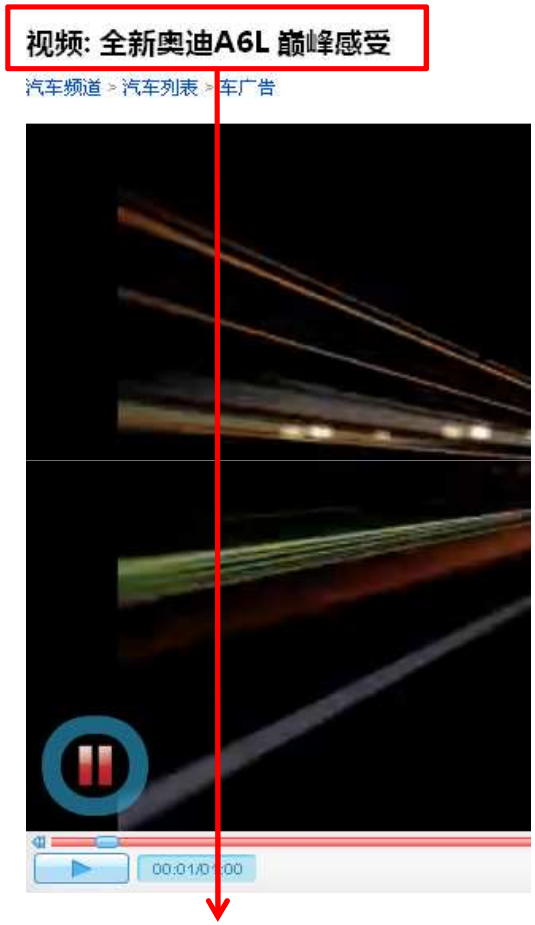
R环境短文本分类器与电商品类数据研究

Bird 张翔

北京•上海•广州•深圳•东京•硅谷•香港

www.iresearch.com.cn

为什么是短文本



1. 提取标题
2. 分词 与自动分类
3. 类目与词频统计

短文本分类的核心目的

2012年5月面部护理类商品销售额标签云

Source：视频数据来自iVideoTracker， 2012.4. 基于对20万名家庭及办公（不含公共上网地点）样本网络视频行为的长期监测数据获得，不包括在线视频客户端数据。
电商数据来自艾瑞对20万名家庭及办公网民网络购物行为监测所得。

短文本分类难点

TFIDF	3G手机	CDMA手机	CD包	CPU	CRT显示器	DIY、模型玩具	GPS系统	GSM手机
TF(microSD)卡	0.19	0.17	0.00	0.00	0.00	0.00	0.14	0.23
直板	0.18	0.22	0.00	0.01	0.00	0.00	0.17	0.21
MP3铃声	0.18	0.17	0.00	0.01				0.19
GSM	0.14	0.12	0.00	0.00				0.21
全新	0.13	0.19	0.00	0.46				0.16
店铺三包	0.13	0.11	0.00					0.15
26万	0.12	0.10	0.00	0.00				0.15
非智能手机	0.11	0.11	0.00	0.00	0.00	0.00	0.03	0.14
无操作系统	0.11	0.11	0.00	0.00	0.00	0.00	0.06	0.15
全国联保	0.10	0.15					0.19	0.14
智能手机	0.10	0.13					0.17	0.05
200万	0.10	0.07					0.03	0.10
2010年	0.10	0.14						
套餐四;套餐三...								
2009年								
1600万								
GSM/WCDMA(3G)								
9.99成新								
6万5千色								
MP3播放...								
WindowsMobile								
2.8英寸								
Nokia/诺基亚	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.01
手写输入;视频播放...	0.04	0.05	0.00	0.00	0.00	0.00	0.02	0.04
电阻式触摸屏	0.0	0.02	0.00	0.00				
LG				0.00	0.00			
套餐三;套餐一...				0.00	0.00			

