火龙果 · 整理
uml.org.cn

大数据思维：新经验主义

火龙果 · 整理
uml.org.cn

大数据的通常定义

- Wikipedia:
  - **Big data** is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications
- 百度百科：
  - 大数据(big data)，或称巨量资料，指的是所涉及的资料量规模巨大到无法透过目前主流软件工具，在合理时间内达到撷取、管理、处理、并整理成为帮助企业经营决策更积极目的的资讯。

- 始终围绕着“大”来定义，没有揭示背后的深层意义

火龙果 · 整理
uml.org.cn

传统思维中的理性主义

- 理性主义
  - 从特殊到一般：相信人能透过现象看到本质
  - 从一般到特殊
- 科学中的实践
  - 从理性或直觉中建立问题的模型，或通过少量样本数据的观察归纳出模型
  - 通过模型判别新样本

火龙果 · 整理
uml.org.cn

为什么需要模型？

- 历史上经验收集和分享的困难
- 模型缓解了经验的不足
  - 从有限的个人经验中得到普遍性的规律
  - 泛化：从已知到未知
- 模型的内在困难
  - 是否总是能从特殊推到一般
  - 复杂模型：股市预测

火龙果 · 整理
uml.org.cn

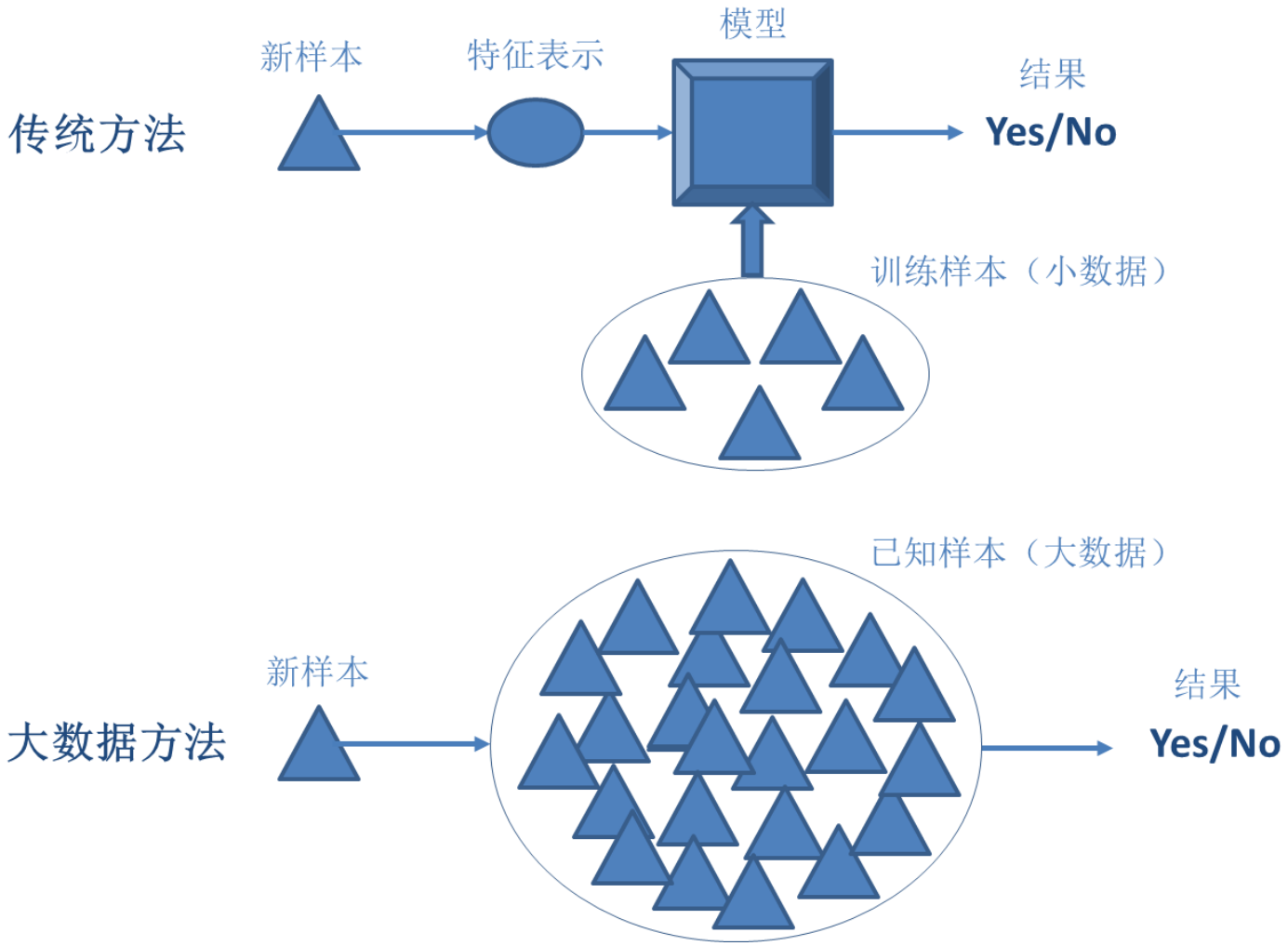
大数据时代

- 新技术使得经验数据的收集和分享变得容易
  - 传感器
  - 互联网
  - 移动设备

- 数据越多，就越不需要模型
  - 覆盖度：对所有或大部分事件，我们有样本来覆盖
  - 精度：对高频事件，我们有足够多样本来提升精度

传统方法 vs. 大数据方法

火龙果 · 整理
uml.org.cn



火龙果·整理
uml.org.cn

例子一：机器翻译

- 问题：将一种语言（如中文）自动翻译为另一种语言（如英文）
- 传统解决方法：语料库+翻译模型
- 大数据方法：平行语料挖掘
  - 从互联网上自动发现大量的双语语料
  - 统计词语、短语、甚至句子之间的对照关系
  - 非常显著的性能提升，目前最好的方法

火龙果 · 整理
uml.org.cn

例子二：查询结果排序

- 问题：给定一个查询，对网页进行相关性排序
- 传统解决方法：排序模型
  - 概率模型
  - 语言模型
  - 神经网络（Learning to rank）
- 大数据方法：用户点击数据挖掘
  - 给定一个查询，根据用户对网页的点击率排序
  - 需要大量的数据：查询数*网页数
- 效果明显提升：最强的feature

