淘宝网
Taobao.com

淘宝网语义分析产品、技术介绍

王天舟（空海）

淘宝网-交易线-语义分析

个人介绍

- 花名空海

淘宝网
Taobao.com

淘宝网中的文本

□ 淘宝主站：

- 30亿店铺、宝贝浏览
- 10亿计的在线宝贝数
- 千万量级交易笔数

□文本数据：

•用户评论

•商品标题、详情页

•用户query数据

•SNS、论坛等其他数据

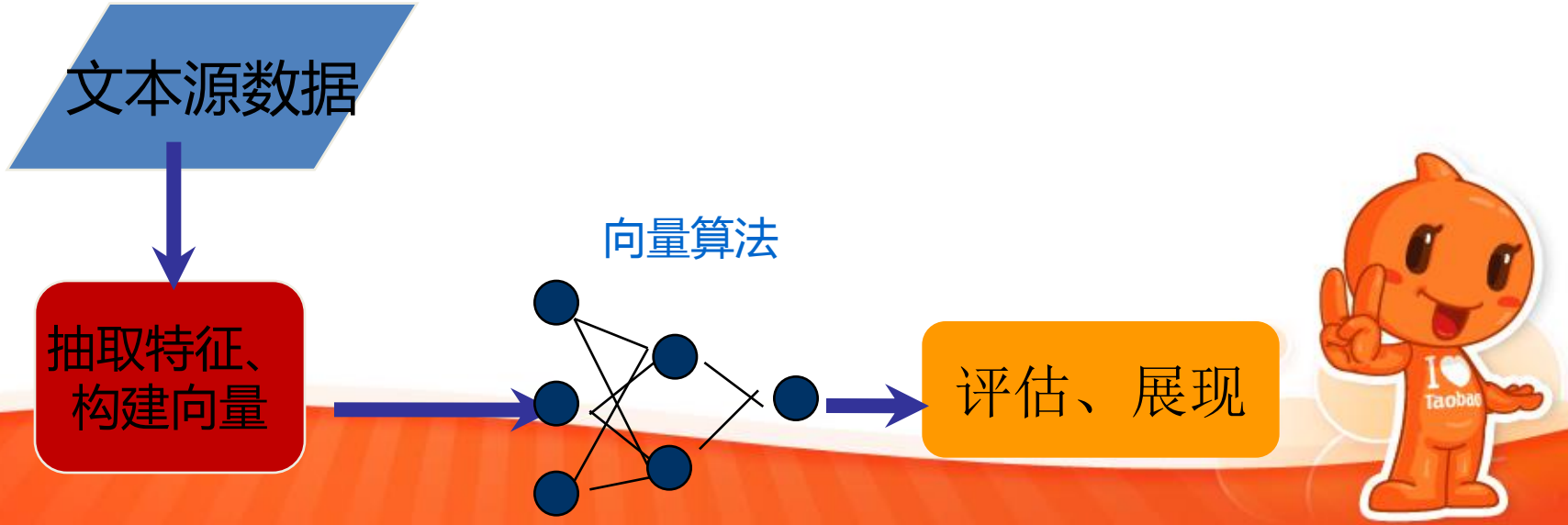
淘宝网
Taobao.com

海量文本数据带来的价值和挑战

有限框架数据的补充、用户UGC信息

如何结构化、高效计算

呈现文本数据挖掘文本价值



语义分析平台架构总览

文本数据

评价 | 商品标题 | 详情页 | ... | SNS/分享数据

特征计算层

分词/新词 | 词之间相关 | 序列标注 | 句法分析 | 实体语义

聚类算法

层次聚类 | 谱系聚类 | KNN | Kmeans

分类算法

SVM/ANN/决策树/贝叶斯

产品

大家印象 | 北极圈 | U站推荐 | UE反馈

淘宝网
Taobao.com

今天的话题

- 相关业务场景: 标签、内容、分类打标
- 相关算法和问题
- 文本技术拓展

淘宝网
Taobao.com

文本标签服务

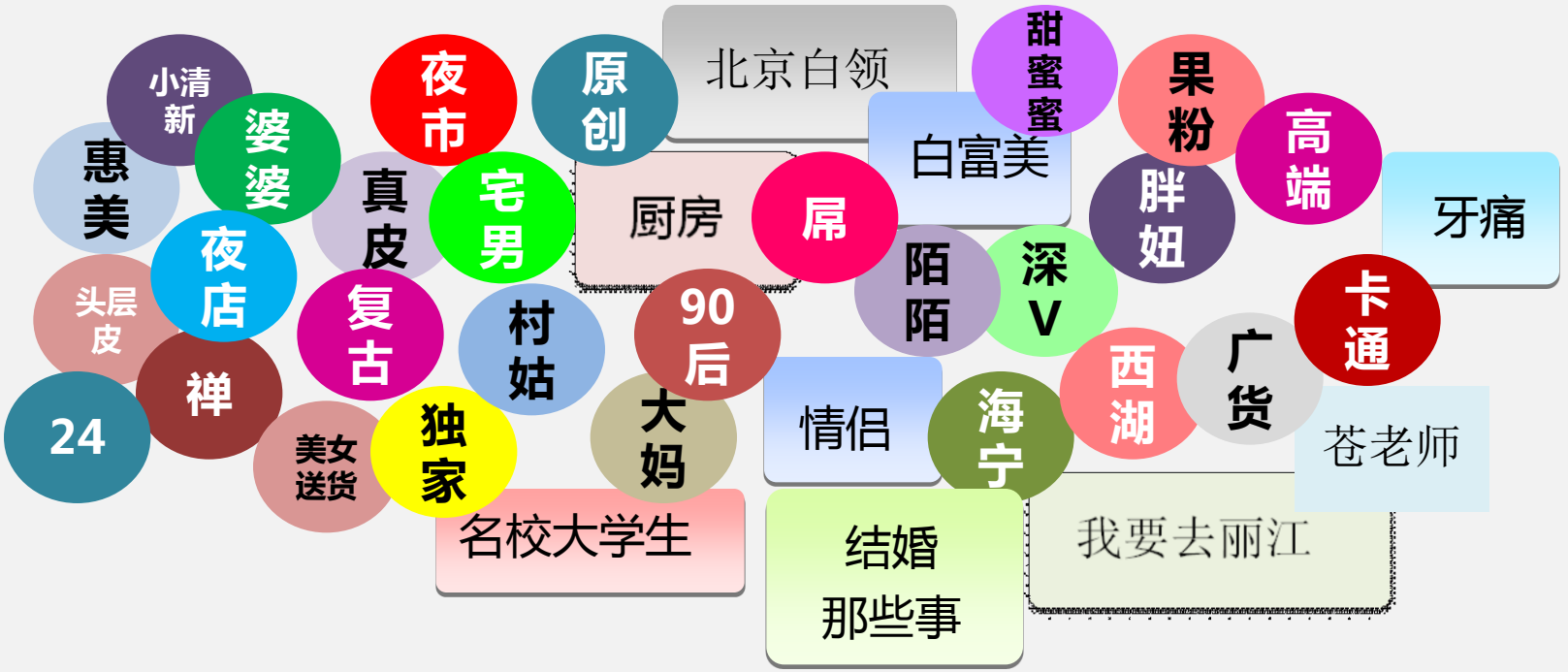
内容相关、相似框架提取

文本分类

淘宝网
Taobao.com

单品标签 “大家印象”

商品导购标签



怎么样的标签的规模是合适的，长尾标签怎么处理？

需要BI提供数据模型，对标签的规模、数量提供界定范围；

根据标签覆盖的商品数来确定需要增加的标签。

店铺导购标签

逛店铺.业务框架:

产生标签

赞同标签

标签偏好

和好友互动

标签重组
找到店铺

达成

消费目的

消费者

语义分析+词库

数据转换文本

自发现

机器挖掘
标签

淘宝网
Taobao.com

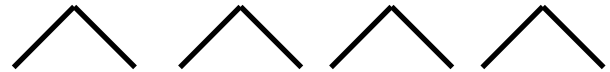
标签当中的问题

标签的来源
标签的属性词义
标签的关系

标签来源、分词问题

- □一元分词
- □二元分词
- □CRF分词

夏季 新款 女装 雪纺 连衣裙



夏季 新款 女装 雪纺 连衣裙
B E B E B E BE BM E

淘宝网
Taobao.com

标签中的短语、新词来源

- 互信息、聚合度、左右熵发现二元
- 前缀树发现长字符串模式

1. Accessor Variety 语义实体的上下文独立性()

AV(ab)=min(|XL|,|XR|)

其中XL ={x|xab为文档中的连续汉字串}

其中XR ={x|abx为文档中的连续汉字串}

|XL|,|XR|分别为集合XL, XR包含的元素个数

AV值越高说明ab上下文独立性越强，越可能成为实体

2.熵

衡量后缀的混乱度

$$H(X)=\sum_{x\in X} p(x)\log\frac{1}{p(x)} \qquad H(X)=x\times\frac{1}{x}\times\log(x)\equiv\log(x)$$

