

BIG DATA

8.大数据分析

8.1 数据分析概述

- 数据分析是以下三部分行为组成的过程：
 - 搜集数据
 - 处理数据
 - 获取信息
- 目的：
 - 对杂乱无章的数据进行集中、萃取和提炼，进而找出所研究对象的内在规律
- 意义：
 - 找到隐藏在繁乱数据中的信息

数据分析的基本方法

- 统计学

- 描述统计学

- 给定一组数据，可以摘要并且描述这份数据的统计学

- 推断统计学

- 观察者以数据的形态建立一个用以解释其随机性和不确定性的数学模型，以此来推论研究中的步骤和母体

- 快速傅里叶变换

有限长序列可以通过离散傅里叶变换(DFT)将其频域也离散化成有限长序列。但其计算量太大，很难实时地处理问题，因此引出了快速傅里叶变换(FFT)。1965年，Cooley和Tukey提出了计算离散傅里叶变换的快速算法，将其运算量减少了几个数量级。从此，对快速傅里叶变换算法的研究便不断深入，数字信号处理这门新兴学科也随其出现和发展而迅速发展。根据对序列分解与选取方法的不同而产生了快速傅里叶变换的多种算法，基本算法是基2DIT和基2DIF。快速傅里叶变换在离散傅里叶反变换、线性卷积和线性相关等方面也有重要应用。

快速傅里叶变换是离散傅里叶变换的快速算法，它是根据离散傅里叶变换的奇、偶、虚、实等特性，对离散傅立叶变换的算法进行改进获得的。对于在计算机系统或者说数字系统中应用离散傅立叶变换，可以说是前进了一大步。

数据分析的基本方法

• 平滑滤波

平滑滤波是低频增强的空间域滤波技术。其目的有两个：一个是模糊；另一个是消除噪声。空间域的平滑滤波一般采用简单平均法进行，就是求邻近像元点的平均亮度值。邻域的大小与平滑的效果直接相关，邻域越大平滑的效果越好，但邻域过大，平滑会使边缘信息损失的越大，从而使输出的图像变得模糊，因此需合理选择邻域的大小。

• 基线和峰值

基线是项目储存库中每个工件版本在特定时期的一个“快照”。它提供一个正式标准，随后的工作基于此标准，并且只有经过授权后才能变更这个标准。建立一个初始基线后，以后每次对其进行的变更都将记录为一个差值，直到建成下一个基线。

峰值是在所考虑的时间间隔内，变化的电流、电压或功率的最大瞬间值。

峰值功率，就是最高能支持的功率 接近或者超过峰值功率。电源的峰值功率指电源短时间内能达到的最大功率，通常仅能维持 30 秒左右的时间。一般情况下电源峰值功率可以超过最大输出功率 50%左右，由于硬盘在启动状态下所需要的能量远远大于其正常工作时的数值，因此系统经常利用这一缓冲为硬盘提供启动所需的电流，启动到全速后就会恢复到正常水平。峰值功率其实没有什么实际意义的，因为电源一般不能在峰值输出时稳定工作。

数据分析的基本方法

• 分类

分类就是找出一个类别的概念描述，它代表了这类数据的整体信息，即该类的内涵描述，并用这种描述来构造模型，一般用规则或决策树模式表示。分类是利用训练数据集通过一定的算法而求得分类规则，分类可被用于规则描述和预测。分类过程是：首先从数据中选出已经分好类的训练集，在该训练集上运用数据挖掘分类的技术，建立分类模型，对于没有分类的数据进行分类。类的个数是确定的，预先定义好的。例如：

(1) 信用卡申请者按低、中、高风险分类；

(2) 对某种生产的全流程进行质量监控和分析，构建故障地图，实时分析产品出现瑕疵的原因分类，有效提高了产品的优良率。

数据分析的基本方法

• 聚类分析

聚类分析指将物理或抽象对象的集合分组为由类似的对象组成的多个类的分析过程。聚类是将数据分类到不同的类或者簇这样的一个过程，所以同一个簇中的对象有很大的相似性，而不同簇间的对象有很大的相异性。聚类分析是一种探索性的分析，在分类的过程中，不必事先给出一个分类的标准，聚类分析能够从样本数据出发，自动进行分类。聚类分析所使用的方法不同，常常会得到不同的结论。不同研究者对于同一组数据进行聚类分析，所得到的聚类数未必一致。在商业中，聚类可以帮助市场分析人员从消费者数据库中区分出不同的消费群体，并且概括出每一类消费者的消费模式或者消费习惯。它作为数据挖掘中的一个模块，可以作为一个单独的工具发现数据库中分布的一些深层次的信息，或者把注意力放在某一个特定的类上以作进一步的分析并概括出每一类数据的特点。图 8-1 是聚类算法的一种展示，Cluster1 和 Cluster2 分别代表聚类算法计算出的两类样本。打“+”号的是 Cluster1，而打“○”标记的是 Cluster2。

数据分析的基本方法

• 因子分析

因子分析是指研究从变量群中提取共性因子的统计技术。因子分析就是从大量的数据中寻找内在的联系，减少决策的困难。因子分析的方法约有十多种，如重心法、影像分析法，最大似然解、最小平方法、阿尔法抽因法、拉奥典型抽因法等。这些方法本质上大都属近似方法，以相关系数矩阵为基础。在社会学研究中，因子分析常采用以主成分分析为基础的反覆法。

• 相关分析

相关分析是研究现象之间是否存在某种依存关系，并对具体有依存关系的现象研究其相关方向以及相关程度。相关关系是一种非确定性的关系，例如，以 X 与 Y 分别记一个人的身高和体重，则 X 与 Y 显然有关系，而又没有确切到可由其中的一个去精确地决定另一个的程度，这就是相关关系，而不是因果关系。