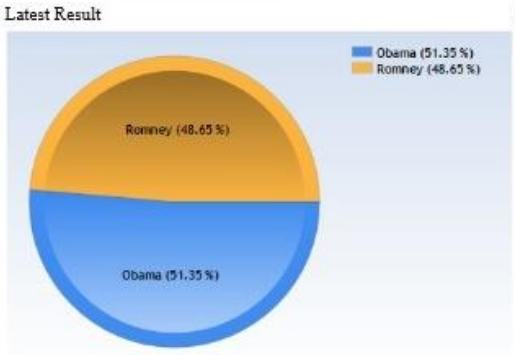
火龙果 · 整理
uml.org.cn

例子三：微软小冰

火龙果 · 整理
uml.org.cn

例子四：预测美国大选

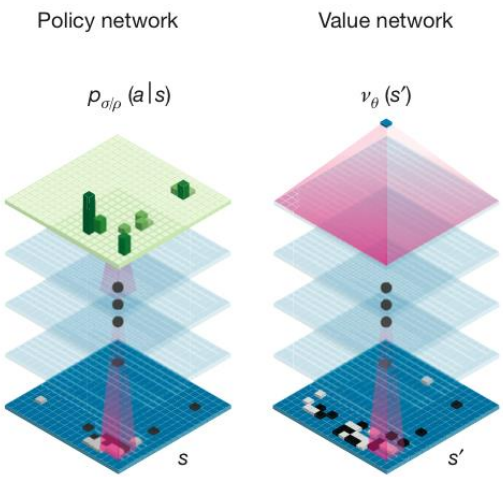
- 问题：2012年美国大选，奥巴马和罗姆尼谁会赢
- 传统的解决方法：民意调查，专家意见
- 大数据方法：网络数据舆情分析
  - http://research.microsoft.com/en-us/projects/websensor/election2012.aspx
  - 从公开的网络数据源（论坛，新闻评论，社交媒体）中收集大量相关数据
  - 分析和统计网民的民意
  - 与真实大选结果非常接近



火龙果 · 整理
uml.org.cn

例子五：AlphaGo

AlphaGo



火龙果 · 整理
uml.org.cn

什么是大数据？

- 大数据是现代社会在掌握海量数据收集、存储和处理技术基础上所产生的一种以群体经验进行判断和预测的**能力**，代表了一种**新经验主义**。

- 内在含义
  - 经验主义 > 理性主义
  - 数据 > 模型
  - 相关关系 > 因果关系
  - 数据 > 平台 > 模型

火龙果 · 整理
uml.org.cn

多大的数据是大数据？

- 当数据多到能对**整个**样本空间进行**充分**覆盖，这样的数据就足够“大”了
  - 对预测掷硬币这种问题，样本空间为{0，1}，1000个样本就足以得到置信度很高的预测
  - 对于机器翻译，样本空间的数量级就大很多：所有可能的句子？

火龙果 · 整理
uml.org.cn

模型真的没有用吗？

- 数据总是不够
  - 样本空间太大
    - 机器翻译例子中所有可能的句子
  - 样本空间变化
    - 查询结果排序例子中，新的查询和新的网页在不停出现
- 模型需要和数据结合，提供适当的泛化能力
  - 如何结合？

火龙果 · 整理
uml.org.cn

大数据应用开发的典型流程

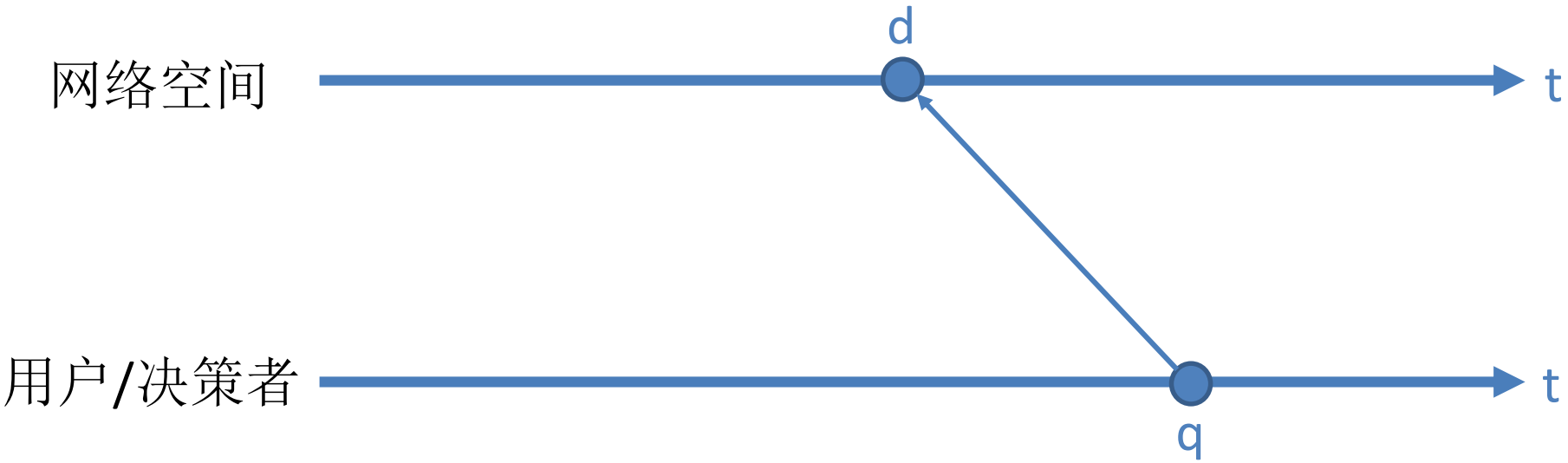
1. 找到一个重要的应用，确定问题的数据类型和样本空间
2. 收集到尽可能多（或足够多）的相关数据来覆盖样本空间
   - 不要特别在意数据质量和格式
3. 选择（或搭建）一个合适的大数据处理平台
4. 针对应用对数据进行预处理
   - 格式转换、数据抽取、数据集成、数据质量控制
5. 处理数据
6. 结果解读和应用

火龙果 · 整理
uml.org.cn

大数据技术：大数据分析引擎

火龙果 · 整理
uml.org.cn

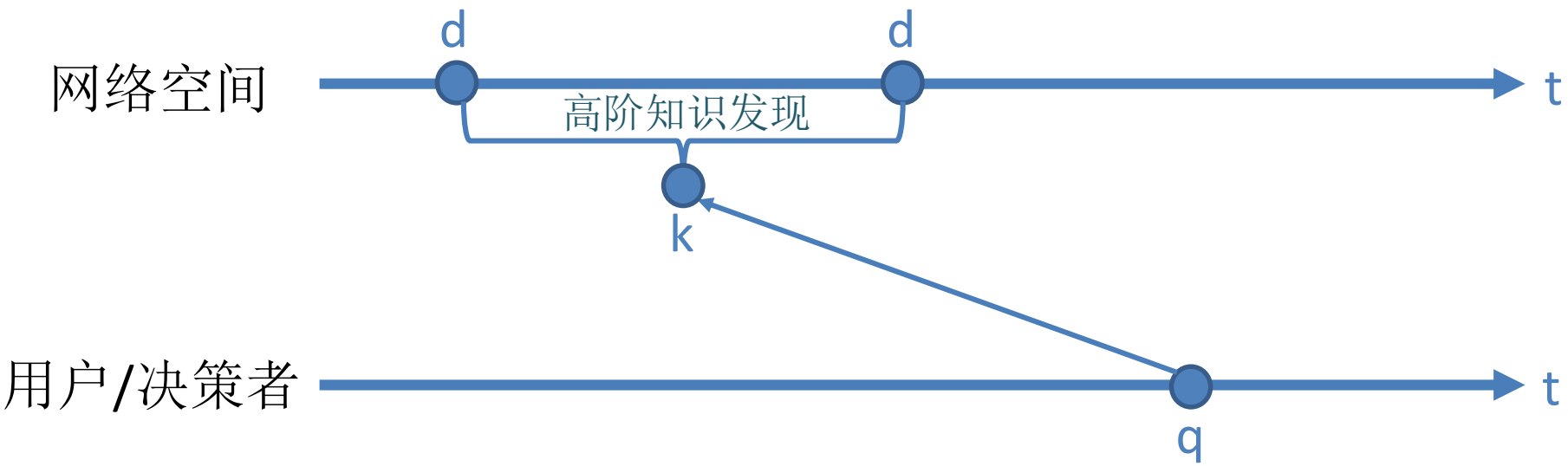
搜索



- **假设一：**文档是基本的检索单位
- 假设二：文档是独立的
- 假设三：文档是静态的
- 假设四：文档是存在的

火龙果 · 整理
uml.org.cn

分析



- **目标：**更好、更全面地感知现在
- 突破假设一（文档是基本的检索单位）、二（文档是独立的）
- **挑战：**
  - 多源异构多噪数据→高阶知识？
  - 数据质量

