

大数据与质量决策

大数据应用案例 – 质量部



对大数据的共识

获取新的认知，创造新的智慧，产生有价值的决策

全体数据
复杂性
是什么

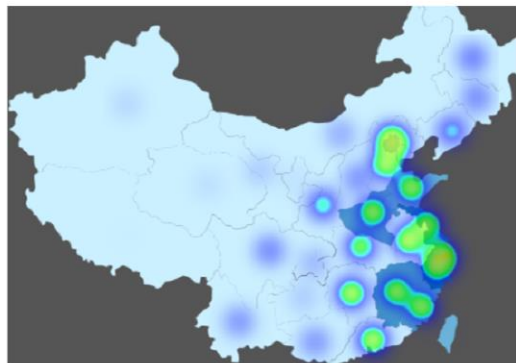
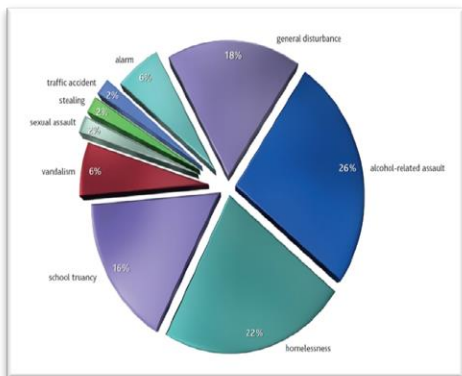
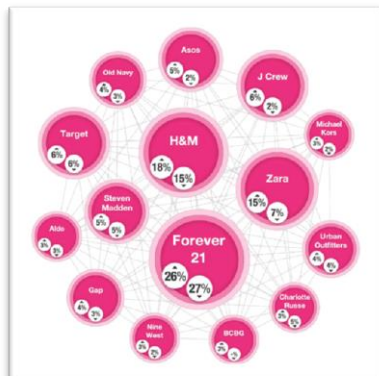
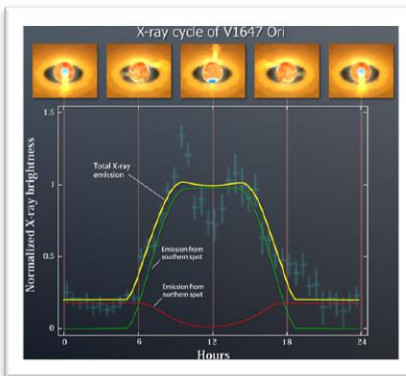
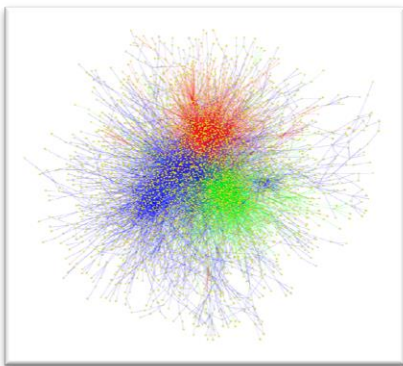
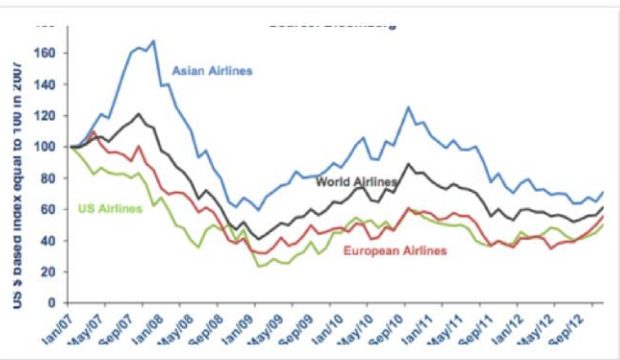
YES

