知乎

如何构建知识图谱?

DataFun...
DataFun社区，公众号ID：datafuntalk，欢迎关注！
关注他

13 人赞了该文章

本文根据转转张青楠老师，在DataFun AI+ Talk中所分享的《二手电商知识图谱构建以及在价格模型中的应用》编辑整理而成。

一、知识图谱概述

这次的分享主要从以下四个部分：知识图谱概述、知识图谱构造、转转二手电商知识图谱、在价格模型中的应用。

1.1 什么是知识图谱

知识图谱是谷歌在2012年提出来的，最初的目的是优化其搜索引擎。在现实世界中是存在很多的实体的，各种人、物，他们之间是相互联系的。知识图谱就是对这个真实世界的符号表达，描述现实世界中存在的一些概念，以及它们之间的联系。具体来说是一个具有**属性**的**实体**，通过关系连接而成的**网状知识库**。

▲ 赞同 13 ▼ 添加评论 分享 收藏 ...

知乎

在电商的知识图中，包括用户、商家、商品，他们带有各自的属性，彼此之间又互相联系。知识图谱的基本组成三要素：**实体、属性、关系**。实体-关系-实体 三元组；实体-属性-属性值三元组，在电商的知识图谱中，用户和商品都是实体。

在知识图谱中，有一类特殊的实体叫做本体，也叫做概念或语义类。它是一些具共性的实体构成的集合。比如说，比尔盖茨和乔布斯都是人，微软和苹果都是公司。

二、知识图谱构建

目前的知识图谱分为两类。一类是开放域的知识图谱，另一类是垂直领域的知识图谱。比如谷歌为搜索引擎所建立的知识图谱就属于开放域的。垂直领域的知识图谱，比如说金融的，电商的。

首先就是要先处理数据。互联网上的数据基本上都是结构化的，非结构化的和半结构化的。结构数据一般就是公司的业务数据。这些数据都存储到数据库里，从库里面抽取出来做一些简单的预处理就可以拿来使用。半结构化数据和非结构化数据，比如对商品的描述，或是标题，可能是一段文本或是一张图片，这就是一些非结构化数据了。但它里面是存储了一些信息的，反映到的是知识图谱里的一些属性。所以需要对它里面进行一个抽取，这是构建知识图谱中比较费时费力的一个工作。

从数据里需要抽取的其实就是之前所提到的实体、属性、关系这些信息。对于实体的提取就是NLP里面的命名实体识别。这里相关的技术都比较成熟了，从之前传统的人工词典规则的方法，到现在机器学习的方法，还有深度学习的一些使用。比如说，从一段文本里面，我们提取出来比尔盖次这个实体以及微软这个实体，然后再进行一个关系提取。比尔盖次是微软的创始人，会有这么一个对应的关系。另外还有属性提取，比如比尔盖茨的国籍是美国。在这些提取完成之后都是一些比较零散的信息，然后在再加之前用结构化信息所拿到的东西以及从第三方知识库里面所拿到的信息做一个融合。

另外还需要做的是实体对齐和实体消歧。

关于实体对齐。举例来说，比尔盖茨这四个字是中文名称，Bill Gates是他的英文名称，但其实这两个指的是同一个人。由于文本的不一样，开始的时候导致这是两个实体。这就需要我们对它进行实体对齐，把它统一化。

