

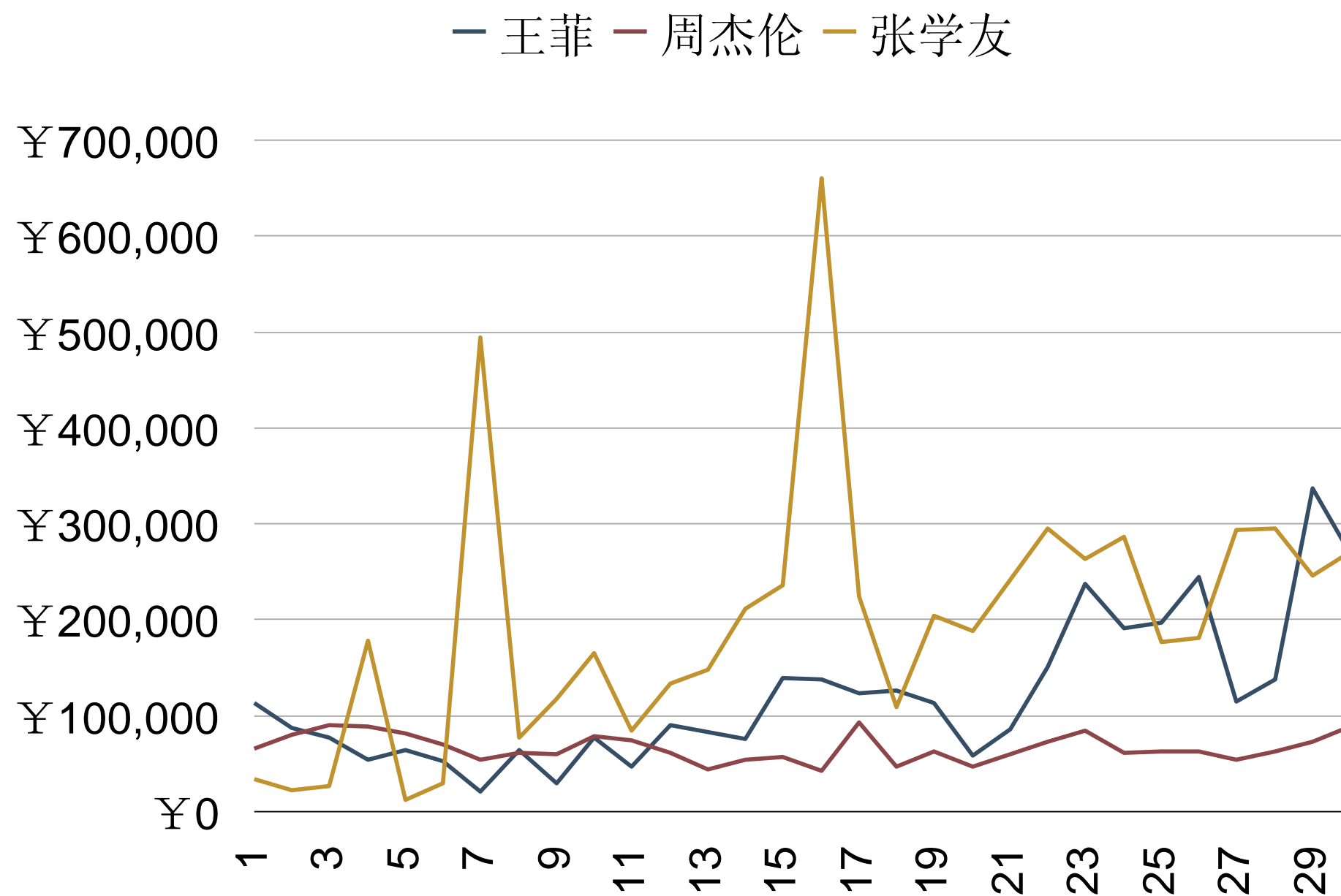
淘宝数据力量

淘宝网数据平台与产品部 贾超（玄澄）

xuancheng.jc@taobao.com

淘宝网

演唱会前一个月门票销售额



粉丝性别年龄构成



年龄分布	百分比
18岁-24岁	32 6.45%
25岁-29岁	191 38.51%
30岁-34岁	193 38.91%
35岁-39岁	59 11.90%
40岁-49岁	15 3.02%
50岁-59岁	5 1.01%
60岁及以上	1 0.20%