没有训练集

约有500个小类目，每个类目

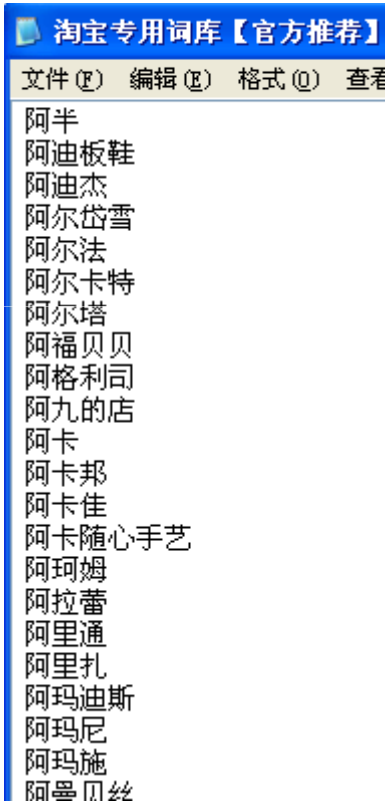
需要200-1000个训练商品

总共需要10-50万的训练集

扩充词库和关键词人工列表

- 淘宝专用词库【官方推荐】

词条样例：阿半、阿迪板鞋、阿迪杰、阿尔岱雪、阿尔法、阿尔卡特、阿尔塔、阿福贝贝、阿格利司、阿九的店、阿卡、阿卡邦、阿卡佳、阿卡随心手艺、阿珂姆、阿拉蕾、阿里通、阿里扎、阿玛迪斯...



Term	Category
保温杯	家居用品/杯具
笔记本电脑	电脑/笔记本
餐桌	家具/桌子
车载充电器	汽车用品/电器/配件
低音炮	音响
电话卡	手机/号码卡-电话卡
儿童座椅	座椅

