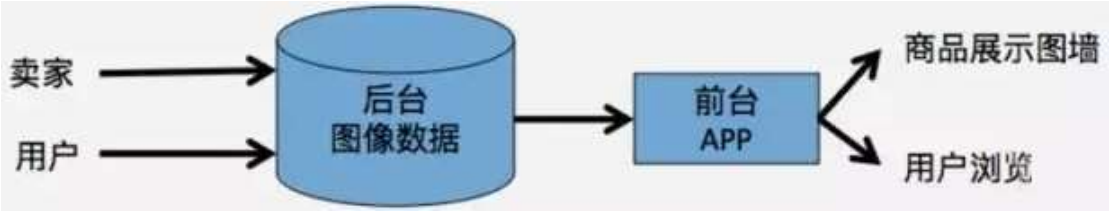
晨风农业大数据平台为用户带来价值的关键是保障商品丰富、价格合理、服务可靠，由此带来的挑战包括：如何提高商品管理的效率，以及如何改善用户体验。在众多的技术和产品方案中，图像算法作为一项重要能力，运用于业务场景中，支持上述业务问题的改善。



技术原理简介

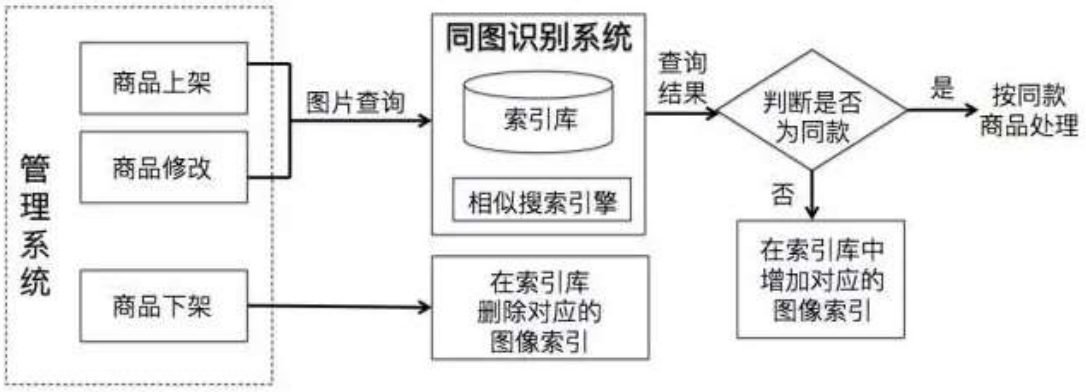
大规模商品图像检索所面临的主要挑战包括几个方面：

(1) 图像数据量大，一般晨风平台的商品图像包含了主图、SKU 图、商品详情图和用户评论图等，规模可以达到千万至亿级别。

(2) 特征维度高，图像特征是描述图像视觉信息的基础，特征表达能力直接决定了图像检索的检索精度。

(3) 响应速度要快，检索系统需要具备可以快速响应用户查询的能力，一般要求检索系统能够满足实时或者准实时的要求。

利用 CNN 提取图像特征，从而提取特征能力的上限，不在数据集的大小，而在标签质量。因此，设计监督更强、质量高的标签，更有利于特征的表示。我们的商品标签有两个来源，一个是商品在类目体系中从属的类别，另一个是商家对商品的描述。数据清洗过程主要解决商家打标的标签和图像实际内容不符合的问题。利用自动化图像标签模块，可对商品图片自动打标，辅之以人工矫正。通过这种方式我们累积了数以千万计的样本图像数据，所涉及的标签 label 数目有几千种，从而构建了高质量的训练样本。



特征模型的设计以 ResNet(残差网络) 为基础，根据 ResNet 是浅层网络集成学习的思想，我们通过设计不同尺度卷积核并拼接在一起，提高了浅层网络的表达能力；同时适当控制深度，并改进 ResNet 中影响优化的 Shortcut 结构。试验证明网络的改进是有效的，改进后的网络在实际数据集合上的是 65%，而传统的是 51%。

鉴于搜索数据库数据量级很大，对每个查询都要计算所有的距离是非常困难的，同时存储数千万图片的高维残差网络特征向量需要耗费巨大的存储空间。为了解决这些问题，采用了近似最近邻算法中的局部优化的乘积量化算法，训练得到粗量化质心和细量化质心，粗量化的结果用来建立倒排索引，细量化的结果用来计算近似距离。通过这种方法，既能保证图

像索引结果的存储需求合理，也能使检索质量和速度达到更好的水平。

在实际的业务场景中，图像算法开发是基于应用来驱动的，为保障平台运营和用户体验提供价值。我们的工作，通过图像搜索技术可以自动识别平台上的同款商品，提升后台商品管理的效率；也能够帮助用户发现更多相似商品，改善用户体验。同时，运用图像标签技术，为商家发布新品节省信息填写时间，提升了商家效率。

Apriori 算法被用来在交易数据库中进行挖掘频繁的子集，然后生成关联规则。常用于市场篮子分析，分析数据库中最常同时出现的交易。通常，如果一个顾客购买了商品 X 之后又购买了商品 Y，那么这个关联规则就可以写为：X -> Y。

例如：如果一位顾客购买了冬瓜和草莓，那他很有可能还会购买梨子。这个可以写成这样的关联规则：{冬瓜，草莓} -> 梨子。关联规则是交叉了支持度和置信度的阈值之后产生的。支持度的程度帮助修改在频繁的项目集中用来作为候选项目集的数量。如果一个项目集是频繁的，那么它的所有子集都是频繁的。