序列标注问题

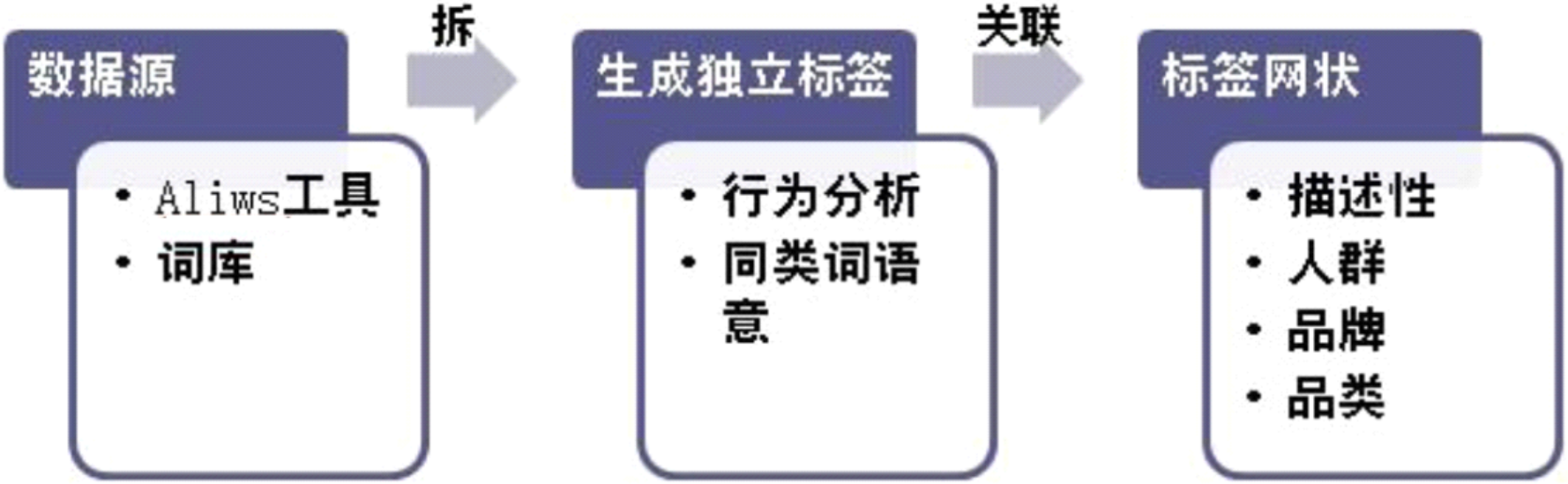
均可以看做概率图模型的不同表现形式

隐马尔可夫模型（Hidden Markov Model，HMM）

最大熵模型（Maximum Entropy Model，MEM）

条件随机场（conditional random fields，CRF）

实体识别的标注问题



异类词（行为分析）

通过商品对应的tag与买家关联；通过tag对应买家重合度算出tag关联度；

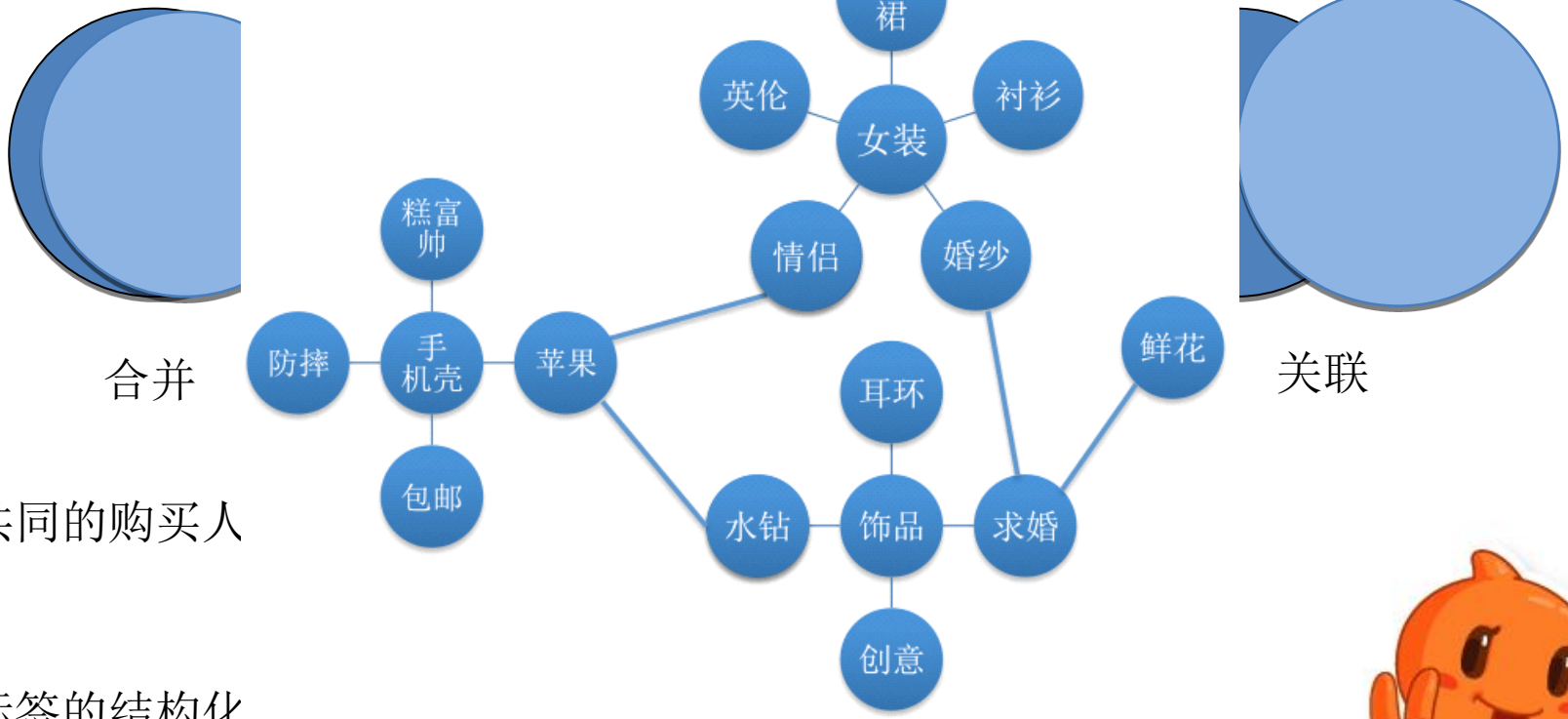
同类词（语意分析）

先词库分析，计算词与词之间的相同出现的概率；

9：佛珠手链是一种流行饰品