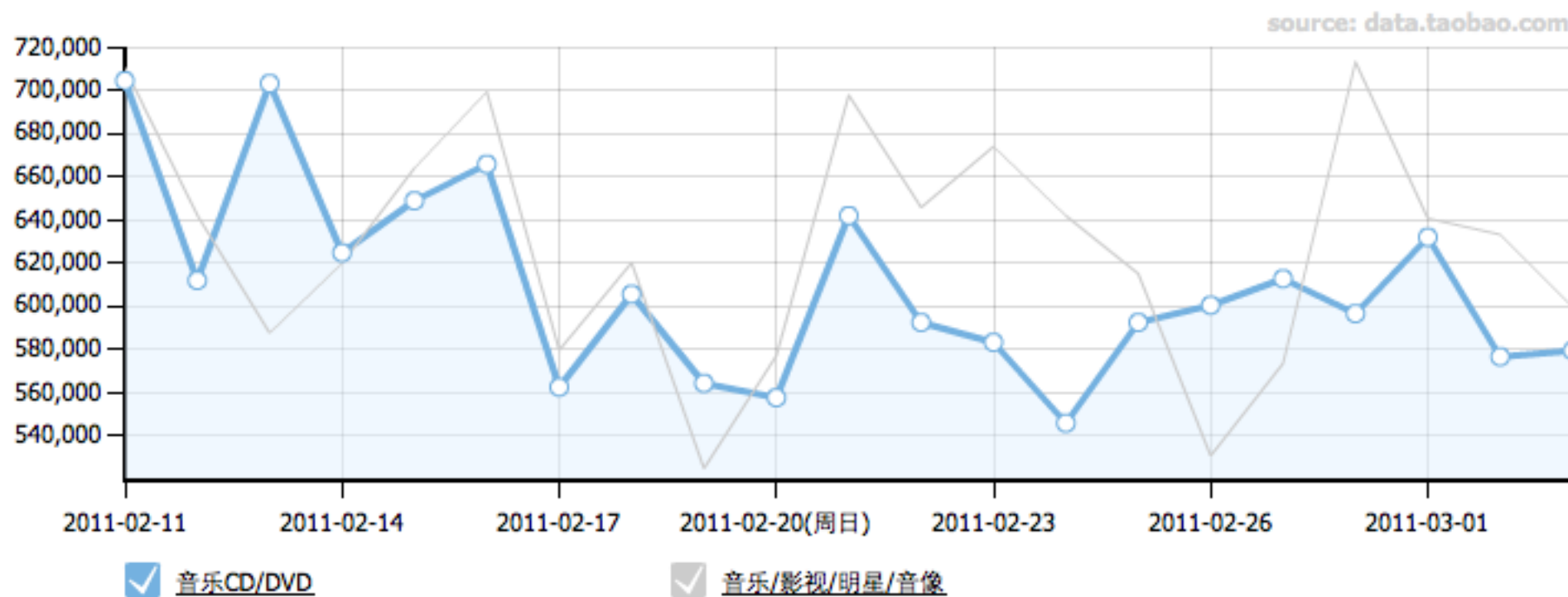
年龄分布	百分比
18岁-24岁	42 7.30%
25岁-29岁	261 45.39%
30岁-34岁	212 36.87%
35岁-39岁	40 6.96%
40岁-49岁	13 2.26%
50岁-59岁	6 1.04%
60岁及以上	1 0.17%



年龄分布	百分比
18岁-24岁	4 5.41%
25岁-29岁	21 28.38%
30岁-34岁	28 37.84%
35岁-39岁	16 21.62%
40岁-49岁	5 6.76%
50岁-59岁	0 0.00%

年龄分布	百分比
18岁-24岁	8 6.20%
25岁-29岁	65 50.39%
30岁-34岁	36 27.91%
35岁-39岁	14 10.85%
40岁-49岁	4 3.10%
50岁-59岁	2 1.55%