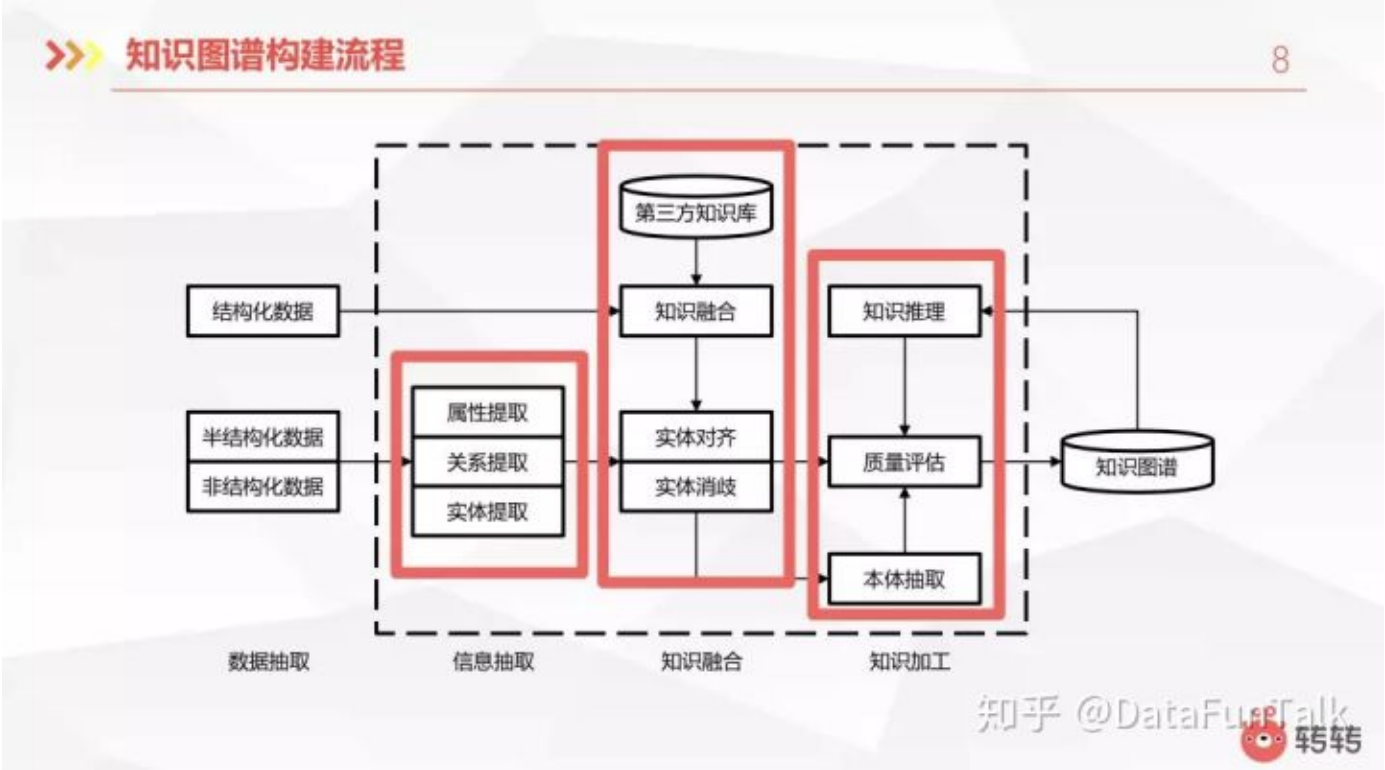
另外是实体消歧。举例来说，苹果是一种水果，但是在某些上下文里面，它可能指的是苹果公司。这就是一个实体歧义，我们需要根据上下文对它进行实体消歧。

在完成了以上步骤之后，接下来就是本体抽取。比如之前提到的微软和苹果，它们的实体是公司。从文本里面可能无法直接提取出来，它们是公司。那么需要一些方法对他们进行抽取。然后搭建出本体库，比如说公司是一个机构，它是有这种上下流的关系的。对于平级的也需要计算一个他们的

▲ 赞同 13 ▼ 添加评论 分享 收藏 ···

知乎

在以上步骤完成之后需要对知识库进行质量评估，这是一个避免不了的人工步骤。在做完质量评估以后，最终形成知识图谱。形成知识图谱以后，有些关系可能是无法直接得到的，然后需要进行知识推理，这可以对知识图谱进行扩展。比如，猫是猫科动物。猫科动物是哺乳动物。这就可以推理出来，猫是哺乳动物。但是这个推理也不是随便就可以推出来的。比如，比尔盖茨是美国人，比尔盖茨创建了一个公司，但这个公司并不一定是美国的。



三、二手电商知识图谱

主要从以下四个部分阐述：业务理解、知识图谱设计、算法、开发。

3.1 二手电商特性

搜索优化和个性化推荐是我们最开始所做的初衷。主要去做一些意图识别或是自动化查询这些。个性化推荐这里，我们利用知识图谱做一些召回源以及推荐排序模型特征。在电商运营这里，主要是帮助后台运营组货。在垂直业务这块，主要是做一些价格模型和供需关系分析。

二手电商不同于一手电商。首先就是数据源的质量。二手电商平台上面的商品都是个人发布。商品的描述信息不像商家那样完整。我们提供给他们的可选项，也都不一定会被完整的填写。

▲ 赞同 13 ▼ 添加评论 分享 收藏 ···

知乎

第三点是具备一些二手属性。二手店电商的商品都有很多二手属性。比如说成色、外观、屏幕划痕、是否换屏、是否翻新等等。

最后是价格差异。商品进行折旧以后，他们的价格会有一些差异。二手商品的价格是具备很强区分度的特征。

3.2 二手电商知识图谱构建

先构建商品的知识图谱。商品的知识图谱是类似树的形态。树由一级一级的节点组成，最后的叶子节点是商品实体，它的下面是一些商品的属性。

遵循业务需求循序渐进。在制作知识图谱的过程中，是边做边用的过程，而不是花费了很长的时间来做的很完整后才去使用。我们是根据具体的需求将知识图谱拆成几个步骤，然后进行持续的输出。

那么怎么拆分？根据之前提到的树的形态的知识图谱，首先要做的是先描点。先把图中的节点标好，然后再去挖掘属性中一些K-V信息，得到一些零散的点边关系，接着再把这些零散的点和边的关系串起来形成一张图，变成知识库。最终，再把商品挂上去。