NO

样本数据
精确性
为什么

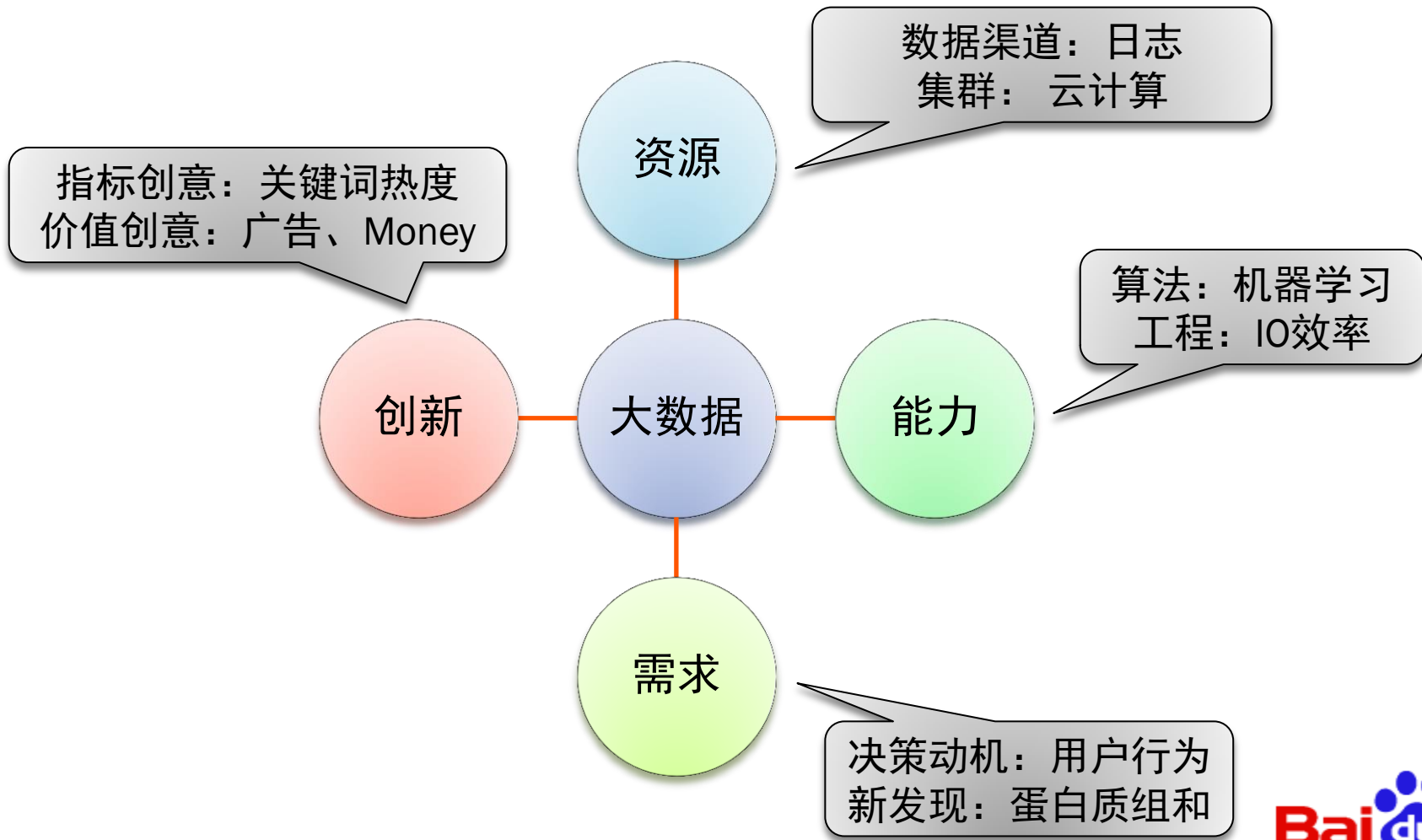
谁让我们的数据变“大”了？ 时间、维度、碎片化

一些大数据应用案例的思考

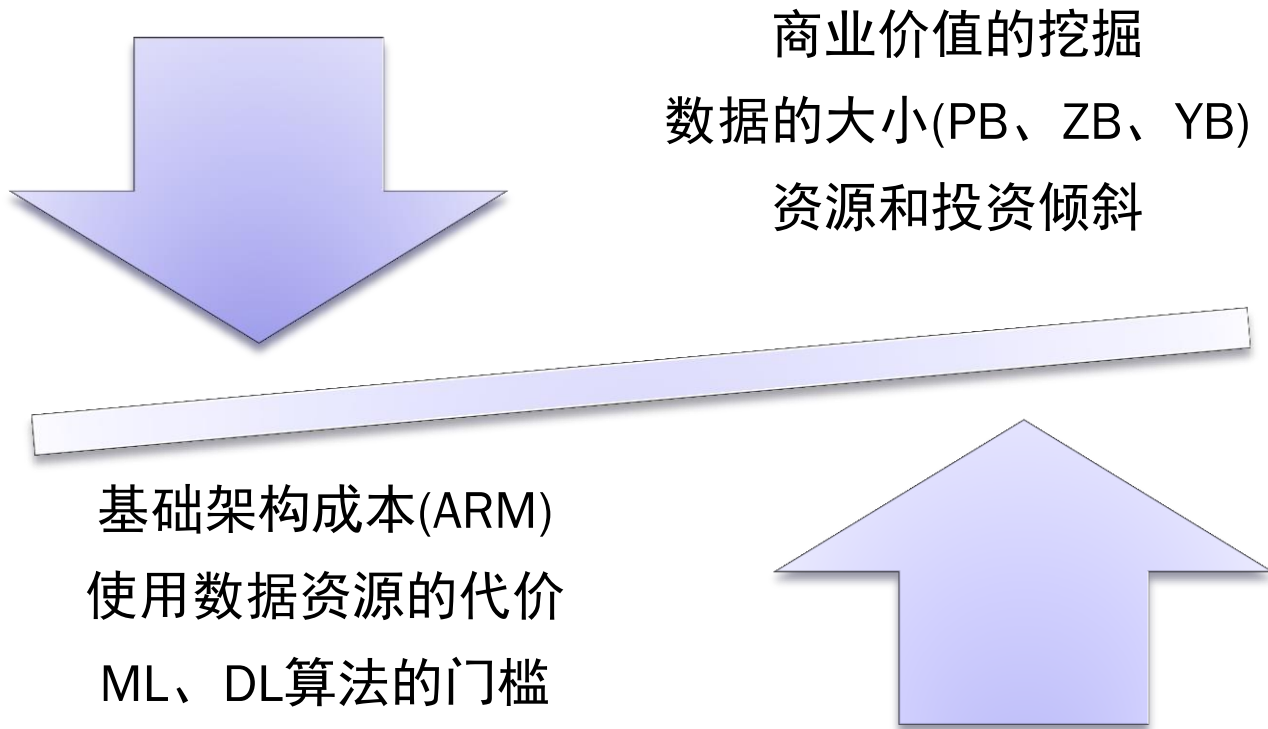


机票价格预测
蛋白质组学
发现横行
青少年犯罪模型
时尚搭配
H7N9

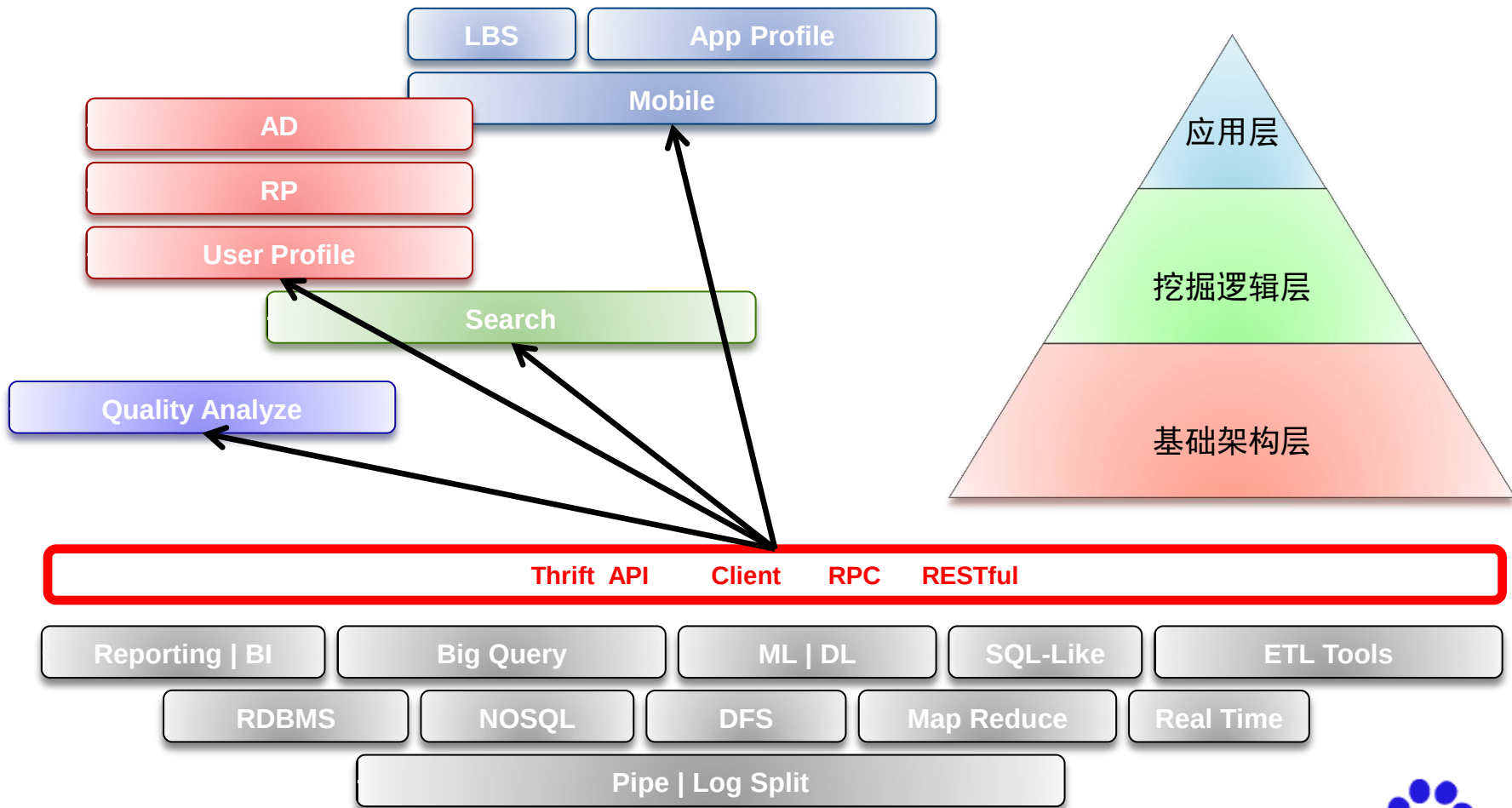
大数据的关键要素



大数据未来



大数据环境



大数据与产品质量视角

基于大数据的产品问题挖掘

- BadCase挖掘、Crash定位、隐性问题与未知Bug

海量用户的体验问题挖掘

- 满意度、情感分析、产品取向与需求预测

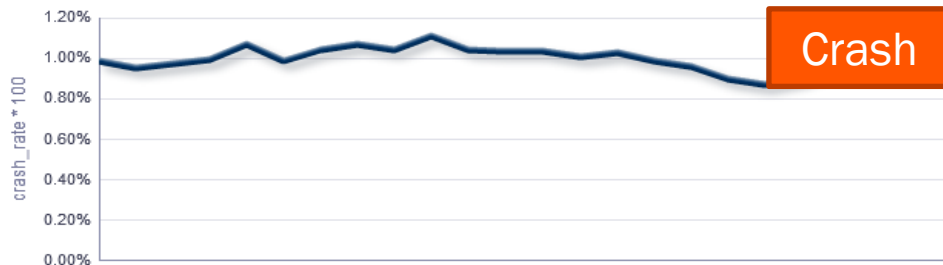
大规模数据的质量

- 死链、图库缺位、云数据保障

App异常监控

version 3-0-4-11

crash_rate * 100



function	exception	cnt
at com.baidu.browser.explorer.BdExploreViewClient.checkWebAppJavaScriptEnable(Unknown Source)	java.lang.NullPointerException	65
at android.util.Pools\$SimplePool.release(Pools.java:112)	java.lang.IllegalStateException	32
at android.os.Stats\$Native.setup(Native Method)	java.lang.IllegalArgumentException	27
at com.baidu.browser.rss.BdRssPictureGalleryLoader.load(Unknown Source)	java.lang.NullPointerException	25
at android.webkit.WebViewCore.contentDraw(WebViewCore.java:1896)	java.lang.NullPointerException	24
at com.baidu.browser.bbm.location.BdLocationWorker.requestLocation(Unknown Source)	java.lang.NullPointerException	24
at android.database.sqlite.SQLiteStatement.native_execute(Native Method)	android.database.sqlite.SQLiteDiskIOException	20
at android.webkit.WebView\$PrivateHandler\$1.run(WebView.java:7603)	java.lang.NullPointerException	16
at android.webkit.WebView\$PrivateHandler\$1.run(WebView.java:7592)	java.lang.NullPointerException	13
at android.webkit.WebView\$PrivateHandler\$1.run(WebView.java:7501)	java.lang.NullPointerException	11
at android.webkit.WebView\$PrivateHandler\$1.run(WebView.java:7518)	java.lang.NullPointerException	10
at android.webkit.WebViewCore.contentDraw(WebViewCore.java:1892)	java.lang.NullPointerException	
at com.baidu.webkit.sdk.BWebView.getHitTestResult(BWebView.java:547)	java.lang.NullPointerException	

