厦门大学校长基金项目

申　请　书

项目名称： 基于在线商品价格大数据的分析与挖掘技术研究

申 请 人：

所在院系：

办公电话：

手　　机：

电子邮箱：

申请日期：

基本科研业务费项目管理办公室

二〇一六年制

一、简表

中文摘要	信息技术的快速发展，产生了海量的对于经济学研究意义重大的、深入个体层面的微观互联网数据，同时催生了对海量互联网数据的采集、管理、挖掘理论和实践，这为许多经济学和统计学问题提供了全新的工具。本项目基于中国官方CPI编制标准和海量数据的采集、管理、清洗、挖掘技术，参考Billion Prices Project at MIT在高频线上价格数据方面的研究成果，探究中国高频线上价格指数的编制方法。本项目探索高频微观价格数据库的建设方法，为计算机科学界通过大数据技术支持社会学科研究提供了样本，有助于为中国经济学界提供更多的数据采集、管理、挖掘、分析方案，推动中国实证研究的进步，同时也为经济学界提供基于微观数据研究宏观经济学问题的范例。
关键词(不超过 5 个，用分号隔开)	
线上价格指数；K-means 聚类；线上线下价格差异；数据挖掘；在线商品价格	

二、报告正文

（二）项目情况

1. 立项依据

1.1 研究意义

1.1.1 理论意义

信息技术的快速发展，产生了海量的对于经济学研究意义重大的、深入个体层面的微观互联网数据，同时催生了对海量互联网数据的采集、管理、挖掘理论和实践，这为许多经济学和统计学问题提供了全新的工具。本项目基于传统的中国官方CPI价格指数设计和编制方案，结合大数据采集、管理、挖掘技术，参考Billion Prices Project at MIT现有的相关研究成果，采集、管理、挖掘海量线上商品数据，构建中国线上每日价格数据库，编制中国每天线上价格指数，为价格指数的编制这个宏观经济学和经济统计学的基本问题，提供大数据背景

下新的思路、解决方案和实证研究成果。

1.1.2 实践意义

由于种种原因，当今中国经济学界的数据获取来源仍然相对较少。同时，经济学界对海量数据的处理需求也越来越复杂。在发展迅速的大数据理论和技术的帮助下，改进经济学界的数据收集方案和数据分析方案有了更多的可能。本项目探索高频微观价格数据库的建设方法，为计算机科学界通过大数据技术支持社会学科研究提供了样本，有助于为中国经济学界提供更多的数据采集、管理、挖掘、分析方案，推动中国实证研究的进步，同时也为经济学界提供基于微观数据研究宏观经济学问题的范例。

1.2 国内外研究现状及发展动态分析

1.2.1 经济学、统计学相关文献

(1)微观价格数据在价格指数编制中的应用

在大数据的技术下，可以获得的微观价格数据来源越来越多，这为非官方机构编制价格指数和官方机构改进价格指数编制提供了可能。Cavallo和Rigobon（2016）介绍了Billion Prices Project at MIT的主要工作：他们基于从2007年开始来自全世界的线上价格数据，编制了十几个国家的线上超市价格指数，并提出了价格指数的质量变化调整方案。Cavallo（2013）基于巴西、智利、哥伦比亚、委内瑞拉与阿根廷线上价格指数与对应官方CPI对比，发现巴西、智利、哥伦比亚和委内瑞拉四国价格水平和价格变动和官方CPI保持一致，但是阿根廷的价格水平和价格变化是官方数据的三倍。基于BPP@MIT的工作，孙易冰等（2014）提出了基于线上零售平台作为数据来源、基于网页爬虫技术的价格指数编制模型。各国官方CPI统计也开始加入线上数据。比如，美国官方CPI数据中9%来源于线上数据；荷兰也将线上价格数据加入了CPI统计中。此外，平台层面的价格变动研究也有了一些成果。比如，Brynjolfsson 和 Smith （2009）， Ellison 和 Ellison （2009a），以及 Gorodnichenko 等人（2014）发现，谷歌购物的价格数据比官方CPI的数据变化的更频繁，尺度也更小。

(2)线上线下价格对比

Cavallo（2016）基于BPP@MIT从2007年开始采集的十几个国家的线上超市价格数据和通过线下人工采集的线下价格数据，对比了十几个国家线上线下价格水平和价格变化的特点，发现72%的价格水平保持一致。基于第三方比价软件数据和官方CPI数据的对比发现，线上

数据的价格变动更频繁而且变化幅度更小，比如Brynjolfsson 等（2003），Ellison 和Ellison（2009a），Ellison和Ellison （2009b），Lnnemann and Wintr （2011）， Gorodnichenko等（2014）， and Gorodnichenko and Ta-lavera（2014）。

(3）互联网数据在经济学的应用

互联网数据近年来越来越多地被应用在经济学各个领域的研究中。比如，Ｒｏｔｈ和Ockenfels（2002）对约480 eBay和亚马逊拍卖中的价格数据进行研究；Ellison and Ellison (2009) 收集的消费者在线购买的计算机内存模块的数据进行分析；Choi 和 Varian（2009） 通过谷歌趋势的数据，预测未来的申请失业救济金；Wu 和 Brynjolfsson（2009）通过谷歌搜索数据预测房屋的价格和销量；Freedman and Jin (2011) 在网络上收集个人借贷网站上的借贷用户数据； Baye, Morgan, 和 Schölten (2004）对比价网站上的销售商价格进行审查。

(4)大数据技术在经济学中的应用

大数据技术在经济学中的研究的应用范围也越来越广泛。比如，Liran Einav1和 Jonathan Levin（2014）表明大数据的出现已经允许更好的测量经济效果和成果，并为经济学家在一系列领域提供了新的方案和思路；Linnet Taylor, Ralph Schroeder, Eric Meyer （2014）探讨了跨学科的大数据的兴起，介绍了新类型的数据被研究人员使用来研究经济问题的现状，并分析大数据对经济的潜在影响，比如如何用计量经济学的方法研究经济学之外的问题、大数据是如何改变或改善经济模型，并催生经济学家和其他学科之间的大数据合作；Choi 和Varian（2009, 2012），Goel 等人 （2010），Carri，ere-Swallow 和 Labbe （2011），McLaren 和 Shanbhoge （2011），Arola 和 Galan （2012），Hellerstein 和 Middeldorp （2012）已经表明，谷歌查询数据可以在短期内显著地预测各种经济指标。

1.2.2 计算机科学相关文献

聚类是数据挖掘中的基本算法。它的作用是在无任何监督的情形下，将具有近似属性或特征的对象放到一起。对于具有不同属性的对象，将形成多个不同的类别。对聚类算法的研究开始于1980年代对于低纬度，小数据的分析。最著名的是80年代早期提出的K平均聚类算法。聚类问题在诸多应用需求中普遍存在,比如，如文本/网页聚类、模式识别、图像链接、图像分割、通过向量量化进行的数据压缩以及最近邻检索。虽然到目前为止人们提出了几十种聚类算法，但大多数的聚类算法，比如具有代表性的DBSCAN和mean shift，它们的聚类时间复杂度都是$\Omega(n^2)$以上。在大数据背景下，这些算法显得无能为力。在本项目中，需要对