▲ 赞同 13 ▼ 添加评论 分享 收藏 ···

知乎

首先，是term层面的一些应用。提取物品词，完成本体构建。然后，K-V层面就是连接点和边。提取tag词，完成属性抽取。接着，在图的层面。tag词树结构化，完成知识库构建。最后，商品粒度。将商品挂靠上去，完成实体抽取。

3.3 商品理解——物品词

首先从商品中提取出它的物品词，然后根据用户的行为数据得出用户偏好物品词，接着根据这个用户偏好物品词进行召回或是排序特征。

那么具体的实现方案：

先是物品词库的构建，不断地挖掘当前都有哪些东西，以及以后还打算做哪些东西。这部分的数据大部分是从我们自有的结构化数据那里拿到的，也有一部分是从外部爬去得到的，还有是从命名实体识别得到的。

接着是上下位关系提取，沙发是个实体，布艺沙发也是个实体。布艺沙发也是沙发的一种，它们是一个上下位的关系。

然后是并列相似度计算。像布艺沙发和皮质沙发的相似度是比较高的，而沙发和相机的相似度就比较低。还有是文本对齐。类似于同义词，比如，相机和照相机其实是指代的同一个东西。

当以上完成以后，就构架出了一个物品词库。接着就是商品层面，商品物品词提取，使用到的数据源有：分类信息、标题文本、商品描述、商品图片。

应用场景主要就是：

▲ 赞同 13 ▼ 添加评论 分享 收藏 ···

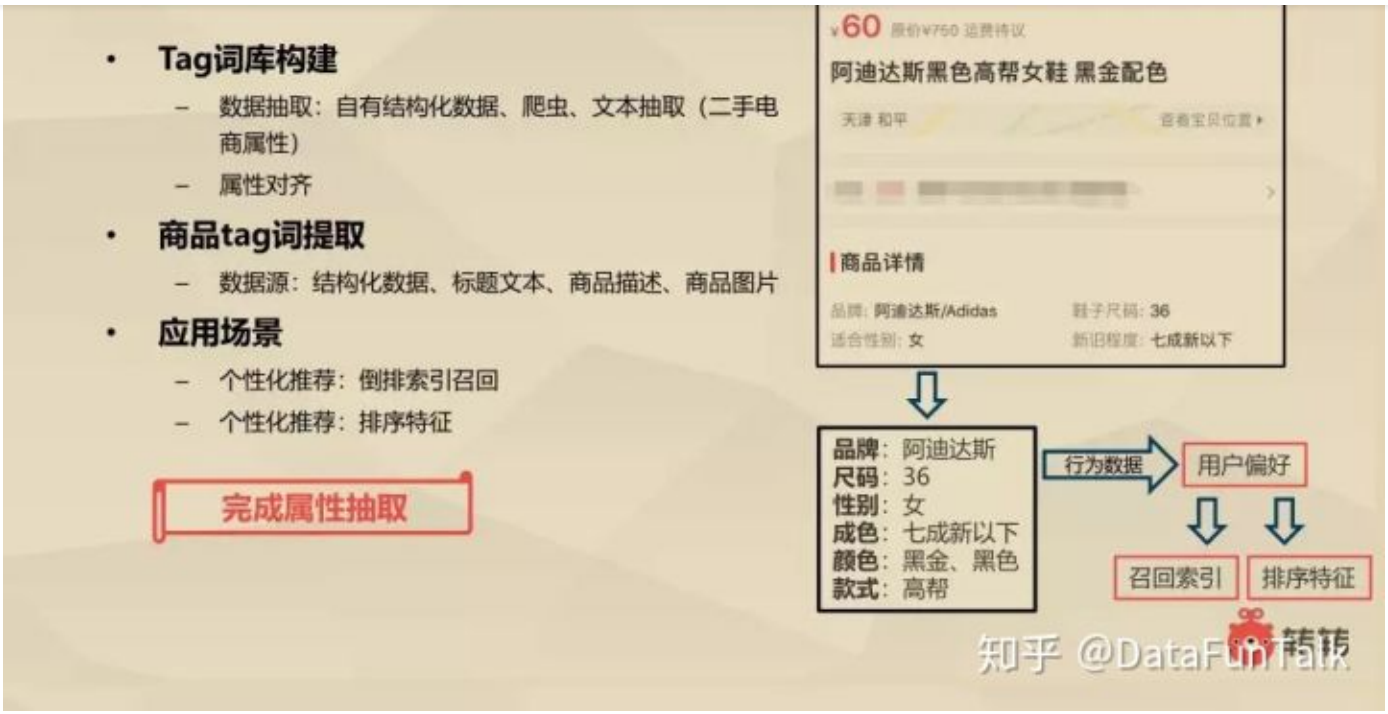
知乎

- 个性化推荐：排序特征

3.4 商品理解——tag词