CD/DVD类目春节后成交趋势

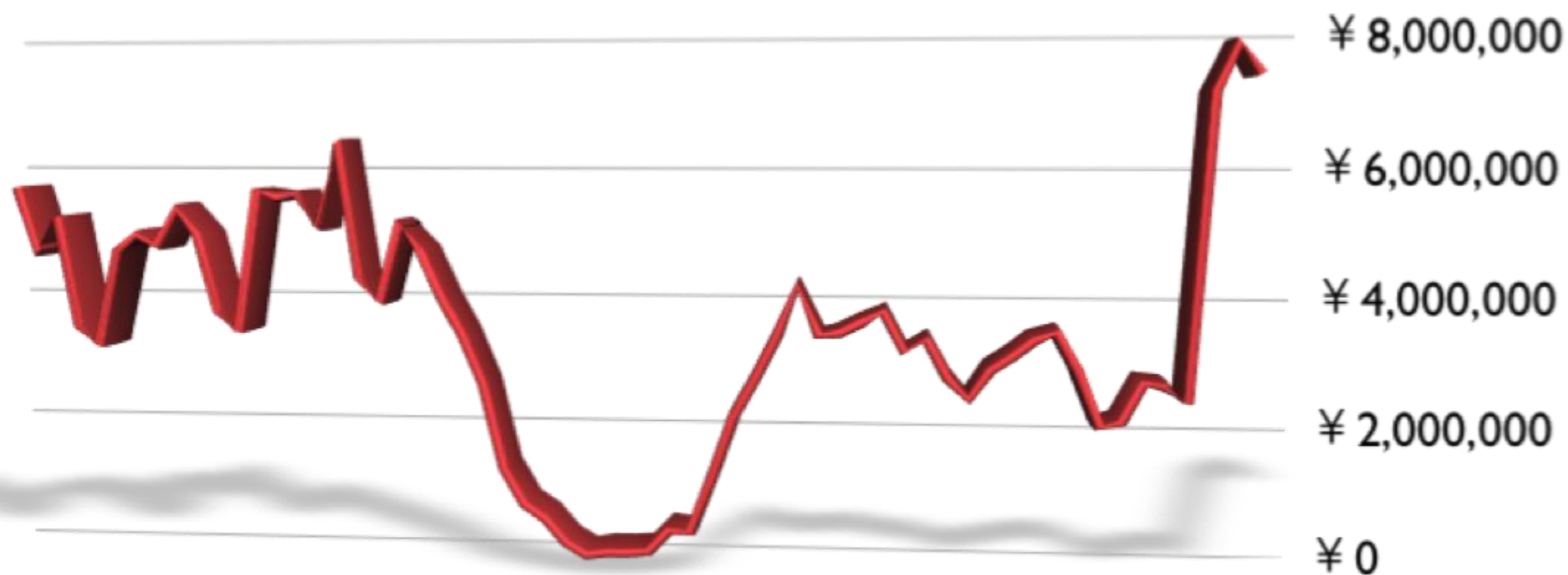


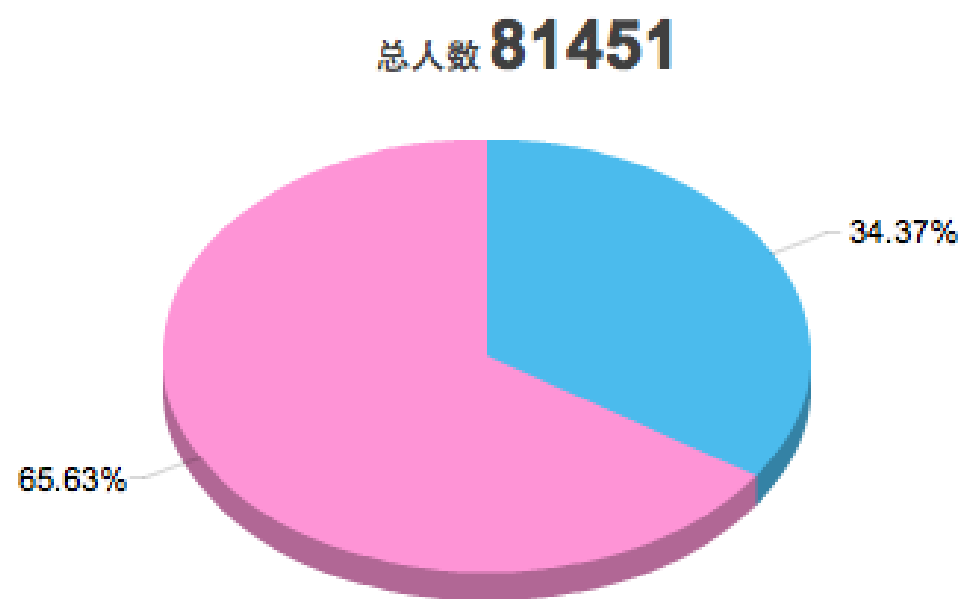
+



演唱会收入 \approx 整个唱片市场销量

2011.1.5-2.11.3.5 iPad成交走势





男性人数
27993

年龄分布	百分比
18岁-24岁	2945 10.52%
25岁-29岁	14047 50.18%
30岁-34岁	7908 28.25%
35岁-39岁	1653 5.91%
40岁-49岁	814 2.91%
50岁-59岁	439 1.57%
60岁及以上	187 0.67%

女性人数
53458

年龄分布	百分比
18岁-24岁	6915 12.94%
25岁-29岁	30282 56.65%
30岁-34岁	11573 21.65%
35岁-39岁	2138 4.00%
40岁-49岁	1344 2.51%
50岁-59岁	987 1.85%
60岁及以上	219 0.41%