Type of Problem Encountered

The mobile app crashed/froze/displayed an error 62%

The mobile app was slow to launch 47%

The mobile app would not launch 40%

The mobile app didn't function as expected 37%

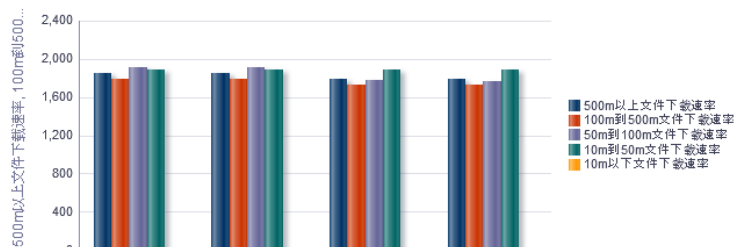
None of these 1%

使用App你能忍受几次闪退？

Online Benchmark

年	日志源	请求类型	500m以上文件	100m到500m文	50m到100m文	10m到50m文	10m以下文件
			下载速率	件下载速率	件下载速率	件下载速率	下载速率
2013	beijing	notice_bcs	1,843	1,788	1,908	1,886	0
		notice_pcs	1,842	1,786	1,902	1,876	0
	nanjing	notice_bcs	1,787	1,727	1,770	1,886	1
		notice_pcs	1,787	1,726	1,765	1,876	1

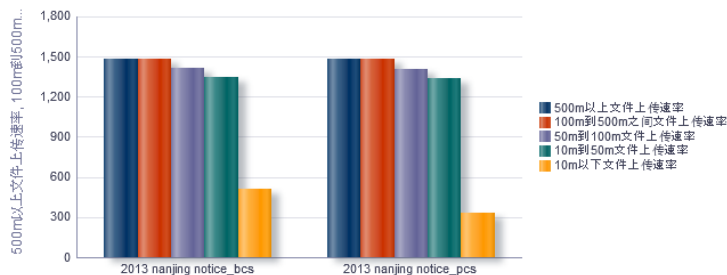
500m以上文件下载速率, 100m到500m文件下载速率, 50m到100m文件下载速率, 10m到50m文件下载速率, 10m以下文件下载速率



version 1.0.0

			500m以上文件 上传速率	100m到500m之间 文件上传速率	50m到100m文 件上传速率	10m到50m文 件上传速率	10m以下文件 上传速率
年	日志 源	请求类 型					
2013	nanjing	notice_bcs	1,481	1,475	1,407	1,340	509
		notice_pcs	1,480	1,474	1,404	1,333	332

500m以上文件上传速率, 100m到500m之间文件上传速率, 50m到100m文件上传速率, 10m到50m文件上传速率, 10m以下文件上传速率



日志

实时传输

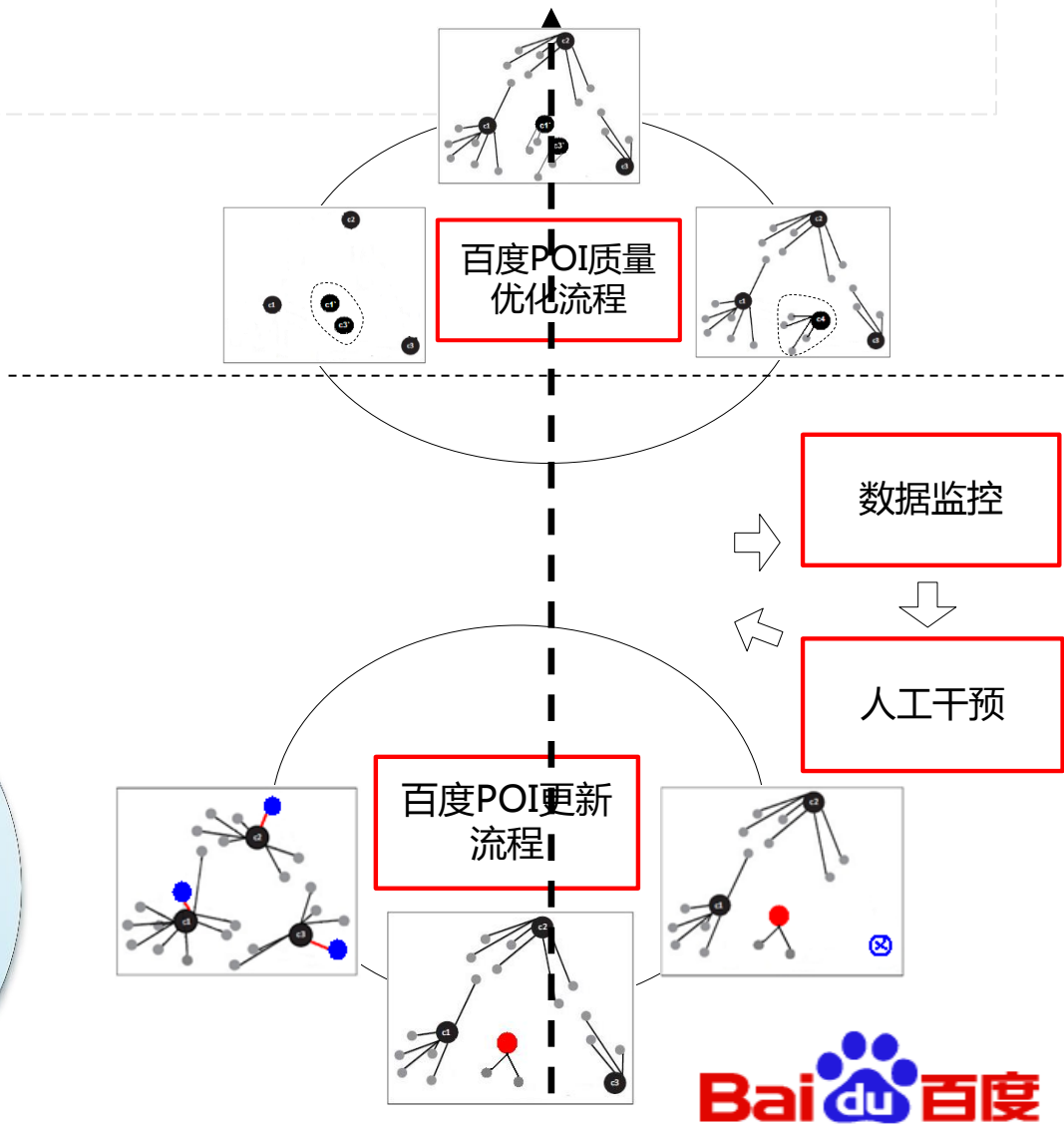
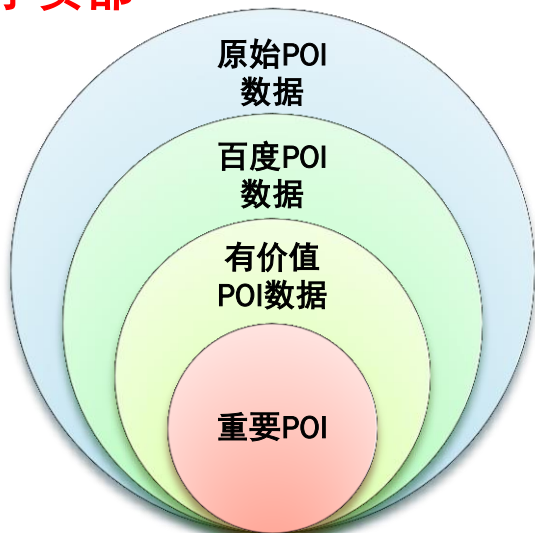
数据清洗

