知乎

推荐算法商品相似度-Item2Vec

语-木三
数据科学家，PADI 潜水员

关注他

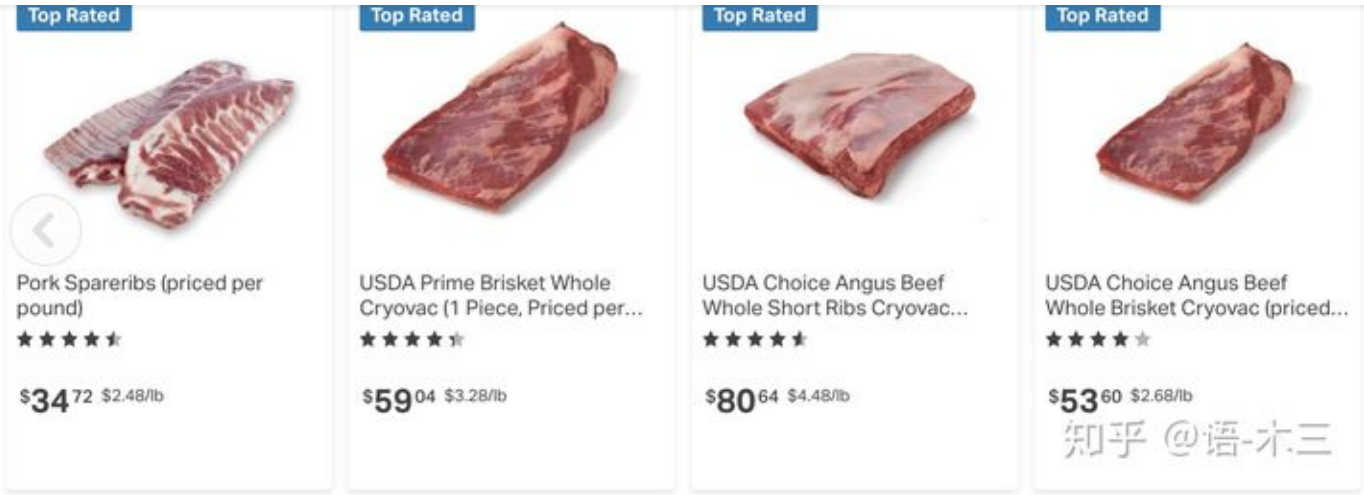
19 人赞同了该文章

这篇文章讲的是如何用Word2Vec的思路来构建推荐算法中的商品相似度。数据用的是线下零售数据，对于线上零售也有借鉴意义。

商品相似度在零售行业的意义

商品相似度分析一直是零售行业中一个非常重要的问题，相信大家都听过沃尔码啤酒跟尿布的故事，虽然我也不知道这是不是真的。对于任何一家零售商来说，能系统化的了解商品之间的关系，并找出substitute item和complementary item，都有非常积极的意义。从库存管理，到货架摆放，到需求预测，选品优化等等问题都会涉及到商品相似度。商品相似度与零售，就像Word Embedding与NLP一样是一个绕不开的话题。

举个很简单的例子，在山姆会员店的网页版中，就有类似"Members also considered"的功能，这里用的推荐算法就涉及到商品相似度的问题。在做选品优化的时候，不同的商品之间有可替代关系，所以商品的实际需求=商品真实需求 + 其他商品的替代需求。这里也会用到商品相似度。



