



大数据成就未来

电子商务网站用户行为 分析及服务推荐

姜鹏辉

2017/11/8



案例背景

Baidu百度

R


百度一下

网页新闻贴吧知道音乐图片视频地图文库更多»

百度为您找到相关结果约100,000,000个

搜索工具

R: The R Project for Statistical Computing




查看此网页的中文翻译, 请点击 [翻译此页](#)
R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNI...
www.r-project.org/ - 百度快照 - 100%好评

R 百度百科

里包恩, 日本漫画及改编动画《家庭教师》中的人物, 是世界顶级的一流杀手, 彭格列九代最信任杀手。受九代首领的委托为了培养沢田纲吉成为彭格列十代首领从意大利...
其他含义:
[《神奇宝贝》中的反派组织](#)
[统计应用软件](#)
[查看“R”全部11个含义>>](#)
baike.baidu.com/

r商标 百度百科




R是REGISTER的缩写, 用在商标上是指注册商标的意思。用圆圈R, 是“注册商标”的标记, 意思是该商标已在国家商标局进行注册申请并已经商标局审查通过, 成为注册商标。R商标具有排他性、独占性、唯一性等特点, 属于注册商标所有人所独占, 受...
[特征 关系 意义](#)
baike.baidu.com/

R 语言的优劣势是什么? - R(编程语言) - 知乎

1. 免费... 开源... (这是最重要的一点好不好, 也是SAS流行于公司, R流行于研究机...
2. 是专门为统计和数据分析开发的语言, 各种功能和函数琳琅满目, 其中成熟稳定的一...
3. 语言简单易学。虽与C语言之类的程序设计语言已差别很大 (比如语言结构相对松散, ...
[查看全部>>](#)
www.zhihu.com/question... - 百度快照 - 1131条评价


百度首页消息设置

其他人还搜




running man

韩国新型态娱乐节目




!

韩国infinite组合成员



节奏蓝调


庾澄庆的音乐风格



r-15

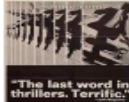
伏见广雪著轻小说

相关字母




s

英文字母第19个字母



z


科斯塔加华斯执导电影



n


拉丁字母

相关符号




d

网络用语为顶



f

字母符号



Rh


化学元素符号

泰迪智能科技
TipDM Intelligent Technology

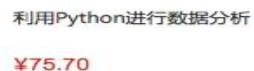
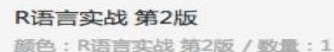
大数据挖掘专家

2

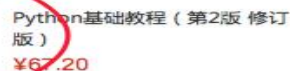
商品已成功加入购物车！



R语言实战 第2版
颜色：R语言实战 第2版 / 数量：1



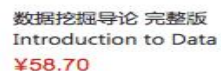
 加入购物车



 加入购物车



加入购



¥58.70



¥69.50



¥58.70

累计评价
400+

排名: 自营 计算机与互联网销量榜 第 60 位

配 送 至: 北京朝阳区 有货, 支持 99元免运费 | 货到付款

服务: 由京东发货, 并提供售后服务。23:00前完成下单, 预计明天(11月26日)送达

选择系列:  R数据可视化  R语言统计  R语言初学  R语言实战 第2版

白条分期: 30天免息 ¥28.21×3期 ¥14.4×6期 ¥7.49×12期 ¥4.03×24期 ?

1

[✈ 加入购物车](#)

——键购

温馨提示: 1. 支持7天无理由退货

55个卖家在售 **¥64.20** 起



案例背景

全部订单

待付款

待收货

待评价3

我的常购商品NEW


近三个月订单

订单详情

2016-10-19 00:13:49

订单号: 39979501033


京东



Python核心编程 (第3版)

x1

找搭配



金字塔原理大全集 (套装共2册)


x1

找搭配

2016-09-10 18:02:42

订单号: 23122141663

河北



小米 (MI) 5000Ah移动电源超薄锂聚合物手机通用充电宝 银色 5000电源+橙色

x1

找搭配

2016-08-02 20:40:18

订单号: 22561711070

吉吉

商品评价

99%

好评度

好评(99%)

中评(1%)

差评(0%)

全部评价(400+)

晒图(96)

好评(400+)

中评(4)

差评(1)

只看当前商品评价


★★★★★


学习R的入门基础书, 不错的选择!


购买21天后评价


2016-08-24 12:47

R语言实战 第2版










点赞(0) 回复(0)

★★★★★

通过这个R语言, 学习数据分析, 满满看


猜你喜欢



用Python写网络爬虫

¥37.70


(已有17652人评价)



利用Python进行数据分析

¥75.70


(已有2968人评价)



Python绝技: 运用Python成为顶级黑客

¥66.20

(已有2820人评价)



影响力 (经典)

¥35.50

(已有41656人评价)

案例背景

- 常见推荐方式
- 热点推荐
- 经常一起购买的产品：打包销售
- 购买此产品的顾客同时也购买了：协同过滤 - 显式需求
- 看过此商品后顾客购买的其他商品：协同过滤 - 隐式需求
- 用户评论（打分）列表
- 亚马逊20%~30%销售额来自推荐系统
- 几乎所有大型电子商务网站10%以上销售额来自推荐系统

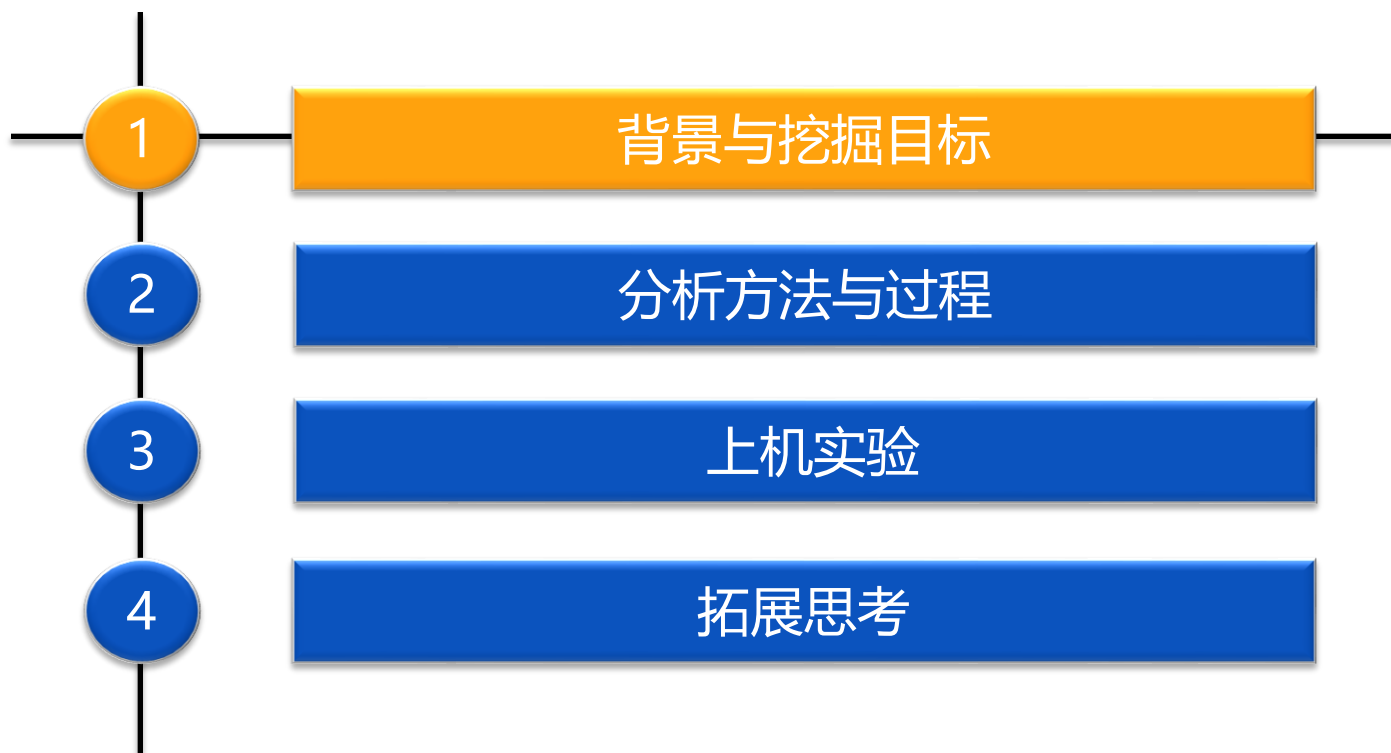


分析方法与过程

用户行为	类型	特征	作用
评分	显式	整数量化的偏好，可能的取值是[0, n]	通过用户对物品的评分，可以精确得到用户的偏好
投票	显式	布尔量化的偏好，取值是0或1	通过用户对物品的投票，可以较精确地得到用户地偏好
转发	显式	布尔量化的偏好，取值是0或1	通过用户对物品的投票，可以精确得到用户的偏好。如果是站内，同时可以推理得到被转发人的偏好（不精确）
保存书签	显式	布尔量化的偏好，取值是0或1	通过用户对物品的投票，可以精确得到用户的偏好
标记书签(Tag)	显式	一些单词，需要对单词进行分析，得到偏好	通过分析用户的标签，可以得到用户对项目的理解，同时可以分析出用户的情感：喜欢还是讨厌
评论	显式	一段文字，需要进行文本分析，得到偏好	通过分析用户的评论，可以得到用户的情感：喜欢还是讨厌
单击流（查看）	隐式	一组用户的点击，用户对物品感兴趣，需要进行分析，得到偏好	用户的单击一定程度上反映了用户的注意力，所以它也可以从一定程度上反应用户的偏好
页面停留时间	隐式	一组时间信息，噪声大，需要去噪，分析得到偏好	用户的页面停留时间一定程度上反映了用户的注意力和偏好，但噪声偏大，不好利用
购买	显式	布尔量化的偏好，取值是0或1	用户的购买行为很明确地说明他对这个项目感兴趣



目录



背景与挖掘目标

某法律网站是一家大型的法律资讯信息网站，它一直致力于为用户提供丰富的法律资讯信息与专业法律咨询服务，并为律师与律师事务所提供卓有成效的互联网整合营销解决方案。



中国法律资讯门户
[城市切换]全国

咨询

请输入您要搜索的关键词.....