数据分析的基本方法

• 对应分析

对应分析也称关联分析，通过分析由定性变量构成的交互汇总表来表示变量间的联系。可以揭示同一变量的各个类别之间的差异，以及不同变量各个类别之间的对应关系。对应分析的基本思想是将一个联列表的行和列中各元素的比例结构以点的形式在较低维的空间中表示出来。

• 回归分析

回归分析是研究一个随机变量 Y 对另一个变量 X 或一组 (X_1, X_2, \dots, X_k) 变量的相依关系的统计分析方法。回归分析是确定两种或两种以上变数间相互依赖的定量关系的一种统计分析方法，应用十分广泛。回归分析按照涉及的自变量的多少，可分为一元回归分析和多元回归分析。按照自变量和因变量之间的关系类型，又可分为线性回归分析和非线性回归分析。

• 方差分析

方差分析又称为变异数分析或 F 检验，主要用于两个及两个以上样本均数差别的显著性检验。由于各种因素的影响，研究所得的数据呈现波动状。造成波动的原因可分成两类，一类是不可控的随机因素，另一类是研究中施加的对结果形成影响的可控因素。方差分析是从观测变量的方差入手，研究各控制变量中哪些变量是对观测变量有较显著影响的变量。

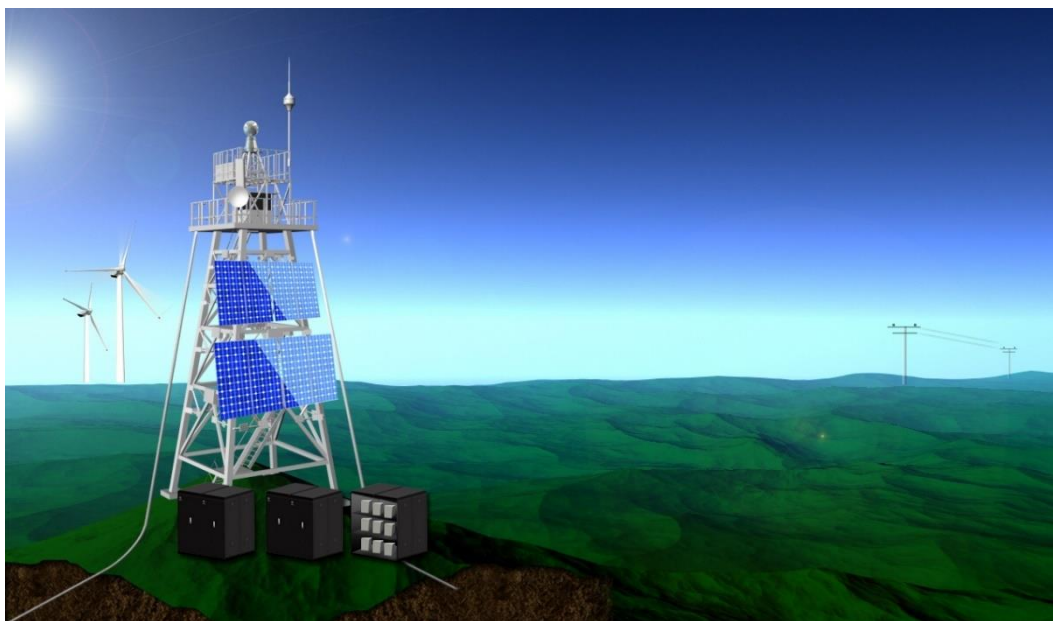
数据分析的类型

- 探索性数据分析
 - 在数据中发现新的特征
- 定性数据分析
 - 定性资料分析、定性研究、照片、观察结果等非数值型数据的分析
- 离线数据分析
 - 复杂和耗时的数据分析和处理
- 在线数据分析
 - OLAP，也称为联机分析处理，处理响应时间要求较高的请求

8.2 大数据分析应用案例

环保监测

- 大数据已经被广泛应用于污染监测领域，借助大数据技术，采集各项环境质量指标信息，集成整合到数据中心进行数据分析，并把分析结果用于指导下一步环境治理方案的制定，可以有效提升环境整治的效果
- 中国水污染地图
- 中国空气污染地图
- 中国固废污染地图
- 汽车尾气污染治理



市场情绪分析

• 市场情绪分析是交易者在日常交易工作中不可或缺的一环，根据市场情绪分析、技术分析和基本面分析，可以帮助交易者做出更好的决策。大数据技术在市场情绪分析中大有用武之地。



信贷风险分析

- 大数据分析技术已经能够为企业信贷风险分析助一臂之力。通过收集和分析大量中小微企业用户日常交易行为的数据，判断其业务范畴、经营状况、信用状况、用户定位、资金需求和行业发展趋势，解决由于其财务制度的不健全而无法真正了解其真实经营状况的难题，让金融机构放贷有信心、管理有保障