零售关键词

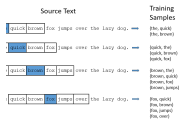
在说具体建模思路之前有几个关键的零售和深度学习的知识点稍微提一下。

- **Utility-based consumer choice model**: 这是在零售业常用的一种建模思路。消费者是否选择购买商品A取决于，Utility(购买A商品) >= -Utility(花费A的价格)。
- **Item Basket**: item basket就是消费者一次性购物所买的东西的集合。通常就是你去超市小票上的全部商品就是一个item basket。这是我们训练Item2Vec的主要数据。
- **商品相似度 (Item similarity)**: 商品相似度是一个比较笼统的概念，根据具体的商业需求可以分为substitute item和complementary item。
  - **Substitute item**: substitute item指的是可以当消费者无法获得他想要的商品A的时候，他可能会选择可替代的产品B，比如可口可乐和百事可乐就是比较理想的substitute item。有人可能会觉得问那超市里卖的各种各样的同类产品是不是都相互可替代呀，当然不是。因为即使是非常相似的产品，也有品牌忠诚度和价格作为影响，比如有很多人他就不和百事可乐，还非说味道不一样。而平时买怡宝矿泉水的人也基本不会去买FIJI。而有些人则基本没有偏好，价格是影响他们的主要因素，打折就好！这类消费者我们称为**价格敏感型**，这个跟穷还是不太一样的。值得注意的是在一个item basket里面的商品通常不是substitute。
  - **Complementary item**: 这类商品是的定义是会一起使用商品，通常是一起购买。比如烧烤架和木炭，牙膏和牙刷，游戏和游戏机等。同样值得注意的是，Complementary item也不一定会一起购买，因为他们的使用周期是不同的。PS4都出了五年了，游戏大作年年有啊。
- **销量预测VS需求预测**：这两个是完全不一样的。销量受限于库存，而需求是可以大于库存的。现实生活中，需求通常是没有可以直接观察的数据的，而销量是能清楚观察到的。大多数情况下零售业的预测是需求预测。因为需求预测可以帮助管理库存， 优化物流，提高货架空间使用率等等。线上零售的逻辑在这里稍有不同，因为在网店消费的客户并不是马上拿到货的，这里给了零售商整理库存的时间。
- **Skip-gram Model**: Skip-gram模型是NLP里面训练word vector的一个主流思路，也是我们这

papers.nips.cc

Word2Vec Tutorial - The Skip-Gram Model

mccormickml.com



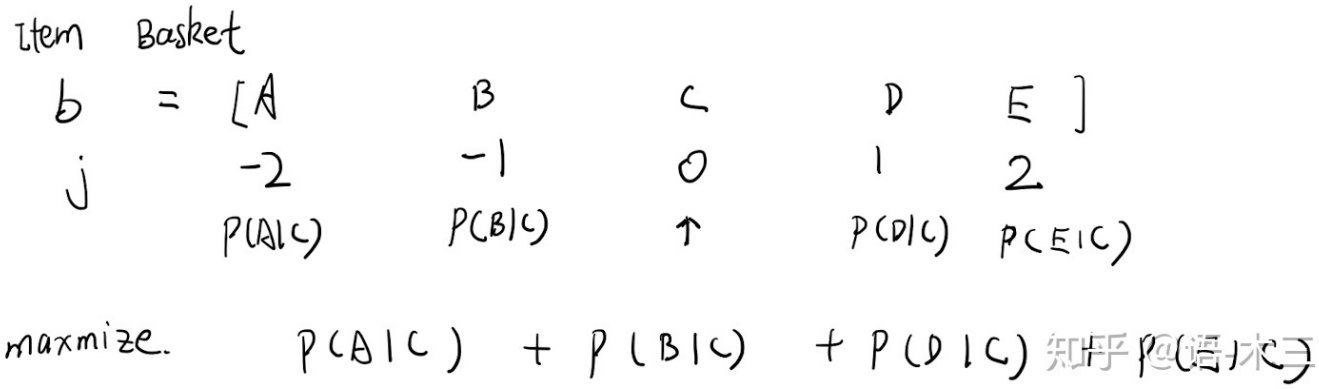
商品相似度建模思路

去年在做Demand forecasting算法的时候，就用过类似Word embedding的方法。我在RNN层之前加了一个item embedding层，这样做显著提高了预测的准确读。然而用端到端训练获得的item embedding质量并不高，因为端到端的方法中损失函数的目标还是提高预测的准确率，也就是一个MAE或者MSE函数。当时我就跟老板提出说可以借鉴NLP里Word vector的思想来研究商品相似度，并构建item vector，后来在看paper的时候又看到几篇有类似思路的研究，就更加肯定了我们这个建模思路。零售数据有以下几个特性能满足Word2Vec的基本思想。

1. Word2Vec假设一个词的意义是通过其所在句子中同时出现的其他词来定义的。在商品相似度这个问题上，一个商品的属性也可以近似的理解为在一个item basket里面的其他商品。
2. Item basket数据是比较容易获得的数据，每个零售商店都可以通过条码扫描器来把一次性付款购买的商品作为一个item basket。这类数据容易获得且准确率高。
3. 因为我们需要区分substitute和complementary，Word2Vec基于单词之间同时出现的概率的思路可以用来区分substitute和complementary。

The Skip-gram Model on Items

我们训练一个模型来预测当商品A出现的时候，其他商品在同一个basket里的概率。



假设我们有 $\{b_1, b_2, b_3, \ldots, b_T\}$ 共 T 组item baskets，window size是 c ，离当前商品的距

这里 $j=0$ ，因为我们不考虑item与自己本身的关系。训练一个神经网络来最大化其他item出现在同一个basket的条件概率，给定当前在这个basket里面的item。这个训练好的模型本身并不重要。我们真正需要的是模型的Embedding层来作为Item2Vec。

