# 数据采集软件系统

## 数据填报系统

### 网上直报系统

建立一个面向茶产业种植、采摘、加工、流通、消费等各环节上数据采集点、相关管理部门的网上直报数据统计系统，实现各项直报数据进行集中式管理，采用主流的 B/S（浏览器/服务器）访问模式，使用先进的Java 编程开发技术，基于传统的 MVC 三层体系架构对整体功能进行分模块设计。

### 电商数据网络爬虫与抽取系统

网络数据采集是通过读取网页获取数据的过程，网页是一种非结构化数据，目前采集方式如下：使用网络爬虫、智能代理、XML 数据压缩、分词系统、正则表达式、XPath 索引、XQuery 查询等互联网搜索引擎常用技术来开发集网络数据抓取、格式匹配、特征值提取、数据压缩存储、数据分析及数据挖掘等功能于一体的网络数据采集系统。

### 数据交换系统

数据交换系统处理与第三方平台/数据库的数据交换，主要通过数据库直连同步，数据库日志解析同步与接口方式实现数据同步。数据交换系统需同时支持结构化数据与半结构化数据的处理。

### 文件系统

文件系统主要针对非结构化数据的处理。包括对电子表格文件，如excel 文件，xml 文件或 csv 文件的导入；以及图片，视频，文本等数据文件的同步。文件系统需提供 GUI 界面，方便系统操作员使用。

## 茶叶全产业链数据库建设

数据库建设是基于数据存储运算系统支撑，将采集系统采集并治理后的数据进行数据建库工作，包括基本信息及价格、成本收益、消费、贸易、管理五类数据，共同建立茶叶全产业链数据库。

## 数据挖掘运算系统

依据经济学理论与统计学理论，结合茶产业实际，开发相关软件，构建决策知识库、模型库、方法库，构建预测分析模型。将茶产业的生产、流通、价格、贸易、成本收益等方面的信息进行数据挖掘分析，并建立智能高效的数据运算系统。数据挖掘运算软件系统基于茶叶全产业链数据库建设，包括数据仓库建设，ETL 任务管理与数据计算模型建设。

## 数据仓库建设

数据仓库建设包括将茶叶全产业链数据库的全量数据按照主题维度与数据分层的要求整合到数据仓库中去，在这个过程中需要完成数据的抽取、清洗、转换、整合、传输、加载等操作。故数据仓库建设包括 ETL。

### 数据计算模型建设

数据计算模型建设包含建成数据分析集成算法库，涵盖常用的机器学习、数据挖掘方法：分类，回归，聚类，推荐、神经网络等。具体如下：

分类算法：（朴素）贝叶斯、决策树、逻辑回归、KNN、SVM、神经网络、随机森林。

回归算法：线性回归、曲线回归、二元逻辑回归、多元逻辑回归。

聚类算法：划分法（K-MEANS 算法、K-MEDOIDS 算法、CLARANS 算法）、层次法（BIRCH 算法、CURE 算法、CHAMELEON 算法等）、密度算法（DBSCAN 算法、OPTICS 算法、DENCLUE 算法等）、图论聚类法、网络算法（STING 算法、CLIQUE 算法、WAVE-CLUSTER 算法）、模型算法（高斯混合模型GMM、SOM）、模糊聚类（模糊 C 均值 FCM）。

推荐算法：基于内容的（tf-idf、cosine）、基于协同的（皮尔森相关系数、余弦相似性、调整余弦相似性）、关联规则、基于效用的、基于知识的。

神经网络算法：ART 网络、LVQ 网络、Kohonen 网络、Hopfield 网络。建成数据分析集成算法库后，通过对底层的分布式算法封装，提供拖

拉拽的可视化操作环境，让机器学习的创建过程像搭积木一样简单。缩短与数据的距离，真正实现了数据的触手可及。同时也提供了命令行工具， 方便您将算法嵌入到自己的工程中。

1、茶叶价格预测模型

开发茶叶价格分析预测模型，收集企业及茶农销售价格、批发市场价格、茶叶进出口价格等历史价格数据，结合气象，区域产量等影响茶叶价格的指标形成多维弱相关的数据集合。利用回归算法开发茶叶价格预测模型系统，形成较为准确的茶叶价格预测模型。

通过对市场茶叶价格的预测与预警，为政府及企业的智能决策服务提供数据依据。

2、气象产量分析预测模型

分析历年茶叶产量与气象关系，通过对气象大数据、农作物生长作物历和各农作物分地区历史产量数据挖掘农业生产与气象各因素的潜在关联关系与特征。通过对大数据的多模泛化关联分析得到茶叶在不同生长期期间单产与各气象要素之间的详细关系。应用 CAMES 系统气象学产量分析方法，开展全国茶叶主产区产量与气象关系的分析。