马上搜索

免费咨询

地区站点：广东站 广州站 深圳站 珠海站 汕头站 韶关站 河源站 梅州站 惠州站 汕尾站 东莞站 中山站 江门站 佛山站 阳江站 湛江站 茂名站 肇庆站 云浮

找律师 按地区 按专业 按律所

法律咨询 HOT 咨询列表 咨询标签 提交咨询

法律知识 婚姻知识 劳动知识 房产知识 法律法规 民事法规 刑事法规 经济法规

公检法 立法动态 法院资讯 司法行政 案件委托 发布委托 法律协助 委托信息

法律资讯 法制要闻 法律生活 法律时评 律师聚焦 律界热点 新案关注 立法草案

专业频道：婚姻法 刑法 劳动法 房地产法 合同法 公司法 交通事故 破产法 合同范本 保险法 著作权法 商标法 专利法 继承法 民事诉讼 刑事诉讼 更多

向全国十万律师提问

请在此输入您的咨询内容，祝您问题早日得到解决!

解答咨询量：1039855条

提交咨询

12分钟前 李佳律师 回复了法律咨询

12分钟前 喻远军律师 回复了法律咨询

广州 肖艳平律师
TEL：185-2064-8558

广州 练武律师
法律咨询：13416225037

广州 梁火榮律师
法律咨询：13480258386



中国最大的法律资讯与律师门户
the award-winning website of law

许胜利律师 TEL：135-3380-6942
承办过大量各类典型疑难案件，积累了丰富的办案经验。

背景与挖掘目标

目前网站上已经存在部分推荐，比如：当访问主页时可以在婚姻栏目发现如下热点推荐。

婚姻法热文

- 01 协议离婚后 反悔 12-27
- 02 离婚后财产纠纷案例及依据 12-27
- 03 离婚时分割房产的几个问题 12-27
- 04 明智女人选择婚前协议 12-27
- 05 我想离婚，但不知道怎么办， 12-27
- 06 分居两年能离婚吗 12-27
- 07 结婚的特征是什么？ 12-27
- 08 罪犯在被管制或缓刑期间能否结 12-27
- 09 婚姻关系存续期间，夫妻一方以 12-27
- 10 事实婚姻还是非法同居 11-30

婚姻法律咨询

我想离婚、要到儿子的抚养权

- 家庭暴力
- 彩礼
- 男方有外遇，提出离婚，财产怎么分割？小孩判给谁？

婚姻法律知识

- 重婚罪 | 重婚罪的犯罪嫌疑人追究刑事责任的程序有两种
- 抚养费 | 抚养费的支付标准和支付方式
- 彩礼 | 彩礼应否返还受关注 应返还的情形
- 感情破裂 | 离婚时如何认定系夫妻感情破裂

当访问具体的知识页面时，可以在页面的右边以及下面发现也存在一些热点推荐和基于内容的关键字推荐

协议离婚后 反悔

作者: JONER 来源: 未知 2013-12-27 18:24

离婚双方在婚姻登记处办理登记手续后，一方又反悔的，能否向法院起诉？这种情形下，一般可作如下几种判断：

- 1、一方又不同意离婚的，法院不予受理。双方要恢复婚姻关系的，重新向婚姻登记处申请复婚登记。
- 2、一方对离婚协议中的内容反悔的，可以在一年以内向人民法院起诉要求撤销，法院在查清签订离婚协议时没有欺诈，胁迫等情形的，裁定驳回起诉。
- 3、一方对子女的要求变更的可以向人民法院起诉。如不具有变更权的正当理由，法院驳回诉讼请求。
- 4、要求增加抚养费的，可以向法院起诉。

相关知识推荐

- 在法国提出离婚会对居留产生不利影响
- 夫妻双方离婚时保险财产如何分割
- 离婚案件中缺席判决的适用
- 公司股份是否应当作为婚前财产进行分割
- 配偶权与离婚精神损害之间的关系
- 离婚必须双方到场吗
- 外地人在北京怎么离婚

推荐阅读

离婚后孩子归谁抚养 有优先条件
离婚两年后，经济条件改善不能成为变更抚养权
判决离婚有哪些法定条件
关于离婚财产分割后逃债的问题【案例详解】
离婚后“夫妻”间给予经济帮助的条件
婚姻登记处办理离婚登记的条件是什么
离婚如何取证

学会主动用法律保护自己的权益
学习法律知识

热文其他离婚知识

- 1 社会抚养费征收程序
- 2 结婚证撕了怎么办
- 3 起诉离婚要提供哪些证据
- 4 非婚生子女户口



背景与挖掘目标

目前情况：

- 目前网页上是基于内容的推荐（通过关键词）以及非个性化推荐。
- 推荐页面位置不明显，长篇的法律知识的末端。

婚姻法司法解释三发布（全文）

2014-11-05 | 作者：z

5096人

核心内容：最新婚姻法司法解释三于2011年7月4日由最高人民法院审判委员会第1525次会议通过，并于2011年8月13日起施行。下面法律快车婚姻法小编为您详细介绍婚姻法司法解释三的内容。

相关文章阅读

咨询专业律师

相关咨询推荐

- 婚姻法司法解释三全文 [来源：国家法规政策]
- 2011新婚姻法全文一婚姻法司法解释三 [来源：婚姻动态]
- 婚姻法司法解释三被误读 [来源：婚姻动态]
- 婚姻法司法解释三发布 多条款涉及房产确权 [来源：婚姻法规]
- 最高法发布19条婚姻法司法解释三 [来源：婚姻法规]
- 婚姻法司法解释三 2014 [来源：婚姻法规]

最新热门法律经验推荐

- | | | |
|----|----------------|--------|
| 1 | 2014驾驶证扣分新规定 | 227952 |
| 2 | 火车上能带酒吗？ | 190369 |
| 3 | 涨工资最新消息2014 | 187821 |
| 4 | 2014年最低工资标准是多 | 187573 |
| 5 | 2014最新劳动法产假规定多 | 164450 |
| 6 | 坐高铁可以带酒吗？ | 152436 |
| 7 | 2014年广东最低工资标准是 | 149284 |
| 8 | 2014天津最低工资标准是多 | 146537 |
| 9 | 2014上海最低工资标准 | 144765 |
| 10 | 江苏最低工资标准2014 | 135710 |



面临以下问题：

- 1.访问用户多，是机会也是瓶颈；
- 2.留住用户，推荐律师；
- 3.自身推荐效果不佳。

背景与挖掘目标

为了能够更好的满足用户需求，依据其网站海量的数据，研究用户的兴趣偏好，分析用户的需求和行为，发现用户的兴趣点，从而引导用户发现自己的信息需求。

挖掘目标：

- 按地域研究用户访问时间、访问内容、访问次数等分析主题，深入了解用户对访问网站的行为和目的及关心的内容。
- 借助大量的用户的访问记录，对不同需求的用户进行相关的服务页面的推荐。



背景与挖掘目标

行为记录

realIP	realAreacode	userAgent	userOS	userID	clientID	timestamp
2683657840	140100	Mozilla/5.0 (Windows NT 5.1) AppleWebKit/537.36 (KHTML...	Windows XP	785022225.1422973265	785022225.1422973265	1422973268278
973705742	140100	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537...	Windows 7	2048326726.1422973286	2048326726.1422973286	1422973268308
3184681075	140100	Mozilla/5.0 (Windows NT 5.1) AppleWebKit/537.36 (KHTML...	Windows XP	1639801603.1422973278	1639801603.1422973278	1422973277375
3184681075	140106	Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 5.1; Tride...	Windows XP	1597050740.1422973305	1597050740.1422973305	1422973282739
2683657840	140100	Mozilla/5.0 (Windows NT 5.1) AppleWebKit/537.36 (KHTML...	Windows XP	785022225.1422973265	785022225.1422973265	1422973290048
207452174	140100	Mozilla/5.0 (Windows NT 5.1) AppleWebKit/537.36 (KHTML...	Windows XP	589522884.1422272394	589522884.1422272394	1422973295258
432282638	140100	Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML...	Windows 7	225597321.1422972988	225597321.1422972988	1422973305001
285097530	140100	Mozilla/5.0 (iPad; CPU OS 8_1_3 like Mac OS X) AppleWebKit...	iOS	2105429197.1422973314	2105429197.1422973314	1422973309154
776247310	140100	Mozilla/5.0 (iPhone; CPU iPhone OS 8_1_1 like Mac OS X; ...	iOS	1577666249.1422457401	1577666249.1422457401	1422973317133
1275347569	140100	Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML...	Windows 7	303317099.1422972785	303317099.1422972785	1422973319480
1768232564	140100	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537...	Windows 7	385670441.1422973098	385670441.1422973098	1422973321566
2891940471	140100	Mozilla/5.0 (Windows NT 5.1) AppleWebKit/537.36 (KHTML...	Windows XP	1468293794.1422972132	1468293794.1422972132	1422973324242
2962015864	140100	Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML...	Windows 7	33209881.1417416558	33209881.1417416558	1422973328316
2977395726	140100	Mozilla/5.0 (Windows NT 6.1; WOW64; Trident/7.0; rv:1...	Windows 7	946774240.1422972832	946774240.1422972832	1422973330427
4207369840	140100	Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML...	Windows 7	256955339.1422876538	256955339.1422876538	1422973333233