关系

徐根宝
加入查询 了解更多

黄光裕
加入查询 了解更多

许荣茂
加入查询 了解更多

徐明
加入查询 了解更多

李彦宏 年度人物
加入查询 了解更多

王石
加入查询 了解更多

杨惠妍
加入查询 了解更多

卢志强
加入查询 了解更多

梁稳根 年度人物
加入查询 了解更多

张近东
加入查询 了解更多

更多

资讯

最近资讯 订阅 王健林 的最近资讯

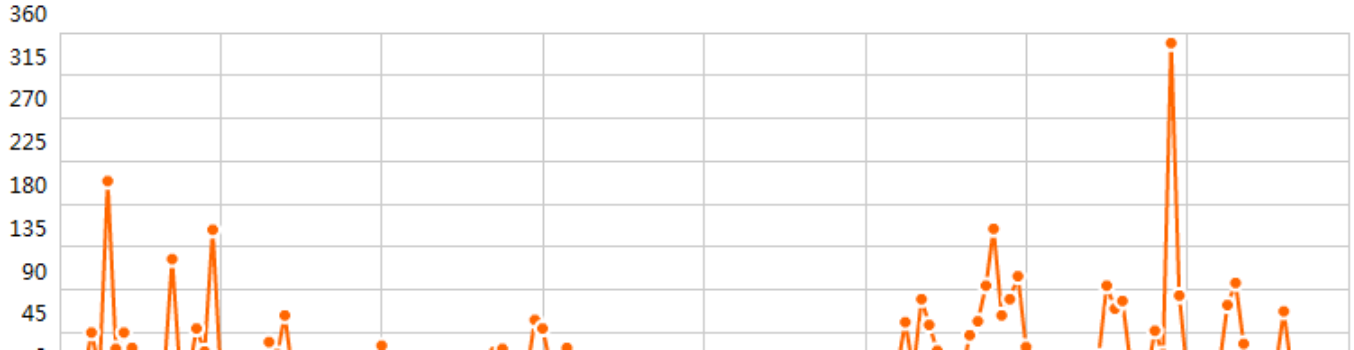
地产女首富吴亚军的细节与淡泊 2010-11-04 10:05 6小时前
胡润近日发布的中国房产富豪榜显示，龙湖地产董事长吴亚军以270亿元的身价排名第二，仅次于万达集团董事局主席王健林(财富280亿元)，吴亚军再次成为内地女首富。 来源：gs.people.com.cn

成功招商只是运营的开始 上海区域万达广场商业管理有限公司总经理王寿庆谈第三代城市综合体 2010-11-04 06:31 10小时前

万达第三代城市综合体的成功首先得益于董事长王健林提出的"订单地产模式"，就是携手知名商业企业，在全国各大城市发展城市综合体，万达负责开发，租售给这些企业经营，并 来源：news.china.com.cn

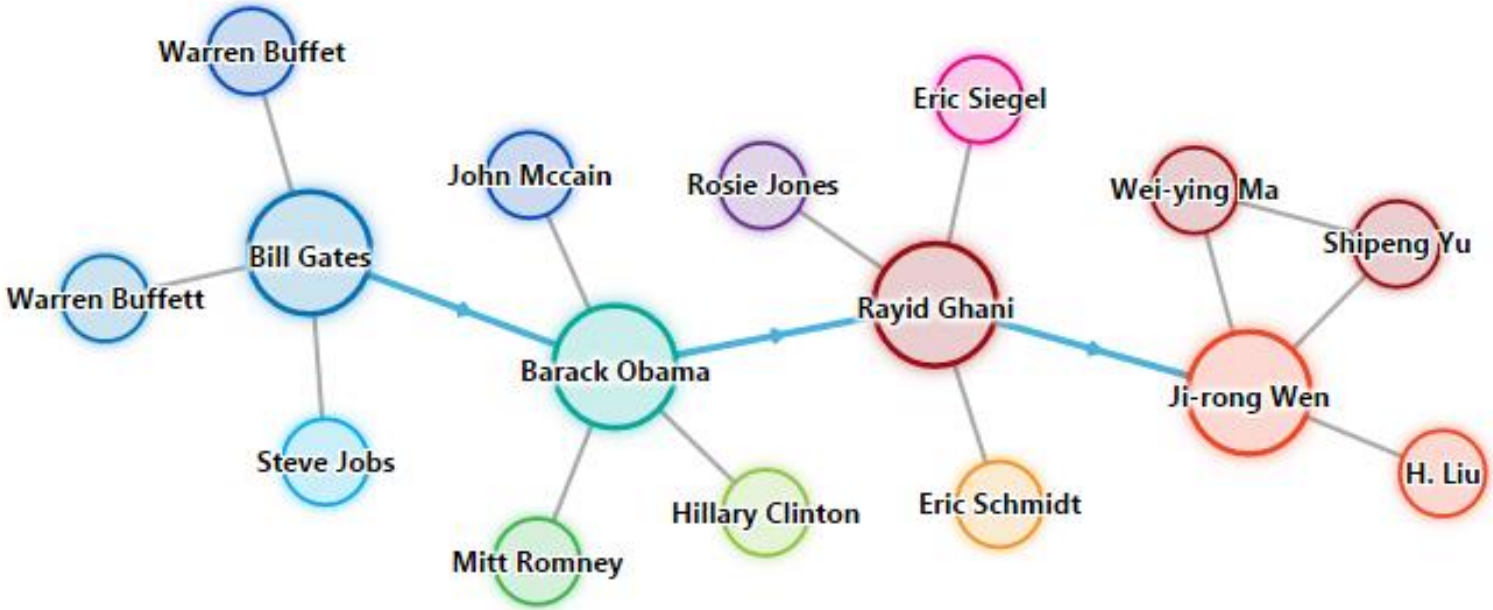
成败孙宏斌 一个地产界争议性人物自我救赎 2010-11-03 12:57 1天前 来源：www.soufun.com
以开发为平台深挖消费潜力 万达迈向世界级企业 2010-11-03 10:35 1天前 来源：jn.focus.cn
商业地产"结构性回暖"的背后逻辑 2010-11-03 10:28 1天前 来源：house.focus.cn
万达集团董事长王健林：商业地产培育需要耐心 2010-11-03 08:53 1天前 来源：cc.focus.cn
万达拟投30亿长安建超级商城 2010-11-02 09:34 2天前 来源：dg.sun0769.com
CIHAF十二年回顾与展望(7)：2005年之亚洲巡礼 2010-11-01 15:25 3天前 来源：house.hexun.com
第四届中国CEO高峰峰会召开 唐骏学历门后现身 2010-11-01 15:00 3天前 来源：news.hexun.com
白国华：许家印的管理模式 2010-11-01 10:09 3天前 来源：sports.sina.com.cn