海量的商品进行聚类分析。传统的聚类方法无法满足这一要求，因此需要探索高效的聚类算法。

在众多聚类算法中，相比较而言，K平均聚类具有线性时间复杂度，过程简单，同时具有适度稳定的性能而被广泛采用。它被认为是数据挖掘领域十大算法之一。然而随着数据规模的急剧胖胀，即使具有线性时间复杂度的算法也不能应对现有的海量数据规模。因此，最近人们对聚类算法的探索，集中在寻找高速而稳定的K平均算法改进方案。鉴于此，我们将着重回顾最近10年人们所提出的各种改进的K平均聚类算法。

目前，改进K平均聚类主要有两种途径。一种主要提高聚类的质量。一个重要的改进来自于Ostrovsky等提出的K-means++。这一改进的主要思想是通过优化K平均所获得的初始聚类中心，使得聚类可以最终收敛于一个较好的局部最优点。而且，由于初始中心的更好选择，聚类收敛速度也会稍有加快。然而，为了找到这些更好的聚类初始点，这个方法需要遍历整个数据集多次。即使最近的改进可以是遍历次数减少，额外的聚类开销仍然无法避免。

众所周知，在K平均聚类过程中，最为耗时的步骤是，在每次迭代过程中，算法需要为每个样本寻找离它最近的那个中心。于是，当样本数目很多（数据集很大）以及样本维度很高的情况下，K平均聚类会变得相当耗时。鉴于此，Kanungo等人提出用K-D树来索引整个数据集，把样本到中心的查询，变成中心到样本的最近邻搜索。从而加快寻找最近中心的速度。Dan等人提出了类似的方法。然而这类方法只适用于样本维度只有几维或十几维的情形。当数据维度增加到几十维甚至上百维的时候，这类方法很难再对K平均聚类进行加速。

为了对聚类进行加速，另外一个常用的算法是，聚类是只是随机的选取一定比例的样本点进行迭代篇。因此聚类的中心实际上是由一小部分样本点生成的。这样做的好处是极大的减小了比较次数，达到加速聚类过程的目的。然而这类方法的缺点是极大的牺牲了聚类质量。这类方法通常得到的聚类结果都比传统K平均算法差。

此外，Karypis等人提出了渐增的K平均聚类方式来提高K平均聚类的质量。这个方法与传统方法最大的不同之处在于，当一个样本从一个聚类移动到另一个聚类时，相应的两个聚类中心将马上进行更新。实验表明，这一方法的收敛速度更快。而且可以获得比传统K平均聚类更好的聚类结果。同时，结合自顶向下层次聚类的方法，Karypis等人还提出用二分层次聚类方式加速，可获得几十倍的速度提升。然而这个方法只适用于经过归一化的向量空间，对于普通欧几里德空间并不适用。

1.2.3 对研究现状的评析

大数据技术一经产生就很快得以应用在经济学领域。在价格指数方面，Billion Prices Project at MIT很早就开始实践基于大数据技术构建高频微观价格数据库、编制高频价格指数的方法，并用于支持相关宏观经济学和国际经济学的研究。世界上许多国家也开始将大数据技术用于价格指数的编制，比如改进价格采集方案、加入线上数据以改进数据来源、改进价格指数的计算和调整方案等。国内经济学界和统计学界也有很多对于价格指数的讨论，其中包括基于扫描数据或者网页爬虫数据编制价格指数的理论框架、在大数据技术下的价格指数编制和改进的理论框架等。

大数据技术也为经济学研究提供了完整的工具链。网页数据抓取和分布式系统提供了高效的数据采集方案；数据库理论可以帮助研究者有效地管理海量微观数据；聚类和分类算法的不断改进为细分数据类别提供了可能；高性能计算工具为应对传统统计和计量计算工具不能解决的海量计算问题提供了帮助。

但是，我们也看到现有研究的不足：

(1) 尽管90%以上的分类商品都是可以在互联网上获取的，由于技术水平的限制，大部分研究项目没有完整地覆盖CPI分类的商品。

(2) 大部分项目没有对于商品做更细致的分类挖掘，因此对细分市场的特征了解不足。

(3) 对于线上和线下价格水平和变化特征的异同，特别是针对细分市场的线上零售平台的价格数据与线下价格数据的水平和变化特征了解不足。

为了改进这些不足，我们将会基于更广泛的数据来源，构建更大、更详细的线上微观价格数据库，并基于高效的商品聚类算法细分商品，对细分小类的价格数据做更深入的特征挖掘，在保证可靠性的前提下，探索更高效的高频价格指数编制、改进和发布方案。

1.3 交叉学科项目涉及的学科及交叉的必要性

1.3.1 项目涉及的学科

价格指数是宏观经济学和经济统计学的传统研究领域之一，也一直是经济学界和统计学界关注的重点。随着技术水平的进步，价格指数的理论和编制实践也在不断进步。经济学和统计学理论与实践为大数据下价格指数的编制更是提供了理论基础和实证基础。

随着互联网技术的不断进步和大数据技术的不断发展，价格指数的编制有了更多新的方法和工具。网页爬虫技术、分布式系统的进步，为研发分布式网页数据抓取程序采集海量微观价格数据提供了可能。数据库理论和实践的进步，为管理海量数据提供了方法和工具。数据挖掘理论、自然语言处理、机器学习，为聚类海量商品、深入挖掘细分市场的价格特征提供了理论基础。高性能计算技术的进步，为处理传统计算工具无法处理的海量计算问题提供了帮助。项目设计的具体学科如下：

（1）经济学

价格指数是宏观经济学的基本指标之一，是用来衡量一个经济体的整体价格（通货膨胀）水平的重要指标。其中，最常用的价格指数是消费者物价指数（CPI）。然而，由于消费者物价指数是固定篮子的拉式指数，因此会高估通货膨胀。同时，由于成本和技术限制，各国统计局均选择代表商品而不是全部商品编制消费者物价指数。在实践中，价格指数的编制和统计学理论和实践也息息相关：修正拉式指数对通胀的高估、代表商品的选择、权重的确定、代表商品的抽样方法以及下个指数误差的测定和调整，离不开经济学实践和统计抽样理论与实践。质量变化调整也大量应用经济学理论和计量经济学方法，比如Hedonic质量变化调整法借助产品的特征对价格变化影响的回归方程，对价格数据进行适当的调整， 然后利用调整后的数据编制价格指数。

（2）计算机科学

随着数据收集技术的进步，数据呈现出越来越强的可拓展性，高维性及非结构性，这些是传统数据分析领域难以处理的部分，在抓取、储存、计算海量数据时，我们使用了分布式系统及非关系型数据库，能够在保证数据完整性的同时提供可靠性、高存取性能、高可用性与可扩展性的数据处理方案。我们将基于自然语言处理相关理论和技术，对商品的名称、描述、评论等文本数据进行前期处理；通过改进的K-means聚类算法，我们将对海量的商品进行自动地聚类，细分商品分类，并基于细分分类支持经济学研究。