原始数据：
1、SQL文件

2、全国6个月数据

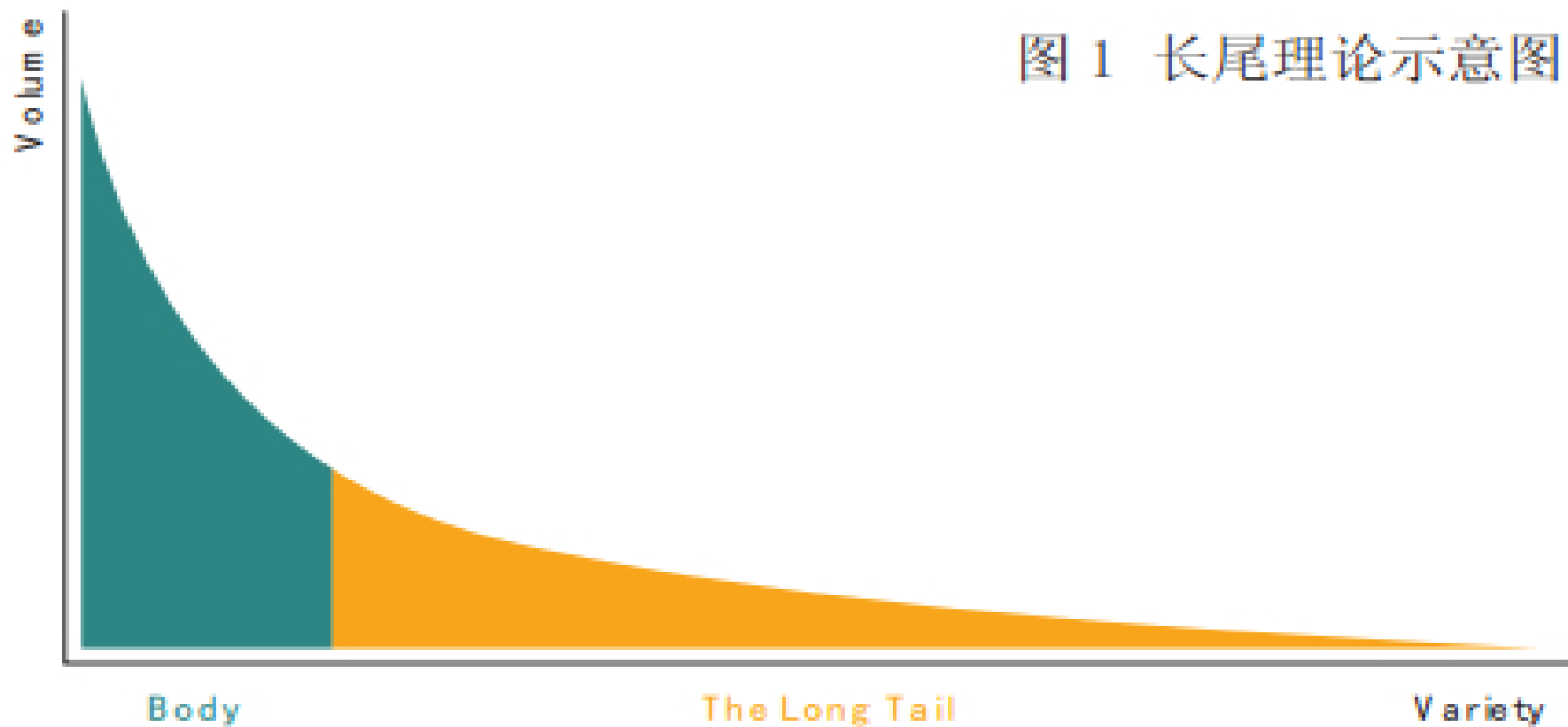
3、数据量250G

背景与挖掘目标

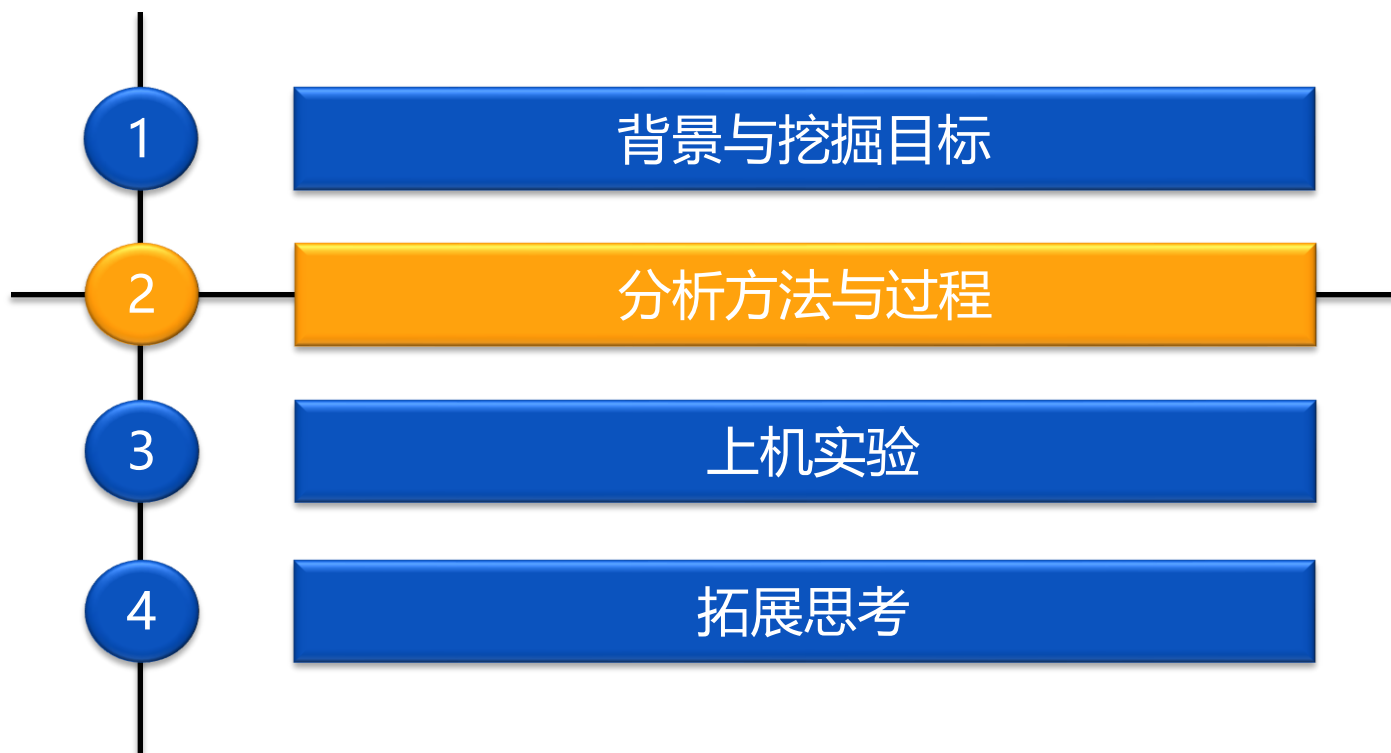
当用户访问网站页面时，系统会记录用户访问网站的日志，其中记录了用户IP（已做数据脱敏处理）、用户访问的时间、访问内容等多项属性的记录，并针对其中的各个属性进行说明见下表。

属性名称	属性说明	属性名称	属性说明
realIP	真实ip	fullURLId	网址类型
realAreacode	地区编号	hostname	源地址名
userAgent	浏览器代理	pageTitle	网页标题
userOS	用户浏览器类型	pageTitleCategoryId	标题类型ID
userId	用户ID	pageTitleCategoryName	标题类型名称
clientId	客户端ID	pageTitleKw	标题类型关键字
timestamp	时间戳	fullReferrer	入口源
timestamp_format	标准化时间	fullReferrerURL	入口网址
pagePath	路径	organicKeyword	搜索关键字
ymd	年月日	source	搜索源
fullURL	网址		