让 v_s 和 v'_s 代表输入层和输出层的item vector of s 接下来我们用softmax的方法来定义 $P(s_{i+j}|s_i)$ 。

$$P(s_{i+j}|s_i)=\frac{exp({v'_{s_{i+j}}}^T v_{s_i})}{\sum_{s\in S} exp({v'_s}^T v_{s_i})}$$

我们可以看到分母的计算量与总商品个数 S 成正比，如果直接用这个定义来训练模型会导致庞大的计算量，所以我们引入negative sampling[3]的概念来帮助训练模型。Word2Vec的paper中同样使用了这个方法。可以用以下公式来近似 $P(s_{i+j}|s_i)$ 。

每一个商品会选 K 个negative sample。 σ 是sigmoid函数。

```
Undefined control sequence \E
```

这个negative sampling公式的第一部分最大化同item basket的其他商品 s_{i+j} 和 s_i 一起出现的概率，第二部分则最小化 s_i 和随机选取的商品 s_k 一起出现的概率。这里假设随机选取的商品与 s_i 无关。

以上部分的逻辑跟Word2Vec完全一样，想详细了解如何训练Skip-gram model的请参考我前面推荐的论文[1]和技术博客[2]。

训练Item2Vec的实操心得

在我们实际训练这个Item2Vec模型的时候，我们需要对原始数据进行一些处理来得到一份更加干净的训练数据。这里有几个简单的心得。

1. **获取干净数据。**Item basket如果只包含单个Item或者远高于平均basket size，如果数据量允许可以删除此类baskets。
2. **价格！价格！价格！**价格是一个非常重要的因素，我在前面提到过买怡宝的人是不会去买FIJI的。所以在实际应用Item2Vec的时候需要吧价格因素加到Item2Vec训练得到的item vector上形成新的item vector。如果在建模中不想直接引入价格概念，可以考虑对商品进行等级划分

细分。

4. **去除不相关**。如果零售数据包含对商品大品类的分类，那么对实际商业中不可能存在相似度的商品可以从item basket里面删除，或者训练中不处理，但是在实际商业应用中进行处理。比如我去商店买菜的时候有时候会顺手买一个PS4游戏回家，但这并不代表PS4游戏与我买的菜有任何相关性。真正与我买PS4游戏相关的是去商店购物这件事（我好像找到了一个很学术的方法来表达每次去商店都想买游戏的事实，手动狗头）。
5. **顺序不重要**。在用以上方法训练Item2Vec模型的时候用消费者购买的**顺序**并不重要，尤其是在线下商店，消费者扫描商品的顺序并不是实际购买的顺序。所以训练Item2Vec模型的时候可以对Item basket随机打乱。Item2Vec的思想与Shopper[4]模型完全相反。Shopper是一个有序的概率模型，而Item2Vec是一个无序模型。我认为有序这个假设并不适用于零售业，因为消费者购买商品的顺序极大的取决于商品的摆放，而且消费顺序数据存在极大的获取难度，并不适用于大多数企业。即使是线上零售，商品加入购物车的顺序数据获取容易，但是有序消费行为依然是一个相当严格的假设，而且并不符合大多数人的消费行为。
6. **非常重要的一点！！！非常重要的一点！！！非常重要的一点！！！Item Basket**这类数据隐藏了两个条件: 1. 消费者花时间去了商店，并且在真实世界里，去商店这个行为本身就有Utility（通常我们认为是负Utility）。2. 消费者去了商店且买了东西，要知道相当一部分人去逛超市是没有买东西的，这部分人没有**Item basket**的数据。通过这两点我们可以认识到该模型是存在固有偏见的，所以对Item2Vec和所有基于Item2Vec的模型进行审核的时候需要仔细思考实际的商业情况然后再做决定。

参考

1. ^ a b Distributed Representations of Words and Phrases and their Compositionality https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf
2. ^ a b Word2Vec Tutorial - The Skip-Gram Model http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/
3. ^ Negative Sampling http://mccormickml.com/2017/01/11/word2vec-tutorial-part-2-negative-sampling/
4. ^ SHOPPER: A PROBABILISTIC MODEL OF CONSUMER CHOICE WITH SUBSTITUTES AND COMPLEMENTS https://arxiv.org/pdf/1711.03560.pdf

编辑于 2020-08-11

零售行业　推荐系统　人工智能算法