标签间的关系

手机套和手机 薄和女士



合并

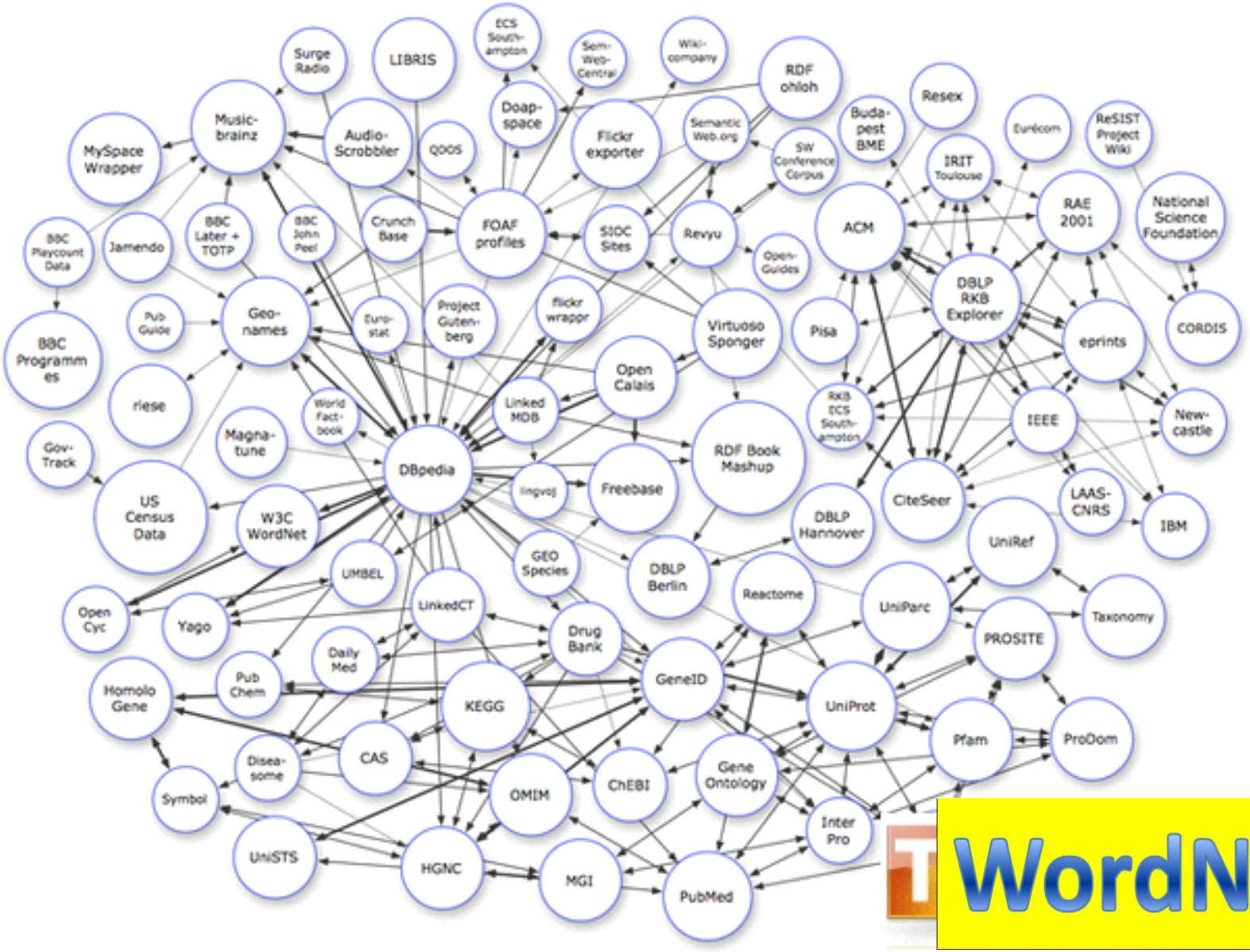
关联

共同的购买人

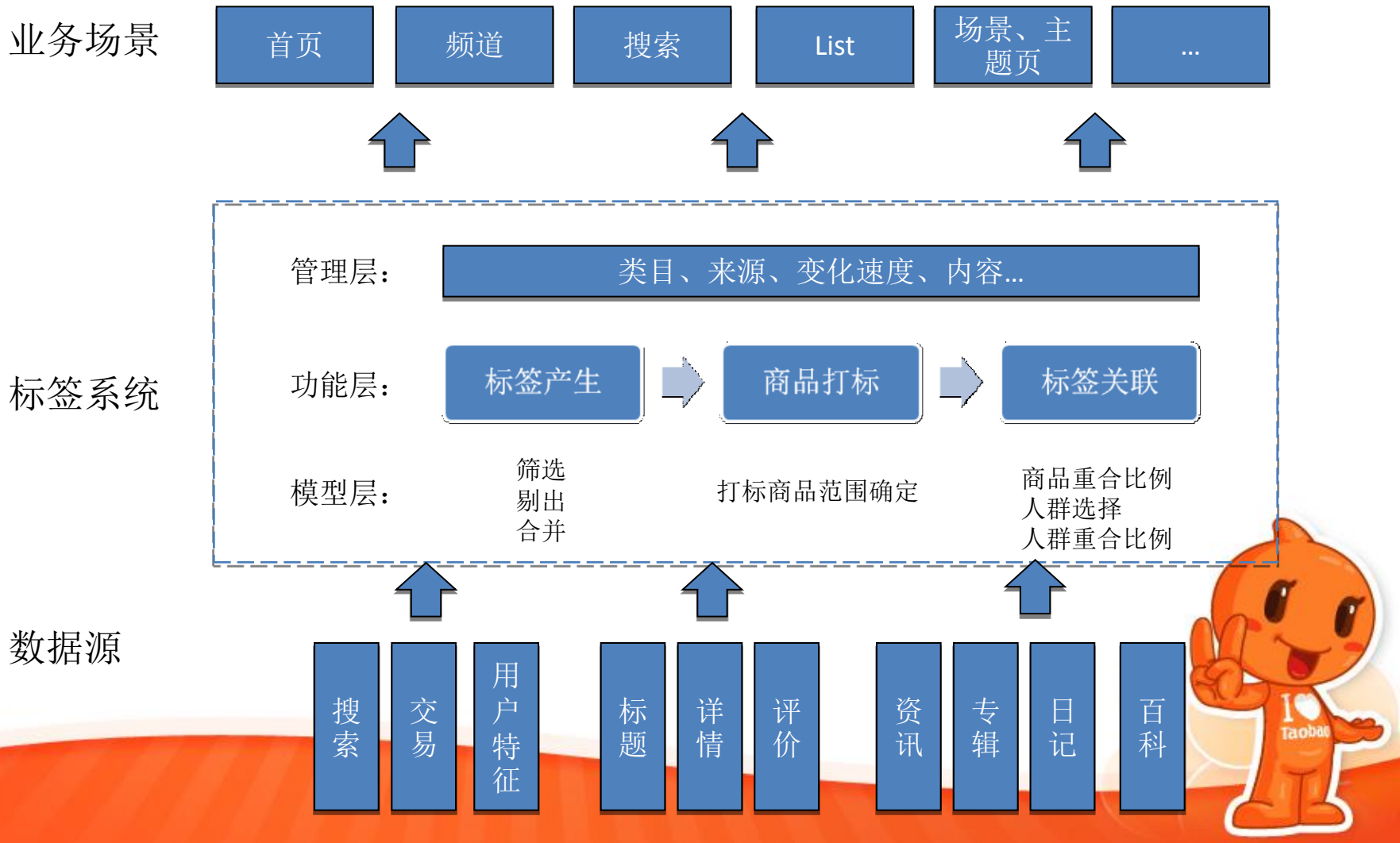
标签的结构化.

类目、来源、变化、内容

标签词语之间关系问题



产品框架


业务场景
首页
频道
搜索
List
场景、主题页
...
标签系统
管理层：
类目、来源、变化速度、内容...
功能层：
标签产生
商品打标
标签关联
模型层：
筛选
剔出
合并
打标商品范围确定
商品重合比例
人群选择
人群重合比例
数据源
搜索
交易
用户特征
标题
详情
评价
资讯
专辑
日记
百科