资讯历史



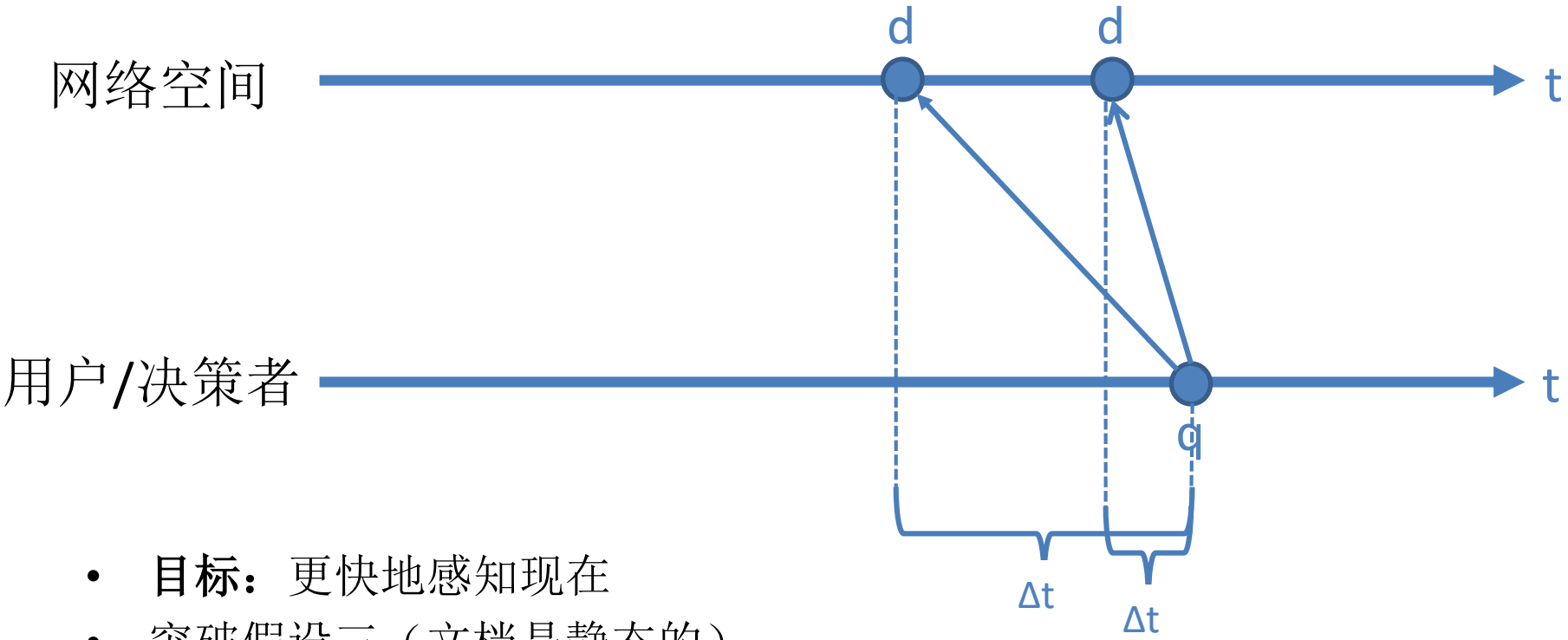
火龙果 · 整理
uml.org.cn

六度联接



火龙果 · 整理
uml.org.cn

实时搜索



- **目标：**更快地感知现在
- 突破假设三（文档是静态的）
- **挑战：**
  - 数据的实时获取和检索

火龙果 · 整理
uml.org.cn

实时搜索

Baidu 百度

明星吸毒

明星吸毒的最新相关信息

曝房祖名供120名艺人吸毒 房祖名朋友圈大揭秘

据香港媒体报道 8月20日晚七点,成龙就儿子房祖名吸毒被抓一事向公众道歉,他在微博中表示自己"教子无方"并愿意和儿子一起共同面对未来之路。可就...

新华网 4小时前

娱评:明星吸毒这件事粉丝别纵容	大众网	4小时前
台媒:警方掌握120位吸毒明星名单 锁定3男星	新华网	4小时前
新京报网声明域名遭篡改 外界猜测与报道...	TechWeb	5小时前
明星吸毒丑闻频爆发 吸毒后神经系统损伤...	海南在线	6小时前

明星吸毒前后对比照颓废吓死人(图)|吸毒|李代沫|张默_新浪娱乐_...

【原创】作者：北京冬雨 - 发表时间：2014年03月19日

文/北京冬雨2014年3月18日下午1点,歌手李代沫因与朋友在出租房吸毒,已被警方带走调查,有消息称李代沫暴瘦被指因吸毒。关于吸毒...

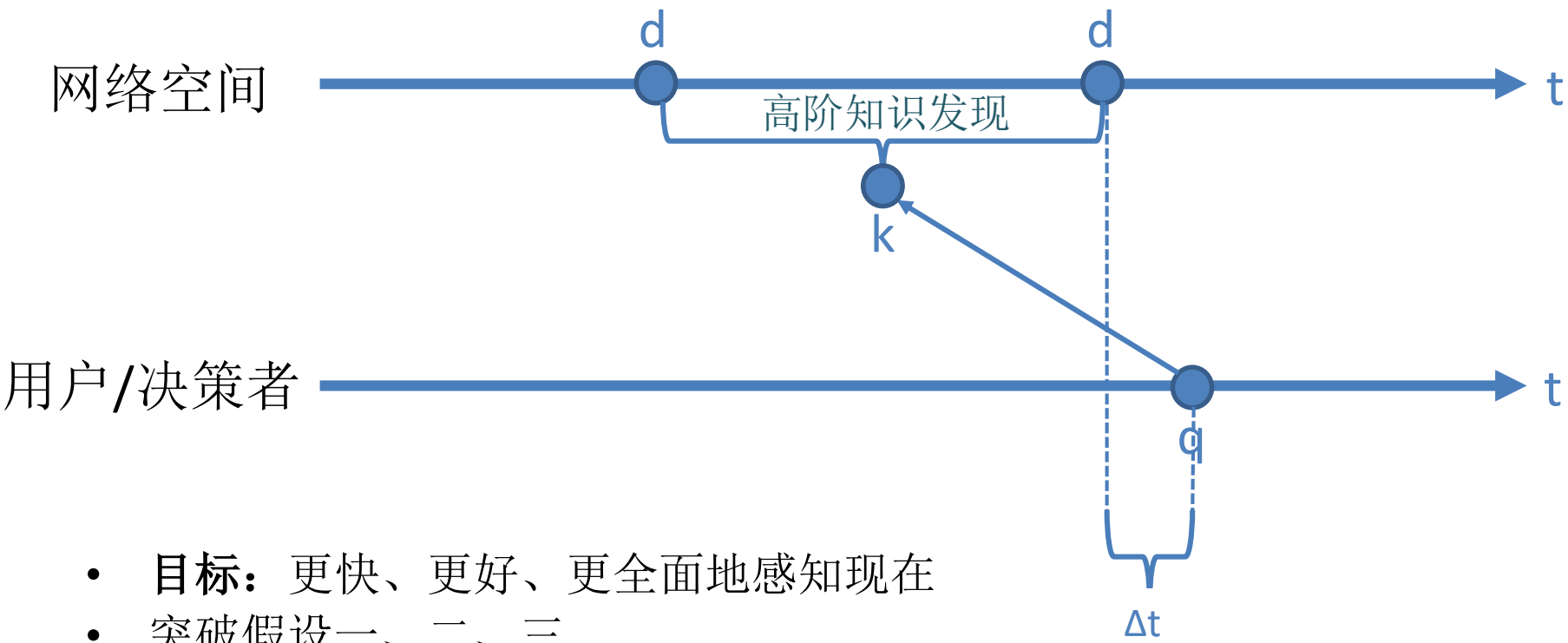
ent.sina.com.cn/zl/bag... 2014-03-19 - 百度快照