后面做了一个商品理解的Tag词，这是物品词的演进，这是服务的升级，刚才我们提取到的是用户感兴趣的东西，但是人往往不会局限于对这个东西感兴趣，还有可能对这类物品有很多的要求。所以需要从属性的角度去挖掘用户的兴趣，比如右下角的例子。对该商品提取出更多的属性。那这个套路和刚才的物品词比较相似。这里需要注意的是，一手化的数据可以从自有结构的数据，爬虫，文本抽取中可以拿到，但是二手数据只能从文本挖掘中抽取。还有属性对齐。还有商品Tag词的提取，他的数据源来源于结构化数据，标题文本，商品描述，商品图片等。应用场景和物品词一致。最后就完成了属性抽取。

▲ 赞同 13 ▼ 添加评论 分享 收藏 ···



3.5 Tag词树结构化

上面做完之后，我们发现提取出的key-value属性，都是各自离散存在的。然后会出现数据质量的问题，所以把之前挖掘出的term给提取出来组成一个树，下面是例子。从这个树里面可以追溯到他的所有信息。

这样的做法还提供了query结构化，对query进行理解，他的应用场景有三部分，个性化推荐和智能搜索，这一块截止，做完了商品库的知识库的构建。后面就是商品挂靠。

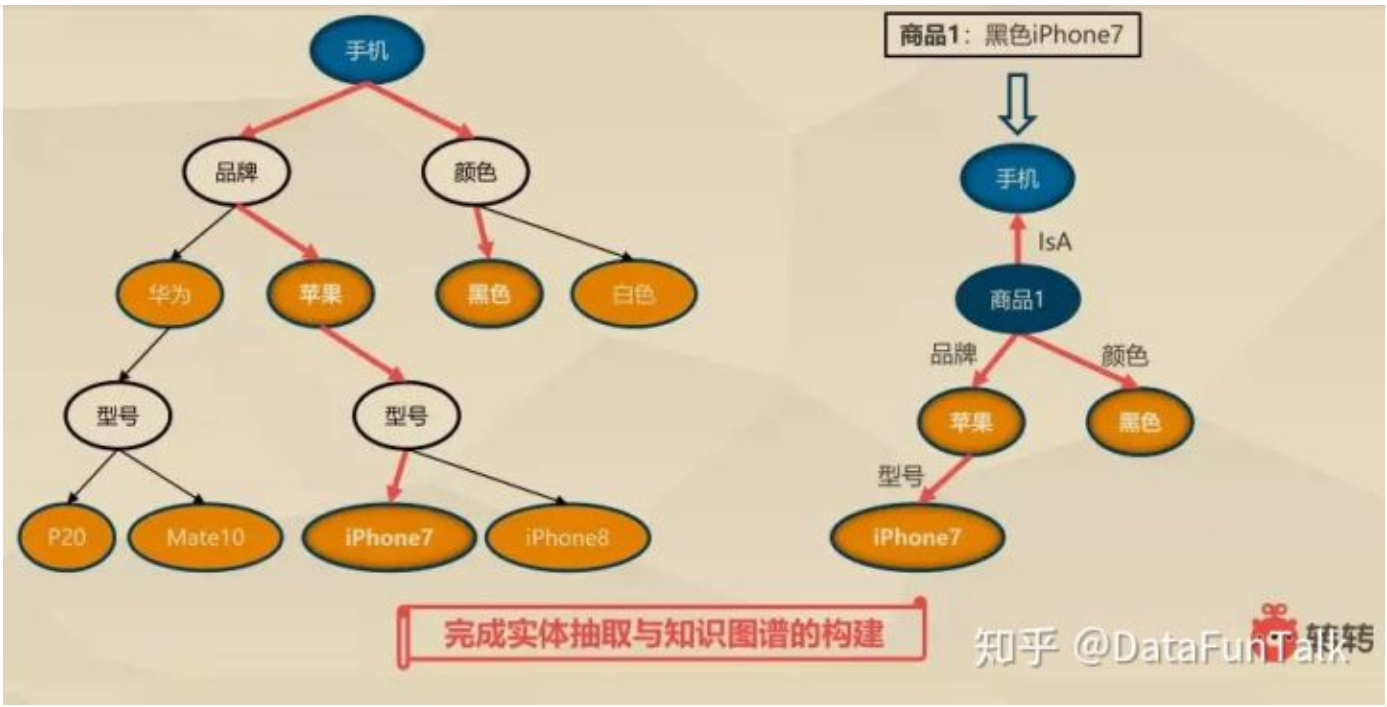
3.6 商品挂靠

商品挂靠指利用分类信息、商品标题、商品描述、商品图片等数据，对本体库（Tag词树型结构）中的节点进行匹配和生成商品知识路径。同时消岐有可能一个商品会匹配到本体库中的多个本体（物品词）和 对属性节点赋予权值，选取匹配权重最高的本体。