大数据征信

- 大数据征信就是利用信息技术优势，将不同信贷机构、消费场景、支离破碎的海量数据整合起来，经过数据清洗、模型分析、校验等一系列流程后，加工融合成真正有用的信息。



发现关联购买行为

- 啤酒与尿布的故事



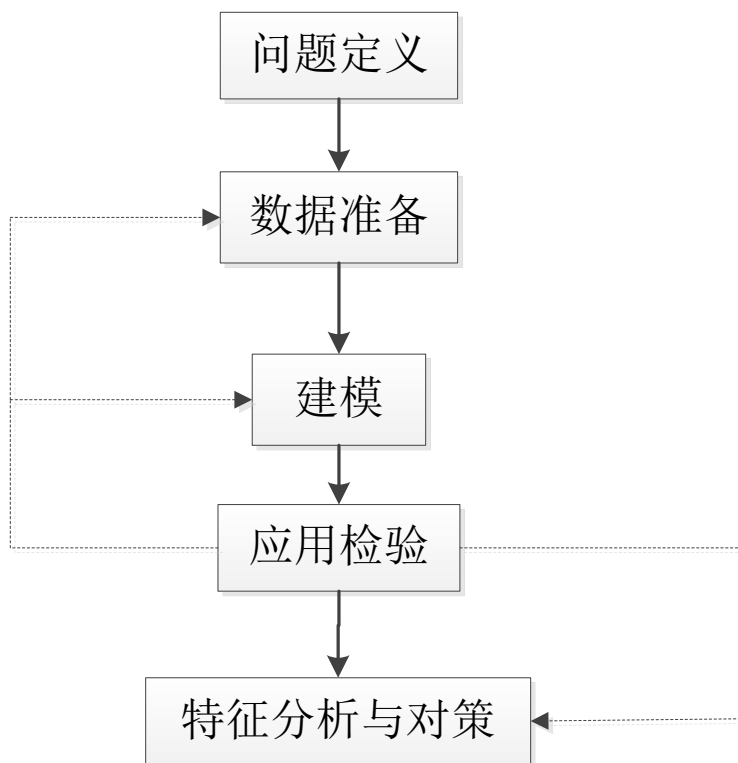
客户群体细分

- 美国Target超市比孩子父亲还早发现他女儿已经怀孕



大数据在电信领域的应用

- 预测客户行为，发现行为趋势，并找出公司服务过程中存在缺陷的环节，从而帮助公司及时采取措施保留客户



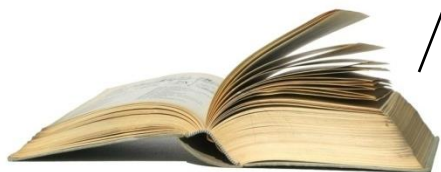
投拍影视作品



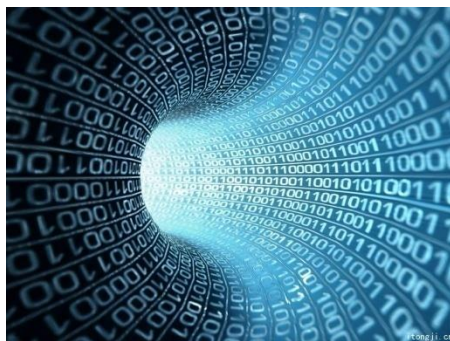
Kevin Spacey



David Fincher



英国同名小说《纸牌屋》



大数据分析



风靡全球的美剧《纸牌屋》

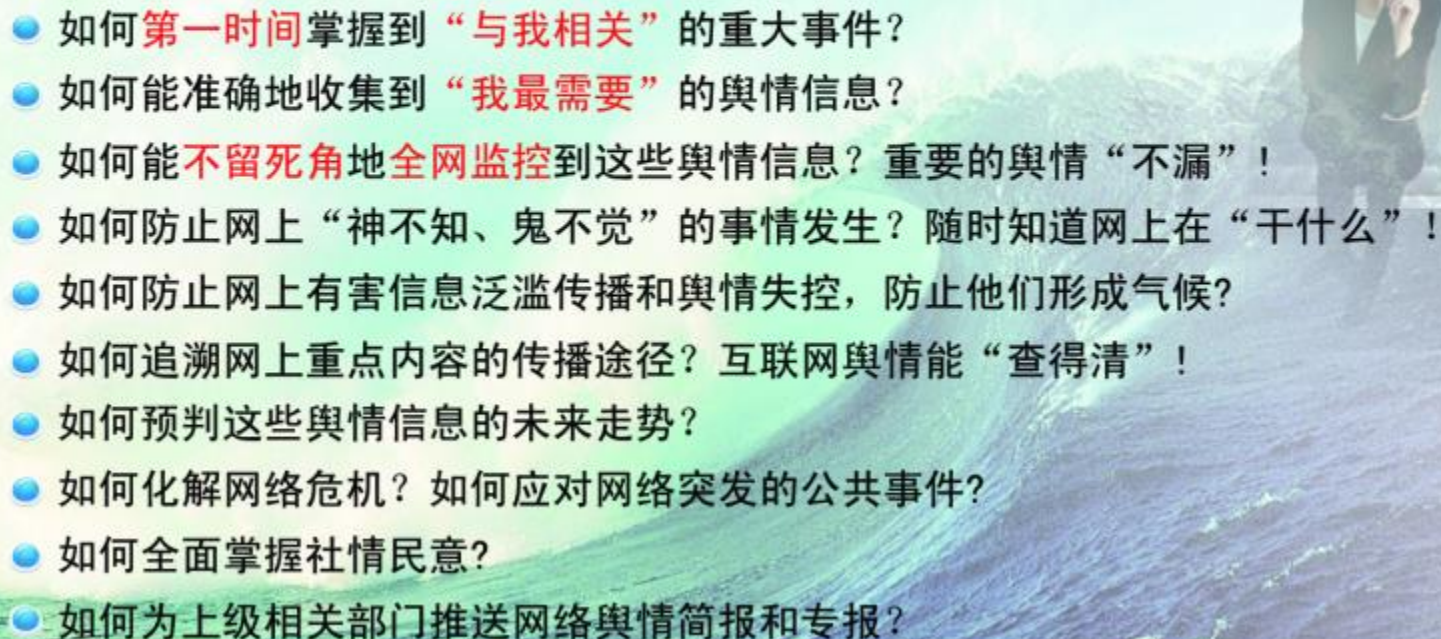
輿情分析

- 什么是輿情？
- 輿情是“舆论情况”的简称，是指在一定时期的一定社会空间内，围绕新闻事件、社会现象和社会问题所表达的信念、态度、意见和情绪的总和。
- 网络輿情（Internet Public Opinion, IPO）特别强调两点：一是新闻事件、社会现象和社会问题主要通过互联网首发或传播，二是表达信念、态度、意见和情绪的公众主要是网民。