香港艺人集体吸毒事件_百度百科

2012年11月28日，轰动香港娱乐的"五艺人吸毒"事件被曝升级，称实为娱乐圈23人相约开"迷幻饭局"，并称已和看过该饭店闭路电视的员工进行沟通，透露23人名单，陈奕迅妻子徐濠萦、赌王女儿何超仪、刚刚大摆婚宴不久的歌手叶佩雯等赫然在列...

火龙果·整理
uml.org.cn

实时分析



- **目标：**更快、更好、更全面地感知现在
- 突破假设一、二、三
- **挑战：**
  - 实时动态高阶知识发现
  - 动态性监测和预警
  - 性能！

实时分析

火龙果 · 整理
uml.org.cn

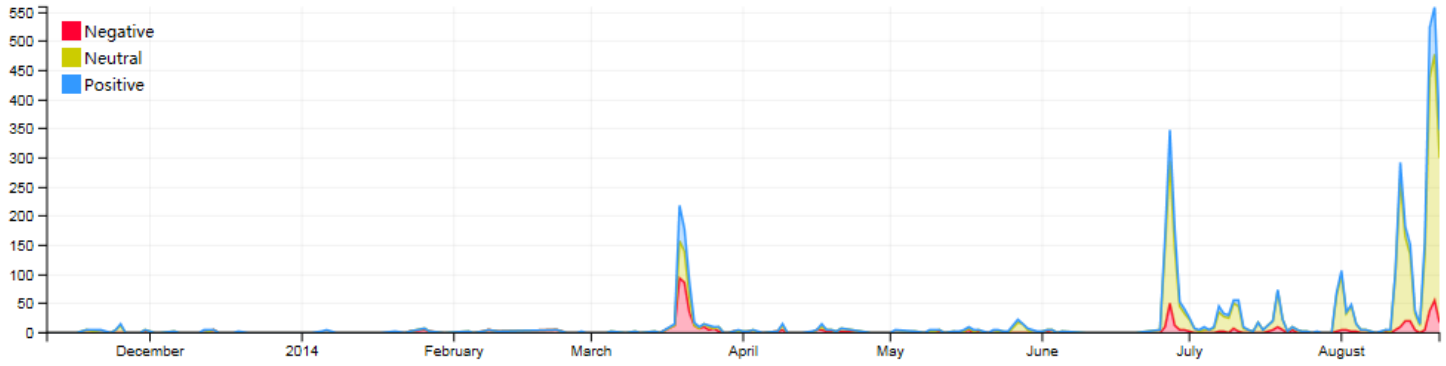
时事探针 首页 热点 日志 您好， jrwen ！ 退出

明星吸毒 搜索

明星吸毒 Multi-Dimensional View Normal View Topic Config Wikipedia

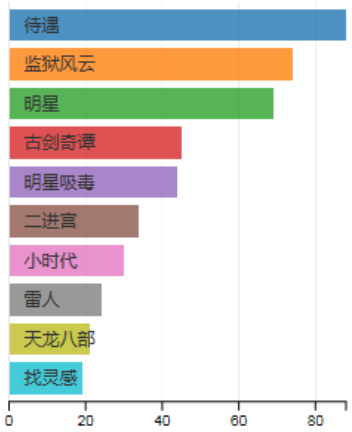
共4863条数据，检索到4863条数据

Trend

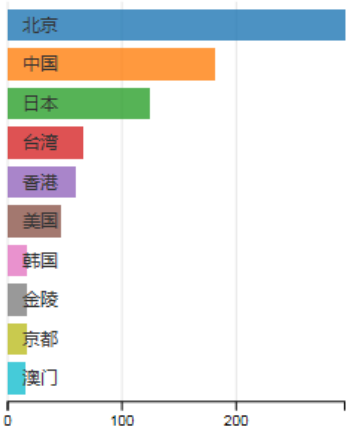


Entity_Person

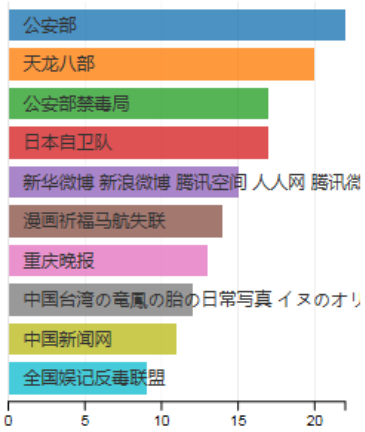
Nugget



Entity_Location

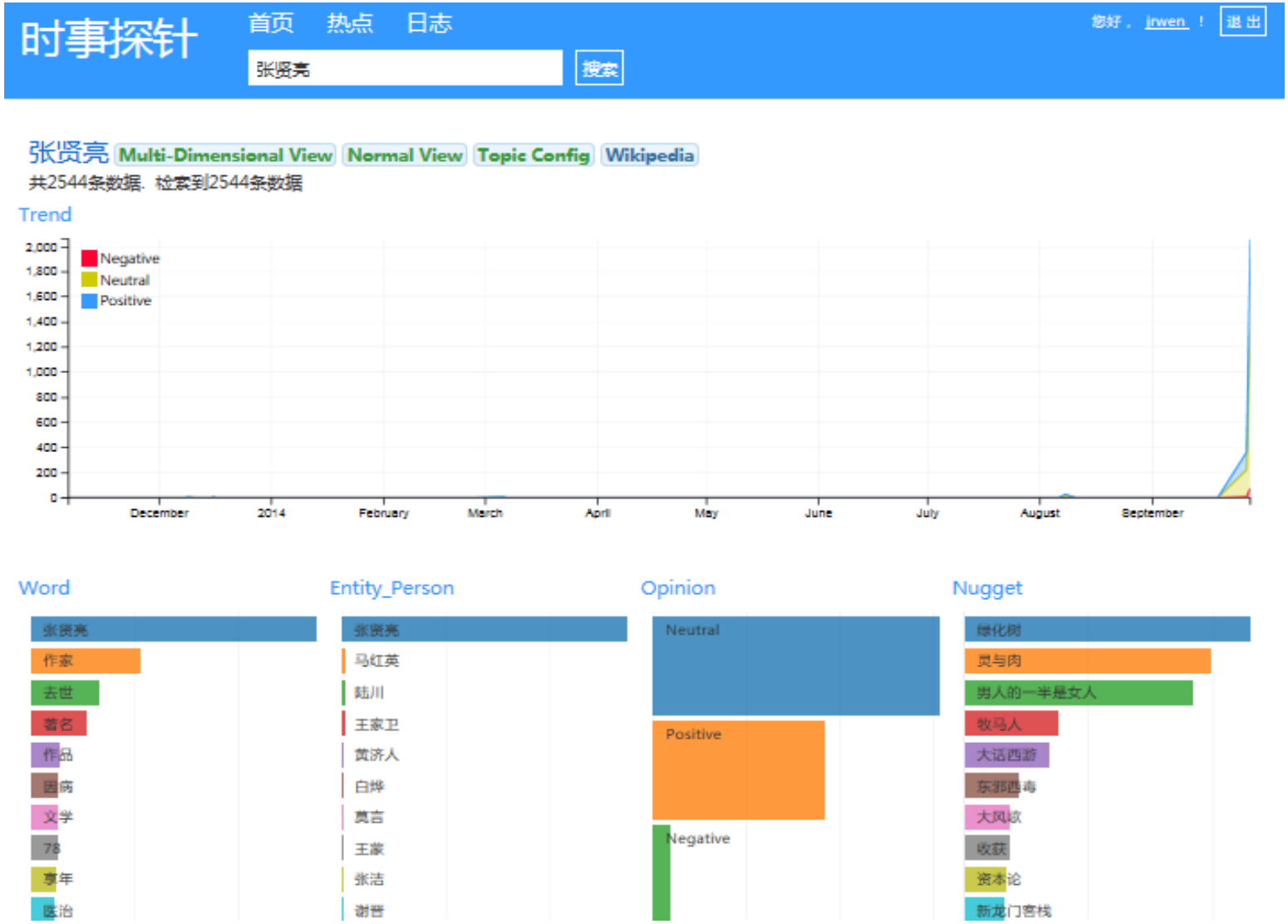


Entity_Orgnization



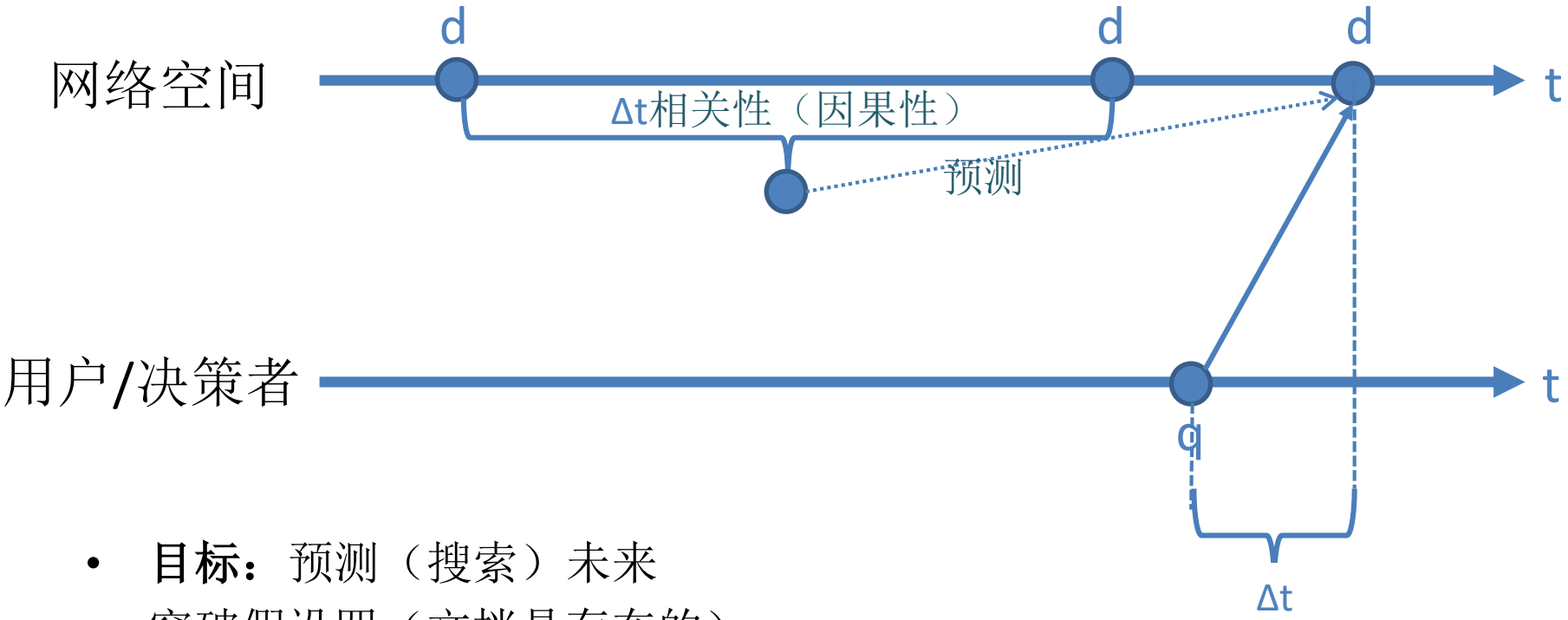
实时分析

火龙果 · 整理
uml.org.cn