遗留的问题

BB霜

美的电器

很美的电器贴图

R文本分析包：

Tm

Rmmseg4j

Smartcn

wordcloud

处理流程（借鉴KDDCUP2005）

谷歌关闭了分类目录
使用人工特征词列表建立
最初训练集

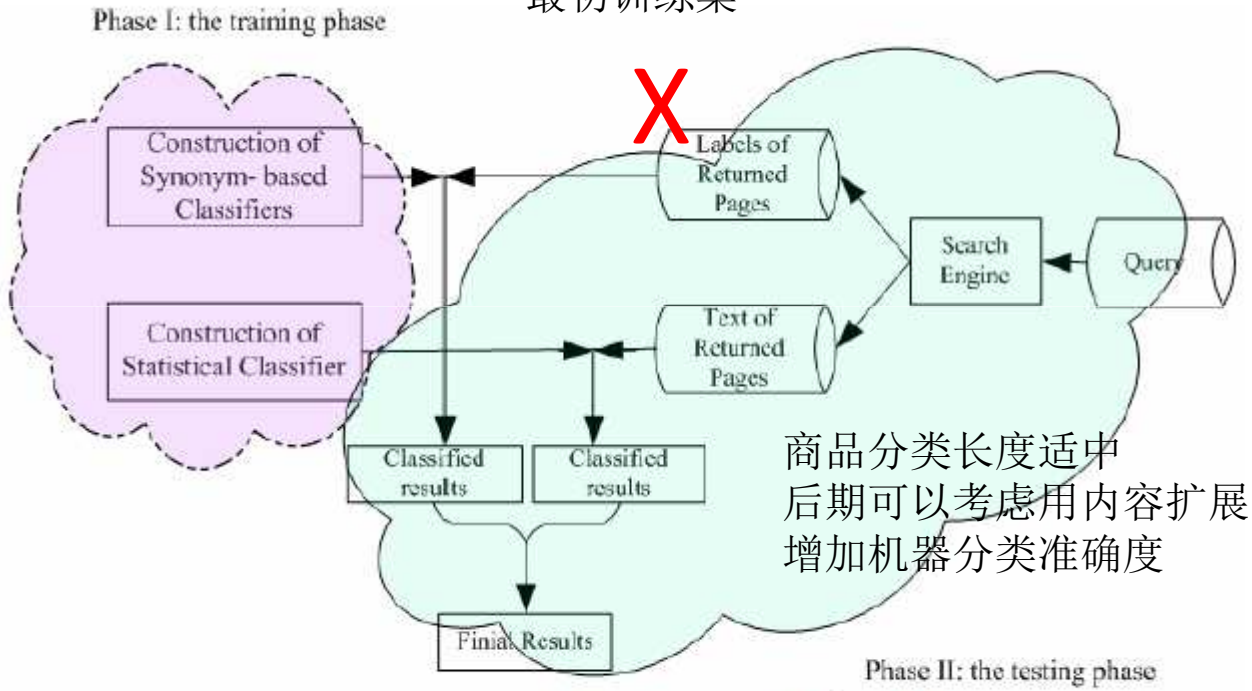


Figure 1. The architecture of our approach $Q^2C@UST$

商品分类长度适中
后期可以考虑用内容扩展
增加机器分类准确度

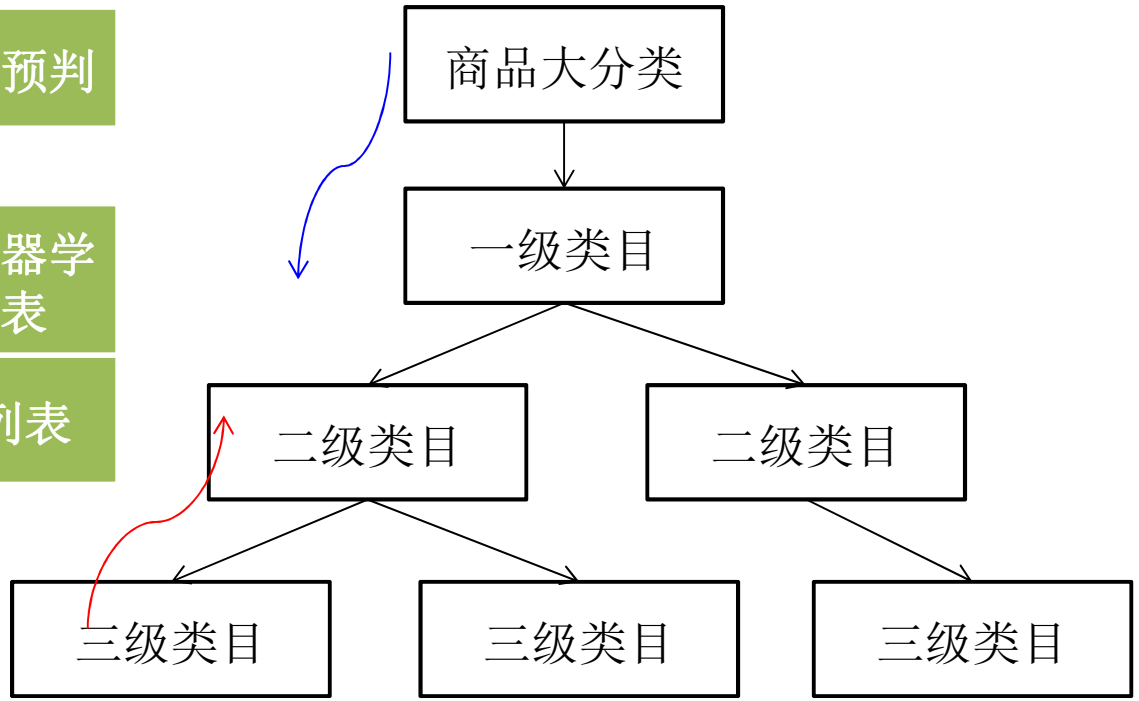
Q2C@UST: Our Winning Solution to Query Classification in KDDCUP 2005

混合分类器流程

交叉类目预判

关键词机器学习打分表

分类词列表



有人工分类表参与的分类准确性接近90%，但是词库建立耗时，目前积累的词库商品覆盖率达70%

只依赖机器自学习的分类准确性为60%-80%，但是在做大类预判的时候覆盖率高

贝叶斯从朴素到“奢华”

- 朴素贝叶斯的假设条件
  - 词之间没有任何依赖关系，即条件独立；

$$P(d \mid c_i) = P(w_1, w_i, ..., w_n \mid c_i) = \prod_{1 \le k \le n} P(w_k \mid c_i)$$

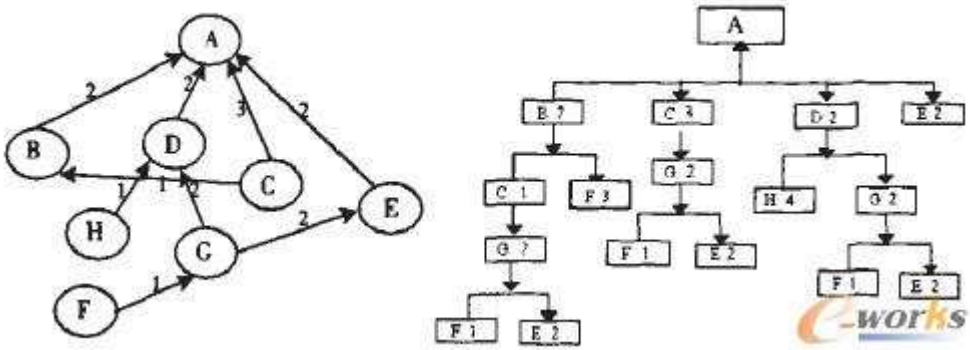
- 贝叶斯网络
  - 人的思维是有依赖关系的
    - “_____苹果__iphone__”
    - “_____苹果__红富士__”
  - 用一个有向无环图和一个条件概率表集合来表示一组变量的联合概率分布