淘宝网
Taobao.com

其他文本标签问题

- 标签质量判定
- 标签排序、相关性展现
- 标签合并去重

淘宝网
Taobao.com

内容相关、提取问题

U站内容推荐

相似Query查询

特定内容提取

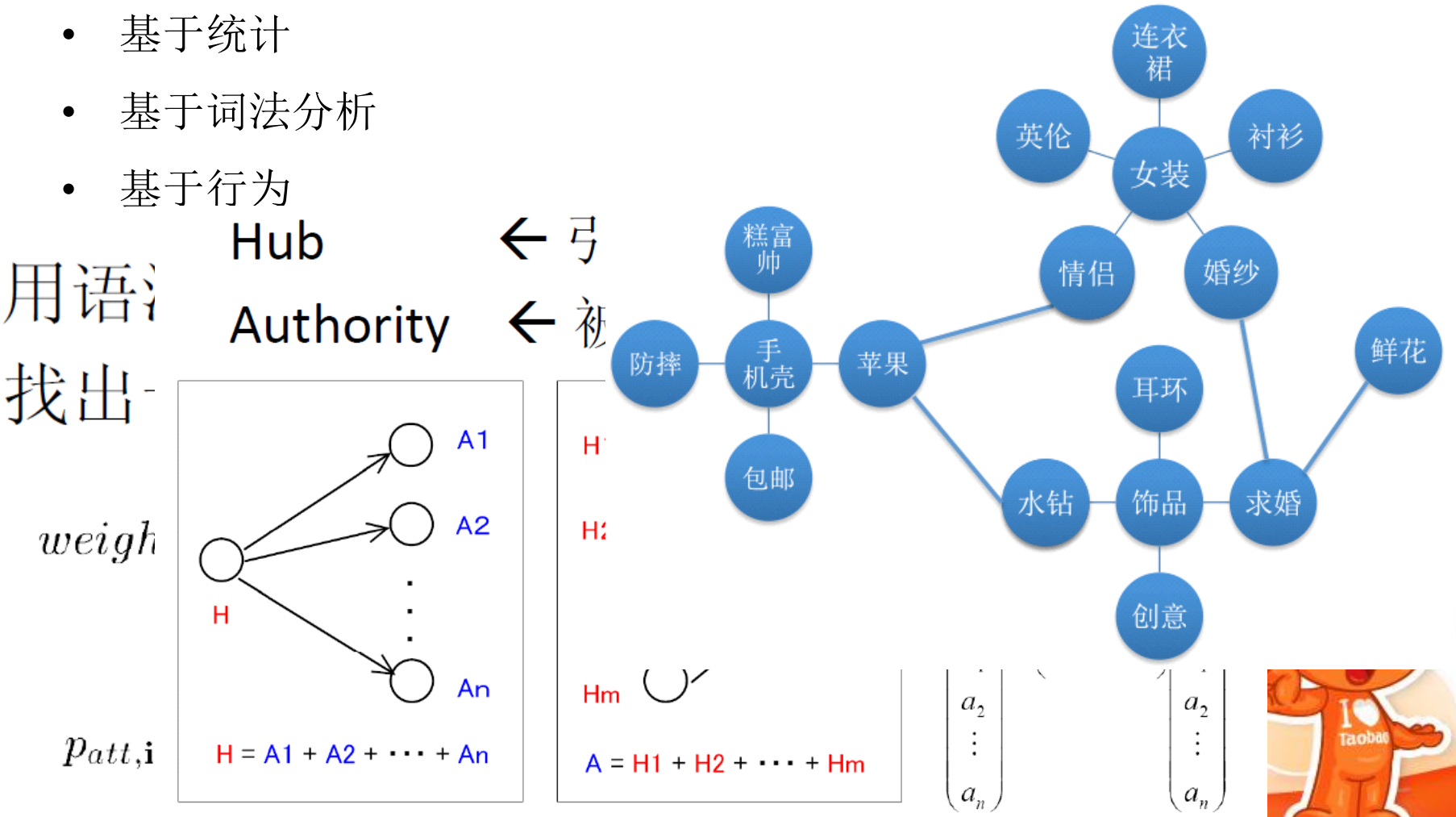
淘宝网
Taobao.com

如何分析相关内容

- 人的行为
- 词之间相关性
- 句子、段落之间相关性

同义词、近义词、词之间关系

- 基于统计
- 基于词法分析
- 基于行为



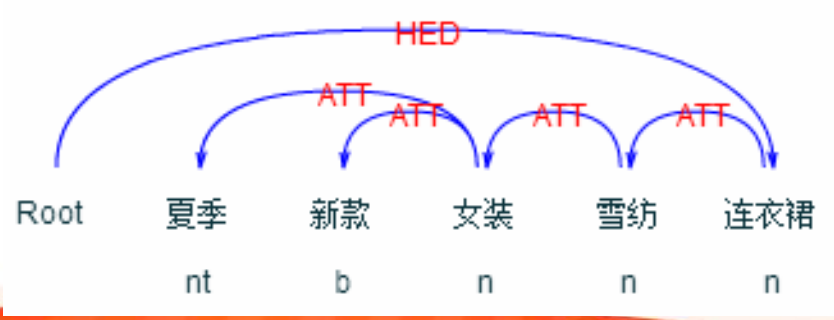
句法分析

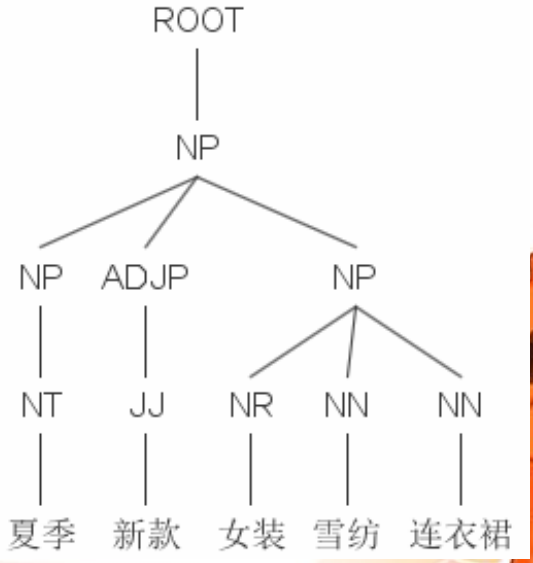
- 短语句法分析

  head-rule, lexicalize, grammar-based un-lexicalize

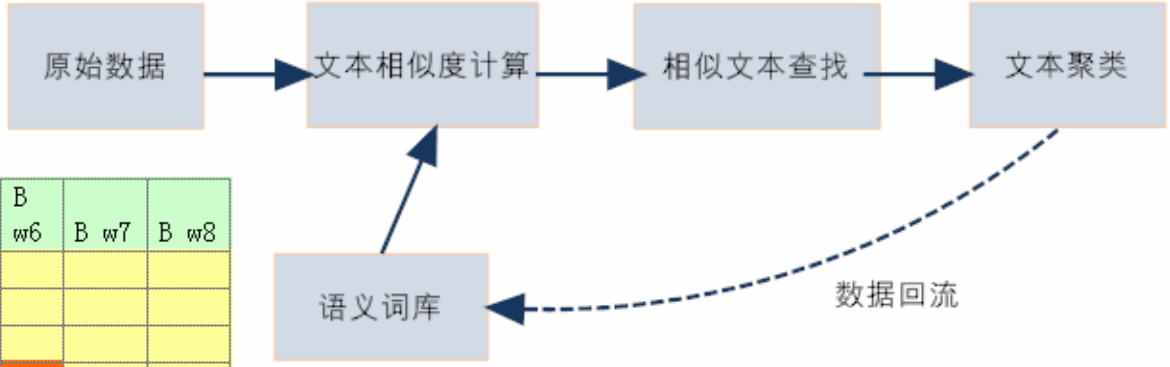
- 依存句法分析

  memory based，classifier-based, feature-verification

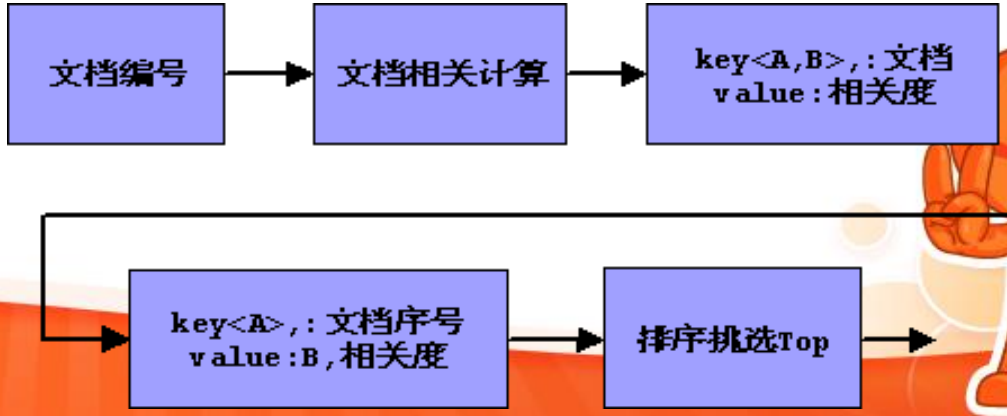




语义推荐



	B w1	B w2	B w3	B w4	B w5	B w6	B w7	B w8
A w1								
A w2								
A w3								
A w4								
A w5								
A w6								
A w7								
A w8								
A w9								