1.3.2 交叉的必要性

价格指数理论来源于宏观经济学。在宏观经济学中，价格指数是衡量经济体整体价格水平的重要指标，其中最常用的指标是消费者物价指数；价格数据指数的编制、计算必须建立在对价格水平的合理描述的基础上。价格指数编制实践离不开统计学的支持。经济统计学为价格数据的采集提供了方案：在统计实践中，代表商品的选择、采样需要统计抽样技术，商

品分类的确定、商品权重的确定需要国民统计学的理论和实践。网页数据抓取技术为海量价格数据的采集提供工具；数据库理论和实践为海量价格数据的管理提供了方案和工具；数据挖掘理论和技术为聚类海量商品、细分商品分类提供了方案；高性能计算工具为解决传统统计和计量计算工具无法解决的大数据问题提供了可能；分布式系统为突破单个硬件瓶颈，连接、管理、维护大量机器提供了方法。因此，本项目需要结合经济学、统计学和计算机科学的诸多领域的理论和实践，为研究线上价格数据提供更好的工具和方法，并在大数据技术改进传统方法，以更适合大数据背景下的经济学和统计学研究。

2. 研究目标，研究内容，学科交叉点，拟解决的关键科学问题，拟采取的研究方案及可行性分析

2.1 研究目标

基于中国官方CPI编制标准和海量数据的采集、管理、清洗、挖掘技术，参考Billion Prices Project at MIT在高频线上价格数据方面的研究成果，探究中国高频线上价格指数的编制方法。研究目标可以概括为以下三个方面：

1.构建中国线上每日价格数据库，并改进官方CPI编制方法编制中国线上每天价格指数。

2.基于改进k-means聚类算法，聚类海量商品，为挖掘更详细的商品价格水平和变化特征提供支持。

3.基于更广泛的价格数据来源，对比中国线上线下价格水平和变化特征，并探究造成差异的可能原因。

2.2 研究内容

2.2.1 编制中国线上每日价格指数

(1) 数据来源选择

我们选择电商而不是大型超市线上商品作为主要数据来源原因如下。首先，中国B2C电商平台能够覆盖官方CPI篮子中50%的分类；其次，我们能够利用电商获得商品的销量数据及其他详细信息；再次，我们能够通过关键词列表，非常方便的从搜索页面获得符合官方

CPI分类的价格、销量等信息。在选择具体的电商凭条时，主要参考的是各大电商平台在中国B2C网络零售市场（包括开放平台式与自营销售式，不含品牌电商）所占的市场份额，根据中国电子商务研究中心（100EC.CN）发布的《2016年(上)中国网络零售市场数据监测报告》　　，国内各大电子商务平台中，天猫排名第一，占53.2%份额；京东名列第二，占据24.8%份额；唯品会位于第三，占3.8%份额。因此CPP选择的线上数据来源为天猫和京东两大平台。

诚然，电商平台也存在相应的劣势。很多促销活动比如618，双十一，双十二等会对价格造成巨大的影响；商品的运费也不在考虑范围内，这些可能会导致结果的偏差。

(2) 数据采集、管理、清洗、规整

在收到网页数据采集系统所爬取的数据后，数据库系统将把数据解析为较为统一的格式并在可接受的时间内按数据库设计范式完成数据的存储。

数据的清洗与规整将在数据解析与存入后进行，其主要任务是采取合理的方案处理获取到的数据中的空值与异常值（如空值修改为NULL、异常值发回爬虫系统尝试重新获取）。对于不能自动处理的记录可以输出到日志系统进行人工处理。

(3) 基于中国官方CPI编制方法的线上价格指数的编制

i. 线上数据的采集

针对目前国内电商平台上商品种类及数量非常大的问题，我们在采集的过程中，需要按照符合官方CPI的8个大类263个基本分类来选择商品，我们利用关键词进行搜索，结合网络爬虫技术，抓取了天猫，京东等电商平台全部的商品价格数据。

ii. 数据的处理

由于资金限制，现在使用的服务器可能会导致一部分数据缺失，对于缺失值采取当天弥补或者用前一天数据弥补的方式进行；由于商家对部分规格商品的调整，可能会造成异常值，因此我们处理过程中剔除掉了价格上升超过500%或价格下降超过90%的商品。

iii. 价格指数的计算

为了计算出每天价格指数，我们参考商品的销量数据来计算权重，计算每一个商品的价格变化，之后按照官方给出的大类的权重计算价格指数。

第一步要计算小类j中商品i的每天价格指数：

$$R_{t-1}^{ij} - \frac{P_t^{ij}}{P_{t-1}^{ij}}$$

第二步计算小类j的价格指数：

$$R_{t,t-1}^{k,l} = \frac{\sum_i S_i^{t,t} \times R_{i,t-1}^{k,l}}{\sum_i S_i^{t,t}}$$

第三步利用Laplace's formula计算未加权重的中类k和大类l的指数：

$$R_{t,t-1}^{k} = \frac{\sum R_{i,t-1}^{l}}{n_k}$$

$$R_{t,t-1}^{l} = \frac{\sum R_{i,t-1}^{k}}{n_l}$$

第四步为加权重计算价格指数：

$$R_{t,t-1} = \sum w^{l} \times R_{i,t-1}^{l}$$

最后计算基期价格指数：

$$I_T = \prod_{0}^{T} R_{i,t-1}$$

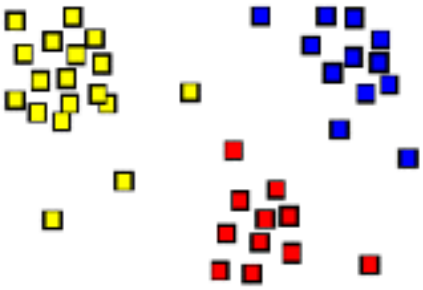
2.2.2 基于改进k-means算法的海量商品聚类

尽管基于我们的方案，我们很容易把对应关键词转换成小类，从而把商品分到CPI小类下。但是，即使是CPI小类下的商品范围也非常广泛，不完全是替代商品，比如“蛋制品”、“固体饮料”、“修理服务”、“家庭手工工具”等等，而许多价格数据来源的分类常常不太可靠或者不详细，这对我们进一步挖掘微观市场的特征造成了阻碍。此外，电商平台本身的搜索引擎会搜索出许多不相关的商品，这些不相关的信息也会被我们的数据抓取程序抓取，比如“酱油”关键词下常常出现酱油瓶等商品，这些商品需要从原始数据中过滤出来。海量商品聚类为解决这些问题提供了可能。

目前数据挖掘技术的主要任务可分为预测任务与描述任务两类：预测任务是根据某些属性预测特定属性的值；描述任务则是探索性地发现数据中的潜在模式，如关联分析、发展趋势、聚类分析、异常检测等。其中聚类分析是指将观察值分割为不同的簇，同一簇内的观察值与属于其他簇的观察值相比尽量相似的一种分析方法。

图1 典型的聚类分析结果

K-means算法是聚类分析中最为广泛使用且行之有效的一种无监督学习算法，它将观察



值分为k个不同的簇，一般采用k-dimension树作为算法数据结构，其基本算法思想如下：

1. 对于d维数据，选取k个聚类中心点；
2. 对于每个数据，将其归到离该数据最近的聚类中心点所属的类；
3. 完成分类后，根据每一个类中的数据值确定新的聚类中心点；
4. 重复步骤（2）-（3），直到每次聚类中心点的变化小于设定阈值。