婚庆喜糖类目 买家性别年龄分析
2010.12.1 - 2011.2.26

更多真相：@数据魔方



United States
308,745,538人

Italy

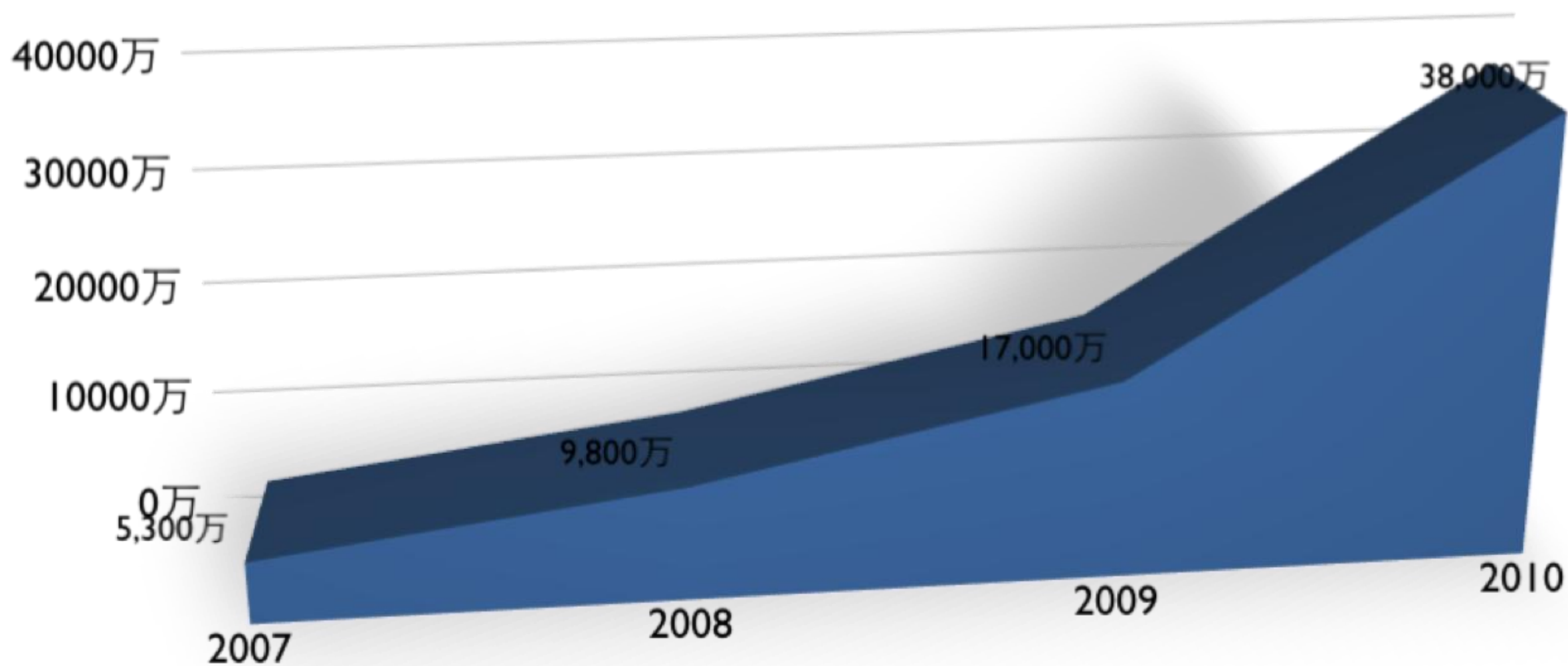