网络舆情信息的主要来源

- 网站新闻评论
- 论坛与社区BBS
- 聚合新闻RSS
- QQ
- MSN
- 博客BLOG
- 微博MicroBlog
- 微信

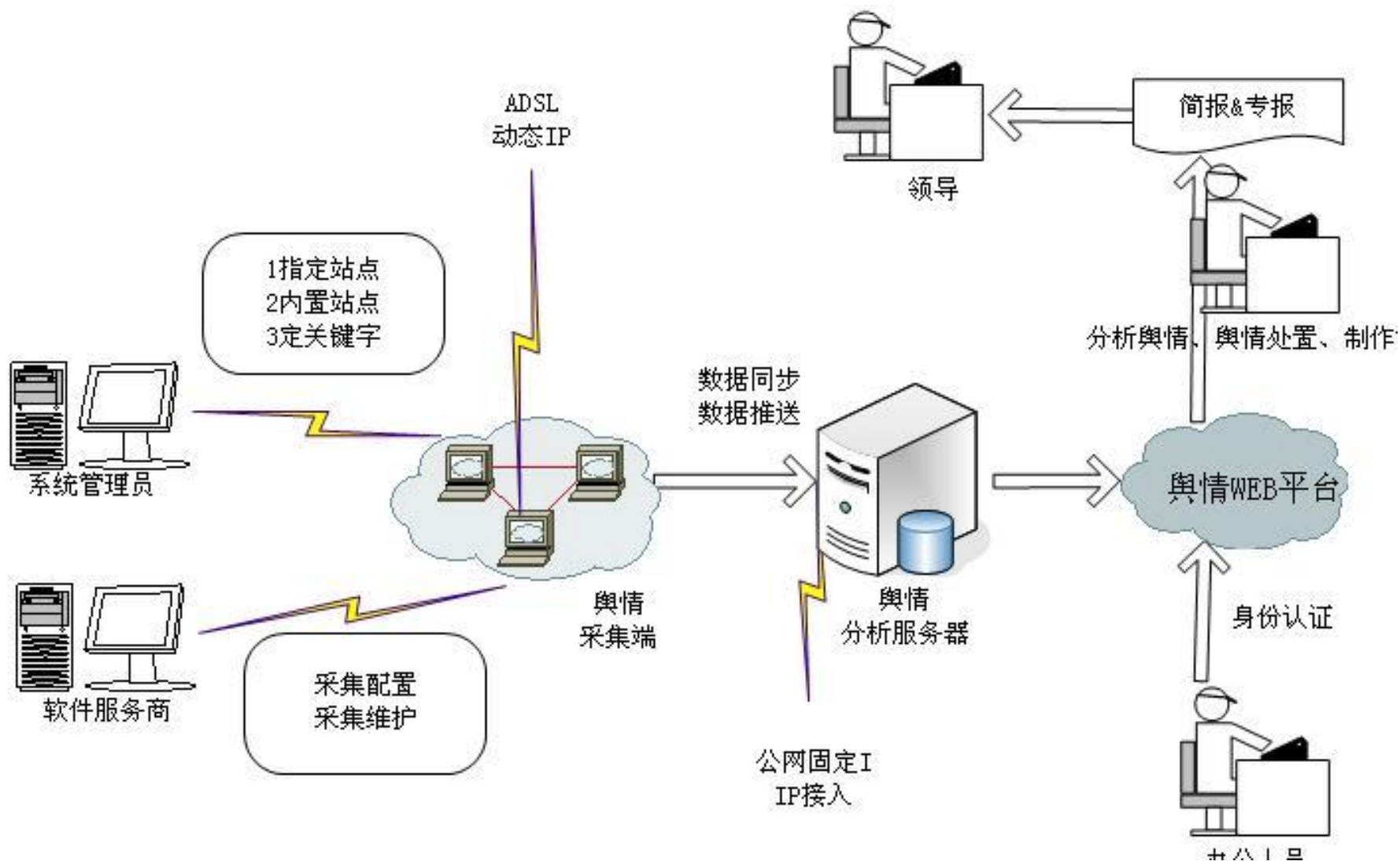
网络信息综合监控的难点

- 
- 如何**第一时间**掌握到“**与我相关**”的重大事件？
 - 如何能准确地收集到“**我最需要**”的舆情信息？
 - 如何能**不留死角**地**全网监控**到这些舆情信息？重要的舆情“不漏”！
 - 如何防止网上“神不知、鬼不觉”的事情发生？随时知道网上在“干什么”！
 - 如何防止网上有害信息泛滥传播和舆情失控，防止他们形成气候？
 - 如何追溯网上重点内容的传播途径？互联网舆情能“查得清”！
 - 如何预判这些舆情信息的未来走势？
 - 如何化解网络危机？如何应对网络突发的公共事件？
 - 如何全面掌握社情民意？
 - 如何为上级相关部门推送网络舆情简报和专报？