对于一个规模为n的数据集而言，k-means算法的时间复杂度为*O(ndk)*。

K-means算法能简单有效地在能够接受的时间复杂度内实现聚类分析，但同时也存在以下缺陷：

1. 需要根据先验知识人为指定目标聚类个数n的值，然而在实际应用中k值很难确定，多数时候事先难以知道将观测值聚为多少类是较为合适的；
2. 需要事先指定初始聚类中心，且聚类结果可能因为初始聚类中心的不同而大不相同，不能产生稳定的结果。

项目将就传统的k-means算法的缺陷提出改进的k-means算法，并应用于海量商品聚类分析，从而发现不同种类商品各自的价格特征，进而支持后续经济学研究。

2.2.3 比较中国线上、线下价格水平及变化特征

虽然B2C线上零售平台、线上超市与线下超市和其他线下交易的商业模式有很多共同点，但是由于线上交易和线下交易固有的一些不同特点，比如线上交易存在运输成本、搜索成本、信息不对称带来的风险酬金等，线上价格水平和价格变化可能会有不同的特征。因此，我们需要收集不同来源的数据，并基于聚类算法结果，对齐每个细分小类，深入挖掘不同来源的价格水平和变化特征。

(1) 数据

为了进行线上和线下数据的对比，我们需要采集多渠道零售商作为数据来源。为了和

Billion Prices Project at MIT的研究结果进行对比，除了B2C线上零售平台之外，我们还需要线上超市价格数据及其对应的线下价格数据，并进行对齐比较。对这部分数据要求为：零售商为多渠道销售（即线上app和线下超市同时销售），所占市场份额在中国排名前列。根据2016年8月凯度消费者指数发布的中国零售商市场份额报告，在排名前十的零售商，线上线下同时销售的为高鑫零售旗下的大润发（线上为飞牛网），华润万家和家乐福。因此我们选择了大润发，E万家和家乐福作为线上超市价格的主要数据来源。同时为了获取这些零售商在线下销售的价格，我们选择了条码扫描类手机应用“我查查”来采集零售商线下销售的数据，“我查查”覆盖了全国32个省市492个城市站点的数据，零售商和商品种类相对较齐全。

数据采集方法和第一部分类似，即使用项目编制的针对每个数据来源对应CPI商品分类的关键词作为输入，抓取每个数据来源的关键词搜索页面的商品数据，并分别储存到数据库中。

数据对齐分为两步：第一步是基于官方CPI的分类标准对其关键词对齐，把不同数据来源的商品汇总到对应分类下；第二步是基于第二部分的聚类结果，对商品进行更细致的分类，基于生成的分类把不同数据来源的数据对齐；然后，根据文本信息标准化价格到单位价格，比如，各种规格的某品牌酱油的价格统一成500ml的价格。

数据预处理的主要工作是过滤每条商品信息的无关信息，只保留ID、时间、价格等主要信息，并进行重复值、异常值处理。对每个数据来源的每个细分小类，计算价格平均值以代表该细分小类的价格水平，并用于第一部分相同的方法计算价格变化以代表该细分小类的价格变化。

(2) 价格水平特征描述

对每个数据来源对应细分小类，以下描述性统计：

- 计算价格水平差异的百分比；
- 计算线下价格水平高于和低于线上价格水平的百分比；
- 对线上线下价格水平是否相等做假设性检验；
- 对没有通过假设性检验的，检验是线上价格偏高还是线下价格偏低。

(3) 价格变化特征描述

对每个数据来源对应细分小类，以下描述性统计：

- 线上线下平均价格变化频率；
- 变化频率是否相等的假设性检验；
- 线上线下绝对变化范围；
- 绝对变化范围是否相等的假设性检验；

- 线上线下价格的是否联动的假设性检验。

(4) 讨论造成差异的可能原因

造成线上线下价格特征差异的原因有很多。比如运输成本、搜索成本、信息不对称带来的风险酬金等。我们主要通过查询超市配送的价格政策和快递公司的定价政策，研究运输成本对价格水平和价格变化的影响；量化搜索成本，并通过计量方法检验其对价格特征的影响；通过量化电商平台商品销售的风险，探究风险酬金对线上线下价格水平和波动的可能影响。

2.3 学科交叉点

价格指数理论来源于宏观经济学。统计学为价格指数理论的编制和计算实践提供了支持。在实践中，代表商品的选择、采样需要统计抽样技术，商品分类的确定、商品权重的确定都需要国民统计学的理论和实践。在这个过程中，网页数据抓取为海量价格数据的采集提供工具；数据库理论和实践为海量价格数据的管理提供了方案和工具，保证了数据的完整性、可靠性、存取性能、可用性与可扩展性。

通过以上步骤采集得到的海量数据在数据挖掘理论和技术的支持下，可以提取出一定的数据模式，这为聚类海量商品、细分商品分类提供了方案，解决了以往技术中出现的可扩展性、高维性和非结构化数据方面的难题，提高了对数据噪音的容忍性；数据挖掘技术中的自然语言处理使我们得以对商品的文本数据进行分析；聚类分析中的K-means算法能够更有效地实现聚类分析，但传统的K-means算法存在k值难以确定、结果不稳定等缺陷，项目就这些缺陷改进后的K-means算法，将可以更好地被应用于海量商品聚类分析，对经济学研究提供支持。

2.4 拟解决的关键科学问题

2.4.1基于中国官方CPI编制方法设计适合大数据的中国线上每天价格指数编制方案

(1) 价格调查频率问题

中国官方在进行样本采集的过程中，对大部分商品采取每月2-3次价格取样，对少数价格变化频繁的5天进行一次价格取样。因此价格指数存在滞后问题。

(2) 高效率数据采集系统的构建

(i) 大部分的网站为了应对过多的网页数据抓取系统在自己网站上抓取数据而导致影响正常用户的使用,都采取了一定的反制措施，需要解决的主要问题如下:

- 访问过于频繁会导致验证码机制的验证，或者是针对IP的封锁。
- 网页采取大量的动态内容导致Web数据难以爬取。

(ii) 在单台服务器上搭建网页数据抓取系统势必造成效率的低下，必须采用多台服务器构建分布式采集系统，需要解决的关键问题如下:

- 节点之间的通信问题，即节点之间如何知道相互的状态并加以调用。
- 操作的原子性问题，即防止一个节点在对一个请求处理的过程中另一个节点也进行处理导致一个请求有多个节点重复处理。
- 节点异常的处理，当节点集群有部分出现异常时，能够在无数据损失的情况下将剩余任务分配给剩余节点，并能迅速通知维护人员加以修复。

(3) 海量数据管理系统的构建

海量数据管理系统的构建主要需要解决以下问题:

(i) 扩展性：数据库系统需要足够的可扩展性以确保日后如果出现数据库系统过载或储存空间不足等情况时能够简单、迅速地扩展数据库系统;

(ii) 合理冗余：对于分布式的数据管理系统而言，系统内每一个节点都应该存储有一定比例的其他节点的数据的冗余备份，以保证在部分节点出现宕机、损坏、不可达等问题时整个数据库系统仍然能够保持高可用性与数据完整性;