这还是刚才的例子，商品挂靠之后生成一个实体（右侧）这一块做完之后完成实体的抽取与知识图谱的构建。目前我们有一些关于知识推理和知识图谱的应用，优先级并不是这样的，目前还没有发力去做。

▲ 赞同 13 ▼ 添加评论 分享 收藏 ···

知乎

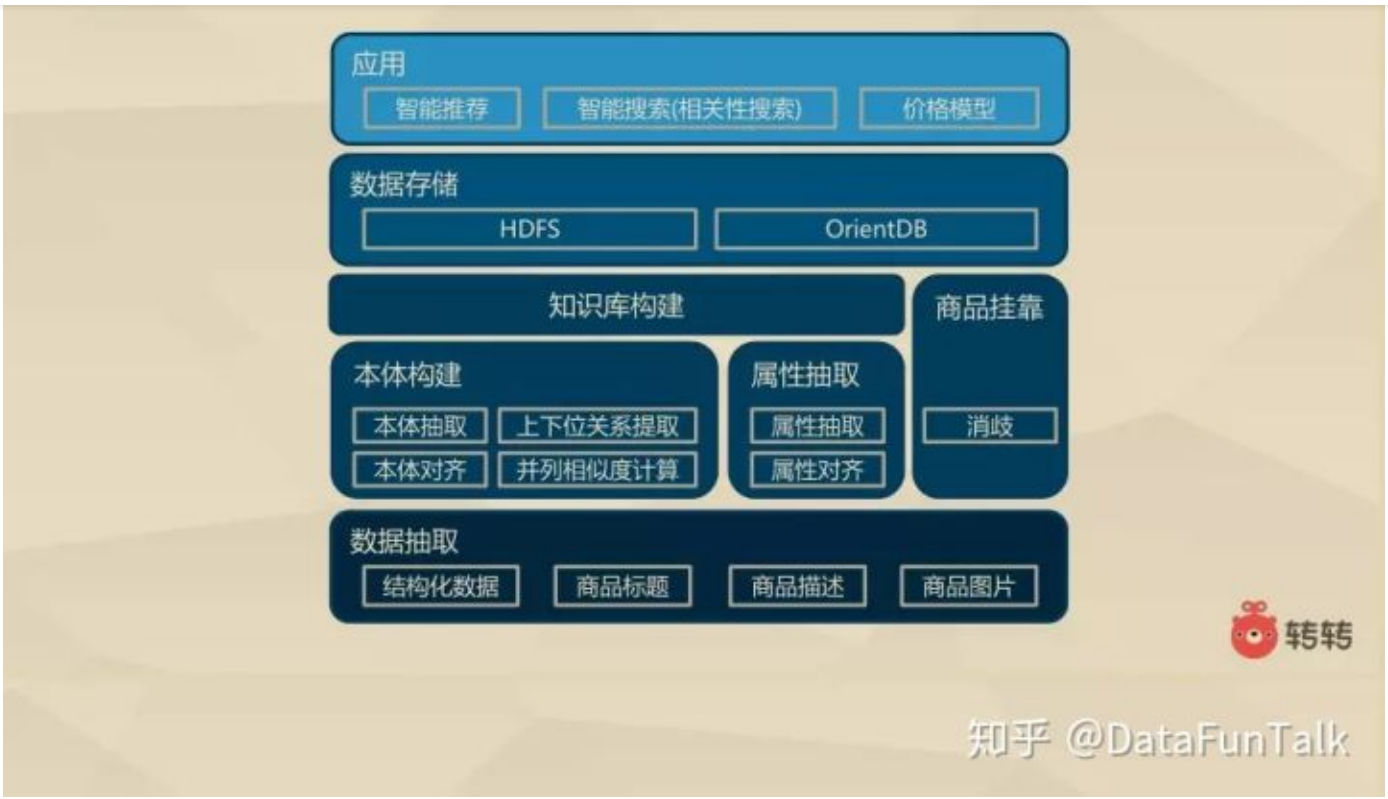


3.7 二手电商知识图谱构建

根据场景去介绍就可以构建出下面的架构，首先是数据抽取，在进行本体构建和属性抽取，在进行知识库的构建，最后完成商品的挂靠，把这些数据存储在HDFS或者OrientDB中，就可以进行智能推荐和智能搜索以及价格模型的构建。这里有一个消岐的概念，他主要是做根据树的权重的加和，权重较高的路径他的置信度就越高。消除一些无效的路径和属性。