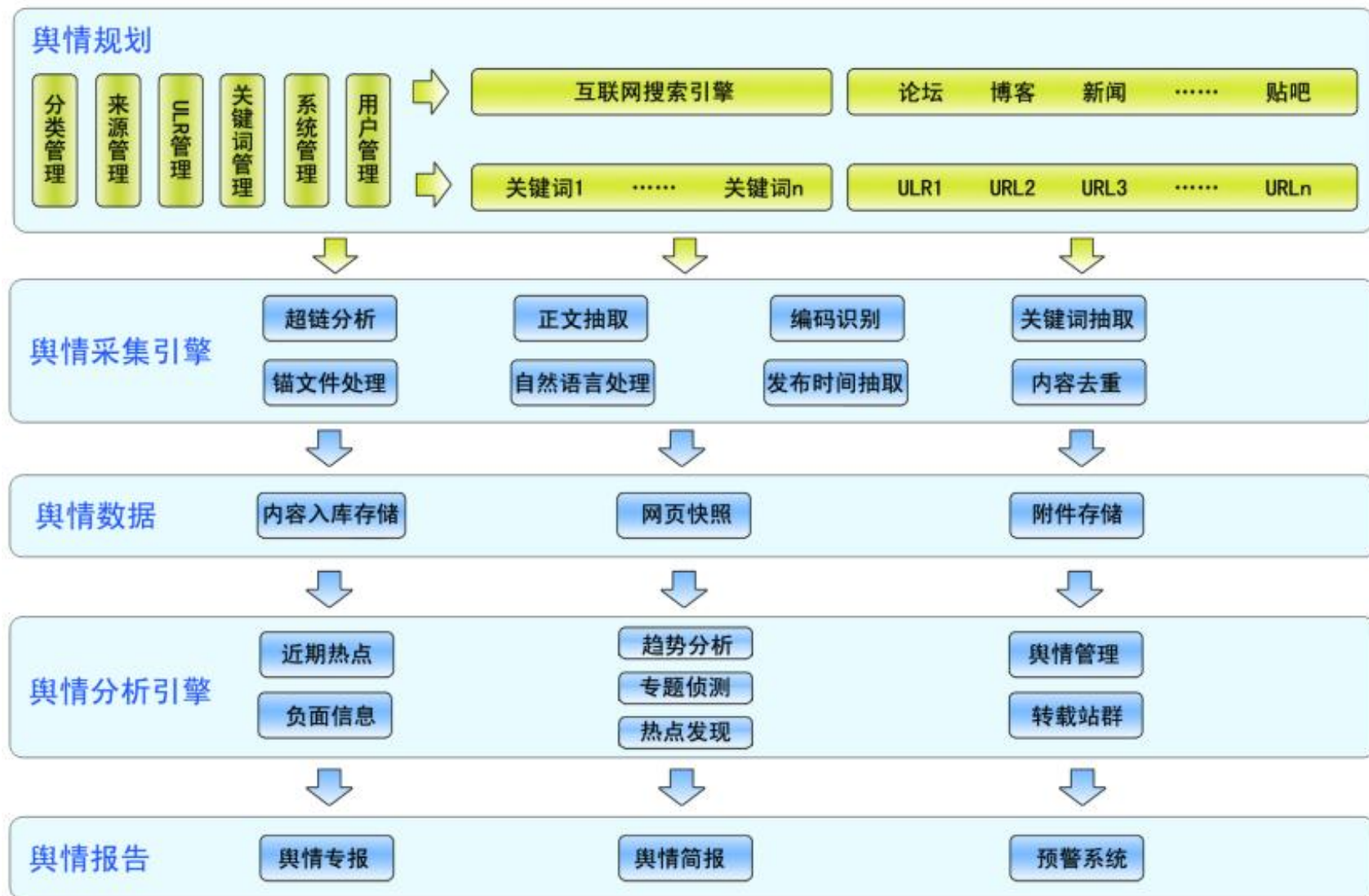
网络舆情监控系统可以实现

- 如何实时去抓取互联网里面的海量数据，这种包括国内网站、境外网站以及一些电子报刊等
- 逐条分析是否与“我”相关
- 逐条分析是否属于“舆情”
- 逐条分析是否属于“负面”
- 分析各条舆情“舆情热度”，评估其影响力，分析“重大舆情”“重点事件”
- 分析各条舆情的传播路径、传播时间，做到舆情能“查得清”
- 每日生成“网络舆情简报”，重大舆情生成“舆情专报”
- 时时进行舆情预警
- 24小时不间断监测，监测时差保障在30分钟内

网络舆情监控的系统工作流程



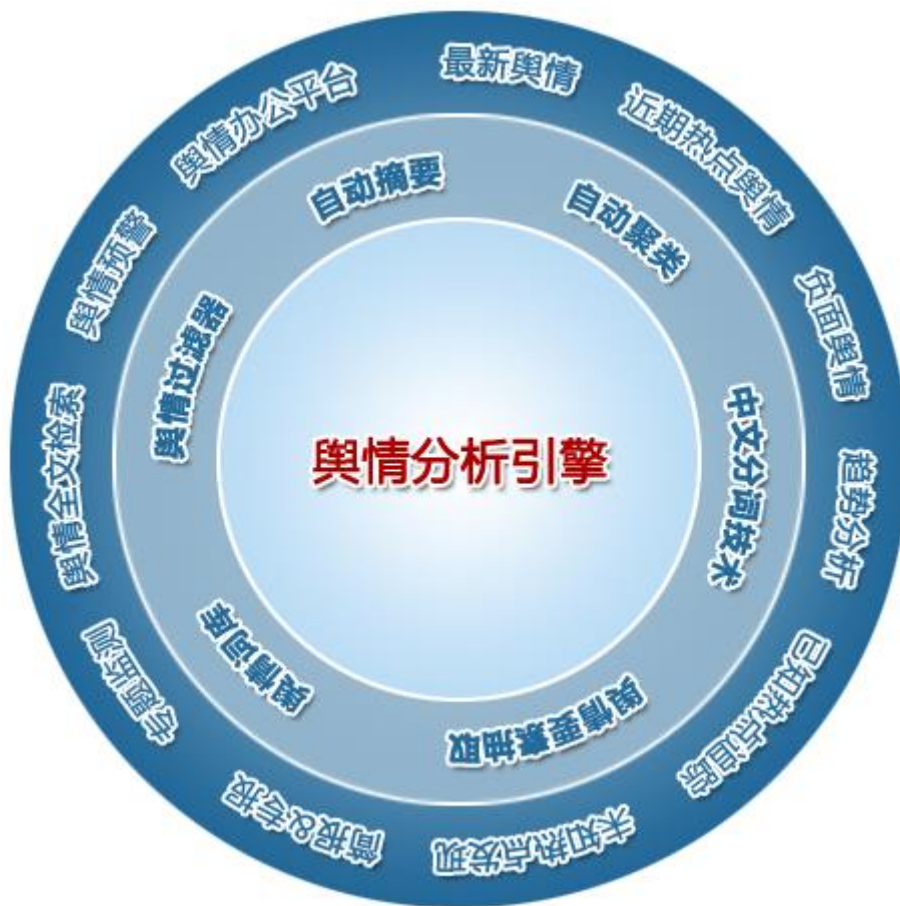
网络舆情监控的系统架构



网络舆情监控的“舆情漏斗”模型



网络舆情监控的“舆情罗盘”



网络舆情监控的“舆情研判”模型



分析一：分析“与我相关”？

1、集团，合资公司，子公司...