(iii) 便于管理：数据库系统应提供方便快捷的数据备份等管理功能。

2.4.2 探索高准确度、高效率的海量商品聚类算法

鉴于传统K平均算法无法满足海量商品的聚类需要，本项目提出一种增强K平均聚类算法。算法改变了传统K平均的聚类过程。使聚类过程更为简洁高效。具体描述如下。

在传统K平均聚类算法中，算法的迭代优化过程可以建模为对如下目标函数进行优化。

$$\min \sum_{q(x_i)=r}^{} \|C_r - x_i\|^2$$

(公式1)

其中$q(x_i)$得到的是x_i所属的聚类，C_r 表示第r个聚类的中心。对上述目标函数进行一系列的等价变换，我们可以得到

$$\text{Max}\sum_{r=1}^{k}\frac{D_r^t D_r}{n_r}$$

（公式2）

其中$D_r=\sum_{x_i\in S_r}x_i$，S_r表示属于第r类的数据集合，x_i表示第i个数据对应的向量表示；D_r表示属于第r个类的所有数据对应向量的和；n_r表示属于第r个类的向量的数目。与公式1所示目标函数对比，新的目标函数把K平均聚类算法变成一个纯优化问题。D_r是属于第r个类内的所有数据对应向量的复合向量。优化这个新目标函数的过程是探索将一个样本从一个聚类移动到另一个聚类能否使得目标函数增大的过程。至此，新的K平均聚类算法设计为：

步骤一：给定***n***个待聚类数据，目标聚类数为k个

步骤二：初始化聚类中心。初始化k个聚类中心采取的策略a或b：

a. 完全随机策略。给定待聚类的***n***条数据，每条数据分配一个随机的从*1*到k的类标签。

b. 传统策略。随机中心采用传统K平均初始化方式，先随机从数据中选取k个作为初始的聚类中心，为每个数据在k个中心中寻找最近的聚类中心，该中心所代表的类的标签将赋予该数据。

步骤三：将一个数据的类标签换做另一个类标签使得目标函数（公式2）的优化函数值增大。采取c或d策略来更换一个数据的类标签：

c. 快速策略。随机选取一个数据，尝试改变当前数据所属的类到另一个类，如果能使目标函数的值变大，则更新这个数据所属的类到另一个类。重复此尝试，直到***n***条数据都被尝试更改类标签一次且仅一次。

d. 最优策略。随机选取一个数据，尝试改变当前数据的类标签为另一个类，如此尝试*k-1*次，找到能使目标函数获得最大增益且增益为正。如此，更新这个数据所属的类到获得最大增益的那个类的类标签。重复此尝试，直到***n***条数据都被逐一尝试更改类标签一次且仅一次。

步骤四：重复步骤三直到公式一中的目标函数无法再获得更高的函数值，或者达到指定的迭代次数。

从上述算法描述可以看到，与传统K平均相比，算法存在三处显著不同。首先，在聚类初始化时，算法并不要求为每个样本计算离它最近的类中心，并将该样本点赋到该中心所代表

的类。其次，在每次聚类迭代过程中，算法不一定要为每个样本寻找离它最近的类中心。由于不需要找最近邻，可以使得聚类时间耗费显著降低。另外，在目标函数（公式2）的直接驱动下，算法并没有显示的生成类中心，一旦发现将一个样本从当前所属类移动到另外一个类可以使得目标函数值增加，算法就将这个样本由当前类移动到另一个类。初步研究结果表明这样的渐增迭代方式显示出更高的效率。

2.5 拟采取的研究方案

2.5.1 基于中国官方CPI编制方法设计中国线上每日价格指数编制方案

(1) 价格调查频率

官方价格指数的滞后性非常强，为了解决这一点，我们通过高频（每天七点到九点抓取商品价格）抓取价格数据，计算每天的商品价格指数。

(2) 构建基于Python及其扩展库构建分布式数据采集系统

（i）基于Python开发网页数据抓取系统可以有效解决数据采集系统遇到的问题.

利用Python的requests第三方库中的方法可以伪造HTTP请求头部报文，达到伪造IP的效果。同时构建代理IP池，每次发送HTTP请求时从代理池中随机选取IP进行伪造，并控制访问频率来防止对IP的封锁。

通过对需爬取网页的人工分析，可以找出部分API接口。直接对API接口发出HTTP请求报文，可以获取直接获取动态内容的json格式数据，有效防止直接爬取动态内容而导致花费大量时间。

（ii）基于Redis的分布式任务调度

Redis是一个开源、支持网络、基于内存、键值对存储数据库。可以用作数据库、缓存和消息队列。Redis作为一个全内存操作的数据库，Redis的读写性能十分优良。此外，Redis支持的数据结构种类丰富，能够处理数据爬取系统中绝大部分的情况。并且Redis内部的操作都保证了原子性，可以有效解决分布式采集系统在这方面的问题。

我们利用Redis做消息中间件构建消息队列。消息中间件可以利用高效可靠的的消息传递机制进行平台无关的数据交流，并基于数据通信进行分布式系统的集成。Redis作为消息中间件有两种模式。第一种是PUB/SUB机制，即发布-订阅模式。这种模式生产者和消费者是1-M的关系，即一条消息会被多个消费者消费。不过这种模式不符合该数据采集系统的要求。第二种是PUSH/POP机制，利用Redis的列表数据结构构建消息队列。生产者push消息到消息队

列，消费者从消息队列中pop消息，并设定超时时间。数据采集系统可以采取这种机制，Master将所需要抓取的URL通过Redis push到消息队列，由Slave从该队列获得URL并在Slave上对获得的URL进行网页数据抓取。

(3) 构建基于MongoDB构建亿级数据的海量数据管理和挖掘系统

在数据库系统方面，不同的NoSQL数据库针对不同应用情景，一般采用列储存、文档储存、K-V键值对储存或图储存等储存模型，因此不同的数据库管理系统都有不同的擅长领域。综合考虑下列原因，我们将选取MongoDB来进行数据库系统构建：

(i) 扩展性：MongoDB为分布式架构提供了自动分片功能与副本集功能，提高分布式系统的可用性与可扩展性。

(ii) 合理冗余：分布式环境下的MongoDB会在分片上维护一定比例的冗余数据，以保证在一定数量的子分片发生宕机等问题后，数据库系统仍然是高可用的、数据完整的。同时，针对主节点发生宕机的情况下，MongoDB存在一套完备的选举机制，整个数据库系统将从剩余分片中推选出最为合理的一个节点作为临时的主节点，保证整个系统的正常运行。

(iii) 便于管理：分布式环境下的MongoDB对外表现是一致的，数据库使用者不用关心数据库是运行在单节点模式下还是集群模式下，有利于减少系统的复杂性。

2.5.2 构建基于改进k-means算法的海量商品聚类系统

首先，商品将通过商品的一系列属性进行描述。比如商品名、商品价格、用途、材质等进行描述。各个属性将根据具体情况实现量化描述，最终将商品表示成一个高维的向量。

根据实际商品类别，尝试采用不同的聚类参数K调用增强K平均聚类算法，进行聚类。由于数据量巨大，增强的K平均聚类算法将和层次聚类相结合，采用自顶向下的层次聚类。具体过程如下。