多维模型

LBS与POI数据纠正



回龙观某某小卖部

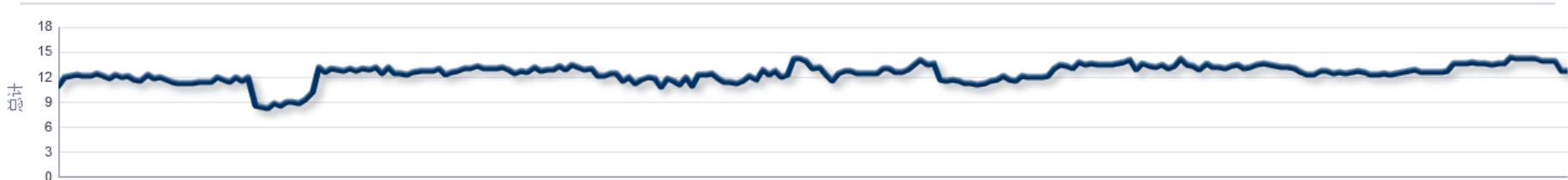




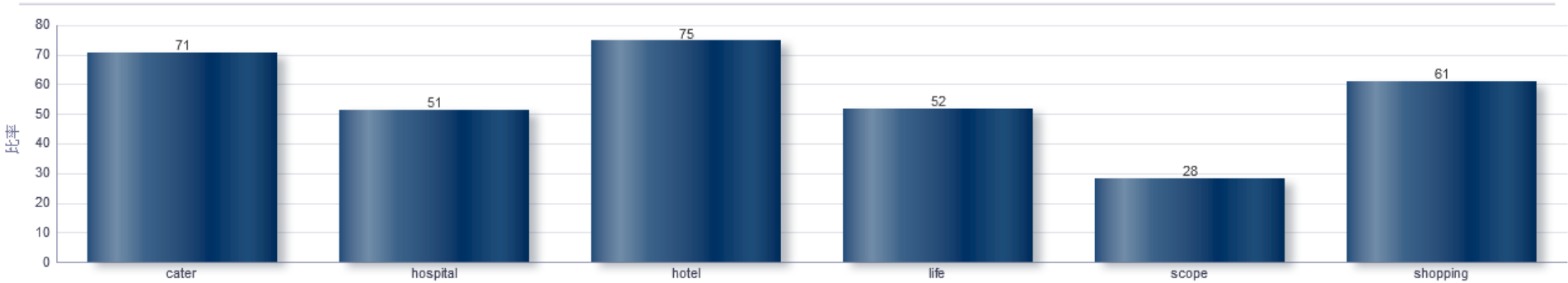
LBS与POI数据纠正

行业 life ▼

平均详情覆盖深度历史曲线



各行业平均详情覆盖率



搜索相关性与Badcase挖掘

[林岚 Tencent Weibo](#)

[林岚](#) Repost: 美容小秘书:【美女养生:洗脸小秘招】1、在水里放上点盐:盐不光可杀菌,还可祛除油脂;2、如果你的脸是干性皮肤,可以在您的脸盆里适当的加一点...

[t.qq.com/linlanxiaoge?lang=en...](#) 2013-1-12 - 百度快照

[正元养生,林岚老师你个臭买药的,DSB - 球迷沙龙 - 北方论坛 - ...](#)

19条回复 - 发帖时间: 2011年2月17日

北方论坛 论坛导航 兴趣圈子 球迷沙龙 正元养生,林岚老师你个臭买药的... 什么TM正元养生,都是骗人的,也不... 铃兰老尸讲一场是300块 打电话的都...

[bbs.enorth.com.cn/thread-3137718-1...](#) 2013-1-3 - 百度快照

[林岚的微博 腾讯微博](#)

[林岚](#) 转播: 修心养生:【胳膊晒黑后让胳膊变白的方法】1:买一瓶白醋,每次... 推荐收听社交金典百科 2月3日 14:04 来自皮皮精灵 全部转播和评论(53) 2...

