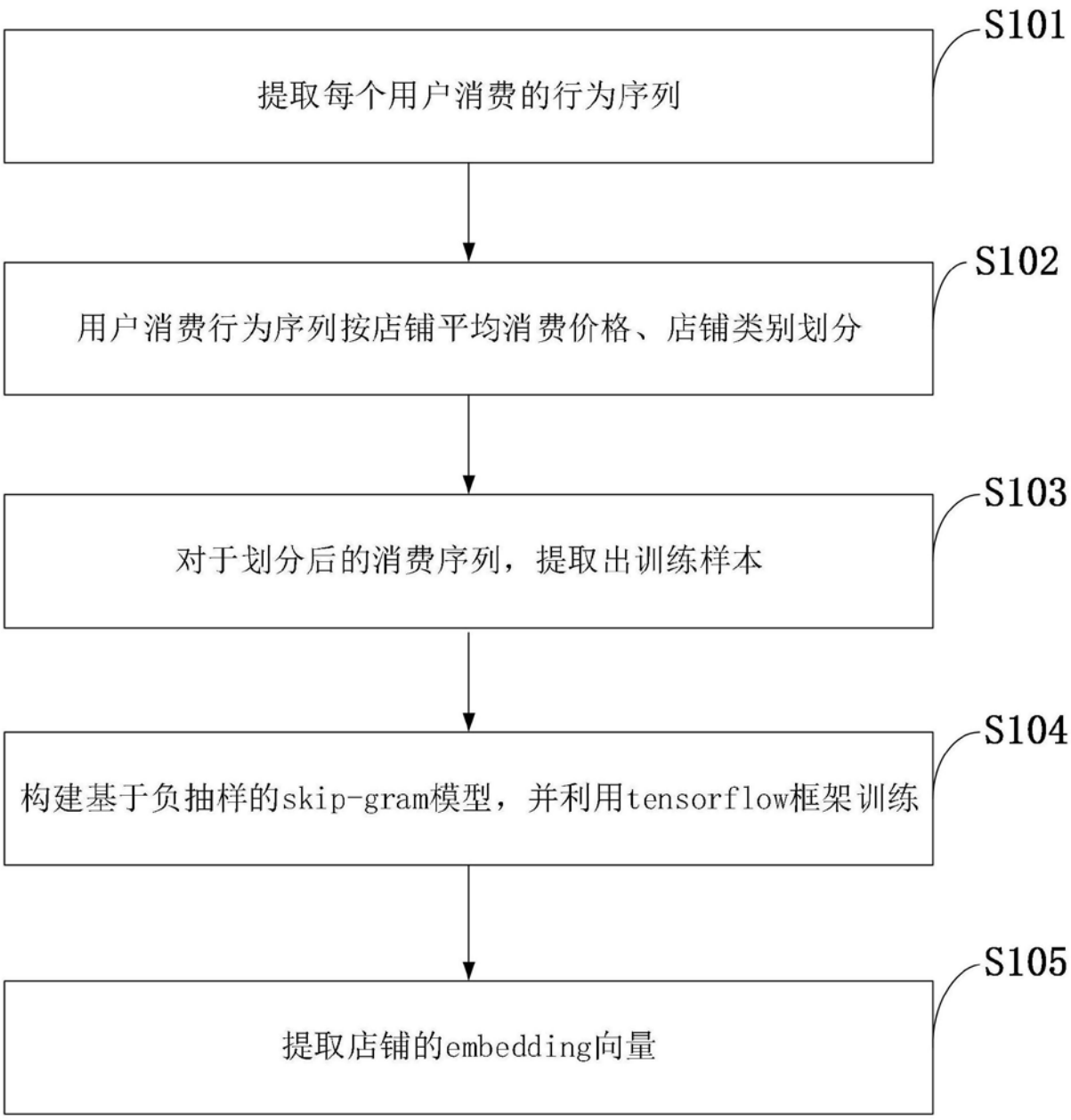

S101
提取每个用户消费的行为序列
S102
用户消费行为序列按店铺平均消费价格、店铺类别划分
S103
对于划分后的消费序列，提取出训练样本
S104
构建基于负抽样的skip-gram模型，并利用tensorflow框架训练
S105
提取店铺的embedding向量