$$P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | Parents(x_i))$$

- 现实是残酷的
  - 对于文本的词量，条件网络过于复杂，可以直接使用“词对”预测（半朴素贝叶斯）

网状类目结构

- 普通目录结构是树状结构固定深度
  - 目前艾瑞使用的是3+1级目录结构，列表形式储存
- 真实类目前存在交叉，是网状关联结构
  - 计算机语言中可以做面向对象结构设计
  - R语言中可以使用sna或igraph的图矩阵格式



结果应用-店铺数据

➢ **剃须护理、汽车电子淘宝店铺垄断性强**

- 马太效应显著的类目有剃须护理及汽车电子，两类销售额10%的店铺占据了约9成的业绩份额。
- 相比之下，部分家具、饮料类目店铺竞争相对充分。

2012年5月淘宝部分类目销售额在店铺的分布情况

	淘宝集市			天猫		
	类目	TOP10%店铺销售额占比	TOP30%店铺销售额占比	类目	TOP10%店铺销售额占比	TOP30%店铺销售额占比
相对垄断品类	剃须护理	92%	98%	剃须护理	96%	99%
	音响播放	89%	98%	汽车电子	94%	99%
	汽车电子	88%	98%	休闲食品	86%	97%
相对竞争品类	住宅家具	69%	92%	办公家具	63%	90%
	碳酸饮料	67%	91%	含乳饮料	63%	91%
	办公家具	66%	89%	矿泉水	39%	83%

注释：表中列举了此次研究品类中，10%店铺销售额占比最多及最少的三个商品类目。
来源：基于情报通抓取的公开交易数据。

细分商品-电视品类分布

➢商城买家电　　C店淘配件

- 在主要B2C网站中，用户购买的电视类商品以电视机本身为主，电视机品牌（索尼、创维、TCL）、配置说明（全高清、LED）等描述的商品名称销售额较高。
- 淘宝集市中，除了电视机外，电视周边商品也有较高的关注度：机顶盒、电视开关、遥控器、贴纸等品类也占据了较多的份额。

2012年5月电视商品销售额标签云

淘宝集市

网上商城

来源：基于艾瑞对20万名家庭及办公网民网络购物行为监测所得。

细分类目-厨房&生活电器

- 年轻、高收入者是厨房生活电器类主要访问者

2012年5月厨房生活电器销售额标签云

2012年5月特定类目访问者基本属性

		厨房电器 比例	厨房电器 TGI	生活电器 比例	生活电器 TGI
性别	男	49%	89	54%	98
	女	51%	113	46%	103
年龄	18岁及以下	5%	35	7%	43
	19-24岁	21%	74	25%	89
	25-30岁	29%	134	29%	133
	31-35岁	17%	139	15%	121
	36-40岁	13%	142	12%	125
	40岁以上	13%	111	12%	98
个人月收入	无收入	8%	44	11%	60
	1000元以下	6%	70	6%	78
	1000-2000元	23%	109	22%	106
	2000-3000元	28%	119	28%	119
	3000-5000元	22%	127	21%	118
	5000元以上	13%	116	12%	102