2、部门与机构,及其主要领导

3、按辖区地域名

4、主要产业项目

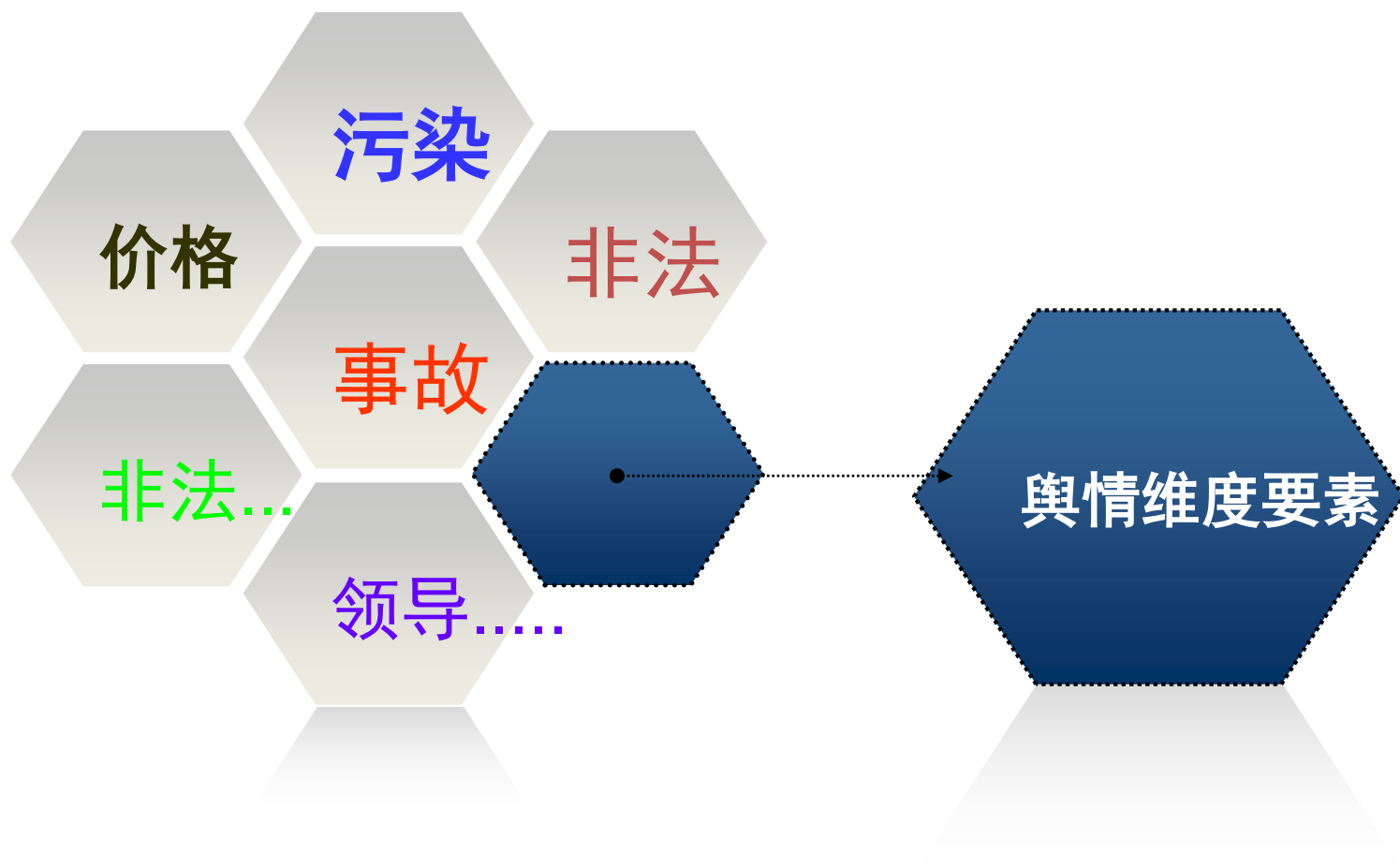
5、重点项目

6、重点活动与事件