内容提取问题

- 关键词、中心词
- 特定场景地址、礼物

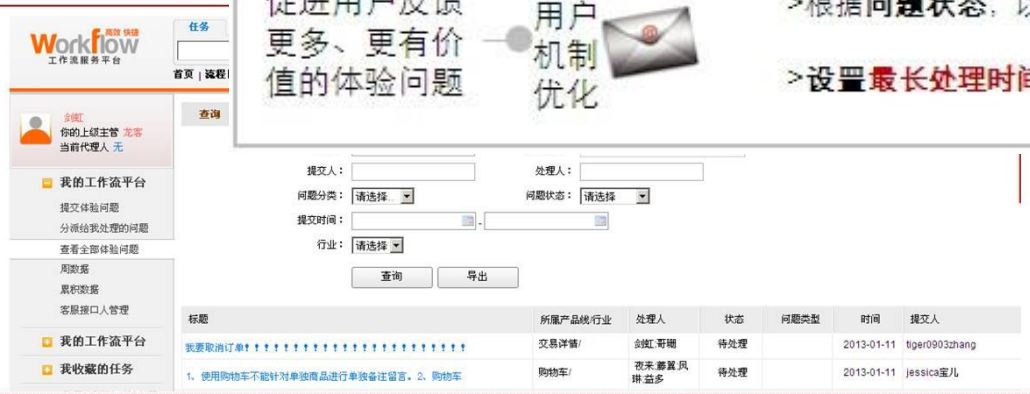
淘宝网
Taobao.com

分类打标问题

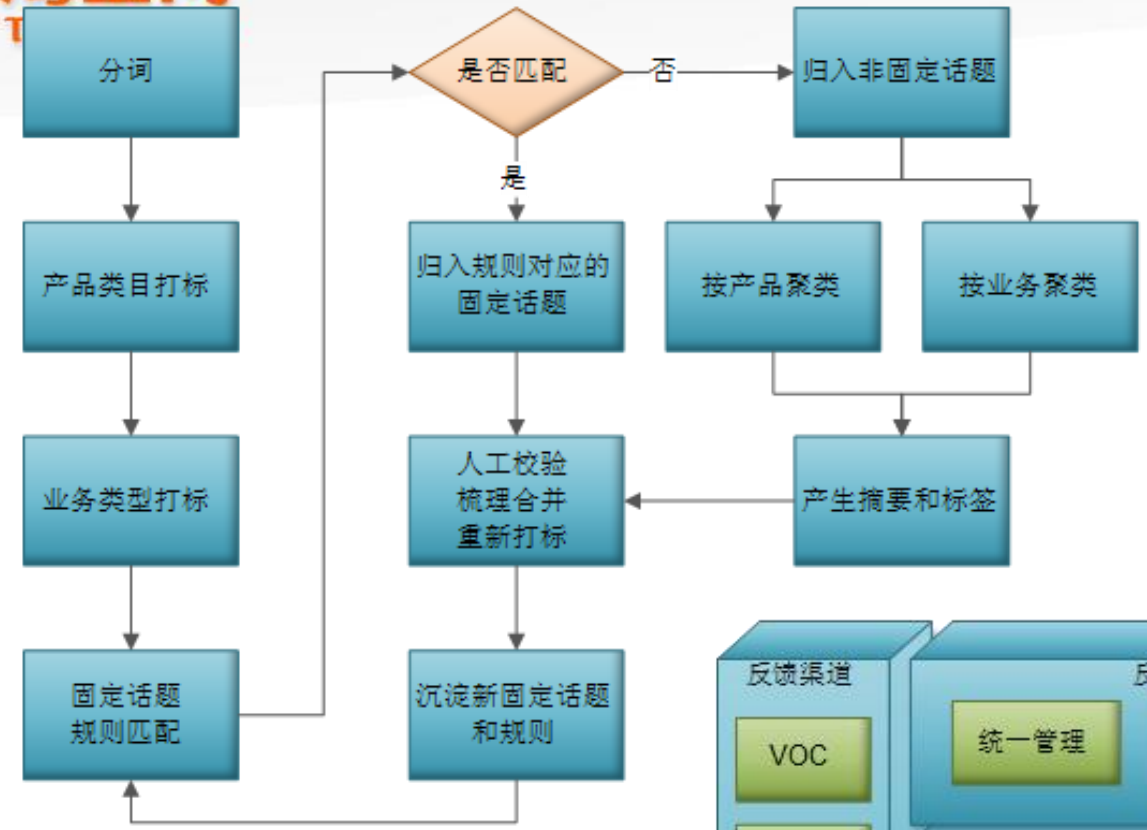
淘宝网
Taobao.com

优化需求点		典型优化点
更多小二通过UE平台及时获取体验问题	更多小二	>宝贝发布等产品反馈**满足运营小二需求**，区分类目展示；
	及时	>UE反馈实现**每小时**同步一次。
有效展示体验问题	问题识别	文本分析介入： >根据**字符有效性**与**字符长度识别无效反馈**； **宝贝体检中心/购物车相似句子分析**；**女装/母婴行业内容识别**。
	问题定位	> **用户反馈问题时所在URL**、**浏览器**、**操作系统**等信息对接。
提高小二在UE工作流平台上的操作效率	可用性提升	>优化**问题展现方式（视觉）**； >同一业务可设置**多位处理人**； >优化**批量处理问题的操作方式（交互）**。
促进用户反馈更多、更有价值的体验问题	回复用户机制优化	>根据**问题状态**，以**格式化方式**引导接口小二系统回复用户； >设置**最长处理时间**，最长两星期系统将**自动回复用户**。

1、UE反馈页面布点的入口

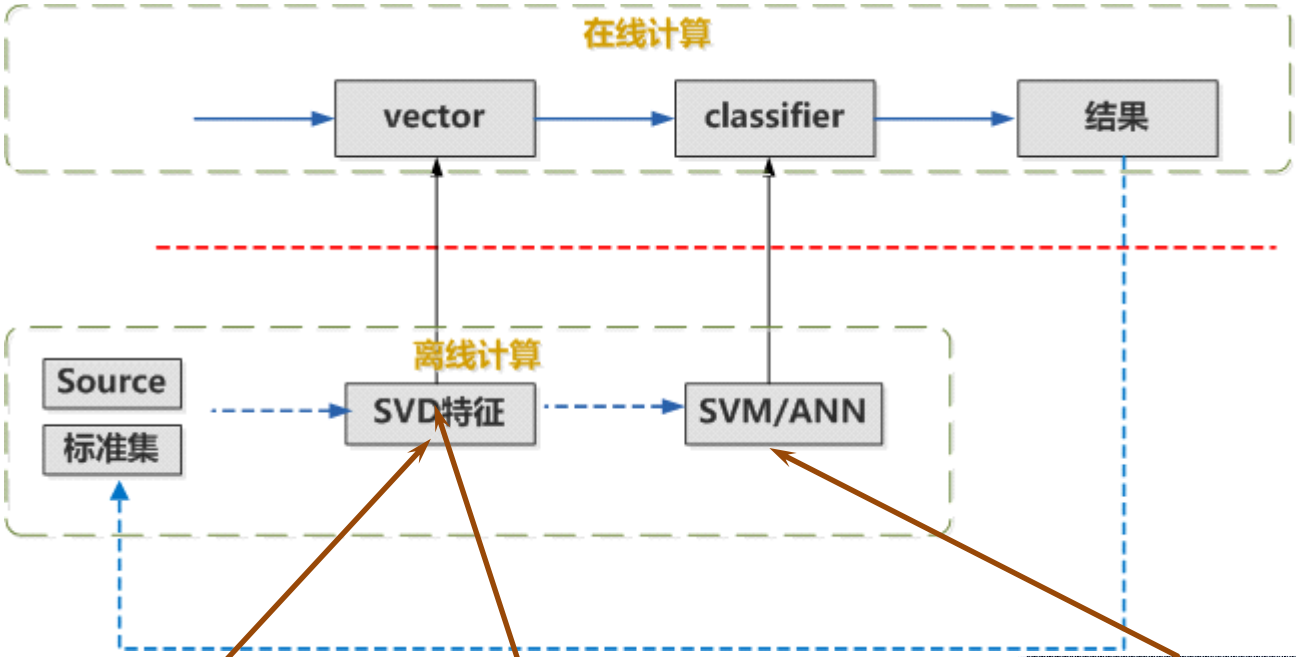


5、问题解决情况邮件给用户


分词
产品类目打标
业务类型打标
固定话题
规则匹配
是否匹配
否
是
归入非固定话题
归入规则对应的
固定话题
按产品聚类
按业务聚类
人工校验
梳理合并
重新打标
产生摘要和标签
沉淀新固定话题
和规则


反馈渠道
VOC
UE
论坛
客服
其他
反馈统一管理平台
统一管理
聚合展现
集中处理
文本分析平台
分词
词性标注
句法分析
语义标注
分类聚类
信息抽取
实体识别
规则引擎
模板匹配
自动摘要
特征抽取
词库管理
固定话题识别
规则管理
非固定话题聚
类规则管理
业务方
产品经理
运营
客满小二
开发人员
其他