▲ 赞同 13 ▼ 添加评论 分享 收藏 ···

知乎



四、在价格模型中的应用

它的应用很多，这里说一下在价格模型中的应用，为什么做这件事情，对于二手商品的来说，很难去定一个合理的价格，所以我们这边希望提供一个定价的能力。

4.1二手标品化

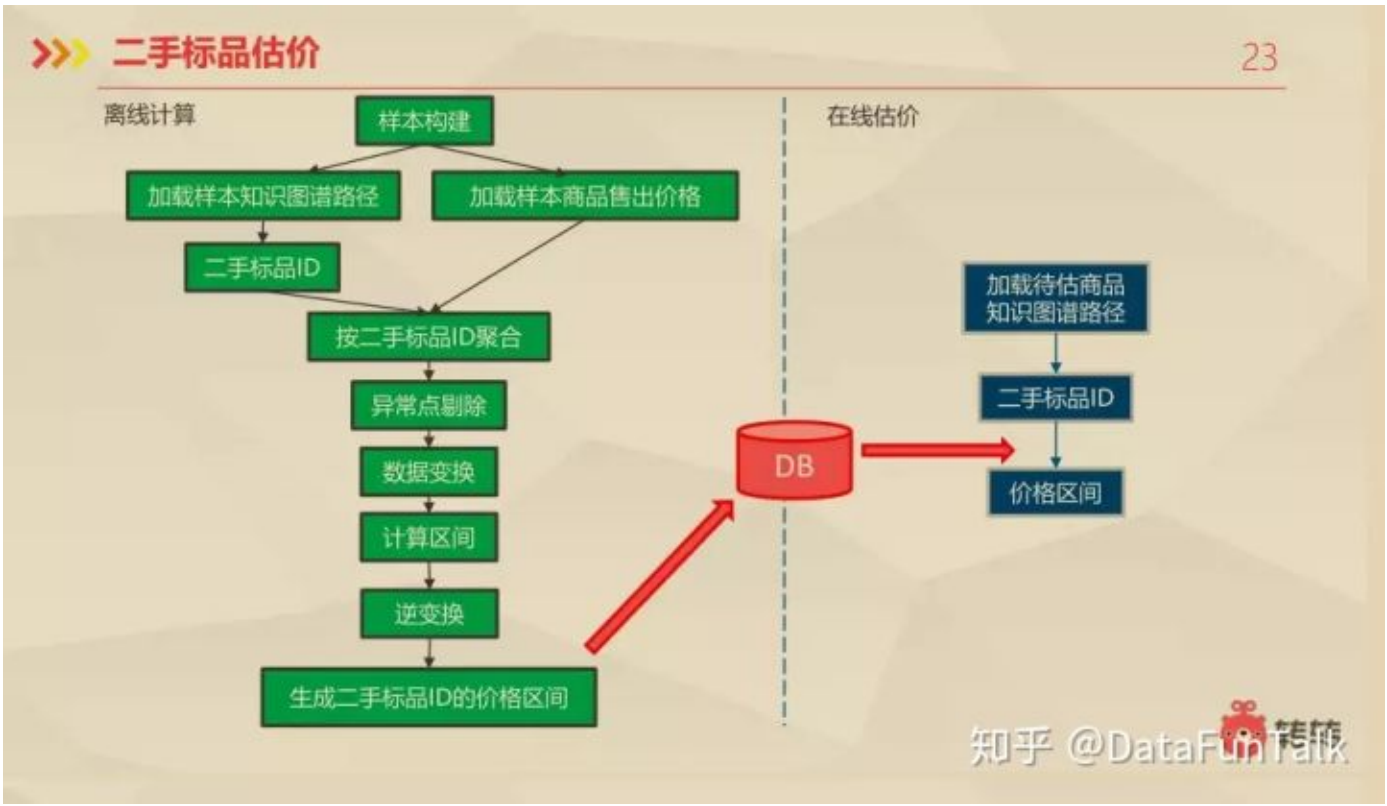
首先需要二手标品化，先做知识谱图商品挂靠，然后去筛选出价格敏感的二手属性。举个例子，我的二手手机屏幕碎了，这是很影响定价的一个因素。但是另外一个手机仅仅是划痕，这个属性对二手手机的定价不是明显的。所以需要去筛选一些对价格影响的属性。在同本体、 同一手属性值和同价格敏感二手属性值下归纳二手标品，把这个ID作为这个实体新的属性打到知识图谱图谱上。我们假设这个标品的商品价格是同分布。针对这个假设，我么做了统计方法做估计价格区间和生成二手标品ID到价格区间的的映射。最后得到的结果是可以支撑这个假设的。

4.2二手标品估价

这块就是整个流程，前面要进行样本构建，然后在加载样本知识图谱和样本商品售出价格，在开始离线计算二手标品商品ID的价格区间。由于我们也没有二手商品的真正的价格，所以这里需要另外一个假设，我们认为大部分成交的二手商品的成交价是合理的，因为这是买家和卖家讨价还价之后的结果，并且基本上满足了双方的心理预期，所以我们收集已成交商品的价格，在按照二手标品ID聚合，对异常点删除，在进行数据变