[t.qq.com/linlanxiaoge?pagesetMyhom...](#) 2012-3-8 - 百度快照

[大家记得不记得一句经典的话 是林岚的妈妈说的 梦里花落知多少吧 ...](#)

1条回复 - 发帖时间: 2006年6月12日

大家记得不记得一句经典的话 是林岚的妈妈说的 只看楼主 收藏 回复 我爱春春_87crz "这是一个女的呀,长的像黄鼠狼似的,看的我腰疼""我看了...

[tieba.baidu.com/p/106401...](#) 2012-10-29 - 百度快照

[林岚老师气血三生法\(转\) 四维风水 新浪博客](#)

更具体的方法请收听长春FM91.6养生讲座,由林岚老师主讲.也可电话咨询4007010927. 老师气血三生法:作者:林岚老师 拍手法:左右手指对手指,掌心对掌心,使劲...

[blog.sina.com.cn/s/blog_71df20850101...](#) 2012-12-16 - 百度快照

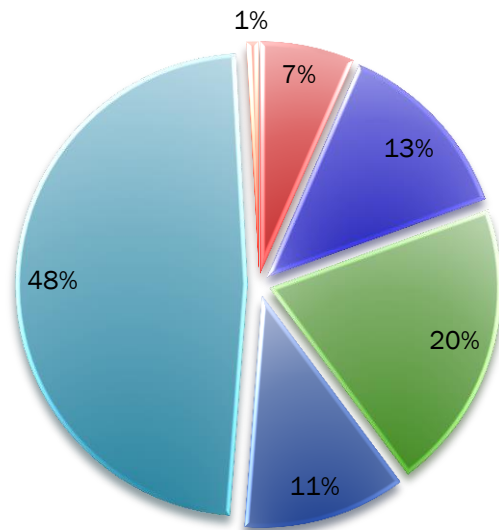
[淘宝店艾米恋旗舰店](#)

[aimilian.tmall.com/](#) 2013-1-23

由于该网站的robots.txt文件存在限制指令,系统无法提供该页面的内容描述 - [了解详情](#)

Query刘艾米完全不相关

相关性bad case类型分布



- 结果召回明显少
- 完全不相关
- 散乱命中
- 多term丢失
- 重要term丢失
- term 顺序改变

Query林岚讲高端养生金典未召回

搜索相关性与Badcase挖掘

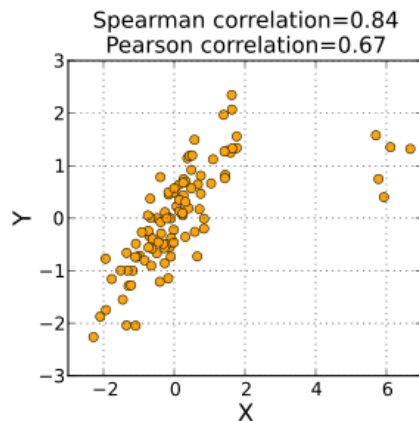
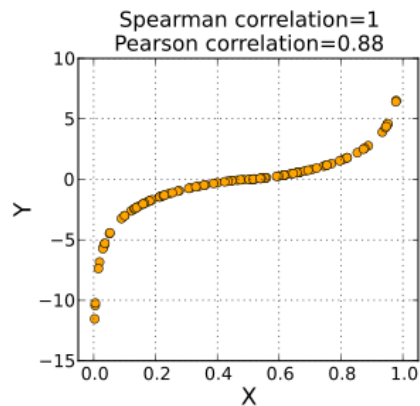
Session切割