分类打标问题

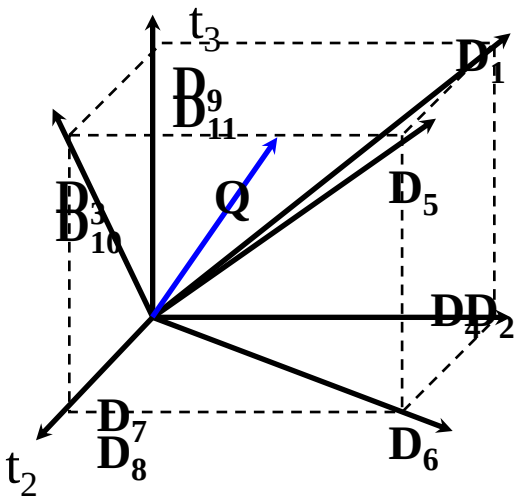


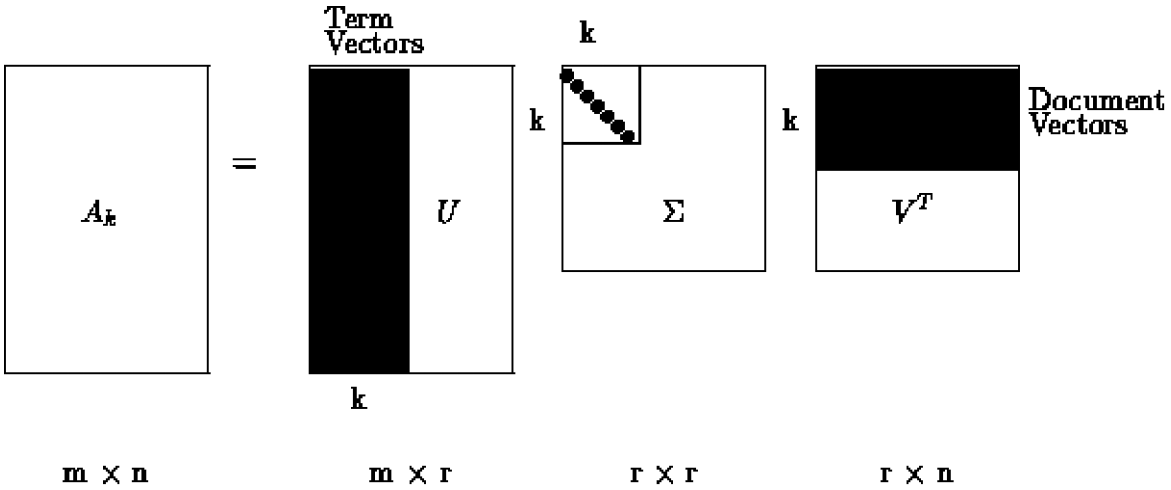
SVD,LDA

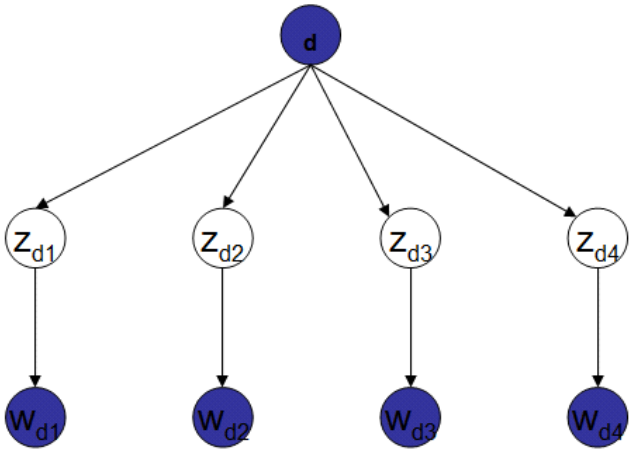
有效特征选取、组合。特征计算

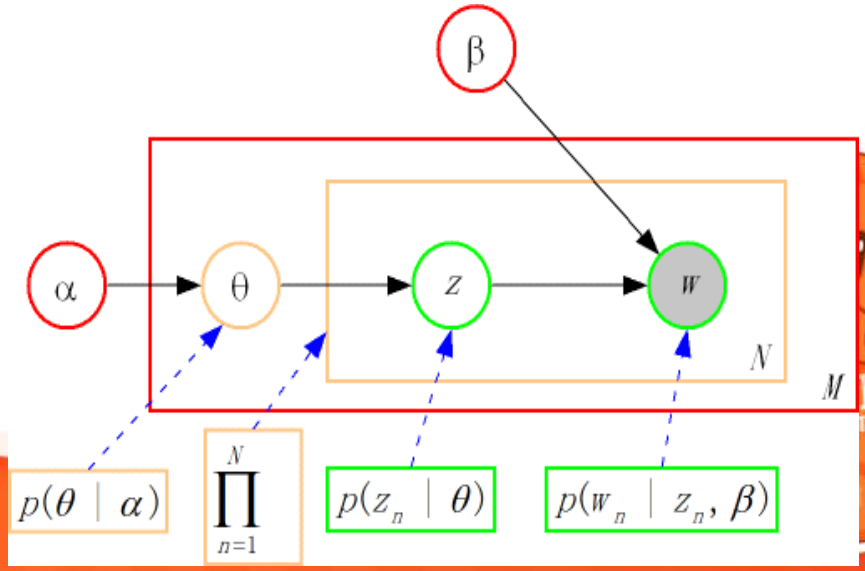
分类器设计。SVM、ANN、贝叶斯、决策树

单词向量化问题

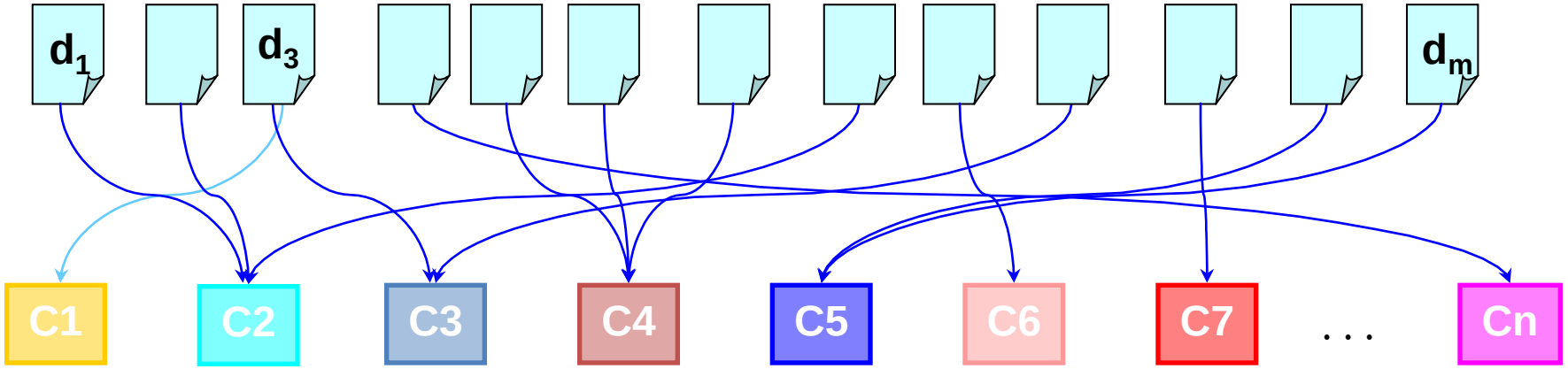


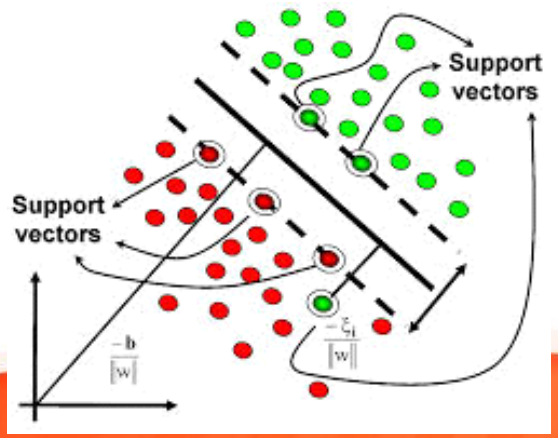


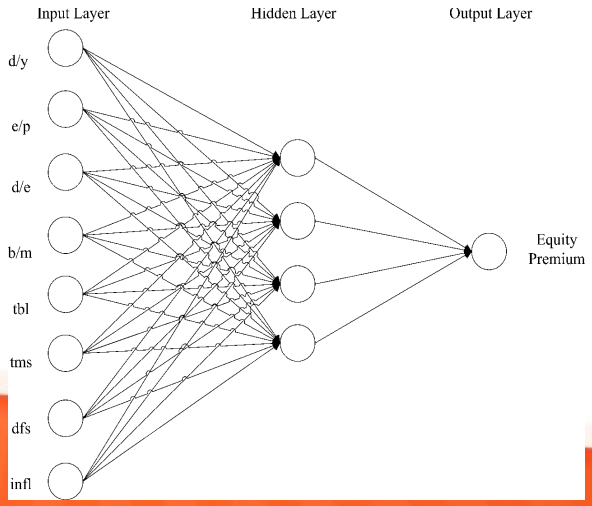


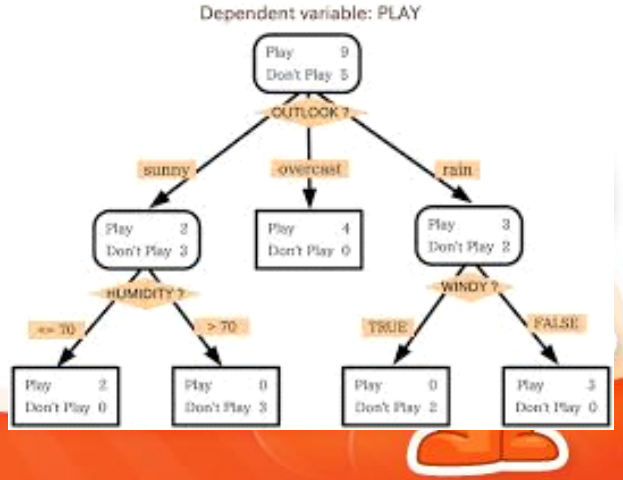


机器分类问题





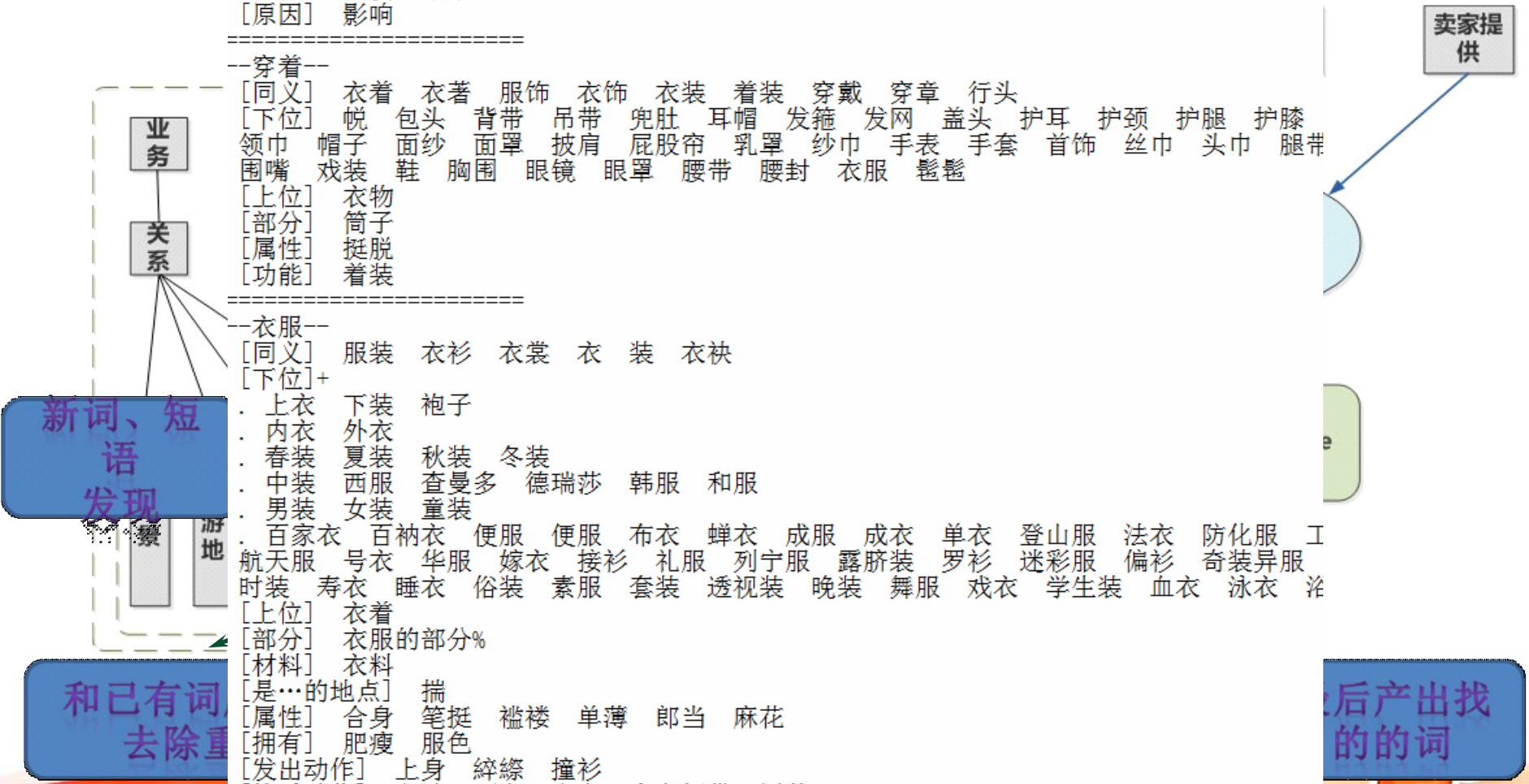




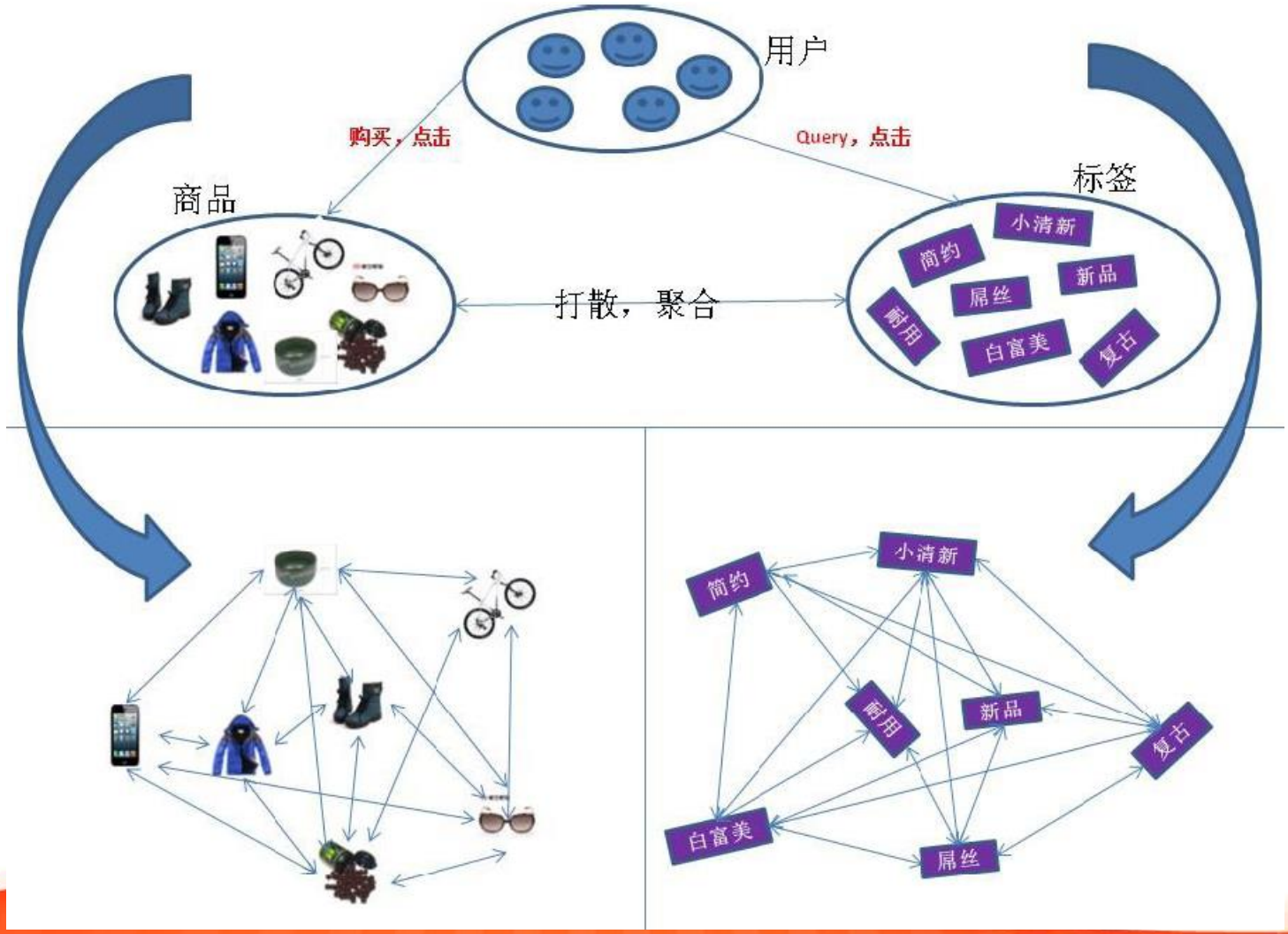
淘宝网
Taobao.com

词库

--效果--
[同义] 作用 影响
[下位] 反馈 关系 害处 好处 后效 效应 效应 余波
[角色] 反作用