60,418,711人



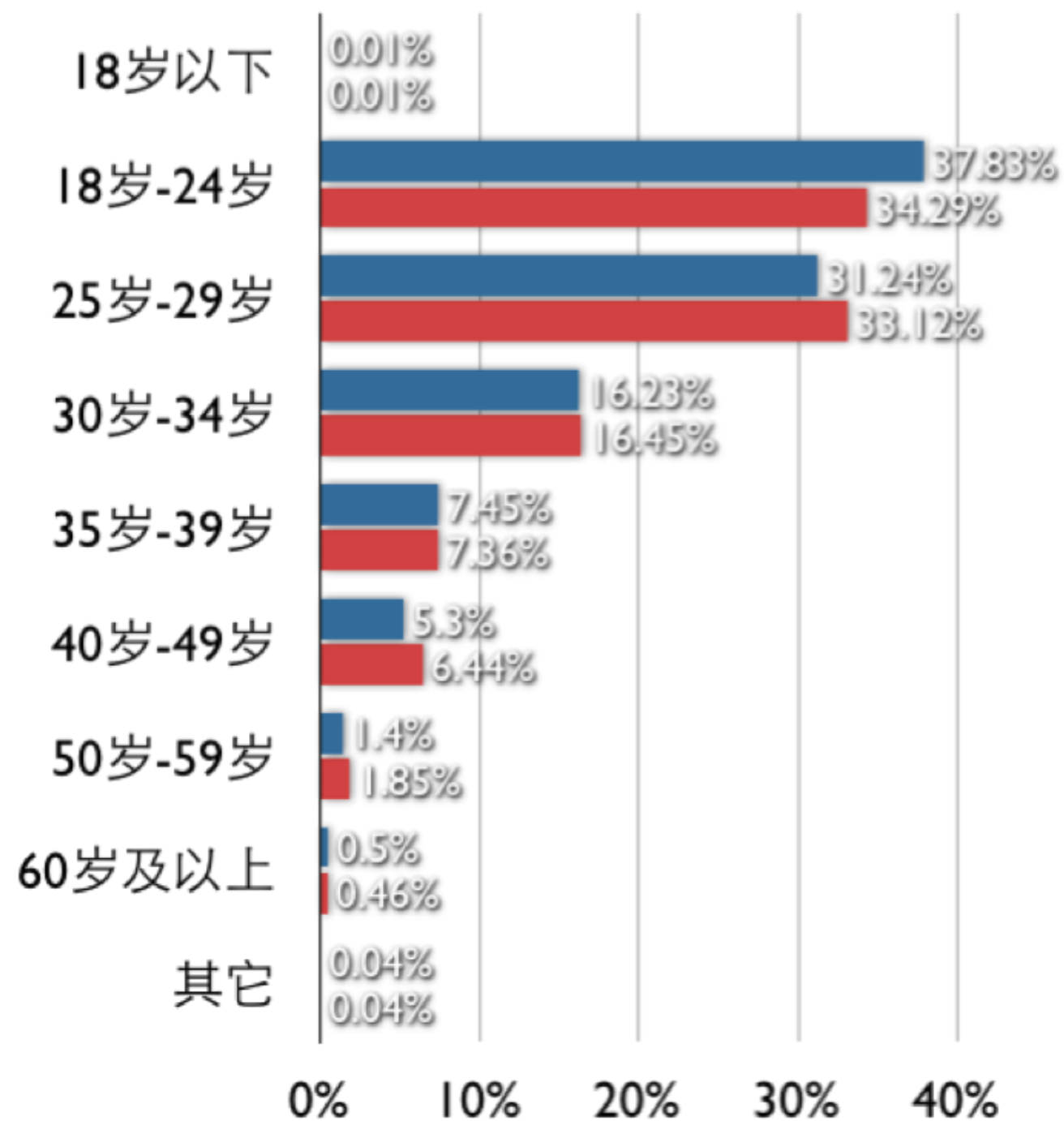
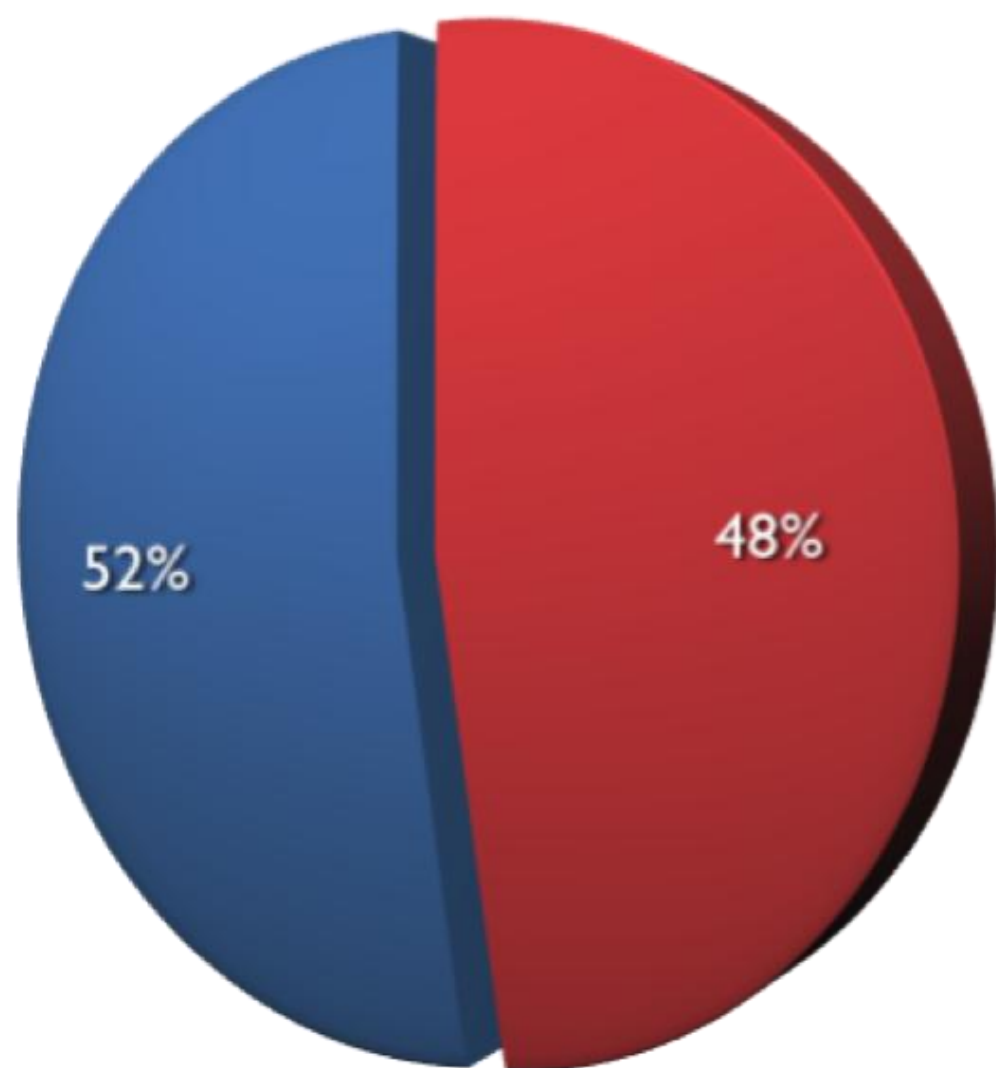
截止2010年底淘宝网注册会员数
3.7亿 > 美国总人口数