火龙果 · 整理
uml.org.cn

相关性 – 预测



- **目标：**预测（搜索）未来
- 突破假设四（文档是存在的）
- **挑战：**
  - 多源异构多噪数据→Δt相关性？
  - Δt相关性→预测？

火龙果 · 整理
uml.org.cn

研究内容

- 数据实时智能感知
- 高阶知识挖掘和分析模型
- 相关性挖掘和预测模型
- 数据和知识可靠性判定
- 大数据实时分析平台

火龙果 · 整理
uml.org.cn

数据质量判定

火龙果 · 整理
uml.org.cn

谣言判别器

- 问题：如何利用网络大数据自动识别谣言
- 挑战：简单应用频率判别不可靠
  - 谣言扩散得很快
- Idea：大数据是自洽的，谣言必然与相关数据不相容

火龙果 · 整理
uml.org.cn

姚明和李宇春是兄妹吗？

Baidu 百度 新闻 网页 贴吧 知道 音乐 图片 视频 地图 文库 更多»

姚明和李宇春是兄妹 | 百度一下

百度为您找到相关结果约**3,020,000**个

姚明和李宇春竟然是兄妹 在线观看 - 酷6视频

时长: 1分钟 - 标清
姚明和李宇春竟然是兄妹
v.ku6.com/show/BpehmT3ycuHqrL...html 2011-10-8 - 百度快照

姚明和李宇春是兄妹- - 阿里巴巴商人论坛
3条回复 - 发帖时间: 2008年4月1日
姚明和李宇春是兄妹,李宇春自选秀功成之后,一路星途平遥直上。俨然内地呼风唤雨式的人物,粉丝“玉米”何止万千,所到之处,无不一片欢呼喝彩,好不气派。 姚明,...
club.china.alibaba.com/threadview/24... 2013-7-20 - 百度快照