长尾理论

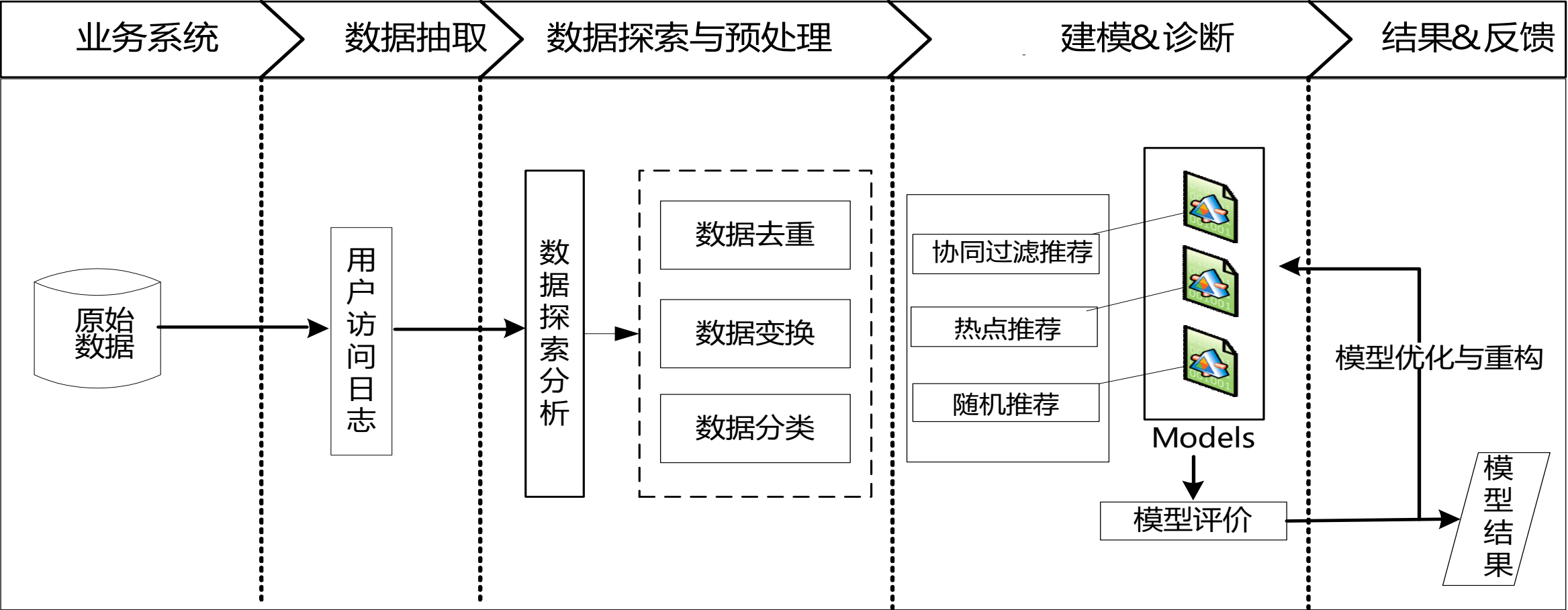


目录



分析方法与过程

总体流程：



分析方法与过程

主要步骤：

- 1.从系统中获取用户访问网站的原始记录。
- 2.对数据进行多维度分析，用户访问时间，用户访问内容，流失用户分析以及用户分群等分析。
- 3.对数据进行预处理，包含数据去重，数据删选，数据分类等处理过程。
- 4.以用户访问html后缀的网页为关键条件，对数据进行处理。
- 5.对比多种推荐算法进行推荐，通过模型评价，得到比较好的智能推荐模型。通过模型对样本数据进行预测，获得推荐结果



1. 数据抽取：

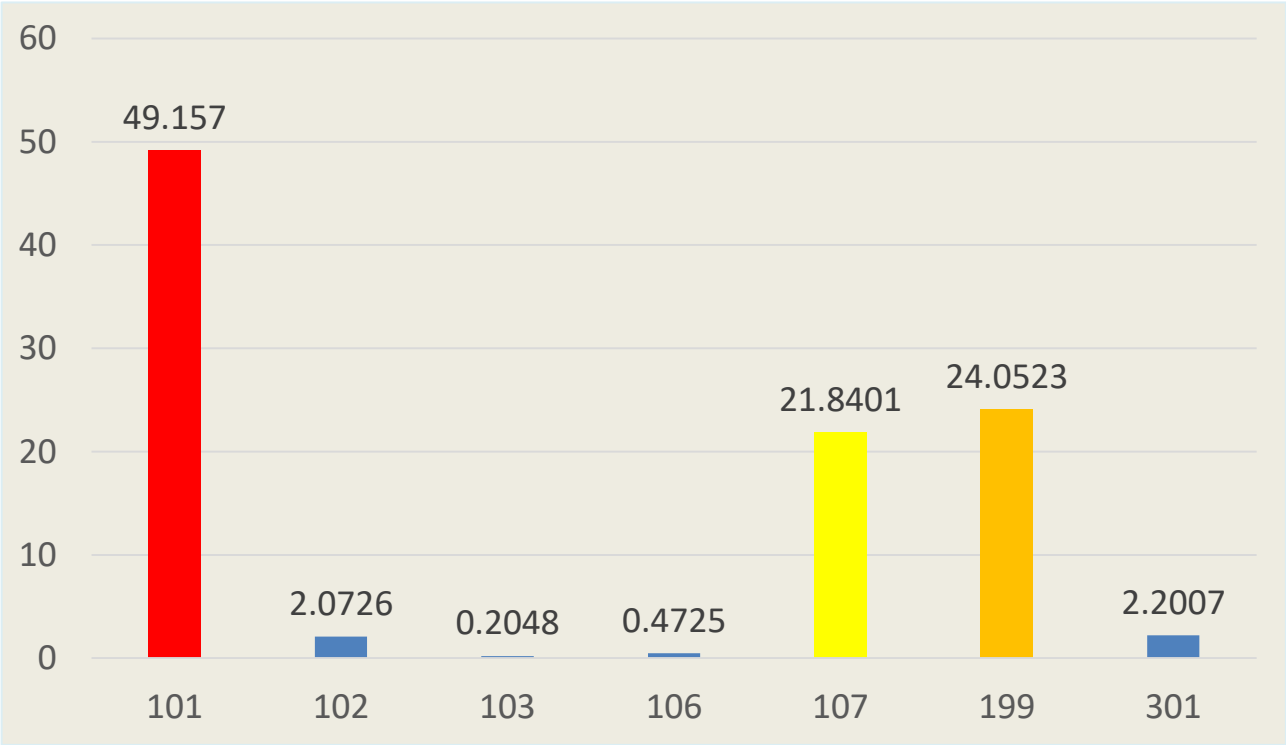
网站每天的访问量有数千万次，非常巨大。为了便于初步分析，选取最近一段时间（2015-02-01~2015-04-29）广州市地区的所有用户访问的详细记录作为原始数据集，总共837450条记录。其中包括用户号、访问时间、来源网站、访问页面、页面标题、来源网页、标签、网页类别、关键词等。

网页类型分析

网页类型	类型说明	记录数	百分比(%)
101	咨询相关	411665	40.1570
199	其他	201426	24.0523
107	知识相关	182900	21.8401
301	法规	18430	2.2007
102	律师相关	17357	2.0726
106	律师事务所	3957	0.4725
103	律师访谈相关	1715	0.2048



网页类型分析



结论：用户浏览的网页大部分为：咨询内容页、知识内容页，法规专题页，咨询经验（在线咨询页）

记录数	百分比	101开头类型
396612	96.3434	101003
7776	1.8889	101002
5603	1.3611	101001
1674	0.4067	其它

记录数	百分比	1999类型
64718	32.13	网址带有?
46993	23.33	Faguizt的网址
36958	19.34	咨询相关
52757	26.2	地区和律师

记录数	百分比	107类型
164239	89.7972	知识内容页
17843	9.76	知识首页
818	0.45	知识列表页

分析方法与过程

网页类型分析

脏数据探索一：对网址分析时，发现网址中存在带 ‘?’ 的情况，一共有65492条记录，占有所有记录的7.8%。

总数	网页ID	百分比
64718	1999001	98.8182
356	301001	0.5436
346	107001	0.5283
47	101003	0.0718
25	102002	0.0382

1999001总数	网页标题	百分比
49894	快车-律师助手	77.0945
6166	免费发布法律咨询	9.5275
5220	咨询发布成功	8.0658
1943	法律快搜	3.0023
1495	其它类型	2.3102

处理：被分享过的网页

http://www.....cn/ask/question_9152354.html?&from=androidqq，对其进行处理，截取?前面的网址，还原其原类型。



分析方法与过程

网页类型分析

1999001数据的处理方式：

- 快车-律师助手这个是针对律师（*区分律师与用户*）；
- 咨询发布成功页面是自动跳转；
- 法律快搜和免费发布法律咨询网址中，*不能直接采用？进行截取*，且大部分网址不是*html形式*，而其本身的类型很混杂，无法还原其原类型，且整个数据集中占比很小，因此可以将这部分数据进行删除；



分析方法与过程

网页类型分析

- 其他类型中的网址情况。存在主网址不包含lawtime关键字的网址，类似的网址为：

<http://www.jifangchina.com/member/Recommend.php>

http://m.sogou.com/ntcweb?g_ut=3&url=http%3A%2F%2Fwww.lawtime.cn%2Finfo%2Fhunyin%2Fhunyinfagui%2F20130115166071_2.html

这类型网址很少101条，因为还原其原来类型很难，故选择删除这类型网页。



分析方法与过程

网页类型分析

脏数据探索二：在记录中，存在一部分这样的用户，他们没有点击具体的网页(.html形式的)，点击的大部分是目录网页，总共有7668条记录。对其进行分析，有一部分是与知识、咨询相关。大部分是地区、律师和事物所相关的。这部分用户很有可能找律师服务的，或者是瞎逛的。在199类型中，律师事物所记录占了1/3，其他大部分是地区的记录。