2010年单日UV峰值
6000万 \approx 意大利总人口数

近4年注册用户数增长趋势



● 女性 ● 男性



超过**8亿**件在线商品

2010年单日成交额峰值**19.5亿**

2010年淘宝商城交易额是09年**4倍**，
单日销售额峰值**9.36亿**

从你看这页Slide的**10s**内，成交商品
8000件，包括：



144



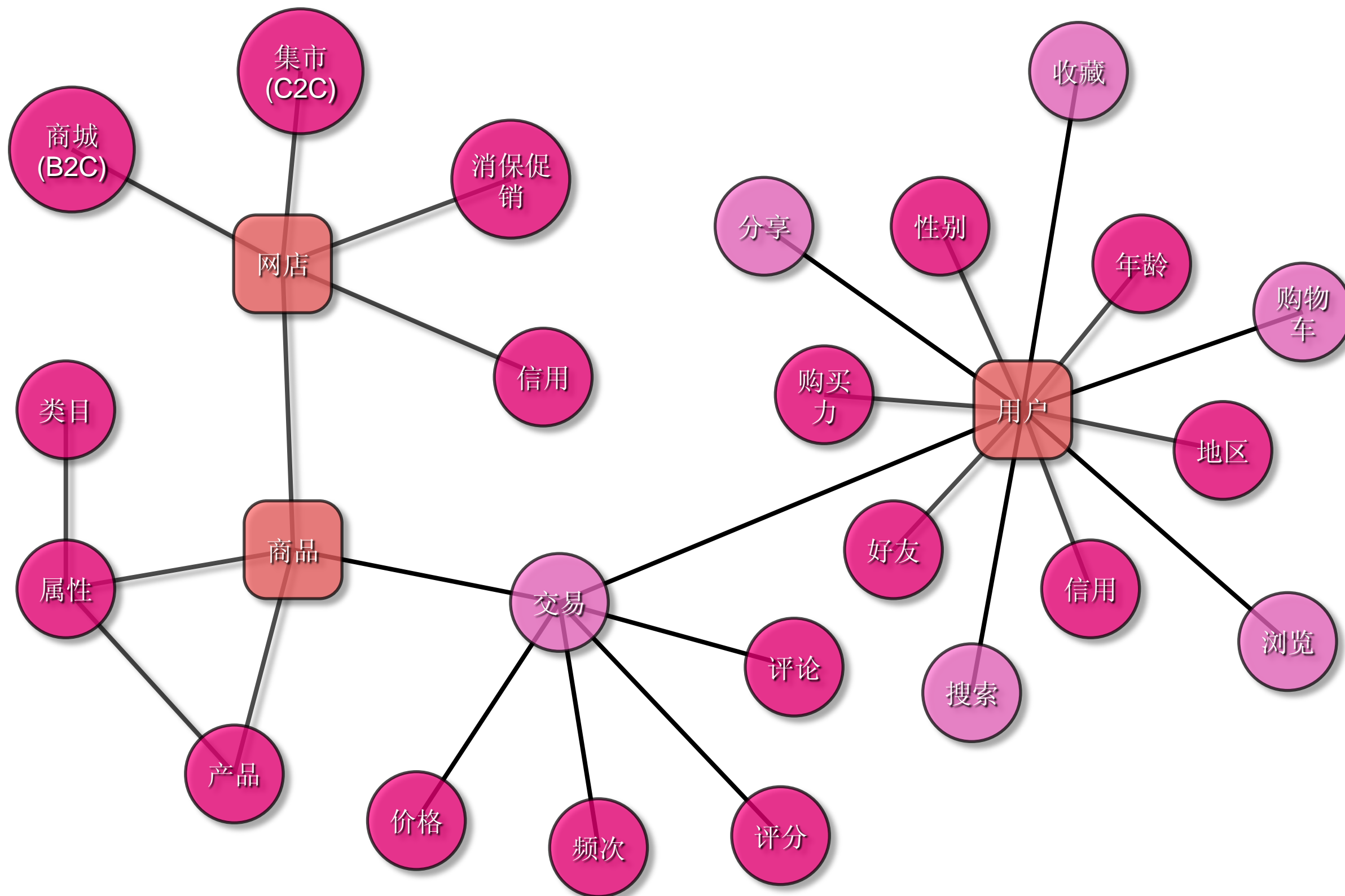
6



147



14



产品**100万**，在线商品**8亿**

归一化后属性**4万**，属性值**100万**，其中包括近**3万**不同的品牌

2010年**66%**用户购买5笔以上，**52%**用户购买10笔以上

5个月内上线商品占全部商品的**86.33%**

在线商品数量庞大，大量相同或近似的商品同时在线，新上架商品所占比例很高

商品数远大于用户数

大量的成交集中在极少量的商品上

仅在购买后可以评分和评价，评分和评价主要针对卖家服务，无法准确体现用户消费倾向

商品的品牌、型号等属性，可视为TAG

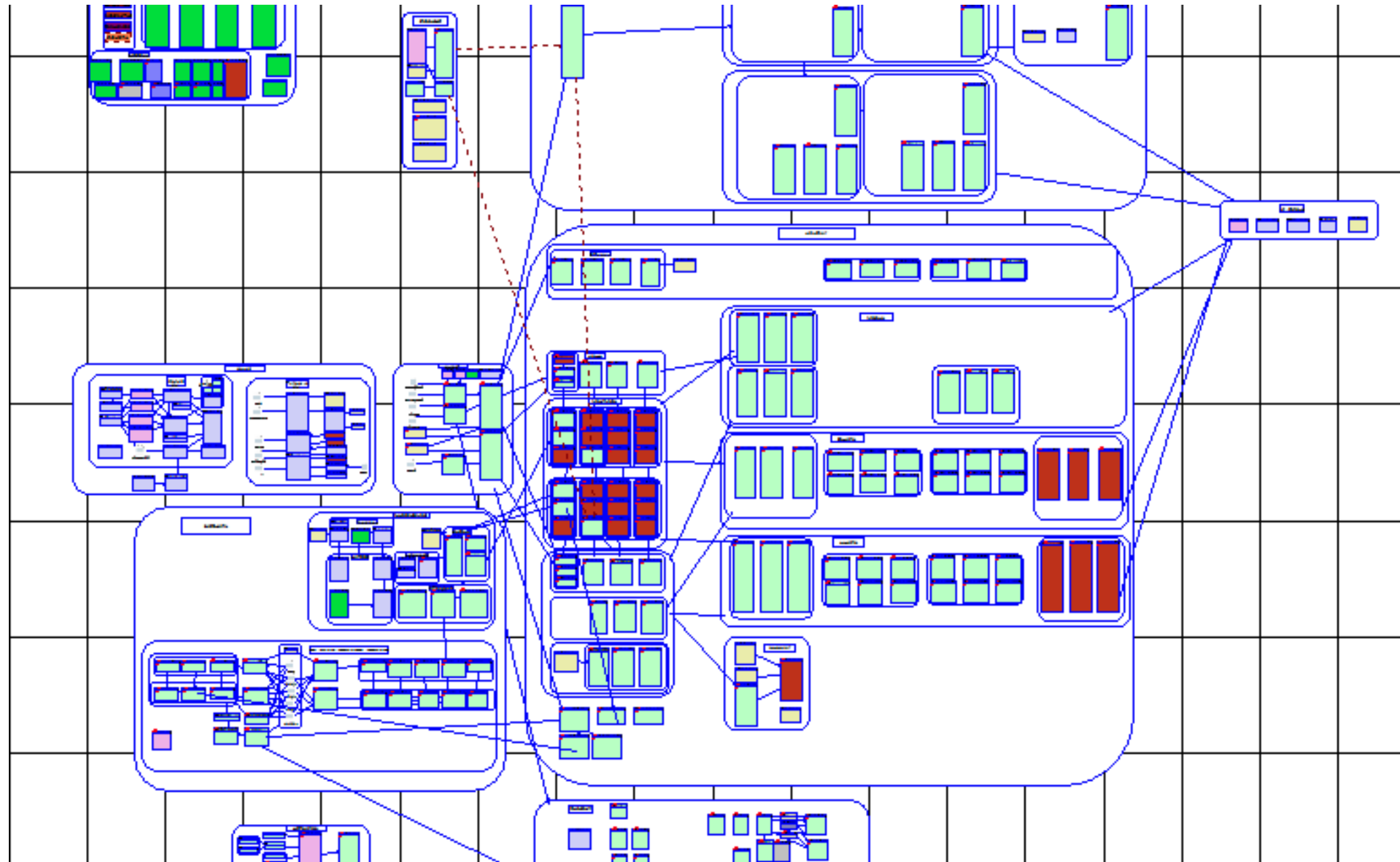
关键属性值的组合（特别是SPU）可用于商品数据的归一化，但大量商品无法标准化，非标类商品占绝大多数