步骤一：将输入数据看做一个类，放入一个优先队列**Q**；

步骤二：按一定标准，从优先队列**Q**中取出一个类，调用增强K平均聚为两类；并且将聚类结果放入优先队列**Q**中；

步骤三：重复步骤二，直到得到K个类。

通过分析，可以发现上述聚类过程的时间复杂度为$O(n*d*\log(k))$。由于，K通常是一个很大的值，比起传统K平均（复杂度为$O(n*d*k)$），其复杂度可以显著下降。

目前，项目组已经获得了一些初步的实验结果。在一个有1百万，128维特征的数据集上，

增强K平均显示出比传统K平均，及其最近的改进方案k-means++较为显著的聚类性能提升。

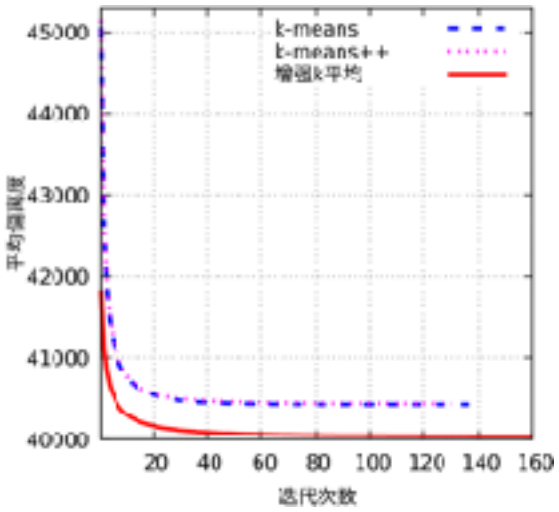


图 2 增强K平均与K平均，k-means++的聚类结果对比。

聚类结果熵值	*k*=5	*k*=10	*k*=15	*k*=20
K平均	0.539	0.443	0.402	0.387
k-means++	0.550	0.441	0.403	0.389
增强K平均	**0.506**	**0.419**	**0.380**	**0.353**

表1 增强K平均与K平均和k-means++聚类质量对比

另外，项目还将该聚类算法在文本聚类上进行尝试。表1显示了在15个标准测试集[1]上的平均熵值（熵值越低聚类结果越好）。从表1可以看出，增强K平均的聚类结果显著高于其他两种普遍使用的K平均或其变种。

[1]http://glaros.dtc.umn.edu/gkhome/fetch/sw/cluto/datasets.tar.gz

2.6 可行性分析

2.6.1 分布式数据采集系统和数据管理系统的初步构建

硬件方面，分布式系统中的主节点由项目组实验室所拥有的性能良好的服务器充当，其他成员节点由腾讯云、阿里云等云服务器提供商所提供的廉价云服务器充当，藉此可以使用较低费用完成分布式系统的构建。

软件方面，我们利用Python和Redis构建了网页数据抓取系统和分布式任务调度系统。并在十几台云服务器上部署了网页数据抓取系统，实现了邮件提醒等模块。进行了部分的压力和稳定性测试，表现良好。 我们采用基于MongoDB的分布式系统的原生的Python语言接口数据库系统与网页数据采集系统、计算架构、实时系统实现对接，初步测试结果良好。

2.6.2 价格指数编制的初步成果

根据获得的商品价格以及购买量，我们根据CPI的计算方法，得到了初步的价格指数。CPI整体稳定在1.00至1.02之间。粘性价格理论是指短期中价格的调整慢于物品市场供求关系的变化。CPI基本符合粘性价格所产生的效应，短时间内的价格指数变动较小。

其中，天猫在6月18日至6月20日，举行了618天猫狂欢节活动。对手机数码家电类，家装家居类，汽车类，美妆类，生鲜类进的部分商品进行30%-90%不等的降价。价格指数产生一定波动，在6月18日明显下降，并在6月20日回升。

我们把价格指数绘制成图表，图中反映了这些信息。

图4 天猫价格指数

范围	个数
小于1.000	4
1.000-1.001	9
1.001-1.002	12
1.002-1.003	16
1.003-1.004	8
1.004-1.005	9
1.005-1.006	5
1.006-1.007	7
1.007-1.008	5
大于1.008	4

表2 价格指数频数统计

其中，八个大类的指数，呈上升趋势。其中，第一大类食品，第二大类烟酒及用品，第三大类衣着，第五大类医疗保健和个人用品的价格曲线和总的价格曲线基本符合。第四大类家庭设备用品及维修服务与总价格曲线差异较大。第六大类交通和通信，第七大类娱乐教育用品及服务，第八大类居住的价格曲线变动较小。这是由于tmall平台上这两个大类的商品较少导致。八个大类的价格指数绘图如下：

