概要介绍

一、 前言

随着网络的迅速发展，数字经济已成热点，作为数字经济的最主要组成部分，电子商务发展飞速。据 CNNIC 调查显示，截止 2021 年 12 月，我国网络购物用户规模达 8.42 亿，用户规模逐年上升。

在网上购物不断增加的同时，一些不正当经营如刷单、虚报价格等行为也随之出现，严重违反电商法，侵犯消费者知情权。近年来，国家开展治理行动，严惩此类不正当行为。但针对如此庞大的商品数量，单靠人工工作量巨大且易出现遗漏。基于此，本项目首创一种高效识别网络零售平台异常商品的方法，不仅可以快速准确筛选出网络零售平台异常商品，还能有效减少人工干预成本和出错率。

团队采用基于内存的 spark 分布式框架，通过 Level 字段对商品进行分类，对于异常价格，采取正态分布进行筛选，通过“桶+二分”方法降低时间复杂度；对于异常销量，通过设置动态区间进行筛选，最终分别得到了价格和销量的异常数据。

二、 创意描述

2.1 我们无法通过传统的工具对大量数据进行处理，所以团队将采取大数据方法。技术选型采用基于内存的 spark 分布式框架，做到高效处理数据。如果数据量过大，则将其部署在多个平台分布式处理。

2.2 团队通过 Level 字段对所有商品进行分类，有效避免商品价值的不同对结果产生的影响。

2.3 对于价格异常的筛选，团队将对数正态分布转换成正态分布，更适合数据的分布趋势。

2.4 团队通过“桶+二分”方法优化聚类算法，降低其时间复杂度。

2.5 对于销量异常的筛选，团队针对不同商品类别设置相应的动态区间进行调整，有更好适应性。

三、 实施方案

四、 功能简介

本项目首创一种高效识别网络零售平台异常商品的方法，不仅可以快速准确定位出价格和销量异常的商品，还能有效减少人工干预成本和出错率。

五、 特色综述

第一步，团队对电商数据各个字段进行观察分析，发现分类的 5 个 Level 较特殊，Level 字段并不是真正的缺失，而是将每一个商品从一个大类层层下分到小类，直到不可细分，但不是所有的大类都可以下分至多层小类，所以就导致了后续部分字段的缺失；

第二步，团队发现不同类别的商品由于价值差异无法直接进行比较，例如小轿车和自行车都属于交通工具类，但价格相差甚大。若先对商品按照 5 个 level 进行分类，我们就可以对同一类商品进行有效比较。

4.1 价格异常：

4.1.1 团队首先分析每一类商品的价格波动，并选取统计学中的变异系数来衡量波动大小，因此团队就得到了存在异常数据概率较大的类别。之后再分析这些类别中商品价格的分布，发现同一类别商品价格相比于其他分布而言更符合对数正态分布。通过资料查阅，发现对数正态分布可以转换成正态分布，再基于统计学的规律，超过 3e 之外的数据为异常数据，因此我们就得到了初步的异常数据。

4.1.2 团队对同一类商品的价格进行聚类，求出每个点到离其最近的 k 个点距离和的平均（简称平均距离），以此作为每一件商品的评分。随后团队取此类中商品个数的 10%为 k 值，这种方法复杂度为(n^2)/10，数据量过大，通过这种方法会花费较多的时间，因此我们对算法进行优化，通过“桶+二分”方法将复杂度优化到 O(m*logm)（m 表示这一类商品中价格的种数，且 m<n），基于此，可以较快处理大规模数据。

4.1.3 团队发现参数字段里面包含了大量的产品信息，因此我们检测所有商品产品信息的完整度，如果该字段缺失程度越高，则代表其异常程度越高。

4.2 销量异常：

4.2.1 首先团队将所有数据对销量进行降序排序，并认为前 30~50%的数据为销量异常数据的概率很大，然后从中选取一条记录，通过它的商品 ID 检索出该商品其他三个月的销量情况。

4.2.2 若其他三个月的数据存在，这个时候就用其他三个月数据的平均值作为一个基值 X，设定一个区间如下图，判断 X 在哪个区间内，从而得到正常的销量范围，如果本条记录的销量数量在范围外则为异常。

图 1-1 销量区间图

若 X 在①内：则 X—50X 算正常，范围外异常

若 X 在②内：则 X—10X 算正常，范围外异常

若 X 在③内：则 X—10X 算正常，范围外异常

若 X 在④内：则 X—5X 算正常，范围外异常

若 X 在⑤内：则 X—2X 算正常，范围外异常

4.2.3 若其他三个月的数量均不存在，则检索出该商品对应类目的其他所有该类目的数据，去掉前后 5%，取中间的 90%取平均值作为基值 X，后续步骤同上 4.2.2。

六、 开发工具与技术

数据处理：Pyspark,Pandas
前端：Vue, Echarts
后端：Flask, Python

七、 应用环境

硬件环境：CentOS 8 核 16 G

八、 成本控制

处理数据量	服务器	购买单价	使用单价	时间	总价（月）
4 个月约 1700 万条数据（7G）	8 核 16G 服务器	2586 元/年	0.45 元/时	约 30 分钟	215.73 元

算法运行时间短，成本低，性能好，经济成本控制合理，性价比高。

九、 结语

网络零售平台不正当行为的出现，不仅是对消费者知情权的侵害，也是对电商法的无视，本团队首创此高效准确的异常商品识别方法，力求发现平台不正当行为，还市场正常秩序，促电商良性发展。