类目繁多，不同类目商品特征不同

用户重复购买率很高（仍无法与内容型网站相比）

涌现了一批拥有很高用户粘性的网店

数据魔方涉及的数据模型图



每日新增数据**20T**

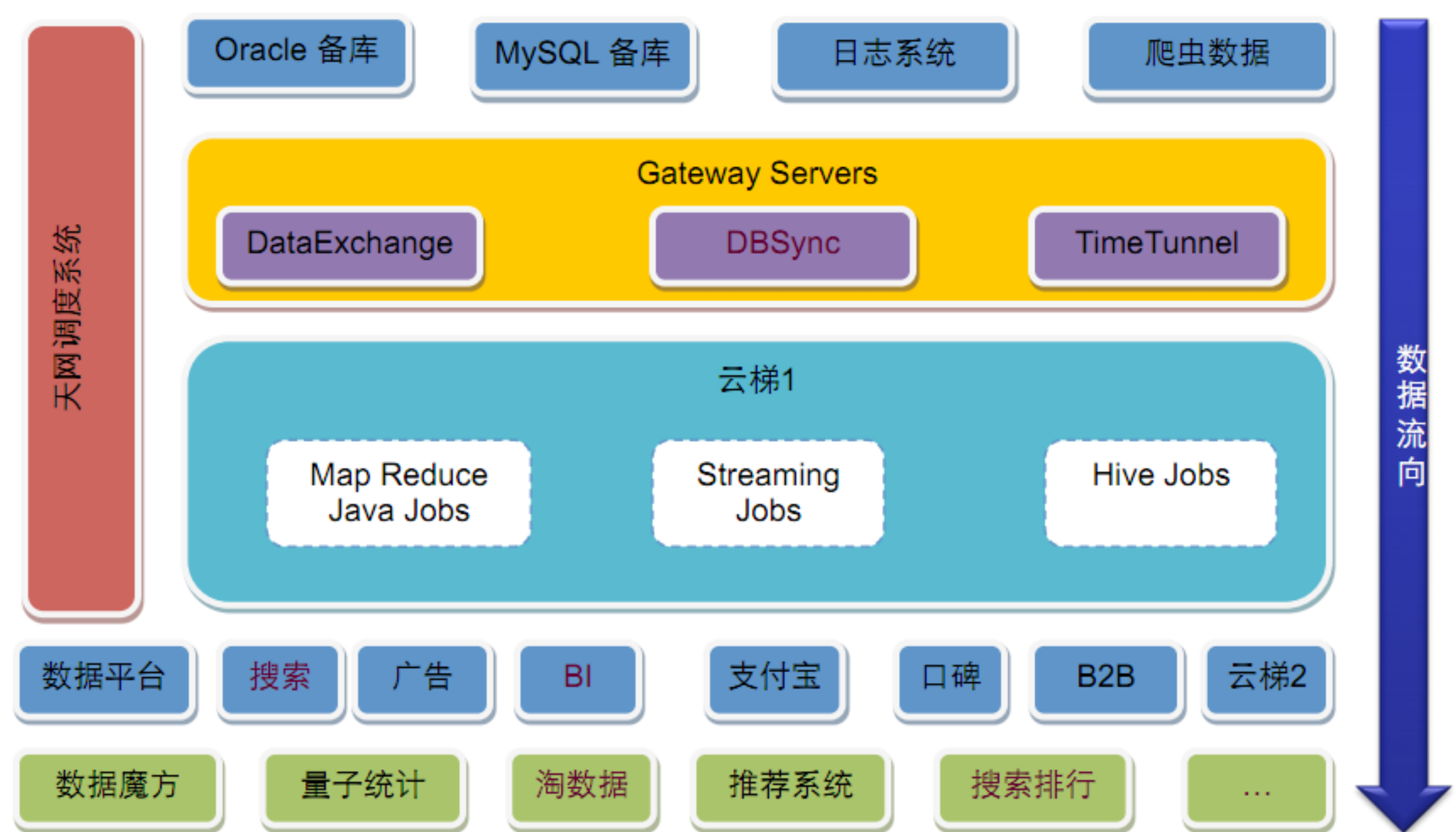
累积数据**14P**

1300+服务器的云计算平台

每天处理**100,000+**作业任务，包括**100+**新增作业任务

每天处理**1P+**数据，包括**0.5%**新增数据

数据平台整体架构



数据化运营

淘数据

数据魔方小二版

自动化数据运营

宝宝用品	纸尿裤	湿巾	奶瓶	奶嘴	洗浴
			    	    	
带匙奶瓶 均价：15元 销量：6537件 整体科学设计，节省使用一般汤匙需反复喂食的时间，给宝宝喂食更轻松，满足宝宝不同阶段成长需要	防胀气奶瓶 均价：69元 销量：5665件 可根据宝宝吸奶时吸力的变化而自动收缩瓶身，保持瓶身内外的压力一致，彻底解决了奶瓶喂奶时胀气的烦恼。	玻璃奶瓶 均价：88元 销量：2563件 能让母亲易于捏握，玻璃奶瓶易于清洗			



Taobao DataCube
数据魔方

淘宝指数

数据可视化 (Data Visualization)