总数	网页ID
3689	199
1764	102
1079	106
846	107
241	101
49	301

网页类型分析

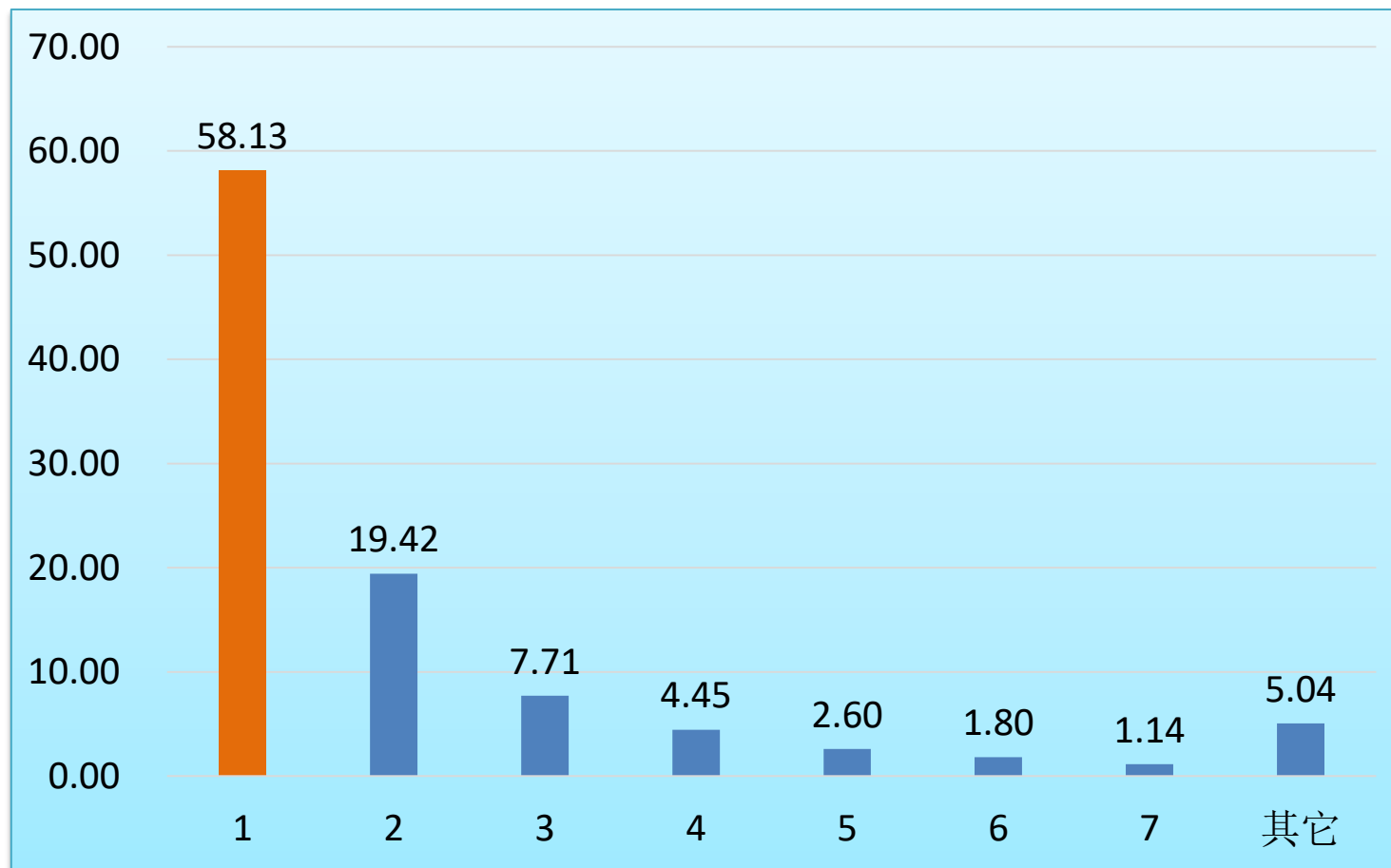
根据上述探索分析，可以设置一些脏数据的规则：

- 1、咨询发布成功页面。
- 2、中间类型网页（带有midques_关键字）。
- 3、？类型中无法还原其本身类型的法律快搜页面与发布法律咨询网页。
- 4、重复数据（同一时间同一用户，访问相同网页）。
- 5、其它类别的数据（主网址不包含lawtime关键字）。
- 6、无点击.html行为的用户记录。
- 7、律师的行为记录（通过法律快车-律师助手判断）

分析方法与过程

点击次数分析

1. 浏览一次的用户占有所有用户58%左右。
2. 大部分用户浏览的次数在2~7次。用户浏览的平均次数是3次



分析方法与过程

点击次数分析

- 1. 浏览一次的用户占有所有用户
58%左右，记录数占20%左右。
20%的用户，为网址提供了接近70%的浏览量
- 2. 点击次数最大值为42790，对其进行分析，是律师的信息，依据律师助手判断。
- 3. 7次以上的情况

点击次数	用户数	用户百分比	记录百分比
1	132084	58.13	20.26
2	44137	19.42	13.54
3	17529	7.71	8.07
4	10112	4.45	6.21
5	5903	2.60	4.53
6	4092	1.80	3.77
7	2597	1.14	2.79
7次以上	12274	5.04	40.84

点击次数	用户数
8~100	12952
101~1000	439
1000以上	19



思考：

用户浏览一次就流失的原因是什么？

如何留住这些点击1 ~ 2次的用户？

分析方法与过程

点击次数分析

- 1. 分析一次浏览次数的用户：问题咨询页，与知识页占比最多。而且这些用户基本上全是搜索引擎进入的。由此可以猜测两种可能：
 - a、有些用户为流失用户，在*问题咨询与知识页*面上没有找到相关的需要。
 - b、用户找到其需要的信息，因此直接退出。
- 2. 如何改善目前这种情况呢？
 - a、用户继续浏览其感兴趣页面
 - b、提供与其相关的页面

网页类型ID	个数	百分比
101003	102560	77.63
107001	19443	14.72
1999001	9381	7.10
301001	515	0.39
其他	202	0.15



网页点击数排名

表中可以看出，点击次数排名前20名中，法规专题占了大部分，其次是法律知识，然后是咨询。

网址	点击数
http://www.....cn/faguizt/23.html	6503
http://www.....cn/info/hunyin/lhlawlhxy/20110707137693.html	4938
http://www.....cn/faguizt/9.html	4562
http://www.....cn/info/shuifa/slb/2012111978933.html	4495
http://www.....cn/faguizt/11.html	3976
http://www.....cn/info/hunyin/lhlawlhxy/20110707137693_2.html	3305
http://www.....cn/faguizt/43.html	3251
http://www.....cn/faguizt/15.html	2718
http://www.....cn/faguizt/117.html	2670
http://www.....cn/faguizt/41.html	2455
http://www.....cn/info/shuifa/slb/2012111978933_2.html	2161
http://www.....cn/faguizt/131.html	1561
http://www.....cn/ask/browse_a1401.html	1305
http://www.....cn/faguizt/21.html	1210
http://www.....cn/ask/exp/13655.html	1060
http://www.....cn/faguizt/39.html	1059
http://www.....cn/faguizt/79.html	916
http://www.....cn/ask/question_925675.html	879
http://www.....cn/faguizt/7.html	845
http://www.....cn/ask/exp/8495.html	726

同网页翻页的点击数

检查发现：同一网页中登录次数最多大部分都是从外部搜索引擎直接入口的网页。平均大概60%~80%的人会选择下一链接，基本每一级都会存在丢失20%~40%的用户，会出现用户衰减的情况。

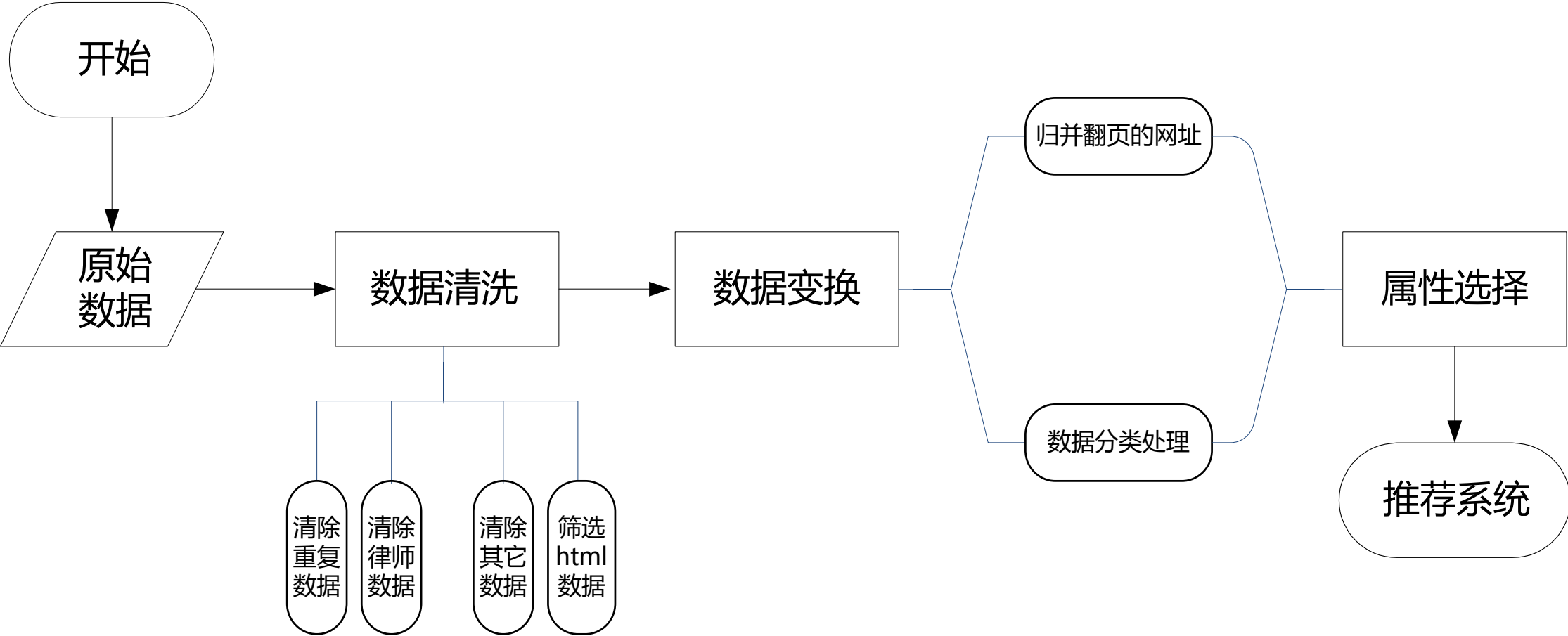
- 1、如果网页需要翻的页数太多，大量的用户基本只会选择浏览了2~5页后，没有搜索到想要的内容，直接就放弃此次的搜索，很少会选择浏览完全部内容。
- 2、通过搜索入口，找到需要的帮助。
- 3、如果翻页情况比较小，可以建议设置全页显示，知识页面无此功能