[上位] 程度 结果
[原因] 影响
==========
--穿着--
[同义] 衣着 衣著 服饰 衣饰 衣装 着装 穿戴 穿章 行头
[下位] 帨 包头 背带 吊带 兜肚 耳帽 发箍 发网 盖头 护耳 护颈 护腿 护膝
领巾 帽子 面纱 面罩 披肩 屁股帘 乳罩 纱巾 手表 手套 首饰 丝巾 头巾 腿带
围嘴 戏装 鞋 胸围 眼镜 眼罩 腰带 腰封 衣服 髻髦
[上位] 衣物
[部分] 筒子
[属性] 挺脱
[功能] 着装
==========
--衣服--
[同义] 服装 衣衫 衣裳 衣 装 衣袂
[下位]+
. 上衣 下装 袍子
. 内衣 外衣
. 春装 夏装 秋装 冬装
. 中装 西服 查曼多 德瑞莎 韩服 和服
. 男装 女装 童装
. 百家衣 百衲衣 便服 便服 布衣 蝉衣 成服 成衣 单衣 登山服 法衣 防化服 工
航天服 号衣 华服 嫁衣 接衫 礼服 列宁服 露脐装 罗衫 迷彩服 偏衫 奇装异服
时装 寿衣 睡衣 俗装 素服 套装 透视装 晚装 舞服 戏衣 学生装 血衣 泳衣 浴
[上位] 衣着
[部分] 衣服的部分%
[材料] 衣料
[是…的地点] 揣
[属性] 合身 笔挺 褴褛 单薄 郎当 麻花
[拥有] 肥瘦 服色
[发出动作] 上身 綷縩 撞衫
[接受动作] 翻改 更衣 和衣 宽衣解带 置装