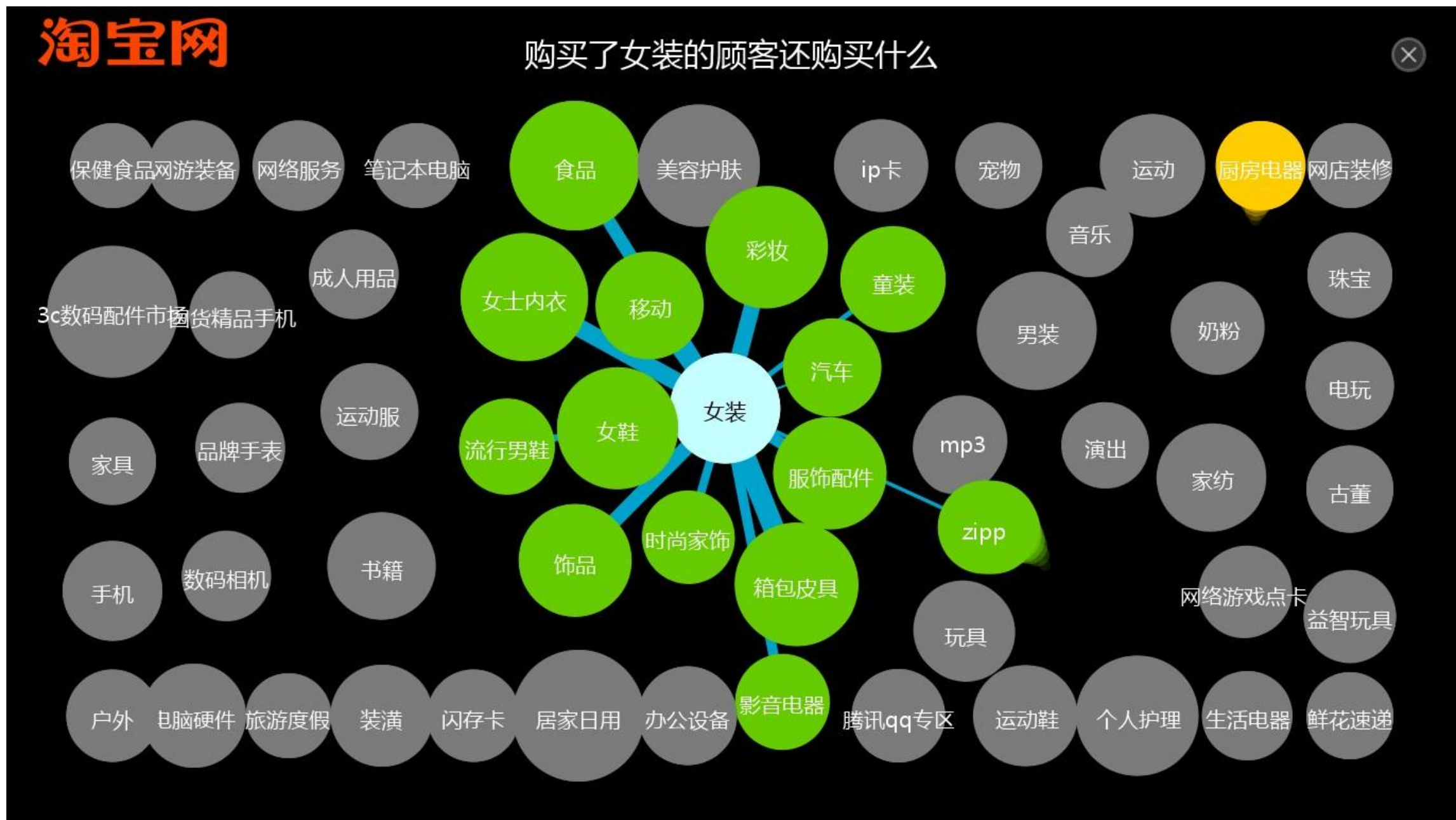
针对海量数据、复杂模式使用图形化的手段

发现数据中隐藏的节奏、趋势及异常

往往结合交互手段

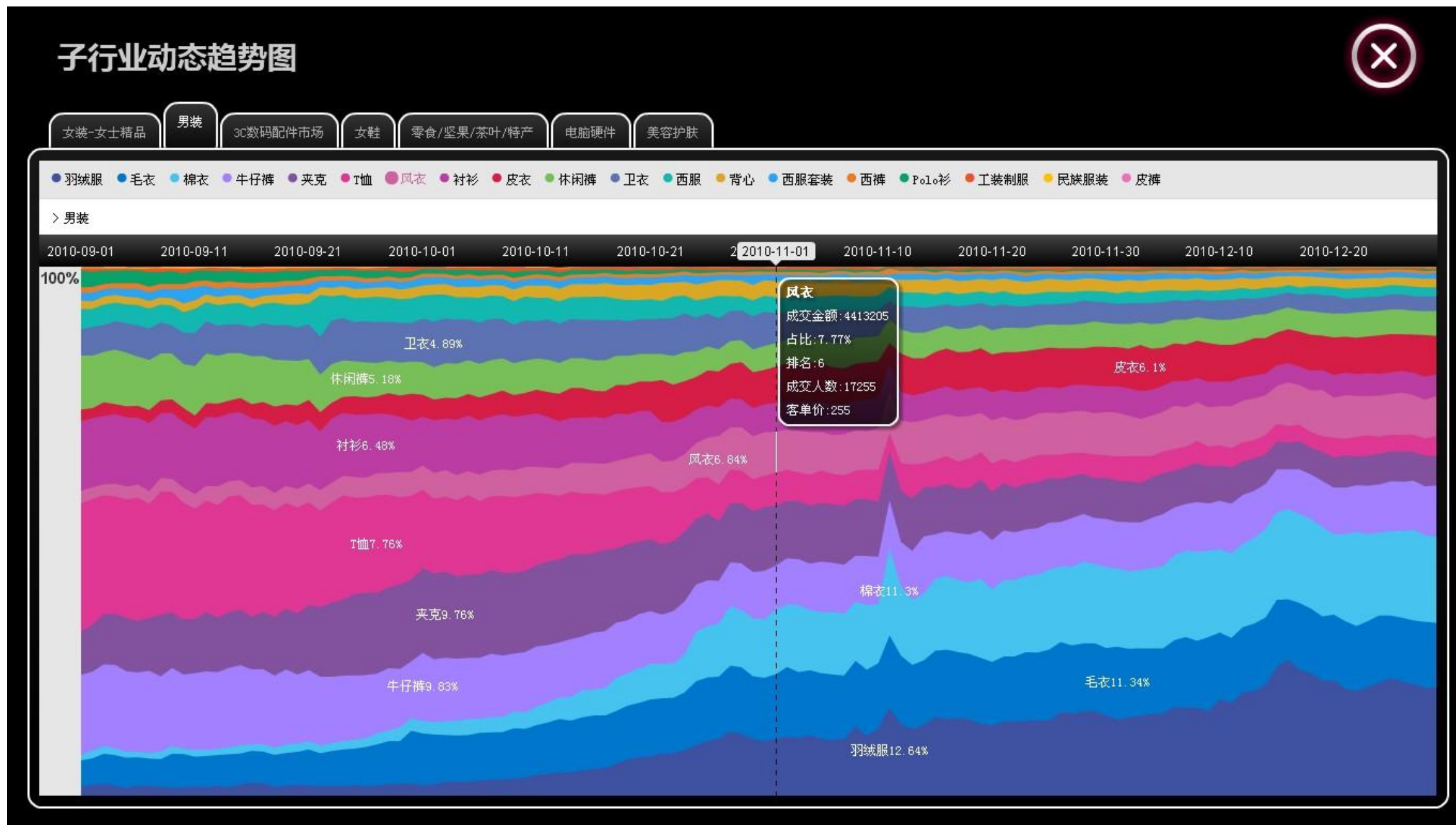
利用人对形状、颜色、密度、速度等图形图像特征的敏感性

可对数据挖掘的结果进行更为有效的呈现



CatLink





TaoTrends

2010年3月31日 **数据魔方**发布，向
全球开放淘宝宏观交易数据

2011年 **淘宝指数**即将上线，进一步
面向消费者开放数据

2011年2月23日 **2011淘宝开放年**
战略发布

四个开放领域：前台开放 数据开放 买家开放 无线开放

合作对象：第三方开发者（ISV） 企业 服务商

截止**2010**年底，淘宝开放平台拥有**11万**注册开发伙伴，**300个**API每天**7亿**次API调用

在**2011**年将重点扶持**100个**“百万俱乐部”成员，并向千万盈利能力迈进

数据开放方针：**有原则 分层次**

数据开放 + 推荐系统 探讨

基于数据开放平台：

提供面向卖家的推荐系统技术产品

深度合作：

淘宝研究院专注于学术研究，同时与高校院所、科研机构开展领域研究的合作

**与第三方合作提供面向垂直市场甚至全网商品的推荐服务
(基于淘宝云计算平台)**

推荐算法竞赛(平台、规则、评价、应用)