China's Prices Project at Xiamen University (CPP@XMU)

项目简介

张果

厦门大学　王亚南经济研究院

2016年12月9日

目录

目录

课题背景：CPI

- 中国官方CPI存在诸多问题
  - 频率低、公布速度慢
  - 不公布原始数据
  - 不公布详细编制方案
  - 官方对于编制方案的解释力度不够
- 大数据技术发展迅速
  - 海量数据采集：高频采集海量数据
  - 海量数据储存：管理、实时发布海量数据
  - 海量数据清洗：处理海量非结构化数据
  - 海量数据分析：高性能/分布式计算、数据可视化

课题背景：线上市场

- 中国互联网市场上，大型活动日（如双十一）越来越多，活动越来越复杂
- 大型活动日提供了一个理想的外生冲击，有利于解决内生性问题
- 基于结构模型(structural approach)的实证文献非常少

目录

课题目标

- 第一阶段(2016.2-2016.10)：基于官方CPI编制标准和电商实时价格数据，编制线上高频价格指数
- 第二阶段(2016.11-present)：基于大型活动日和电商微观数据，研究线上市场的产业组织特点

目录

文献综述：CPI

- 对中国CPI编制的讨论
- 基于线上价格数据编制价格指数
  - Billion Prices Project at MIT：http://bpp.mit.edu/
  - 清数iCPI：http://www.bdecon.com/

文献综述：线上市场

- 消费者：Adda&Cooper(2006)
- 零售商：Ellison&Ellison(2005);Fan(2013)
- 平台：Rysman(2009)
- 互联网数据：Edelman(2012)

目录

感兴趣的方向

- 主题(Topics)
  - Dynamic elasticity estimation and welfare analysis
  - Reputation dynamics
  - Search obfuscation
- 方法(Methodologies)
  - 动态结构模型(Dynamic structural approach)
  - 离散选择模型(Discrete choice model)

目录

数据来源

- 平台选择：市场份额最大的平台——Tmall,JD
- 数据来源选择：搜索页搜索结果(前2-5页)→商品详情页
- 分类选择：基于CPI分类及其分类解释，分别对不同平台分别编制关键词列表

目录

初步成果

- 分布式爬虫+自动化运维系统：海量数据采集[1]
- 数据库方案：海量数据管理[2]
- 初步数据分析[3]
- 初步资料整理[4]

[1] 部署需要，尚未开源
[2] https://github.com/xmucpp/cppdbKit
[3] https://github.com/xmucpp/double11-data
[4] https://github.com/xmucpp/double11-summary

目录

需求分析

- 可以应对网站反爬的爬虫
- 可以调度多台机器的系统
- 可以自动化管理多台机器的系统
- 可以方便地调用数据的数据库系统

解决方案

- 分布式反爬+自动化运维系统
  - 爬虫组件：爬虫模块、代理模块、日志模块
  - 分布式组件：机器调度模块、消息队列模块、脚本调用模块
  - 自动化运维组件：自动化部署模块、自动化管理模块
- 数据库系统(developing)

目录

团队成员

- 导师：茅家铭老师[5]
- 负责人：张果(14)[6]
- 核心成员：林行健(13)、黄玺(14)、刘晓曼(15)、朱星宇(15)、张祎璘(15)、马宁(16)、唐瀚林(16)
- 成员构成：
  - 13级本科生：1人
  - 14级本科生：5人
  - 15级本科生：7人
  - 16级本科生：8人

[5] http://www.wise.xmu.edu.cn/people/faculty/a81c4142-cb73-4f3f-94b4-2b937d4c1acf.html
[6] https://guo-zhang.github.io/

团队成员

- 专业背景：
  - 王亚南经济研究院：7人
  - 经济学院：6人
  - 计算机系：4人
  - 管理学院：1人
  - 外文学院：1人
  - 人文学院：1人
  - 国际学院：1人