通过茶叶生长期气象要素的变化监测，对区域茶叶产量进行预测和预警，及时发布预警结果，实现茶叶气象产量的实时分析与预测。

3、投入产出分析模型

通过对企业及茶农经营数据与销售数据，批发市场茶叶区域流向数据、青茶价格指数等数据的整合，采用聚类算法形成对企业或茶农投入产出情 况的分类，总结形成投入产出比优秀企业或茶农的相似性，及其与其他企 业的差异。通过投入产出分析模型的建立，为政府部门或管理经营者对智 能调控茶叶生产投入提供依据。

4、舆情分析模型

在国家互联网信息监管制度允许条件下，对互联网上涉及茶叶全产业链的政策信息、消费趋势、敏感事件等信息进行全面抓取，分析舆情可能导致的后果，开发舆情分析产品与服务，包括茶叶舆情预警服务系统、茶叶互联网舆情年度报告，为政府部门、生产经营者提供茶叶产业链社会关注焦点，进而为提前研判茶叶产业形势提供依据。

5、情感分析模型

收集消费者口碑评价反馈数据，消费者行为监测，在线问卷反馈等数据，利用 NLP（自然语言处理）技术对这类带有情感色彩的主观性文本数据进行分析、处理、归纳和推理，得到消费者对某一地域、某一品牌或某一茶叶品种的情感分析。情感分析系统可以将消费者对产品的积极或消极评价以数字的形式直观地展现在政府部门、生产经营者面前，同时形成关键字云图，及时为茶叶产业链企业提供消费者对这一产品的评价分析。

6、智能推荐模型

根据消费者数据，包括人口特征、消费偏好、购买偏好、口碑评价反馈数据、各地消费对品种的接受程度数据，客单价数据、消费周期数据、消费人群数据、以及对消费者行为的动态跟踪数据，形成消费者画像，采用基于项目的协同过滤推荐为消费者推送最感兴趣的产品或茶产业相关文章。为企业提供智慧营销及品牌塑造服务。

## 茶产业大数据共享系统

数据共享交换系统作为茶叶全产业链大数据信息资源交换体系建设的核心内容，由交换桥接、前置交换、交换传输、交换管理等子系统组成， 并可结合信息资源目录服务，实现信息资源的编目、订阅、发布、管理、查询等功能，实现对信息资源的管理。通过交换桥接子系统将部门需要交换的信息交换到前置交换信息库，在交换管理子系统的流程控制下，通过交换传输子系统、前置交换子系统，把需要交换的信息定向传输到接收部门。

## 数据共享接入

茶叶大数据中心的数据共享服务模式大致可以分为物理拷贝模式、在线查询模式、共享交换模式、系统接口模式四种。

## 多功能门户网站

定向采购或开发相关系统软件，用于对大数据结果进行共享与呈现，服务政府与茶叶经营主体。建设一个集指数发布、重要市场信息发布、专业信息咨询的多功能门户网站，网站面对所有公众。通过数据可视化，定期发布不同指标的日度指数、周度指数、月度指数、季度指数、年度指数； 定期对指数数据进行统计分析，制定并发布指数分析报告，为行业从业者解读指数深层含义。在全产业链数据库中设立可公开的标准化数据子库， 配置数据下载功能，供高等院校、科研机构的专家、学者们进行数据的深度挖掘和研究。

### 数据视图门户模块

数据视图门户模块提供与各主流前台后台可视化框架相兼容的嵌入 式模块，这些模块可以方便地在各个需要引用的网站页面嵌入并生成服务。对公共服务部分，服务可包含数据的查询，数据报表的显示及预定义的互 动，及数据分发和共享。对专业管理，服务入口则提供数据的录入，数据追溯甚至数据建模调整及触发数据报表的发布及更新。

## 数据多维可视化

### 生产方面应用

采集和整理历史生产数据与环境数据，进行茶叶生产应用服务。如灾情预警防治，地图上可以直观展示各种灾情类型主要发生的地区以及影响程度。自然灾害预警通过将灾害历史发生区域、自然灾害类型、灾害损失、预防措施、易发区域预警、已发生灾害政府补贴发放等多个信息进行展示， 可以把全部区域的自然灾害预警工作进行直观的呈现。

### 流通方面应用

对茶叶价格及销售数据进行定期抓取、过滤、汇总、分析，并通过平台进行集中展示。可进行价格行情的地区选择、每月茶叶报价比较、种类对比；区域价格排行、平均增减率分析、区域价格对比等。以表格的方式展示茶叶在全国各省份当天的价格信息，可实现按不同类别、不同省份、不同市场查询每日的茶叶价格信息。