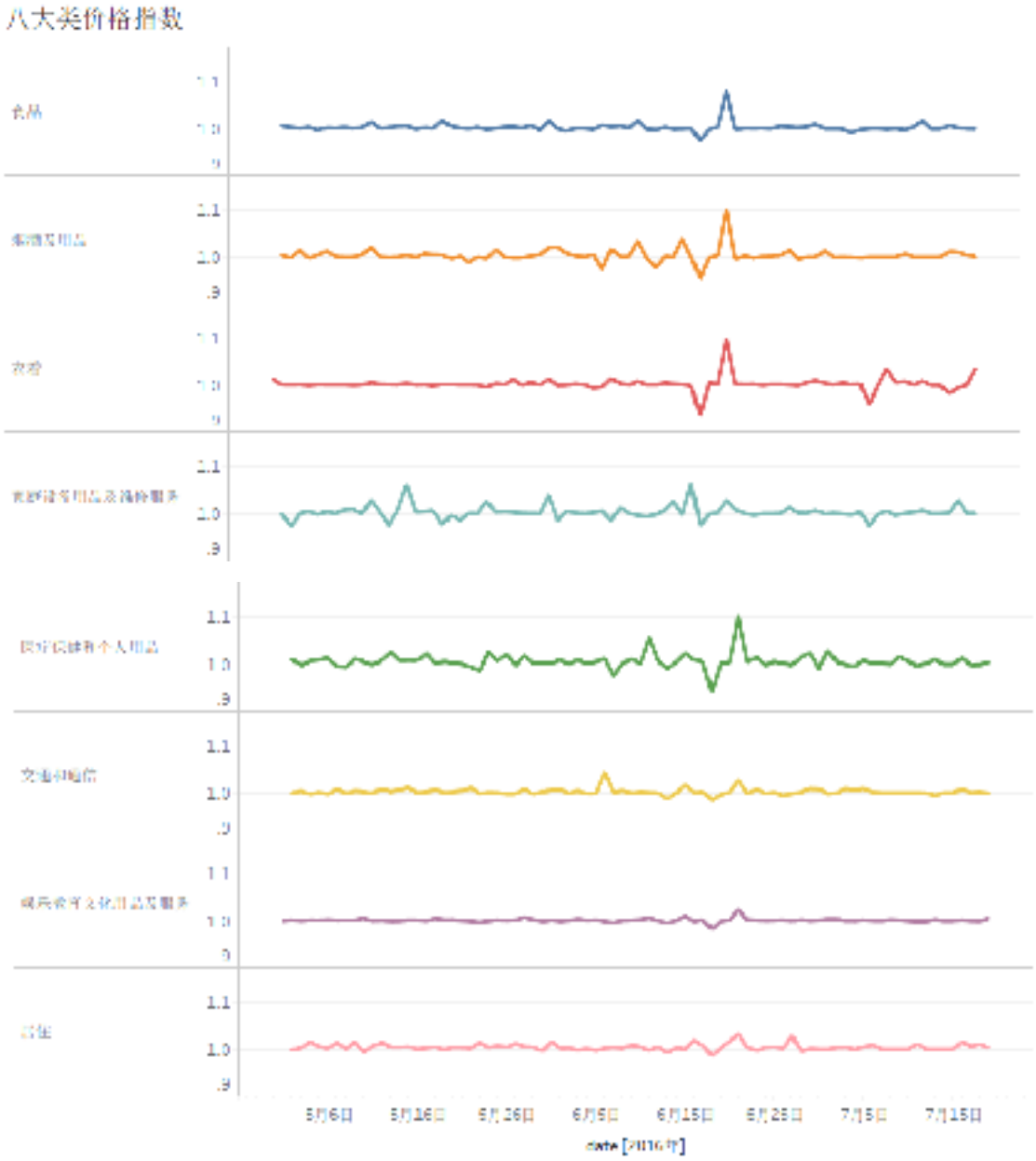


图4 八大类价格指数

针对于第四大类的价格指数与实际价格指数波动差异较大的现象，对第四大类中两个小类，家具和家庭设备的价格指数进行绘图，并且有待后续研究。

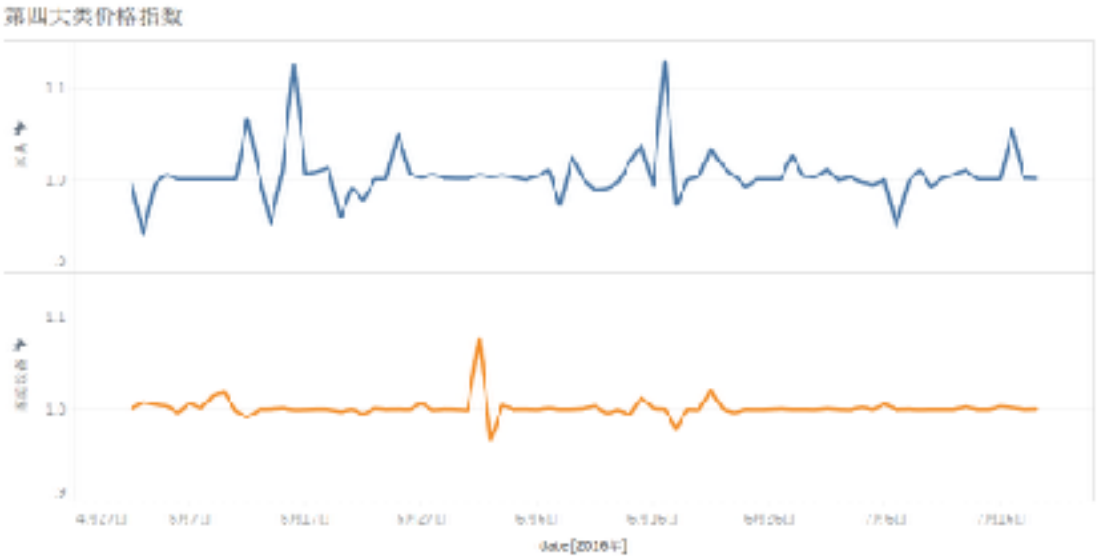


图5 第四大类中两个小类的价格指数

3．本项目的创新之处

在研究内容上，我们将基于各种线上平台，构建更完整地覆盖CPI分类商品、商品信息更详细的高频微观价格数据库，改进了中国官方CPI编制方法以适应大数据背景下价格指数编制的特点，并基于海量价格数据构建了高频线上价格指数。我们将会用海量商品聚类细分商品，以进一步挖掘更细分的微观市场特征。基于商品聚类结果，我们将更深入地讨论纯线上数据的价格水平和变化特点，并与现有各种来源的线上、线下数据对比。

在研究方法上，我们把分布式Web采集技术和分布式数据库应用到数据库的构建中，以构建数量更大、信息更全面、处理更高效的价格数据库。为了应对海量数据的处理和计算问题，我们使用了多种高性能计算工具代替传统的统计或者计量计算工具。我们将把项目负责人最新的研究成果——改进的k-means算法应用到海量商品聚类中，以提高商品聚类的可靠性。

本项目在研究内容和研究方法上为中国的经济学界应用大数据技术支持经济学研究提供了很好的范例，为现代统计学和计算机科学深入变革中国经济学界的数据采集、管理、挖掘和分析方案提供样本。

4.3 已具备的实验条件

实验室拥有六台大规模图形工作站，专业图像（视频）采集装备和 Internet 接入设备

等，有一台24核112G内存的IBM服务器。为在线数据收集提供了必要的硬件保障。同时实验室还配30 多台高性能计算机，还有其他许多实验材料、图书资料和丰富的电子文献资源等，为科研工作的顺利展开提供了基础条件。

4.4 尚缺少的实验条件和拟解决的途径

由于项目研究内容的实际需要，实验室尚缺支持少大容量内存的计算机。需要独立存储系统，用来存储网上价格数据。这将通过获批的项目经费购买，购买设备单项不超过3万。

参考文献：

[1] Y. Zhao and G. Karypis: Empirical and theoretical comparisons of selected criterion functions for document clustering, Machine Learning, vol. 55, pp. 311--331, 2004.

[2] H. Jegou, M. Douze, and C. Schmid: Product quantization for nearest neighbor search, Trans. PAMI, vol. 33, pp. 117--128, Jan. 2011.

[3] M. Muja and D. G. Lowe: Scalable nearest neighbor algorithms for high dimensional data, Pattern Analysis and Machine Intelligence, IEEE Trans. on, vol. 36, pp. 2227--2240, 2014.

[4] A. Babenko and V. Lempitsky: Additive quantization for extreme vector compression, in CVPR, pp. 931--938, 2014.

[5] M. Ester, H. peter Kriegel, J. Sander, and X. Xu: A density-based algorithm for discovering clusters in large spatial databases with noise, in In Proceedings of Knowledge Discovery and Data Mining, pp. 226--231, 1996.

[6] D. Comaniciu and P. Mee: Mean shift: A robust approach toward feature space analysis, Trans. PAMI, vol. 24, pp. 603--619, May 2002.

[7] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg: Top 10 algorithms in data mining, Knowledge and Information System, vol. 14, pp. 1--37, Dec. 2007.

[8] D. Arthur and S. Vassilvitskii: K-means++: The advantages of careful seeding, in Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1027--1035, 2007.

[9] B. Bahmani, B. Moseley, A. Vattani, R. Kumar, and S. Vassilvitskii: Scalable k-means++, In Proceedings of the VLDB Endowment, vol. 5, no. 7, pp. 622--633, 2012.