Spearman's Foot Rule

加权、标注、迭代模型

算法移植Map Reduce

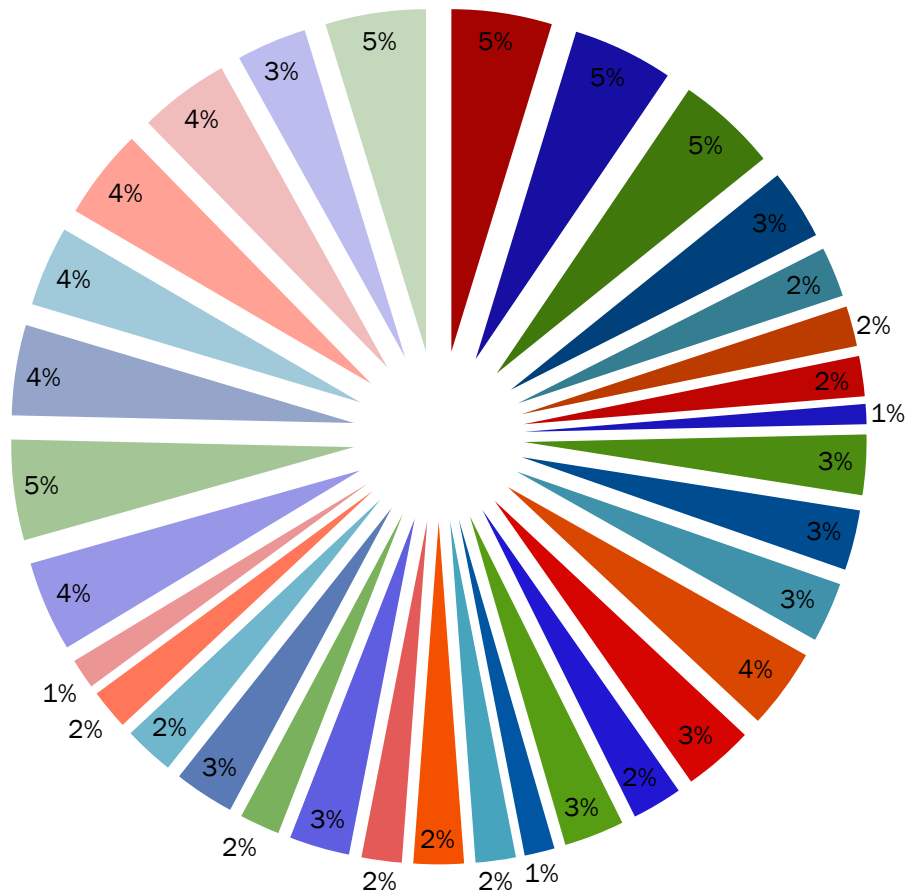
评估准确率和召回率



无线搜索Query满意度



无线搜索Query满意度



- 所在簇中心的置信度为good
- 所在簇中心的置信度为normal
- 所在簇中心的置信度为bad
- 是否session最后一个query
- 下一个query是否有语义包含关系
- 下一个query是否相同
- 下一个query是否有点击
- query的归一化长度
- 同一session内与本query有语义关系的query变换次数
- 平均点击位置
- 是否点击前三
- 逆序点击对占总点击次数
- 翻页次数
- 搜索结果第一页停留时间
- 初次点击发生时间
- 初次点击停留时间
- 最后一次点击停留时间
- 最后一次点击发生时间
- 总平均停留时间
- 是否来自RS

无线搜索Query满意度

每日上亿条访问日志ETL

10 fold cross validation

$T = \{p_1, t_1, o_1, p_2, t_2, o_2, p_3, t_3, o_3, p_4, t_4, o_4, \dots, p_{10}, t_{10}, o_{10}\};$

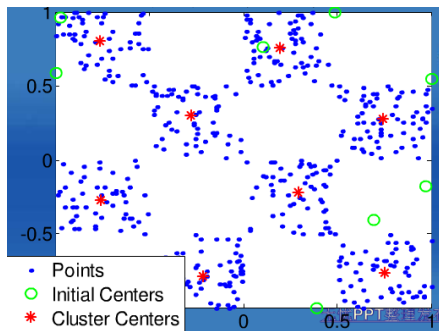
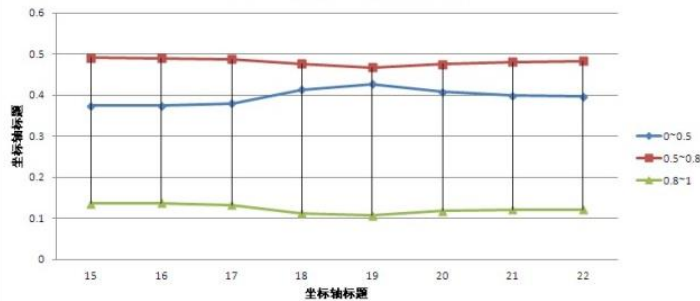
将行为序列信息离散化为0/1特征

$X = \{f_1, f_2, f_3, f_4, f_5, f_6, f_7, \dots, f_n\} = 0/1$

算法移植Map Reduce

评估准确率和召回率

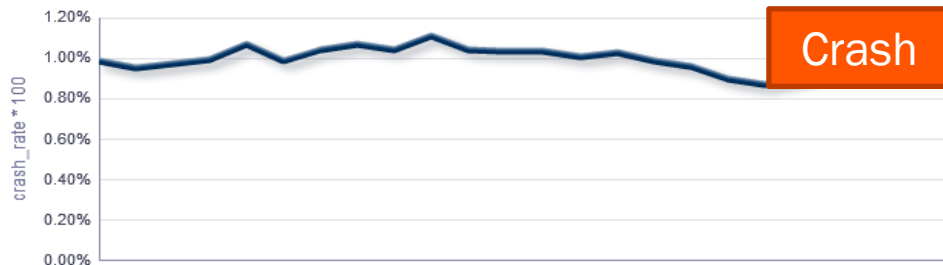
小说一期用户满意度趋势图



App异常监控

version 3-0-4-11

crash_rate * 100



function	exception	cnt
at com.baidu.browser.explorer.BdExploreViewClient.checkWebAppJavaScriptEnable(Unknown Source)	java.lang.NullPointerException	65
at android.util.Pools\$SimplePool.release(Pools.java:112)	java.lang.IllegalStateException	32
at android.os.StatFs.native_setup(Native Method)	java.lang.IllegalArgumentException	27
at com.baidu.browser.rss.BdRssPictureGalleryLoader.load(Unknown Source)	java.lang.NullPointerException	25
at android.webkit.WebViewCore.contentDraw(WebViewCore.java:1896)	java.lang.NullPointerException	24
at com.baidu.browser.bbm.location.BdLocationWorker.requestLocation(Unknown Source)	java.lang.NullPointerException	24
at android.database.sqlite.SQLiteStatement.native_execute(Native Method)	android.database.sqlite.SQLiteDiskIOException	20
at android.webkit.WebView\$PrivateHandler\$1.run(WebView.java:7603)	java.lang.NullPointerException	16
at android.webkit.WebView\$PrivateHandler\$1.run(WebView.java:7592)	java.lang.NullPointerException	13
at android.webkit.WebView\$PrivateHandler\$1.run(WebView.java:7501)	java.lang.NullPointerException	11
at android.webkit.WebView\$PrivateHandler\$1.run(WebView.java:7518)	java.lang.NullPointerException	10
at android.webkit.WebViewCore.contentDraw(WebViewCore.java:1892)	java.lang.NullPointerException	
at com.baidu.webkit.sdk.BWebView.getHitTestResult(BWebView.java:547)	java.lang.NullPointerException	