图1

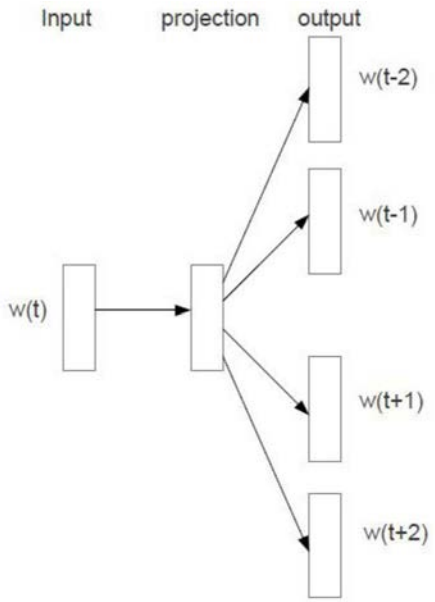


图2

```
  宏泰祥生煎店（梅川路店）23.19  Inhouse有机轻体沙拉（沪定路店） 24.26  唐三义腊汁肉夹馍 25.42  张亮麻辣烫（金沙江路店） 25.49  大娘水饺（梅川路店） 25.00
Neaest to 港粤烧腊（金沙江路店） 19.10:
  蜀蘸串串香（赢华国际店）18.57 港粤烧腊（118广场店）15.75 道味麻辣烫（118广场店）18.67 香哈美食荟 19.66 头料来了 19.32
  巴西秘制烤肉饭 19.20 众品粥店 14.59 简食轻厨（118广场店）15.22 海卤控（古香广东煲仔饭）18.84 张三的粉 18.06
Neaest to 味轩饭店 20.50:
  一家人 24.79 8道菜 29.30 重庆酸菜鱼（巴黎春天店）20.58 骨头饭（118广场店）21.27 肖记水煮鱼（118广场店）26.31
  星期五餐馆 27.03 我爱酸菜鱼 29.30 豪客来家常菜 27.32 一米良品（淞虹路店）21.37 红辣椒麻辣烫 22.31
Neaest to 黄山菜饭骨头汤 26.66:
  私房酸菜鱼（大渡河路店）25.00 百味家常菜馆（中江路店）25.92 五芳斋（杨柳胃路店）26.60 胎林牛蛙@麻饭 25.00 好媳妇脆皮鸡米饭 25.00
  爱吃始蝶的天鹅@麻饭 24.04 壹丰源大食堂（大渡河路店）29.50 一品香特色葱油饼 28.04 如海阁广东潮汕特色美食 29.20 黄山咸肉菜饭骨头汤 26.77
Neaest to 骨头饭（118广场店）21.27:
  重庆酸菜鱼（巴黎春天店）20.58 肖记水煮鱼（118广场店）26.31 豪客来家常菜 27.32 味轩饭店 20.50 一家人 24.79
  我爱酸菜鱼 29.30 8道菜 29.30 红辣椒麻辣烫 22.31 豫申园城隍庙小吃 28.20 问记我家酸菜鱼 25.00
Neaest to 张记油条拉面 27.37:
  张记油条拉面（兰溪路店）26.17 大王锅贴（武宁路店）25.28 段氏龙虾（兰田路店）25.00 阿三生煎 25.00 蜀香冒菜馆（曹杨路店）25.09
  真如小方烤鸭 29.38 天下第一皮（武宁路店）24.52 好粥源（粥面饭）24.99 福建千里香馄饨王 20.53 重庆麻辣烫 22.29
Neaest to 面对面重庆小面 25.77:
  老碗面 28.28 蜀都冒菜 28.45 宏泰祥生煎店 23.87 黄焖鸡米饭 28.87 皇玲粥店（大食焰美食城店）25.00
  文龙正宗桂林米粉（骨头汤菜饭）24.33 真心炒饭小食堂 25.56 小杨麻辣烫（安边路）25.81 老母鸡汤咸泡饭 25.66 张记油条外卖（丰庄店）26.32
Neaest to 詹姆仕韩式简餐（百联中环店）41.75:
  泰妃阁（西郊店）41.57 新贝乐意式休闲餐厅（中环百联店）35.79 食之秘（近铁店）38.73 沙拉先生 34.15 鼎鲜外带寿司（中环百联店）44.04
  品韩舍石锅拌饭炸鸡（百联中环店）32.88 食其家（威宁路店）36.34 芝根芝底（曹杨店）32.68 友丽拌饭（长风景畔店）41.35 掌上韩品（百盛店）38.52
Neaest to D0（上言每百联中环店）38.21:
  满记甜品（上三每真光店）44.74 Royaltea皇茶（中环百联店）33.39 桂源铺（中环百联店）32.45 HEY JU工CE茶桔便（中环百联店）37.16 CoCo都可（真光店）31.59
  R&B世界茶饮（百联店）33.91 1点点（真光店）39.40 禾目章鱼烧（百联店）34.02 幸福侯彩擂（中环百联店）35.17 米芝莲（中环百联店）31.01
Neaest to 味千拉面（农工商118店）55.42:
  必胜客（金沙江店）99.68 达美乐比萨（中江路店）75.23 千拉面（江桥万达店）53.70 米斯特比萨（金沙江西路店）161.78 味千拉面（近铁店）53.24
  PizzaMarzano玛尚诺（宏伊店） 98.94  PizzaExpress马上诺（环球港店） 98.53  达美乐比萨（梅川店） 74.44 必胜客（凯旋店） 99.94 棒约翰比萨（金沙江路店） 60.37
```