注释：左图中仅包括了销售额TOP50的商品；销售额为淘宝集市与网上商城合计值；

右表中TGI是目标群体指数，TGI指数=（目标群体中某一特征的群体所占比例/总体中相同特征的群体所占的比例）*标准数100

来源：淘宝集市数据基于情报通抓取的公开交易数据，网上商城总体数据基于艾瑞对20万名家庭及办公网民网络购物行为监测所得。

热销品牌-厨房电器类

➢淘宝集市与网上商城热销品牌重合度高

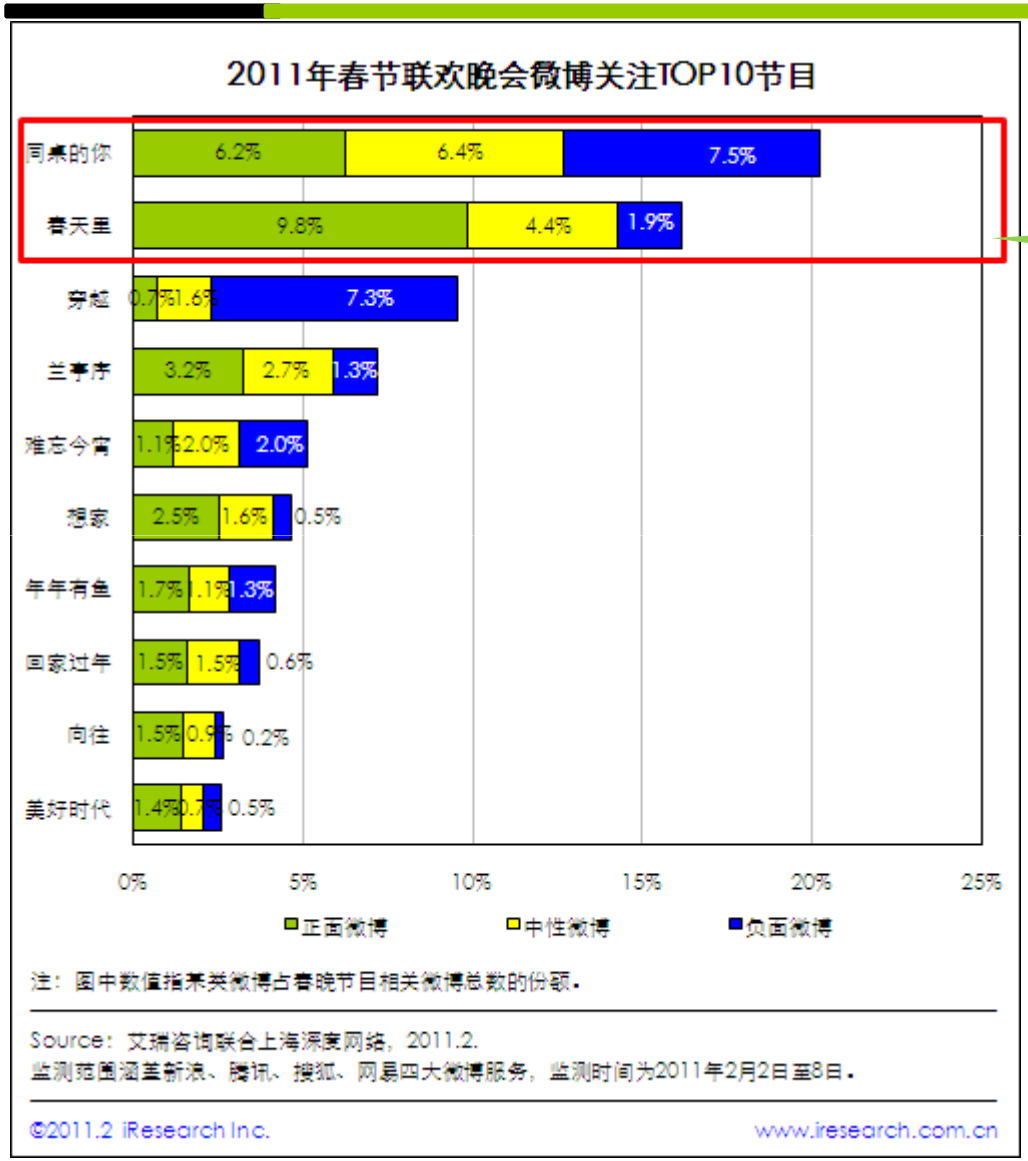
- 厨房电器销售额较高的品牌中，九阳、美的及苏泊尔在淘宝集市及网上商城排名均居前列，两类电商网站的热销品牌存在较高的重合。

2012年5月厨房电器类销售额第一阵营品牌墙

淘宝集市	网上商城总体
Joyoung 九阳 健康食尚家 美的 Midea SUPOR 苏泊尔 沁园 QINYUAN DONLIM 东菱	美的 Midea Joyoung 九阳 健康食尚家 Galanz 格兰仕 SUPOR 苏泊尔 POVOS 奔腾

来源：淘宝集市数据基于情报通抓取的公开交易数据，网上商城总体数据基于艾瑞对20万名家庭及办公网民网络购物行为监测所得。

想要更多



注：图中数值指某类微博占春晚节目相关微博总数的份额。

Source：艾瑞咨询联合上海派度网络，2011.2.
监测范围涵盖新浪、腾讯、搜狐、网易四大微博服务，监测时间为2011年2月2日至8日。

草根节目在网民评论中好评高

感谢聆听，期待合作！

艾瑞 网民风向标 搜索

选择艾瑞，

选择值得信任的合作伙伴！

iResearch 艾瑞咨询集团 | 艾瑞十年 2002-2012