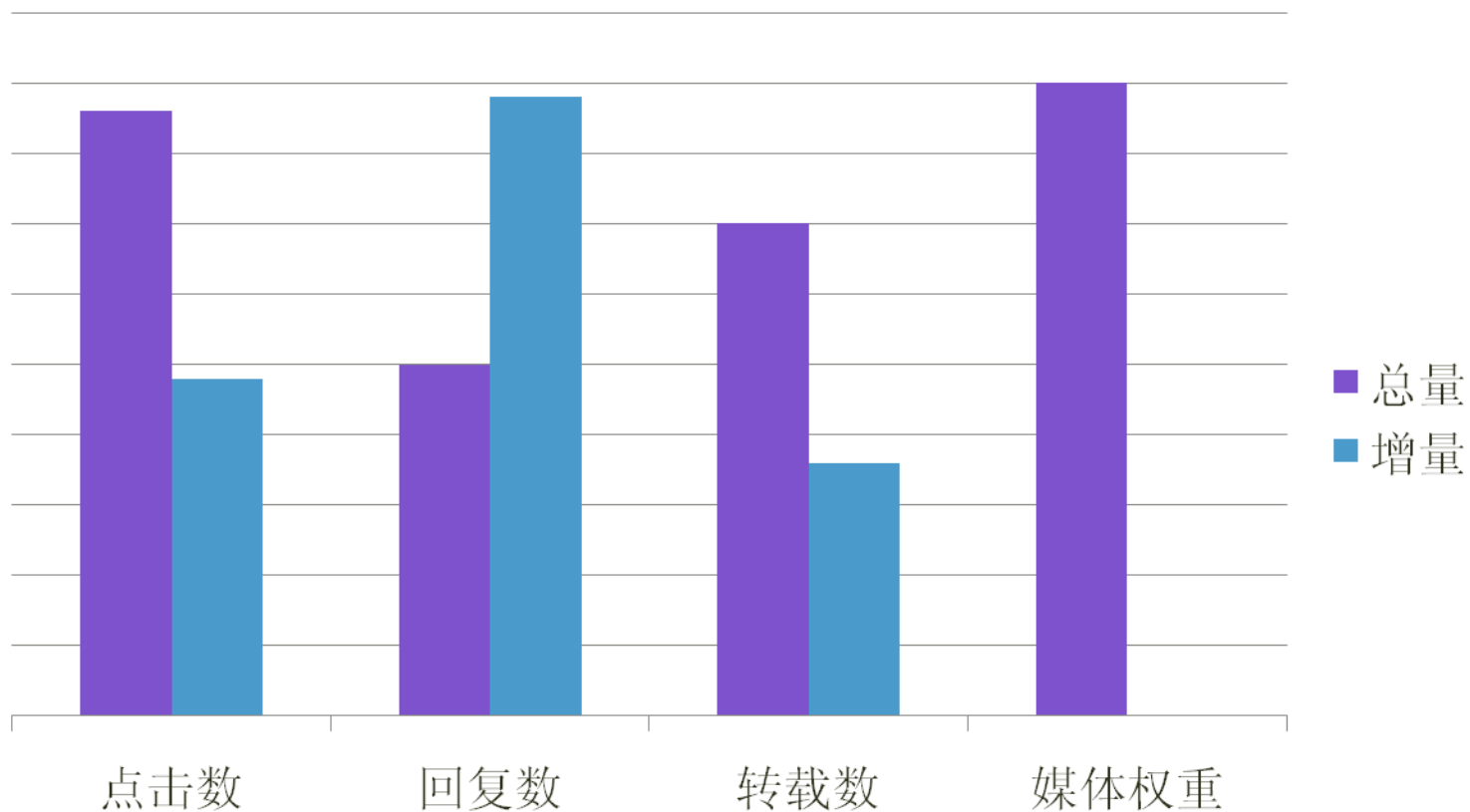
7、其它

分析二：分析是否属于舆情

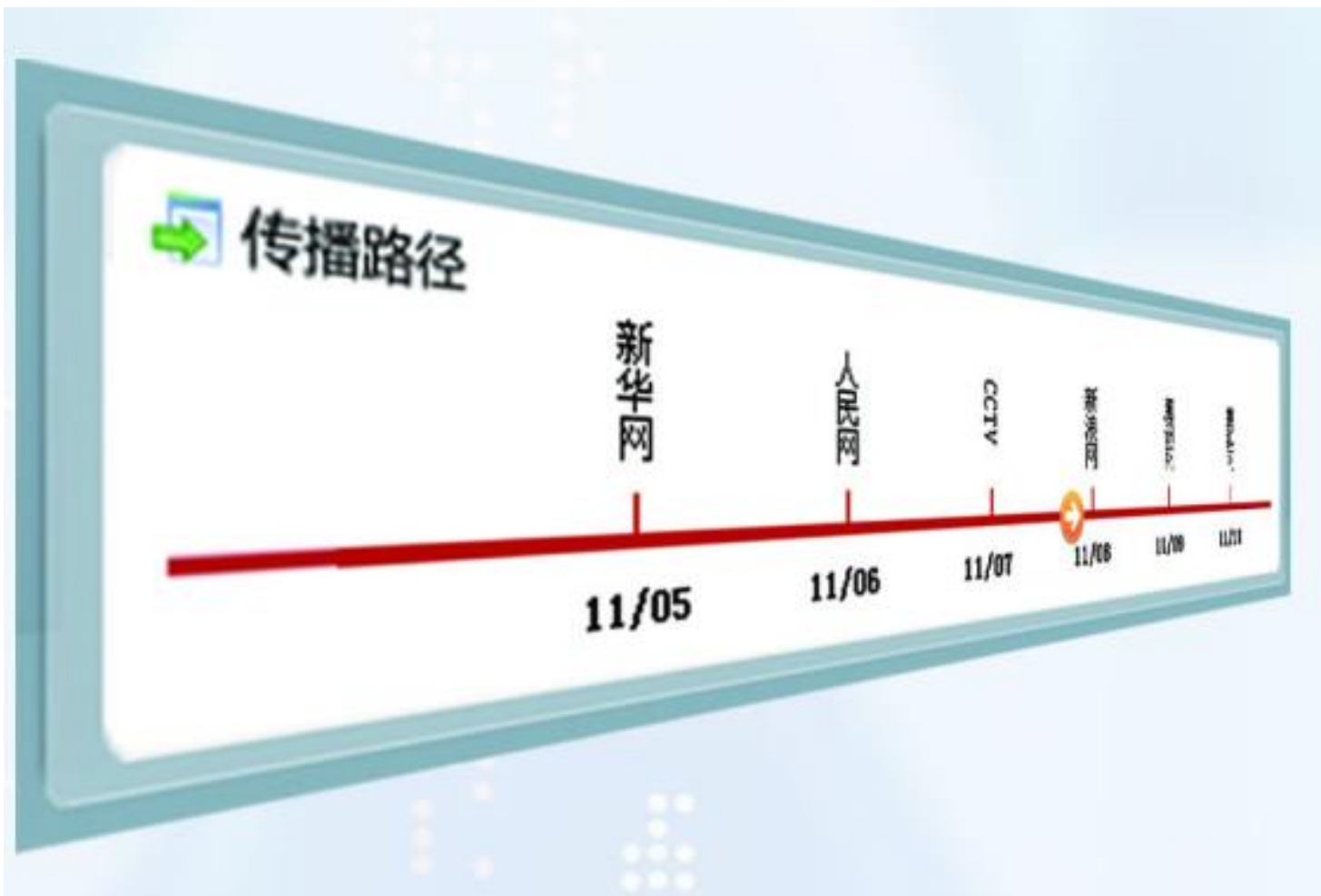
- 3880个舆情维度...