▲ 赞同 13 ▼ 添加评论 分享 收藏 ···

知乎



4.3 非二手标品商品估价

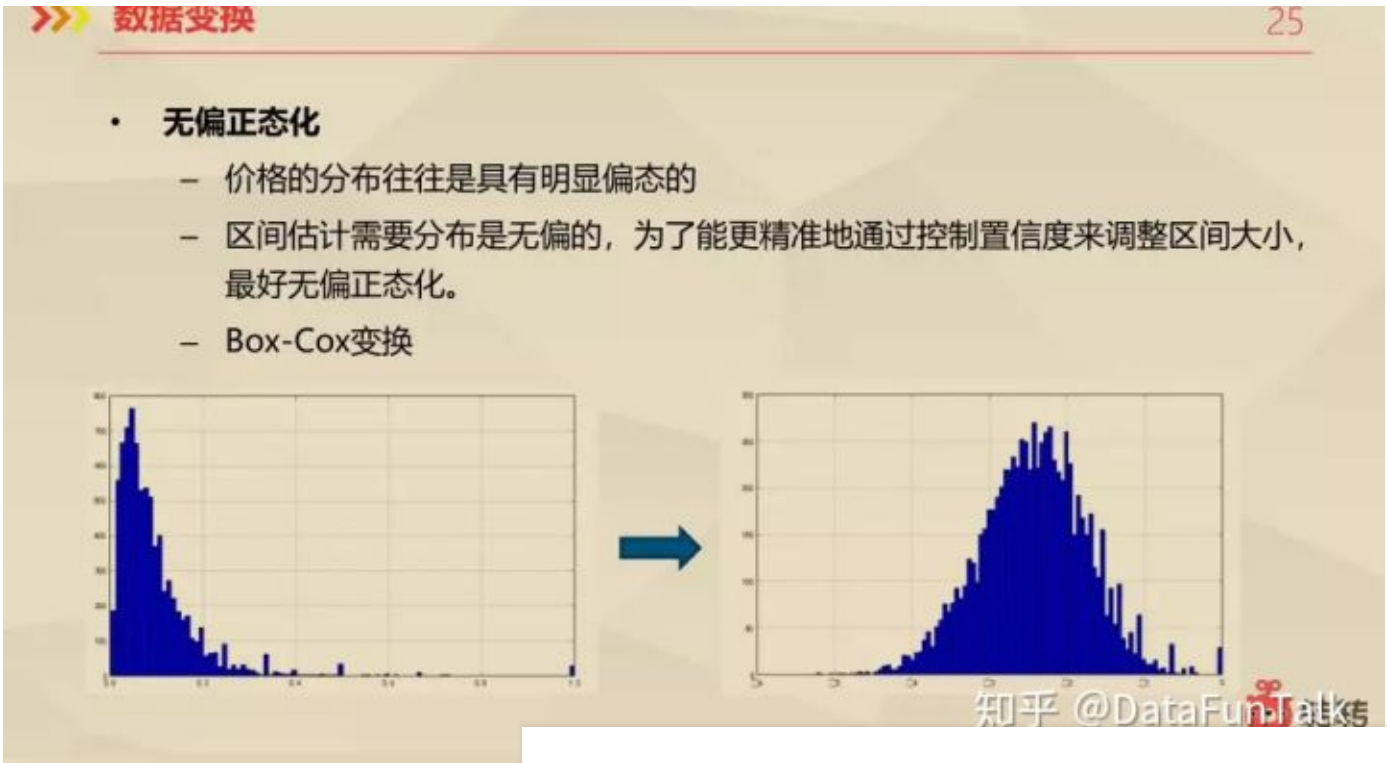
上面仅仅说了二手标品ID的估价，这里还有非二手标品商品估计。手机很好说，但是衣服的话，从一手状态就不太好标品化，这有一套另外的解决方案，首先还是基于知识图谱制作，查找图谱中最近的TopN个出售商品，在聚合出售的价格，删除异常点，进行数据变化，计算价格区间，最后进行逆变换，生成商品价格区间。

▲ 赞同 13 ▼ 添加评论 分享 收藏 ···

知乎

4.4 数据变换

对于价格来说，他的分布有明显偏态的，但是区间估计需要分布是无偏的，为了能更精准地通过控制置信度来调整区间大小，最好无偏正态化。类似于左下角的分布，拿对数变换或者平方根变换就可以变换成近似正态分布，但实际数据的情况会复杂多样一些，为了能很好得无偏正态化，我们采用Box-Cox变换。对数变换和平方根变换是其特例。