姚明和李宇春是兄妹!! - 伍豪传人的日志 - 网易博客
3条评论
姚明和李宇春是兄妹!! ,伍豪传人的网易博客,海纳百川,有容乃大;精诚所至,金石为开。,自信、坚强、豁达、深刻、仁爱
blog.163.com/wuhaochuanren/blog/stat... 2013-1-24 - 百度快照

姚明竟然和李宇春是孪生兄妹相信吗?? - 已解决 - 搜搜问问
10个回答 - 最新回答: 2007年12月4日
最佳答案: 本来相信的,但开了这2张图象后,打死也不信了!!
wenwen.soso.com/z/q227691...htm 2013-1-18 - 百度快照

【号外】姚明和李宇春是兄妹(有图为证!!!) 蚌埠吧 百度贴吧

火龙果 · 整理
uml.org.cn

逻辑

- 兄妹(A,B) => 籍贯(A) = 籍贯(B)
- 兄妹(A,B) => 父亲(A) = 父亲(B)
- ……

- 大数据的作用：你可以提出一个假命题，但不能改变所有的相关命题。

火龙果 · 整理
uml.org.cn

比较籍贯

- 大数据告诉我们：
  - 姚明的籍贯很可能是上海
  - 李宇春的籍贯很可能是成都
- 所以：姚明的籍贯≠李宇春的籍贯

火龙果 · 整理
uml.org.cn

比较父亲

- 大数据告诉我们：
  - 姚明的父亲很可能是姚志源
  - 李宇春的父亲很可能是李升敏
- 所以：姚明的父亲≠李宇春的父亲

火龙果 · 整理
uml.org.cn

逻辑

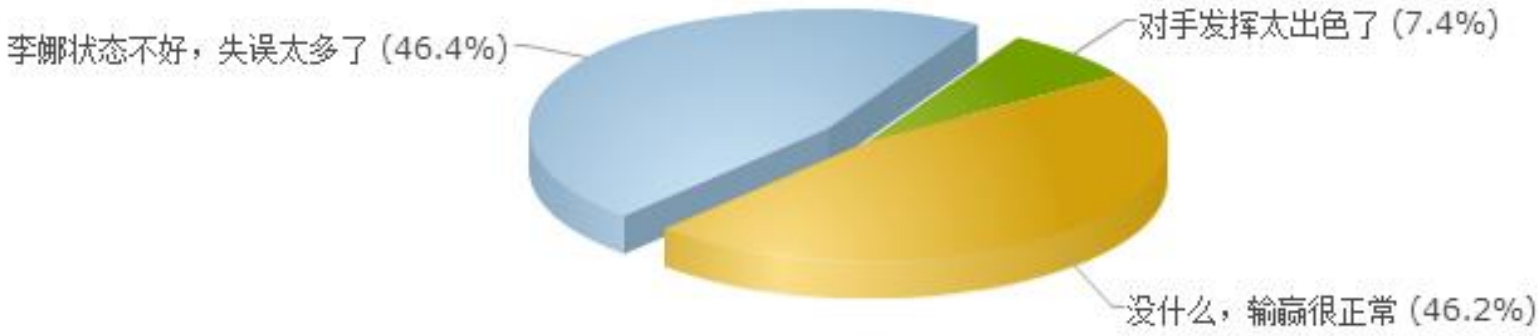
- 籍贯(A) ≠ 籍贯(B) ∧ 父亲(A) ≠ 父亲(B)
  => ~兄妹(A,B)

这种网上调查有多可靠？

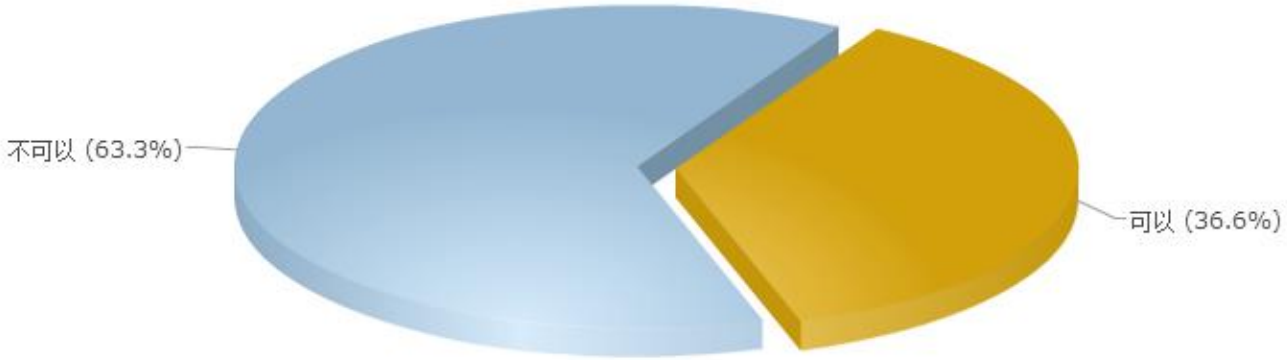
火龙果 · 整理
uml.org.cn

来源：新浪体育 创建时间：2014年05月27日 状态：进行中 参与人数：3,715

1. 如何看待李娜法网一轮游？



2. 您认为李娜还能否登上世界第一宝座？



火龙果 · 整理
uml.org.cn

大数据抽样

- 大数据是全样本，不需要抽样？
- 大数据 = 大样本 ≠ “好”样本
- 大数据抽样
  - 接近于真实的样本分布
  - 随机性

火龙果 · 整理
uml.org.cn

大数据应用：计算社会科学

火龙果 · 整理
uml.org.cn

社会科学应用

外国人对中国的偏见和误解

- 与国学院合作，利用时事探针系统收集国内外数据，针对“外国人对中国的偏见和误解”这一主题进行了大数据分析，揭示了一些主要的偏见和背后的原因。

产业转移

- 与经济学院合作，通过收集互联网数据对我国的“产业转移”这一主题进行了分析，发现了近年来国内产业转移的一些基本模式和变化规律。

农产品价格预测

- 与中国农业科学院合作，通过互联网数据对农产品价格的变化进行预测，我们收集了互联网上各地的农产品价格和媒体上的农业相关信息，发现将这些数据结合可以非常有效地预测农产品价格的变动趋势。