网址	点击数	比率
http://www.....cn/info/gongsi/slbgzcdj/201312312876742.html	243	
http://www.....cn/info/gongsi/slbgzcdj/201312312876742_2.html	190	0.782
http://www.....cn/info/hetong/ldht/201311152872128.html	197	0.468
http://www.....cn/info/hetong/ldht/201311152872128_3.html	293	0.696
http://www.....cn/info/hetong/ldht/201311152872128_4.html	180	0.614



分析方法与过程

数据处理流程：



2. 数据预处理—数据清洗：

针对上述归纳的脏数据类型，对原始数据进行数据清洗。

删除数据类型	删除数据记录	原始数据记录	百分比
中间类型网页（带midques_关键字）	2036	837450	0.24
(快车-律师助手)律师的浏览信息	185437	837450	22.14
咨询发布成功	4819	837450	0.58
主网址不包含lawtime关键字	92	837450	0.01
法律快搜与免费发布法律咨询的记录	9982	837450	1.19
其它类别带有？的记录	571	837450	0.07
无.html点击行为的用户记录	7668	837450	0.92
重复记录	25598	837450	3.06



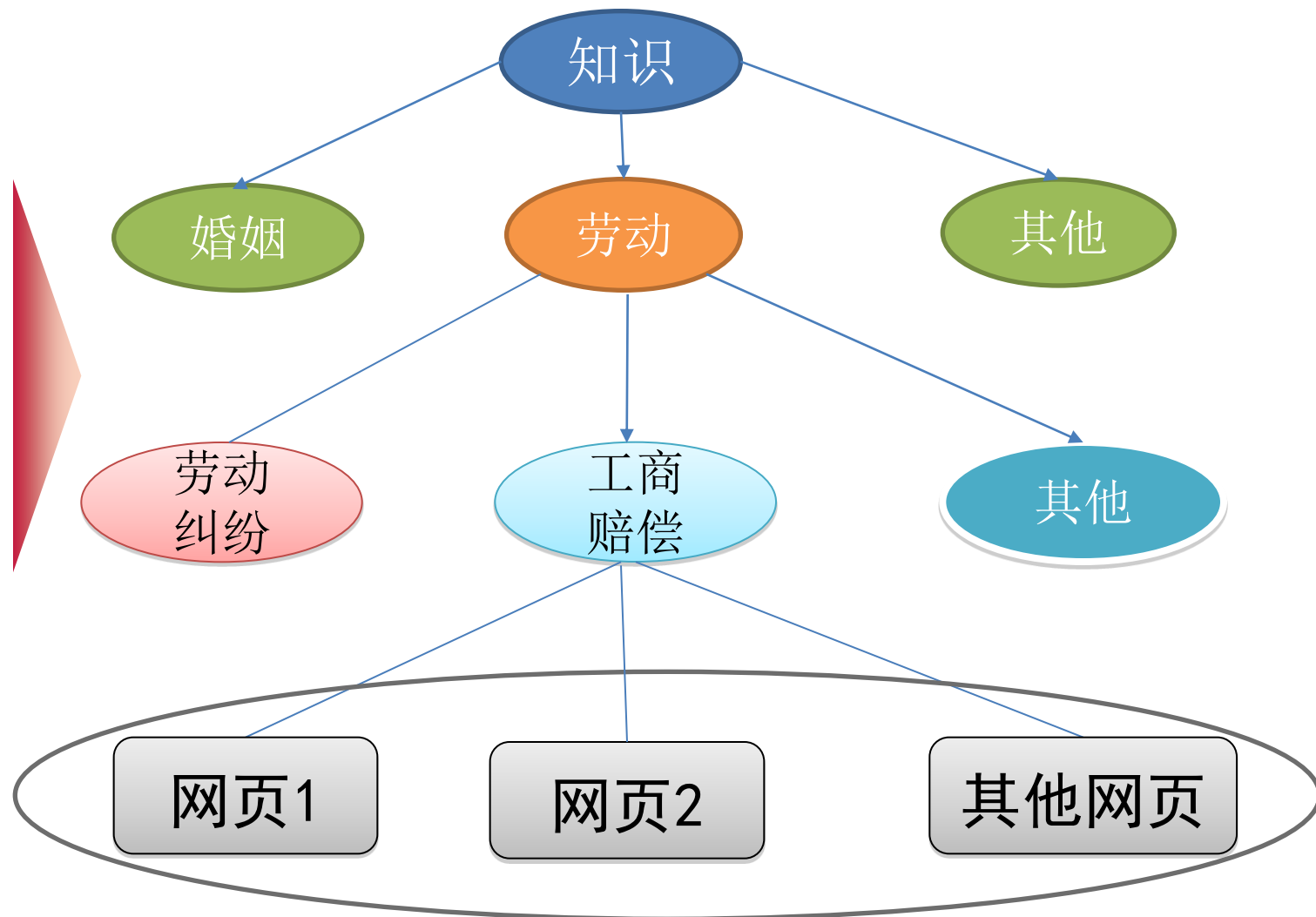
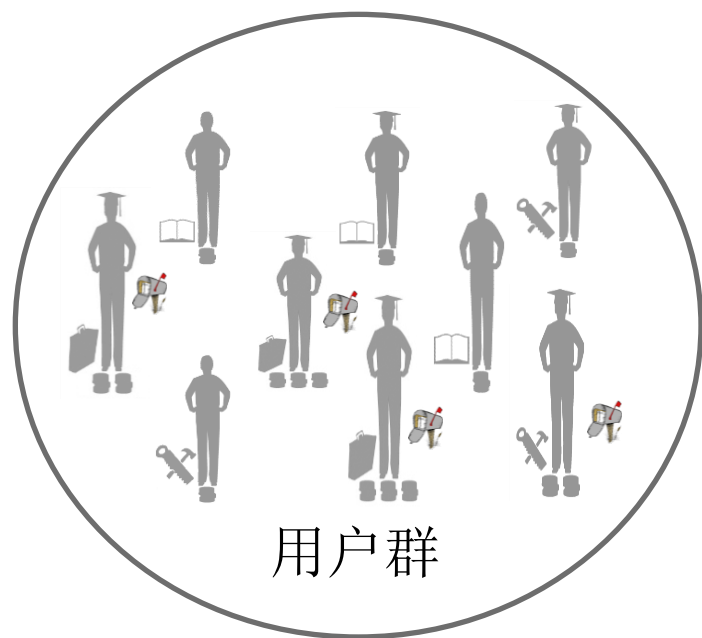
分析方法与过程

2. 数据预处理—数据选择：选择记录中html后缀的记录，并对其进行手动分类。

类型	总记录数	百分比（576722）	说明
ask	384092	66.6%	网址中包含ask,askzt关键字的记录
info	148298	25.7%	知识类
法律法规 (faguizt)	40123	7%	网页ID为301001， faguizt关键字的网址
其他	4209	0.7%	律师访谈、案例、知识库等

分析方法与过程

2. 数据预处理—数据集成：



3. 统计分析

- 选取知识内容中的婚姻类型进行分析。
- 统计知识内容中婚姻中的热门点击排行榜：

网址	内容	点击次数
http://www.....cn/info/hunyin/lhlawlhxy/20110707137693.html	离婚协议书范本（2015年版）	4697
http://www.....cn/info/hunyin/jihuashengyu/20120215163891.html	2015最新产假规定	574
http://www.....cn/info/hunyin/hunyinfagui/201411053308986.html	新婚姻法2015全文	531
http://www.....cn/info/hunyin/jiehun/hunjia/20110920152787.html	广州法定婚假多少天	222
http://www.....cn/info/hunyin/jihuashengyu/201411053308990.html	男人陪产假国家规定2015	211