Type of Problem Encountered

The mobile app crashed/froze/displayed an error 62%

The mobile app was slow to launch 47%

The mobile app would not launch 40%

The mobile app didn't function as expected 37%

None of these 1%

使用App你能忍受几次闪退？

用户反馈数据与产品改进

我是秦皇岛的，怎么一直显示沧州的天气

音量控制翻页加上呗

我读一本，鱼人二代小说。怎么看了一大半，后面没有了？是什么原因。

下方的热点，影视等栏目，最好换成有用户自己选择敢兴趣的信息。

搜索牡丹江天气

太忙太慢了

小说内容太少了

数据抓取

微博，论坛，贴吧，网页库

反馈数据

数据提取

网页分析

结构化

语义分析

情感词典

特征挖掘，褒贬，分类

用户反馈数据与产品改进

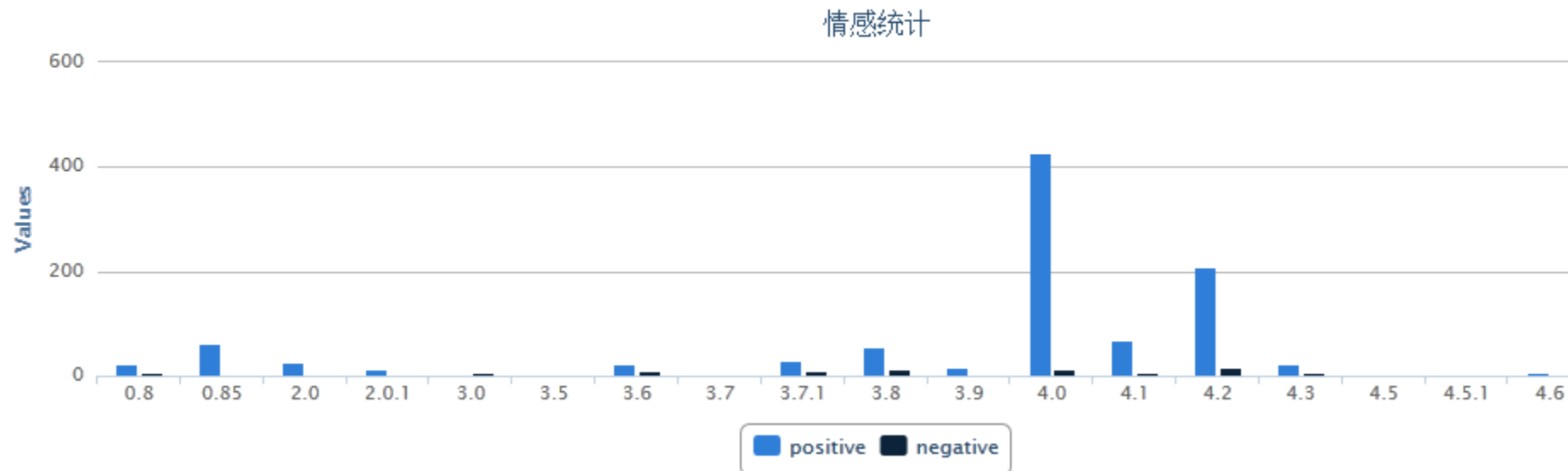
☰ top问题

👤 用户评分

📊 情感统计

👤 多维度统计

百度框Android - 情感统计



数据质量与脏数据监测

Data type
check

Character
check

Limit
check

Uniqueness
check

Logic
check

Range check

Statistics
check

Consistency
check

Existence
check

Data
Zoom

Record
Trace

Dead Link
check

网页数据一致性

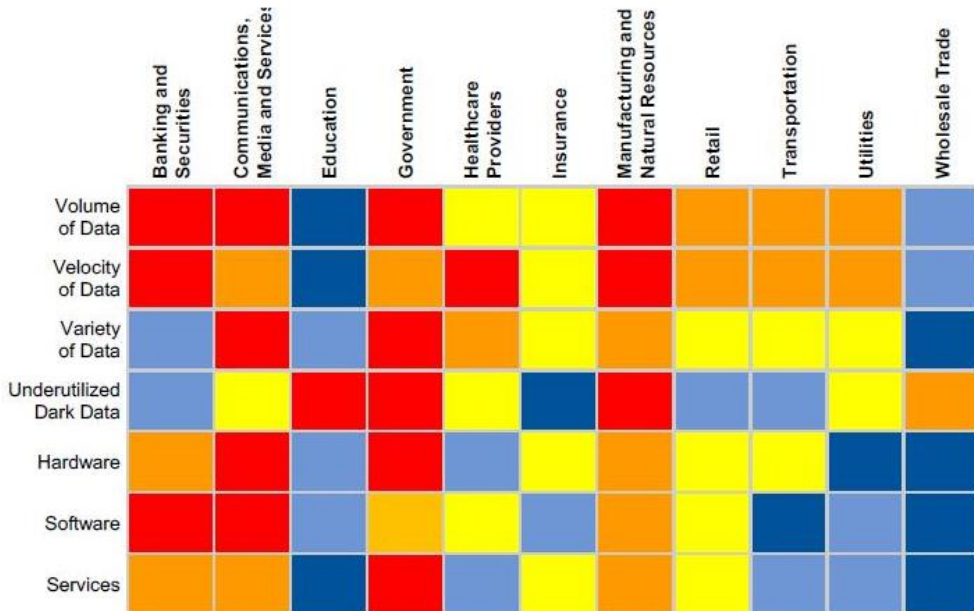
网页数据时效性

位置信息准确性

地图数据完整性

关于大数据的应用

Big Data Opportunity Heat Map by Industry



产品改进决策应用只是冰山一角

复合型人才：基础架构 + 数据挖掘 + 业务和行业知识

百度技术沙龙

畅想

交流

争鸣

聚会

关注我们：t.baidu-tech.com

资料下载和详细介绍：infoq.com/cn/zones/baidu-salon

“畅想·交流·争鸣·聚会”是百度技术沙龙的宗旨。百度技术沙龙是由百度与InfoQ中文站定期组织的线下技术交流活动。目的是让中高端技术人员有一个相对自由的思想交流和交友沟通的平台。主要分讲师分享和OpenSpace两个关键环节，每期只关注一个焦点话题。

讲师分享和现场Q&A让大家了解百度和其他知名网站技术支持的先进实践经验，OpenSpace环节是百度技术沙龙主题的升华和展开，提供一个自由交流的平台。针对当期主题，参与者人人都可以发起话题，展开讨论。

InfoQ 策划·组织·实施

关注我们：weibo.com/infoqchina