高考舆情分析

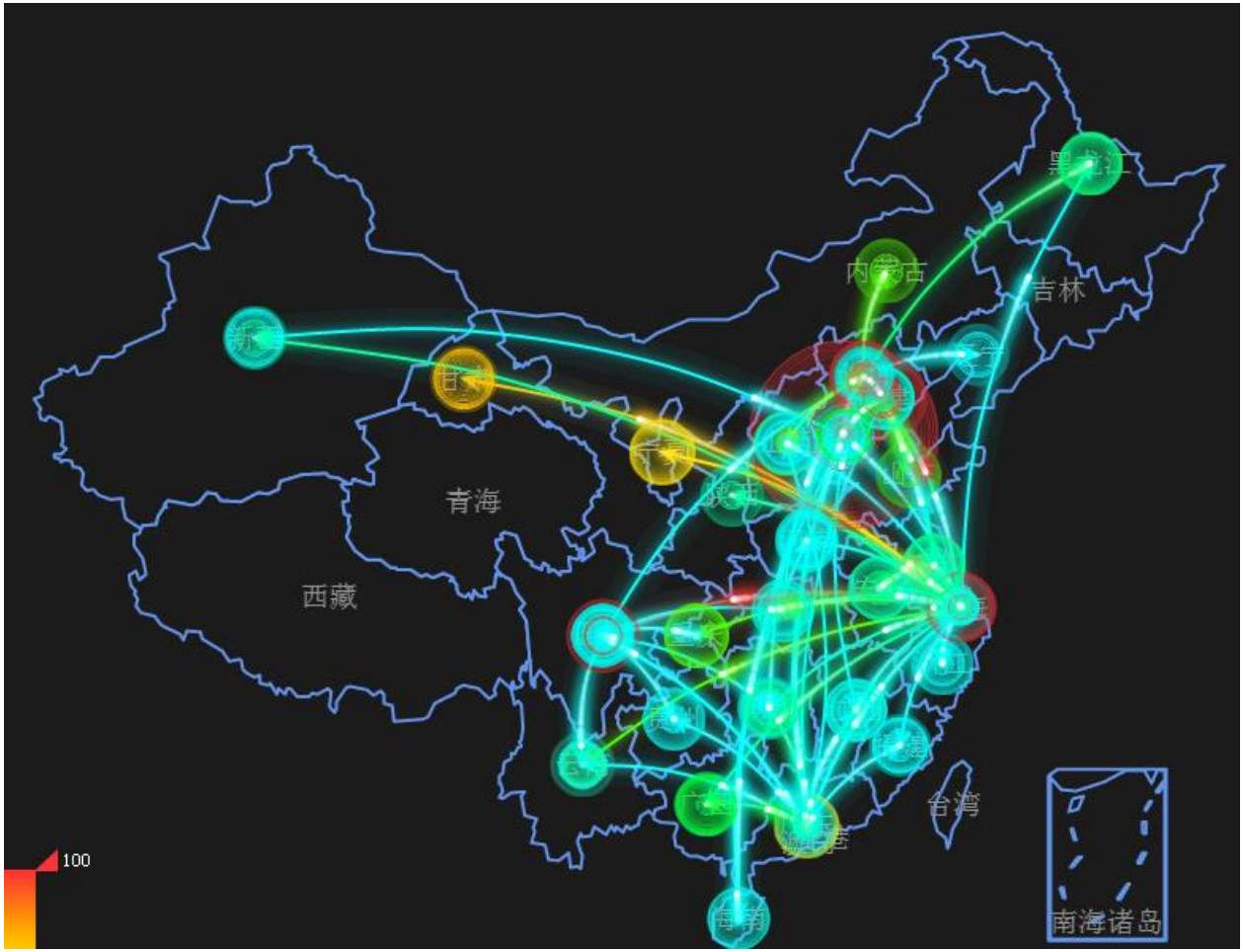
- 与百度公司合作，对高考舆情进行大数据分析，并将分析结果报送教育部。

法律大数据分析

- 与法学院合作，对法律判决文书进行大数据分析。

火龙果 · 整理
uml.org.cn

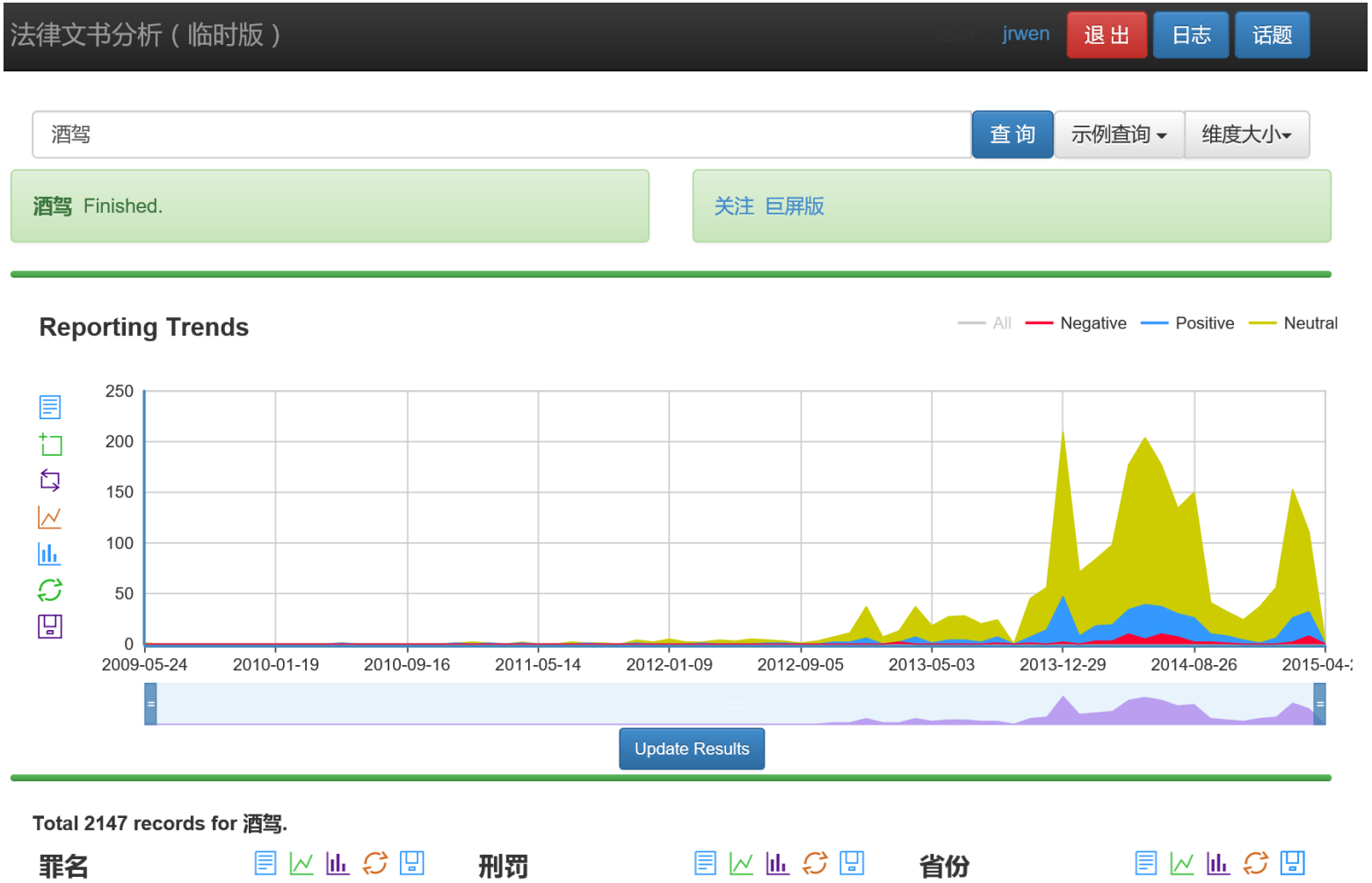
演示 – 产业转移



http://websensor.playbigdata.com/chanyeqianyi

火龙果 · 整理
uml.org.cn

演示 – 法律大数据



http://websensor.playbigdata.com/verdict2/Index.aspx

火龙果 · 整理
uml.org.cn

谢谢！