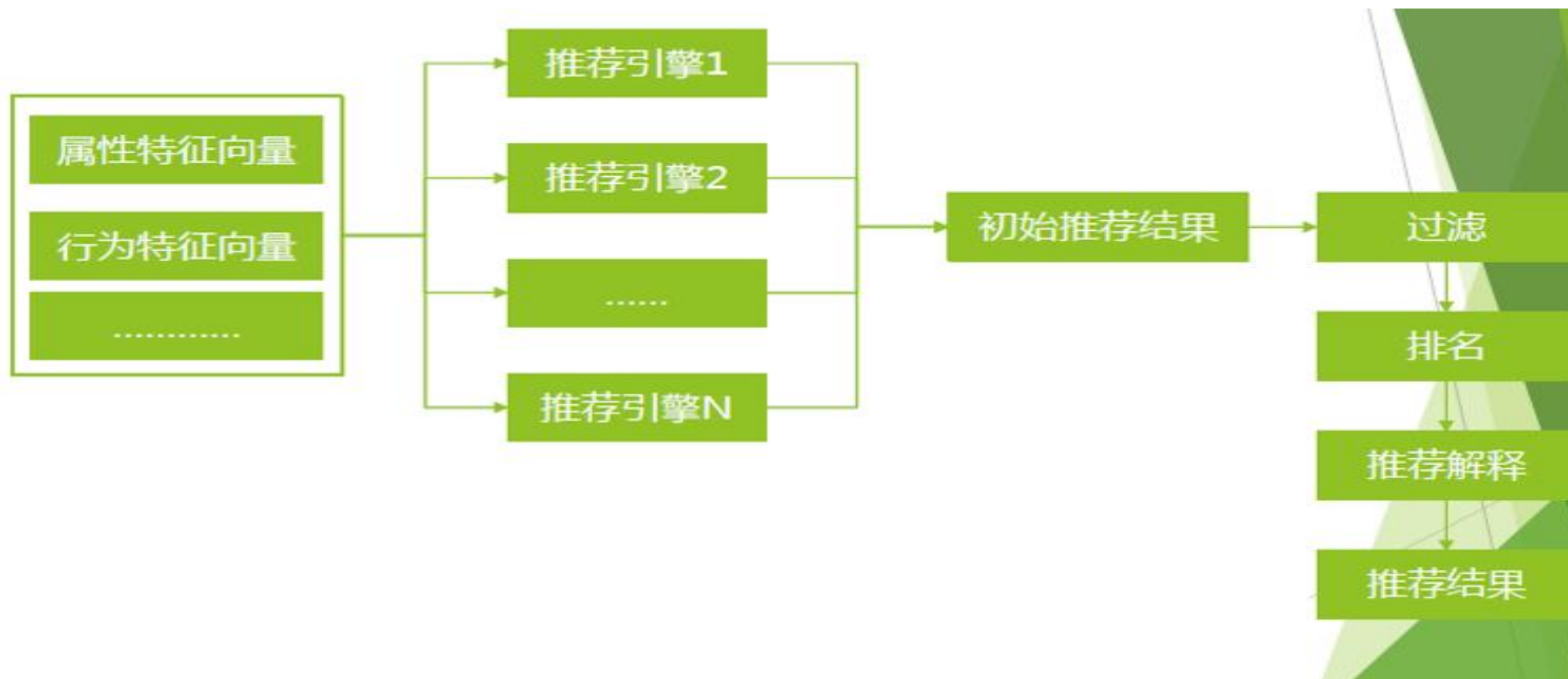
3. 统计分析

- 对其中的点击次数与网页进行分析。

点击次数	网页个数 (3314)	网页百分比	记录数 (16849)	记录百分比
1	1884	56.85	1884	11.18
2	618	18.65	1236	7.34
3	247	7.45	741	4.4
4	151	4.56	604	3.58
5~4679	414	12.49	12384	73.5

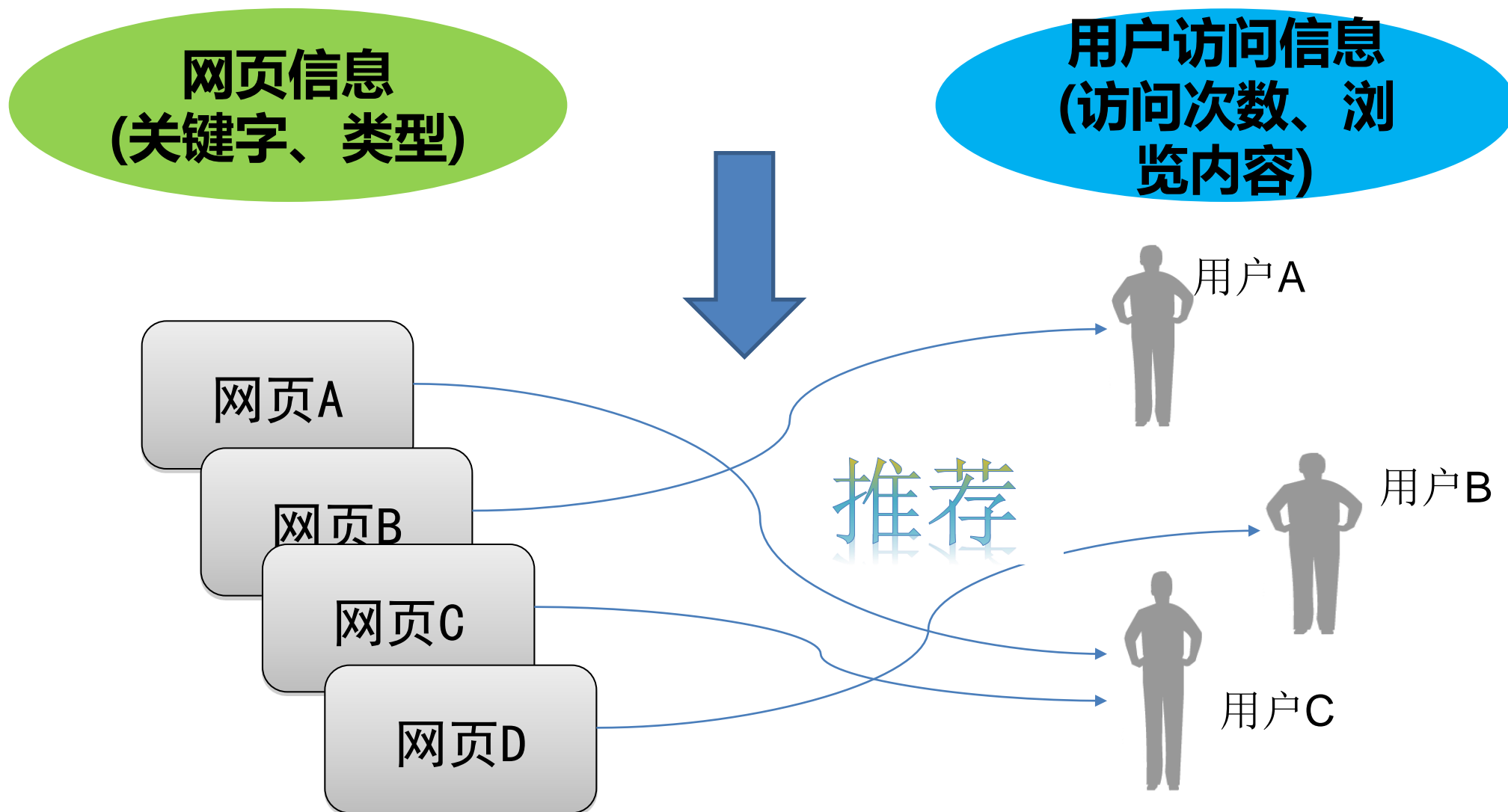
我们可以看出，接近80%的网页只占了浏览量的20%左右，满足二八定律。
因此可以考虑，按点击行为进行分类分析。因其它80%的浏览量来自于比较热点的页面。

模型构建



为了更好的帮助用户从海量的数据中快速发现感兴趣的网页，在目前相对单一的推荐系统上进行补充，采用协同过滤算法进行推荐，其推荐原理如下。

分析方法与过程



由于用户访问网站的数据记录很大，如果对数据不进行分类处理，对所有记录直接采用推荐系统进行推荐，这样会存在以下问题：

- 数据量太大意味着物品数与用户数很多，在模型构建用户与物品的稀疏矩阵时，出现设备内存空间不够的情况，并且模型计算需要消耗大量的时间。
- 用户区别很大，不同的用户关注信息不一样，因此即使能够得到推荐结果，其推荐效果也会不好。。

分析方法与过程

分析思路

数据量大

网址分类

用户分类

模型计算耗时

用户区别大效果不理想

计算快

效果好



分析方法与过程



关联规则：基于支持度和置信度准则，挖掘数据中隐含的规律，得出购买物品A的情况下购买B的概率。

特点：从整体的数据中挖掘潜在关联，与个人的偏好无关，适用于item不多，并且非重度个性化的场景，如超市购物，汽车导购，医疗诊断等。

基于项目的协同过滤

1. 性能：UserCF适用于用户较少的场合，否则计算用户相似度矩阵的代价很大；而ItemCF正好相反，适用于项目数明显小于用户数的情况
2. 实时性：UserCF用户有新行为，不一定造成推荐结果的立即变化；ItemCF正好相反，一定会导致推荐结果的实时变化
3. 长尾物品丰富，用户个性化需求强烈的领域
4. 推荐理由：与UserCF相比，ItemCF利用用户的历史行为给用户做推荐解释，更容易让用户信服

分析方法与过程

基于物品的协同过滤算法主要分为两步：

➤ 计算物品之间的相似度。

计算相似度的方法有：

1、夹角余弦

$$sim_{lm} = \frac{\sum_{k=1}^n x_{k1} x_{km}}{\sqrt{\sum_{k=1}^n x_{k1}^2} \sqrt{\sum_{k=1}^n x_{km}^2}}$$

2、杰卡德 (Jaccard) 相似系数

$$J(A_1, A_M) = \frac{|A_1 \cap A_M|}{|A_1 \cup A_M|}$$

3、相关系数等。

$$sim_{lm} = \frac{\sum_{k=1}^n (x_{k1} - \bar{A}_1)(x_{km} - \bar{A}_M)}{\sqrt{\sum_{k=1}^n (x_{k1} - \bar{A}_1)^2} \sqrt{\sum_{k=1}^n (x_{km} - \bar{A}_M)^2}}$$