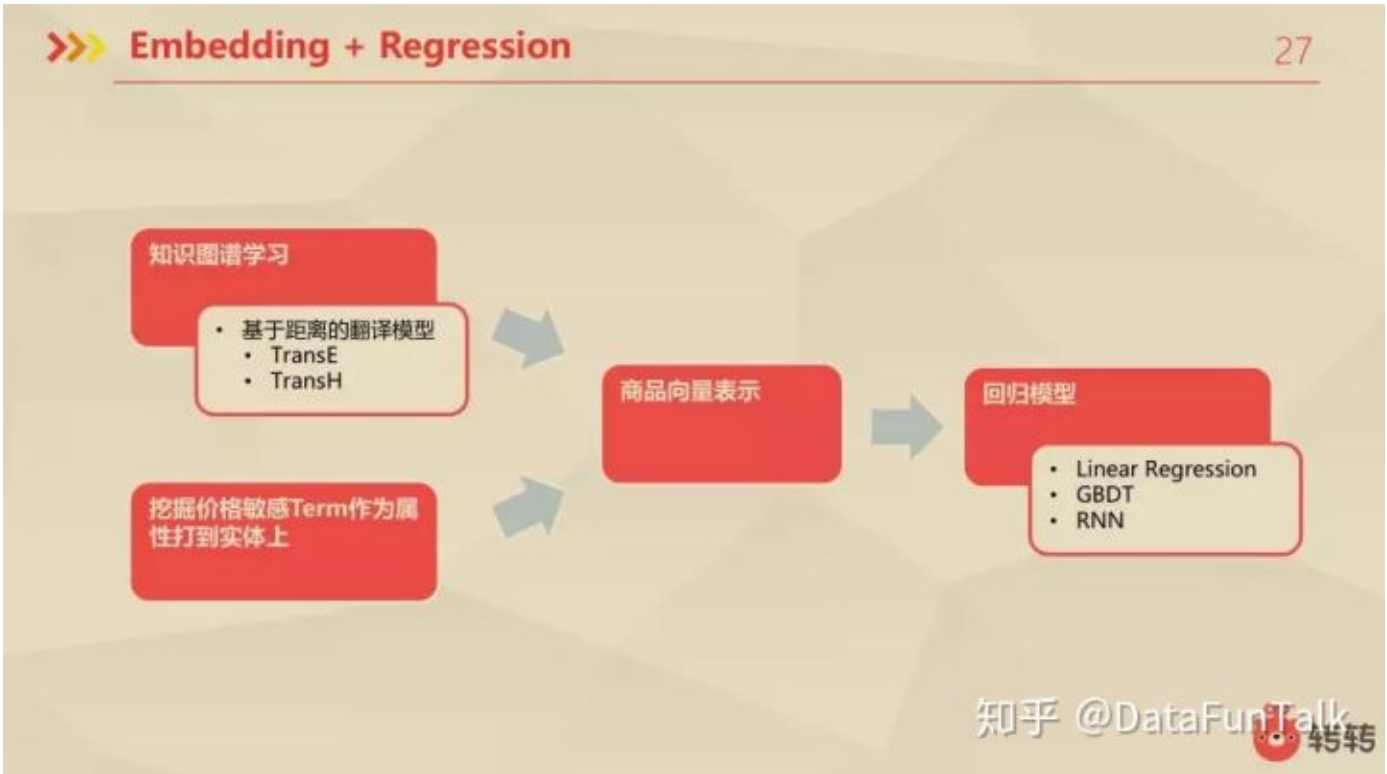
▲ 赞同 13 ▼ 添加评论 分享 收藏 ...

知乎

有了正态分布之后，我们可以做区间的划分，首先我们希望这个区间可以涵盖大多数的商品，可以求均值，标准差，根据不同业务的需要，计算出价格区间，然后将计算出来的区间的上下限，做Box-Cox逆变换。这样才是真正的价格区间。

4.6 Embedding + Regression

刚才所说的是基于统计的方法，后面还有另外一种能够做法，基于回归的方法。先进行知识谱图的学习，挖掘出价格敏感的term作为属性打到实体上，在把商品用向量表示，做回归模型。然后可以用回归的方式去预测出商品的基本的定价。



作者介绍：

张青楠，算法架构师，转转算法部基础模型团队负责人。主导了整套电商基础模型体系的建立。曾就职于当当推荐部，任资深推荐算法工程师。

——END——

发布于 2018-09-22

机器学习 知识图谱 深度学习（Dee

▲ 赞同 13 ▼ 添加评论 分享 收藏 ···

知乎

推荐阅读

知识图谱好文章整理

转自：知识图谱交流圈 欢迎加知识图谱QQ交流群：829449428 1、社会化推荐在人人网的应用 2、金融知识图谱的现状和展望 3、中文知识图谱构建思路是什么？ 4、搜索引擎和知识图谱那些事 (上)....

时海 发表于小k看互联...

美团大脑：知识图谱的建模方法及其应用

美团技术团... 发表于美团技术博...

干货|个热点之

第四范式

还没有评论

写下你的评论...

▲ 赞同 13 ▼ 添加评论 分享 收藏 ...