分析三：輿情热度分析(要素)



分析四：传播路径和转载站群分析

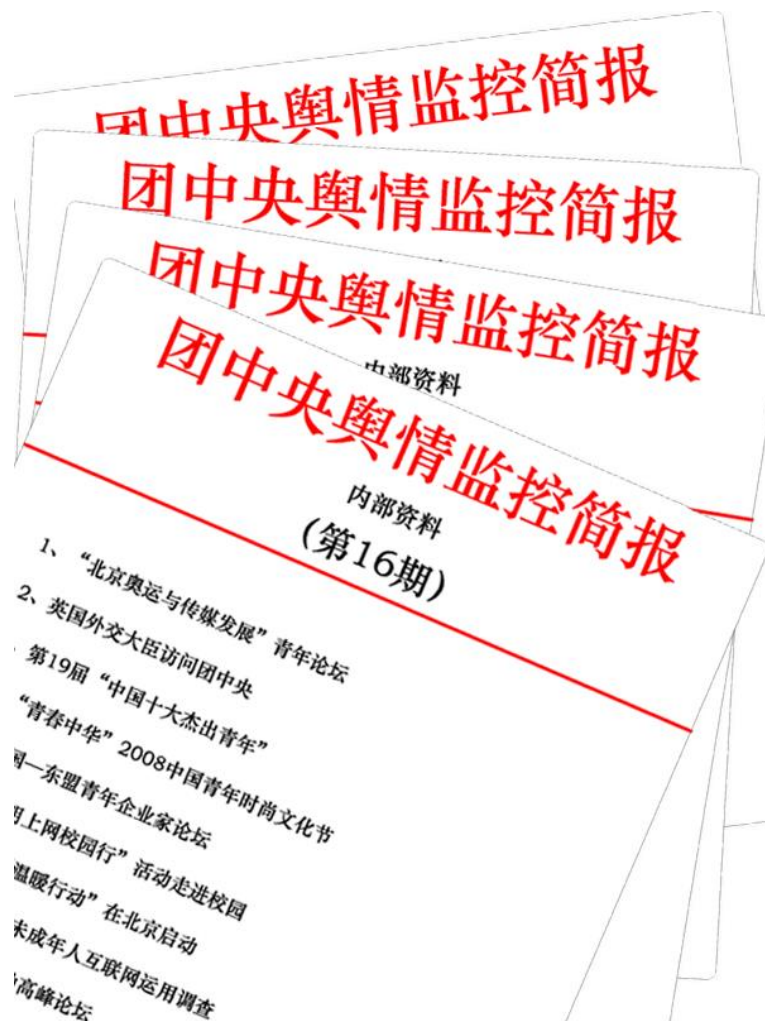


分析五：关联分析(要素)

相关舆情

| | |
|----------------------------|----------------|
| 吕中楼 沁水吕中楼 沁和能源吕中楼 39..... | 07-24 16:45:35 |
| 网贴曝山西沁水数百亿国有资产流失或成“天..... | 07-12 11:32:00 |
| 山西沁水39名老干部检举信续：传吕中楼意..... | 07-15 19:18:17 |
| 吕中楼意欲外逃纪检公安部门应密切关注 | 07-14 12:26:45 |
| 鸿水：谁将改写“中国经济犯罪数额排行榜” | 08-04 01:31:27 |
| 800亿!谁将改写“中国经济犯罪数额排行..... | 08-04 18:20:28 |
| 山西省沁水县39名老干部的一封血泪检举信 | 07-18 14:46:35 |
| [推荐]山西省沁水县39名老干部的一封血..... | 07-08 00:00:00 |
| 沁水39名老干部检举信续：传吕中楼意欲潜..... | 07-10 22:36:05 |
| 他们改写“中国经济犯罪数额排行榜” | 08-03 09:41:24 |

輿情简报&专报



社交网络带来的改变

- 社交网络中的数据，是一部由海量用户UGC的历史著作
- 社会学上的测不准定律：
- 海量用户，实时数据，完整数据 三者同时存在
- 社交网络产生了这些数据
- 社交网络的群众情绪可记录
- 社交网络的二度好友理论
- 群体智慧



社交网络挖掘案例——天气信息蕴藏在社交网络中

- 利用社交网络数据监测天气变化
- 传统方式：卫星云图，实地仪器监测
- 社交网络：通过各个地区的喊热人数进行监测
- 喊热的词有哪些，如何结合语境？

基本算法流程



基于社交网络数据的预测与问题

- 社交网络挖掘面临的挑战
 - 机对语言语境的理解：自然语言处理技术
 - 机器对情感的识别：情感分析算法
 - 垃圾信息检测：水军，僵尸识别
 - 合理的器抽样方法：抽样的准确性
- 微博挖掘的挑战
 - 文本短，特征稀疏
 - 文本口语化，符号化
 - 僵尸，水军多
 - 几个微博传播分析网站
 - <http://www.weiboreach.com/>
 - <http://vis.pku.edu.cn/weibova/weiboevents/index.html>