图3

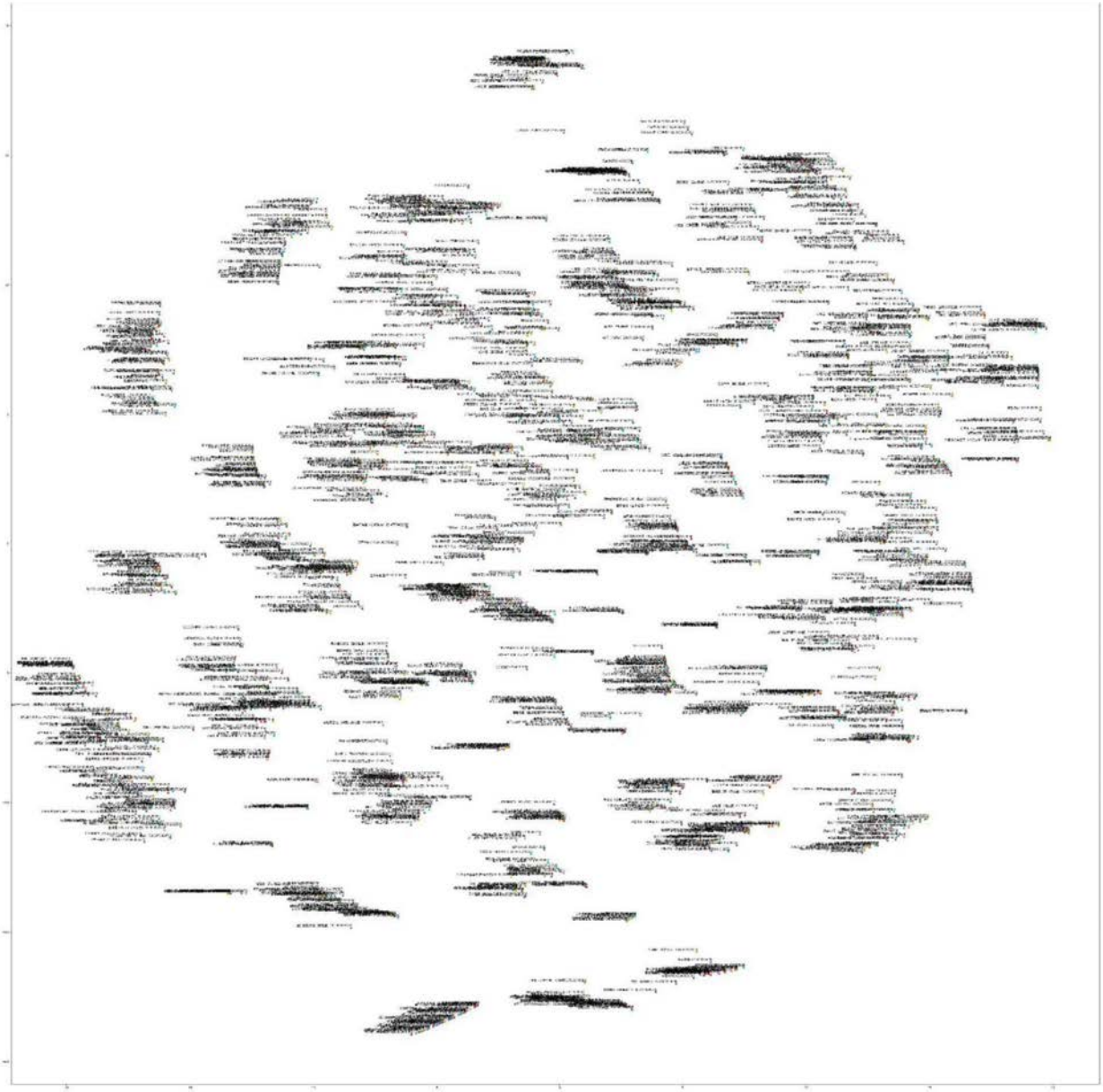


图4

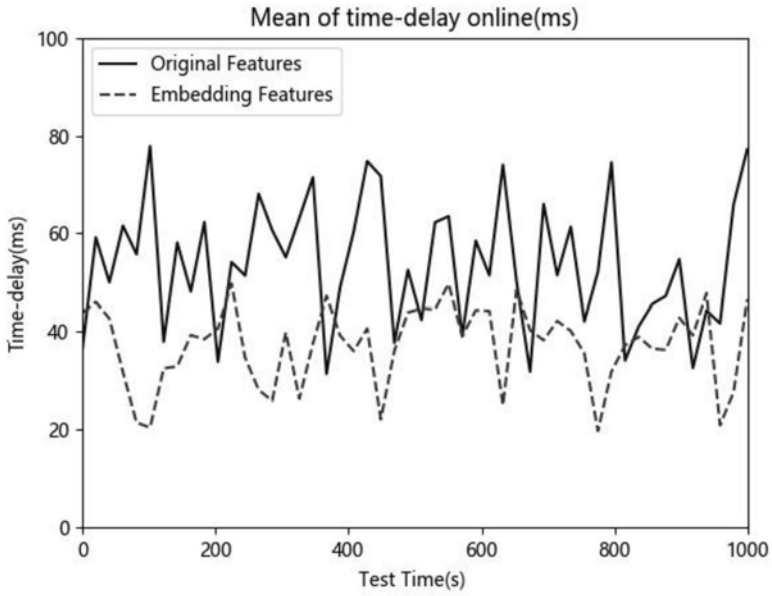

Mean of time-delay online(ms)
Original Features
Embedding Features
Time-delay(ms)
Test Time(s)
0
20
40
60
80
100
0
200
400
600
800
1000

图5