分析方法与过程

基于物品的协同过滤算法主要分为两步：

- 根据物品的相似度和用户的历史行为给用户生成推荐列表

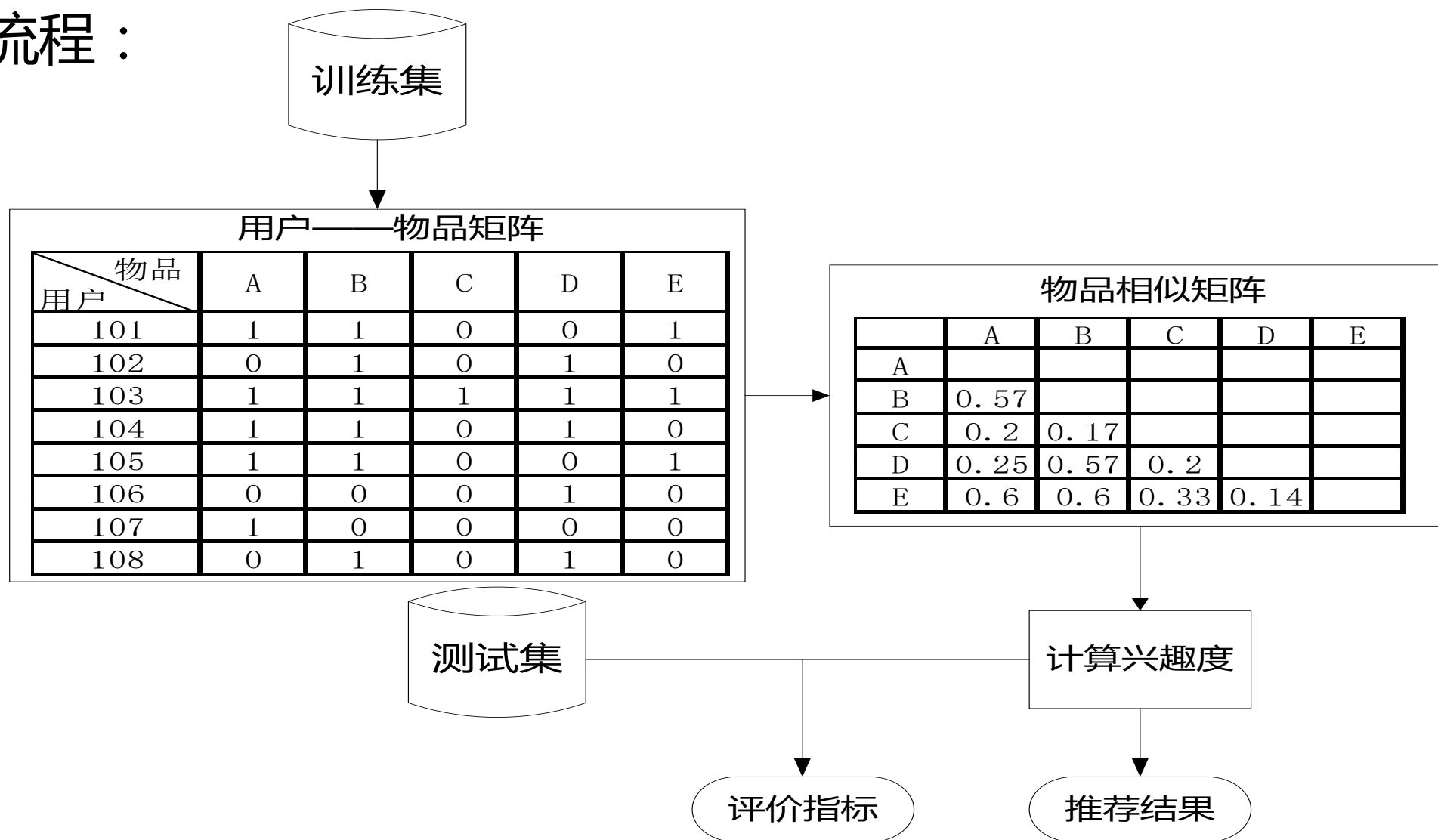
计算兴趣度： $P = R * SIM$

其中R代表了用户对物品的兴趣（即用户是否购买、访问、评分的高低等），SIM代表了所有物品之间的相似度，P为用户对物品兴趣的程度。



分析方法与过程

算法流程：



分析方法与过程

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

		预测	
		推荐物品数(正)	未被推荐物品数 (负)
实际	用户喜欢物品数 (正)	TP	FN
	用户不喜欢物品数 (负)	FP	TN



分析方法与过程

模型评价：准确率与召回率 (Precision & Recall)

Precision 就是检索出来的条目中 (网页) 有多少是准确的。

$$precision = \frac{TP}{TP + FP}$$

Recall就是所有准确的条目有多少被检索出来了。

$$recall = \frac{TP}{TP + FN}$$

		预测	
		推荐物品数 (正)	未被推荐物品数 (负)
实际	用户喜欢物品 数 (正)	TP	FN
	用户不喜欢物 品数 (负)	FP	TN



分析方法与过程

用户	点击网页	推荐网页
116010	"http://www.....cn/info/hunyun/lhlawlhxy/20110707137693.html"	[1] "http://www.....cn/info/hunyun/lihunshouxu/201312042874014.html" [2] "http://www.....cn/info/hunyun/lhlawlhxy/201403182883138.html" [3] "http://www.....cn/info/hunyun/hunyunfagui/201411053308986.html" [4] "http://www.....cn/info/hunyun/jihuashengyu/20120215163891.html" [5] "http://www.....cn/info/hunyun/hynews/201407073018800.html"
11175899	"http://www.....cn/info/hunyun/lhlawlhss/2010120781273.html" "http://www.....cn/info/hunyun/lhlawlhzy/20120821165124.html" "http://www.....cn/info/hunyun/lhlawlhzy/201311292873596.html" "http://www.....cn/info/hunyun/lhlawlhzy/201408253306854.html"	[1] "http://www.....cn/info/hunyun/fuyangyiwu/201404222884700.html" [2] "http://www.....cn/info/hunyun/hunyunfagui/201410153308460.html" [3] "http://www.....cn/info/hunyun/hunyunjiufen/pohuaijunhunzui/20130719167114.html" [4] "http://www.....cn/info/hunyun/jiehuncaili/2011011297291.html" [5] "http://www.....cn/info/hunyun/lhlawlhxy/2011010492149.html"
418673	"http://www.....cn/info/hunyun/lihunfangchan/20110310125984.html"	null

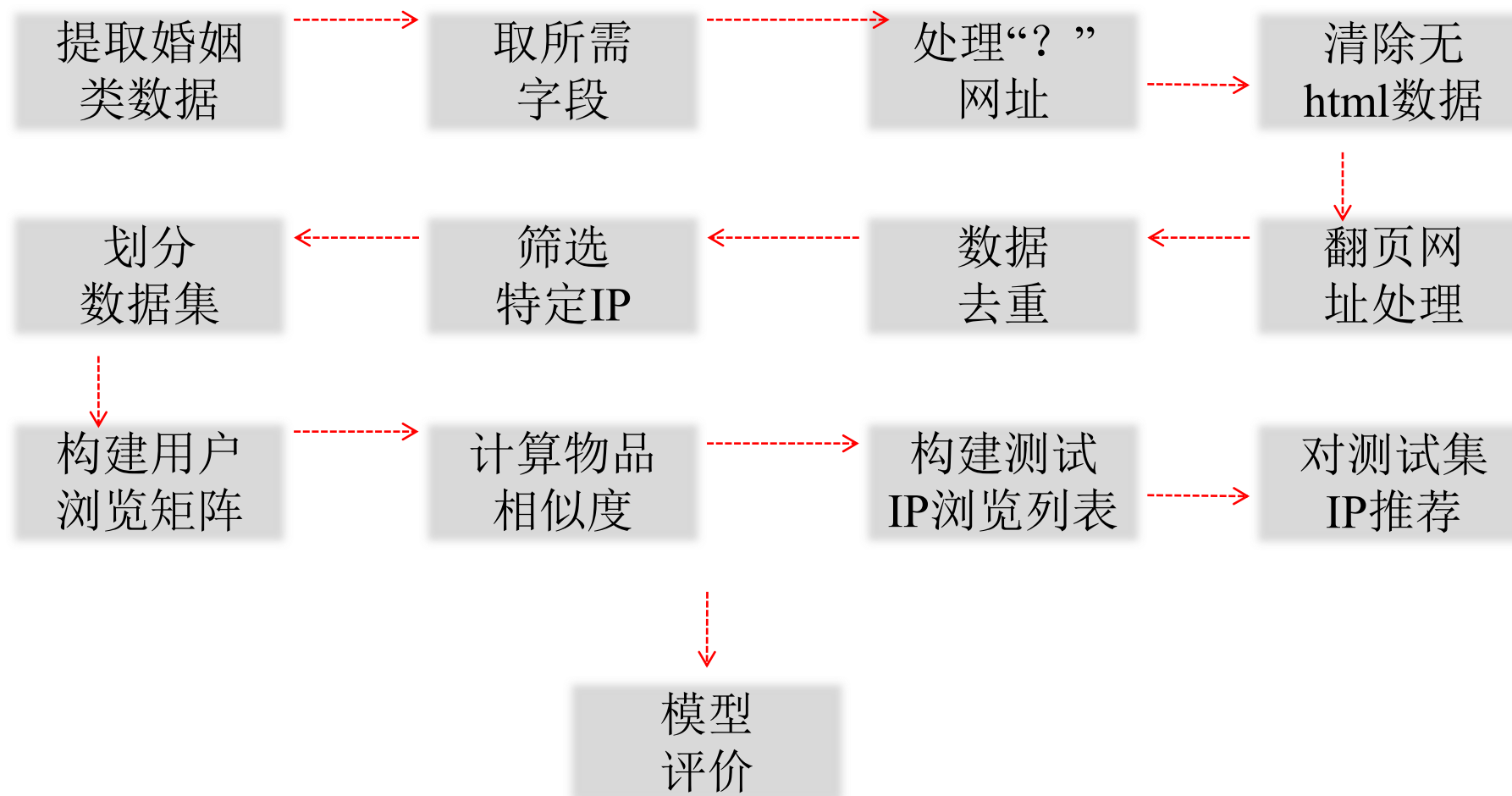


目录



模型构建

➤ 以婚姻数据构建模型



目录





大数据成就未来



Thank you!

泰迪科技 : www.tipdm.com
热线电话 : 40068-40020