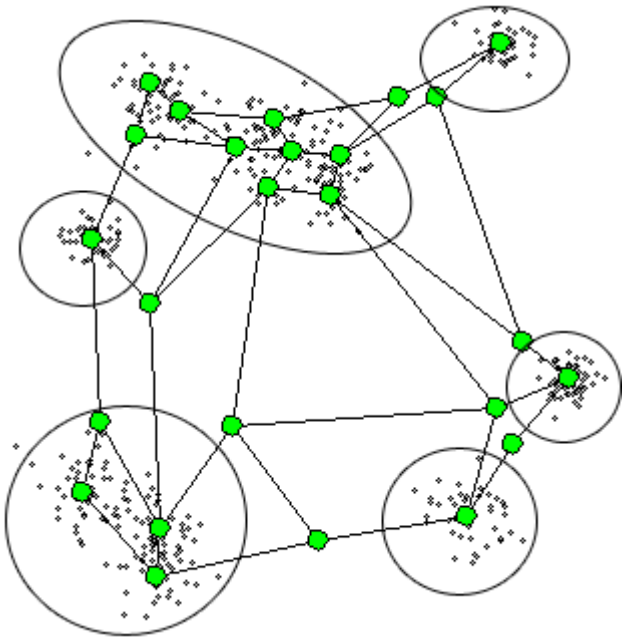
小结

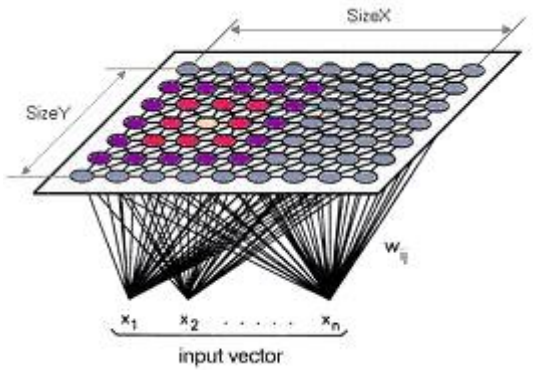

用户
购买，点击
Query，点击
商品
标签
小清新
简约
新品
屌丝
耐用
白富美
复古
打散，聚合
简约
小清新
耐用
新品
复古
白富美
屌丝

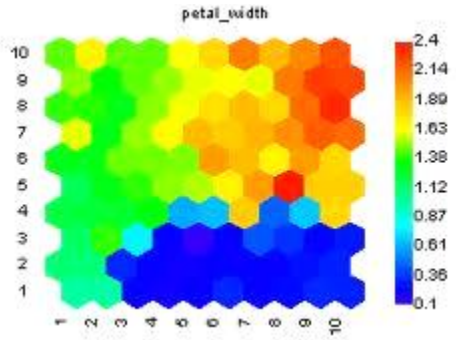
淘宝网
Taobao.com

文本技术相关介绍

数据聚类、天然类目、类目团簇







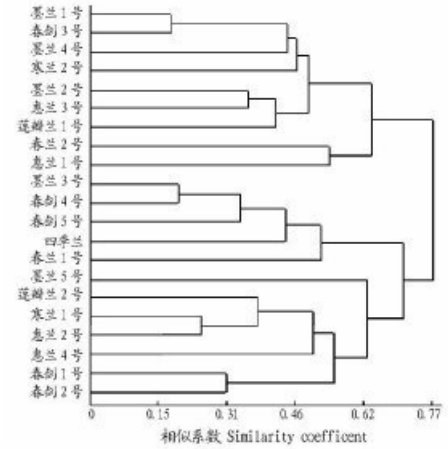
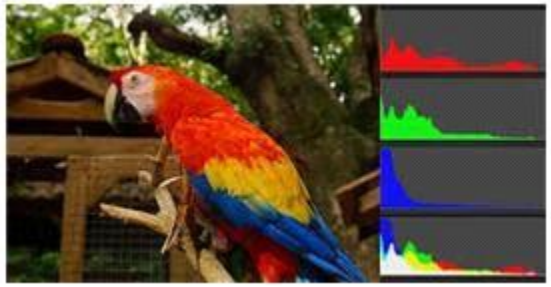


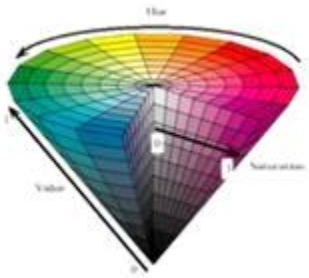
图 2 ERAPD聚类分析结果

谱系、层次聚类

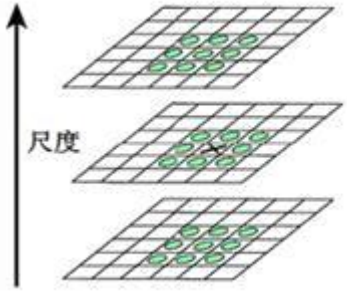

防潮垫
露营/旅游/登山帐篷
睡袋
地布/地席
营地灯/帐篷灯
头灯
登山杖/手杖
防雨罩/背包配件
登山包/双肩背包
腰包/挂包/臂包
头巾/遮耳
雨衣/雨裤/雨披
其他
其他
水壶/水瓶/水杯
急救包/急救箱
干粮/户外食品
打火机/火柴/喷枪/点火装置
工兵铲/工兵锹
求生哨
挂扣
军刀卡/万能刀卡
指南针/温度计/气压计/高度计
苹果数据线
苹果充电器/电源
其它苹果配件
苹果屏幕贴膜
苹果线控耳机
苹果车载充电器
苹果外置电池
苹果手写笔
苹果底座/支架
苹果取卡针
苹果防尘塞
苹果视频线/转换线
苹果键盘保护膜
苹果笔记本包
Touch通讯模块套
移动电源
平板电脑/MID
笔记本电脑
一体机
数码摄像机
普通数码相机
苹果维修服务
其它

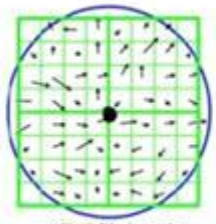
图片语义应用

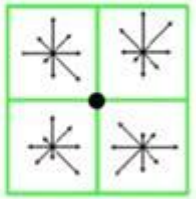


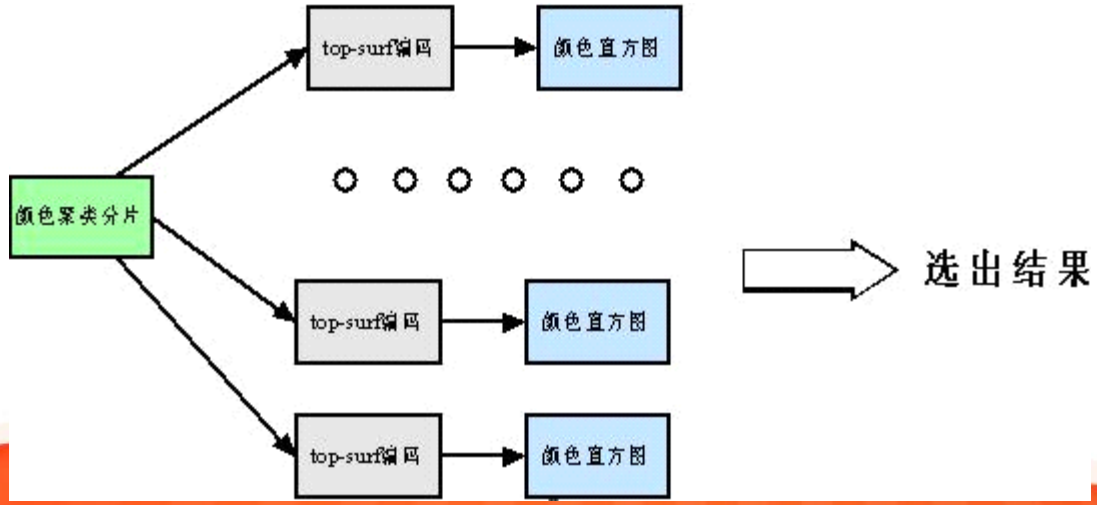


0.60;== 135070, 3.39, 179.25, 184.17 =====
0.18;== 39755, 171.14, 230.79, 232.35 =====
0.13;== 29245, 29.35, 24.46, 16.35 =====
0.10;== 22548, 210.60, 158.77, 115.21 =====









淘宝网
Taobao.com

谢谢