[10] T. Kanungo, D. M. Mount, N. S. Neanyahu, C. D. Piatko, R. Silverman, and A. Y.Wu: An

efficient k-means clustering algorithm: Analysis and implementation, Trans. PAMI, vol. 24, pp. 881--892, Jul. 2002.

[11] D. Pelleg and A. Moore: Accelerating exact k-means algorithms with geometric reasoning, in Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 277--281, ACM, 1999.

[12] D. Sculley: Web-scale k-means clustering, in In Proceedings of the 19th international conference on World wide web, pp. 1177--1178, 2010.

[13] A. Goswami, R. Jin, and G. Agrawal: Fast and exact out-of-core k-means clustering, in Proceedings of the Fourth IEEE International Conference on Data Mining, pp. 83--90, 2004.

[14] A. K. Jain and R. C. Dubes, Algorithms for Clustering Data. Prentice-Hall, Inc., 1988.

[15] Y. Zhao and G. Karypis: Hierarchical clustering algorithms for document datasets, Data Mining and Knowledge Discovery, vol. 10, no. 2, pp. 141--168, 2005.

[16]Alvin E. Roth, Axel Ockenfels: Last minute bidding and the rules for ending second price auctions: evidence from ebay and amazon auctions on the internet, American Economic Review, 2002.

[17]Glenn Ellison, Sara Fisher Ellison: Search, obfuscation, and price elasticities on the internet, Econometrica, 2009.

[18]Hyunyoung Choi, Hal Varian: Predicting initial claims for unemployment benefits, Citeseer, 2009.

[19]Lynn Wu, Erik Brynholfsson: The future of prediction: how Google searches foreshadow housing prices and quantities, ICIS 2009 Proceedings, 2009.

[20]Seth M. Freedman, Ginger Zhe Jin: Learning by Doing with Asymmetric Information: evidence from Prosper. com[R], National Bureau of Economic Research, 2011.

[21]Michael R. Baye, John Morgan, Patrick Scholten: Price dispersion in the small and in the large: Evidence from an internet price comparison site[J], The Journal of Industrial Economics, 52(4): 463-496, 2004.

[22]Liran Einav, Jonathan Levin: Economics in the age of big data[J], Science, 346(6210): 1243089, 2014.

[23]Linnet Taylor, Ralph Schroeder, Eric Meyer: Emerging practices and perspectives on Big Data analysis in economics: Bigger and better or more of the same?[J], Big Data & Society, 1(2): 2053951714536877, 2014.

[24]Hyunyoung Choi, Hal Varian: Predicting the present with Google Trends[J]. Economic Record, 88(s1): 2-9, 2012.

[25]Sharad Goel, et al: Predicting consumer behavior with Web search[J]. Proceedings of the National academy of sciences, 107(41): 17486-17490, 2010.

[26]Yan Carrière-Swallow, Felipe Labbé: Nowcasting with Google Trends in an emerging market[J]. Journal of Forecasting, 32(4): 289-298, 2013.

[27]Nick McLaren,Rachana Shanbhogue: Using internet search data as economic indicators[J]. Bank of England Quarterly Bulletin, 2011

[28]Concha Arola and Enrique Galan: Tracking the future on the web: Construction of leading indicators using internet searches, Technical report, Bank of Spain, 2012.

[29]Hellerstein R, Middeldorp M: Forecasting with internet search data[J]. Liberty Street Economics, Federal Reserve Bank of New York, 2012.

[30]Alberto Cavallo, Roberto Rigobon: The Billion Prices Project: Using Online Prices for Measurement and Research, Journal of Economic Perspectives, Vol 30(2): 151-78, 2016.

[31]Alberto Cavallo: Online vs Official Price Indexes: Measuring Argentina′s Inflation, Journal of Monetary Economics, 60(2), 152-165, 2013.

[32]孙易冰, 赵子东, 刘洪波: 一种基于网络爬虫技术的价格指数计算模型, 统计研究, Vol.31, No.10, 2014.

[33]Brynjolfsson, Erik, Yu Jeffrey Hu, and Michael D. Smith: A Longer Tail?: Estimating the Shape of Amazon's Sales Distribution Curve in 2008, Workshop on Information Systems and Economics (WISE), 2009.

[34]Ellison, Glenn, Sara Fisher Ellison: Search, obfuscation, and price elasticities on the internet, Econometrica, 77.2.427-452, 2009.

[35]Ellison, G., S. F. Ellison: Search, Obfuscation, and Price Elasticities on the Internet, Econometrica, 77(2), 427-452, 2009.

[36]Ellison, G., S. F. Ellison: Tax sensitivity and home state preferences in internet purchasing, American Economic Journal: Economic Policy, 53-71, 2009.

[37]Gorodnichenko, Y., V. Sheremirov, O. Talavera: Price Setting in Online Markets: Does IT Click?, NBER Working Paper Series (20819), 2014.

[38]Brynjolfsson, E., A. A. Dick, M. D. Smith: Search and Product Differentiation at an Internet Shopbot, SSRN Electronic Journal, 2003.

[39]Lnnemann, P., L. Wintr: Price Stickiness in the US and Europe Revisited: Evidence from Internet Prices*, Oxford Bulletin of Economics and Statistics, 73(5), 593-621, 2011.

[40]Gorodnichenko, Y., O. Talavera: Price Setting in Online Markets: Basic Facts, International Comparisons, and Cross-border Integration, NBER Working Paper (20406), 2014.

三、计划进度及预期成果

起止时间	主要计划	预期成果
2017年1 月-2018年 1月	1.收集相关文献，确定研究方向； 2.收集在线商品价格数据	1.完成文献和数据收集; 2 .设计出鲁棒性较高的网络爬虫
2018年1 月-2019年 1月	1. 价格指数初步编制； 2. 线上线下价格差异研究； 3. 商品聚类算法的研究	1. 实现初步价格指数编制； 2. 发现线上线下价格差异及存在原因； 3. 聚类分析海量商品，发现其价格特点
2018年12 月-2019年 12月	1. 改进价格指数编制方法； 2. 完善商品聚类算法	1 .完善价格指数编制方式存在的不足； 2 .应用商品聚类算法，比较线上线下价格变化水平及特征

五、经费预算（万元）

项目名称	关于大规模近似图片检索及链接的关键技术研究	
科目名称	预算	计算依据与说明
1、设备费	6.5	用于购买计算机，独立存储设备，音响设备等。无单个设备在5万元以上
2、材料费	3.5	用于购买耗材如打印纸、硒鼓、大容量内存及硬盘等
3、测试化验加工费	0.0	
4、差旅费	3.5	用于研究团队出差调研、学术交流
5、会议费	0.0	
6、国际合作与交流费	0.0	研究团队参加国际会议支出，包括注册费，和差旅费。一次国际会议0.8-1.2万元
7、出版/文献/信息传播/知识产权事务费	3.0	用于支付出版费，资料费、知识产权公证及专利申请所产生的费用
8、劳务费	10.5	支付参加项目的研究生劳务费。博士700元/月，硕士600元/月，每年工作10个月
9、专家咨询费	2.0	邀请国外专家咨询5次，每次0.4万元
10、其他	0.0	
合 计	29.0	