团队成员

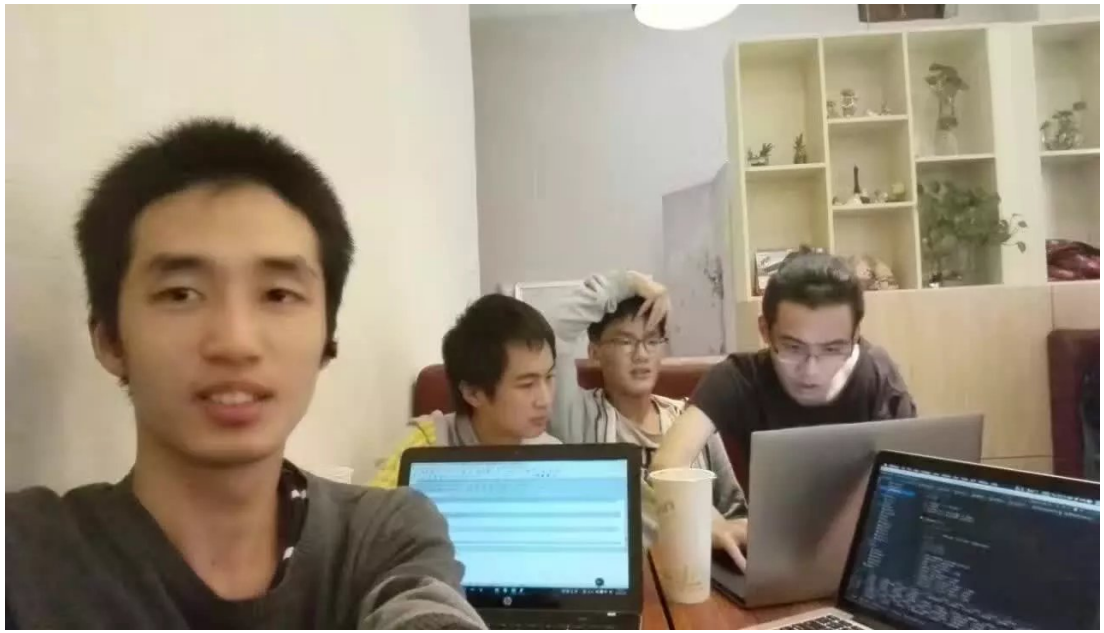

Figure: CPP主力程序员（从左到右：张果、黄玺、唐瀚林、林行健）

目录

团队分工

- 技术部
  - 服务器组：林行健(13)、唐瀚林(14)
  - 数据库组：黄玺(14)、刘理(16)、李蔚然(16)
  - 爬虫组：马宁(16)、岳忠信(16)
- 数据组：张祎璘(15)、周韵丰(14)、吕昕(14)
- 宣传组：朱星宇(15)、姜昊(16)、杜雪旸(16)、林逸伦(15)
- 文献组：张果(14)、刘晓曼(15)、郝泽栋(15)、庄建伟(15)、张伟贤(16)
- 财务组：张晓博(14)、王芷若(15)

目录

当前进度

- 服务器组：修改分布式爬虫系统的bug
- 数据库组：设计、测试数据库方案
- 爬虫组：Tmall、JD评论爬虫
- 数据组：双十一数据描述统计
- 宣传组：项目网站制作
- 文献组：双十一、双十二资料整理；梳理文献

目录

进度计划

- 数据库组：部署数据库
- 数据组：双十一、双十二数据描述
- 宣传组：上线项目网站
- 文献组：双十二资料整理；梳理文献

目录

主要困难

- 技术：Tmall反爬机制不明，无法对应破解，爬虫效率仍然不太理想
- 人员：缺少网络工程师、NLP工程师、机器学习工程师、网站设计、宣传文案等
- 硬件：缺少高配置服务器一台
- 资金：没有资金来源

目录

主要平台

- 项目网站(正在修复)：http://www.xmucpp.com/
- Github主页：https://github.com/xmucpp
- 知乎专栏：https://zhuanlan.zhihu.com/xmucpp
- 知乎账户：https://www.zhihu.com/people/cpp-45-10

主要平台

- 微信公众号：xmucpp2016(XMUCPP)

目录

联系我们

- 项目邮箱(张果)：zhangguocpp@163.com
- 加入我们(刘晓曼)：liuxiaomancpp@163.com
- 知乎：
  - CPP：https://www.zhihu.com/people/cpp-45-10
  - 张果：https://www.zhihu.com/people/zhang_guo
  - 刘晓曼：
    https://www.zhihu.com/people/liu-xiao-man-3-2
  - 朱星宇：https://www.zhihu.com/people/felix-zhu-23


CPP