第六章 移动互联网内容安全

6.1 手机短信内容安全

6.2 网络新媒体内容安全

- 6.2.1 开源情报分析
- 6.2.2 社交网络分析
- 6.2.3 网络舆情分析

6.1.1 手机短信息

- 手机短信息（或者简称手机短信）是现今社会极为普及的一种信息传播方式。手机短信既可以指短信息本身，也就是被传送的客体，由短信发送者编写，包含需要被传输的消息，长度限制在140字以内。
- 手机短信以手机终端为载体，通过无线传输通道在不同的用户间传输。传输既可以是点对点的，也可以是点对多点的，即群发短信。

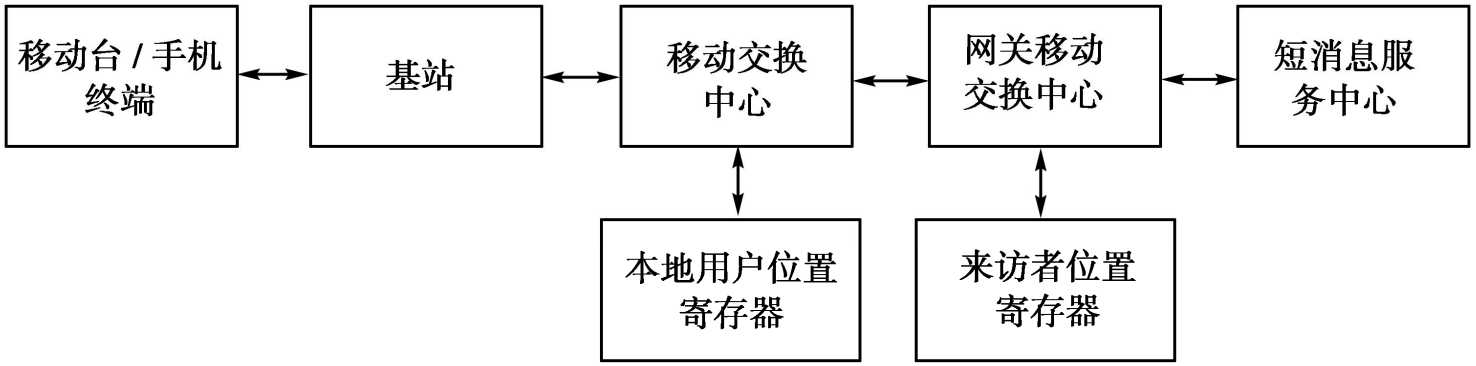
6.1.2 SMS短信

- SMS短信因其易操作性、即时性和廉价性体现了明显的优势，受到广大用户的欢迎。
- SMS采用存储——转发机制。
- SMS的工作原理包括短信传输系统、短信传输结构以及短信发送方式三部分组成

6.1.2 SMS短信

1. SMS短信传输系统

SMS短信传输系统是由移动业务交换中心（MSC）、短消息业务网关移动交换中心（SMS-GMSC）、本地用户寄存器（HLR）、访问者位置寄存器（VLR）和短消息服务中心（SMSC）等部分组成的。



6.1.2 SMS短信

3．SMS短信发送方式

(1) 有线短信发送方式

(2) 发送业务定制方式

(3) 电脑控制手机发送方式

(4) 电脑终端发送方式

6.1.3 手机短信的传播特性

- 为了更好地了解短信传播特性，首先从内容的角度，对短信的构成进行分析。
- 1. 短信构成状况
  - （1）个人短信
  - （2）广告短信
  - （3）互动短信
  - （4）官方短信
  - （5）垃圾短信

6.1.3 手机短信的传播特性

- 2. 短信传播特性
  - (1) 便捷性
  - (2) 交互性
  - (3) 定向性
  - (4) 个性化

6.1.4 不良内容短信简介

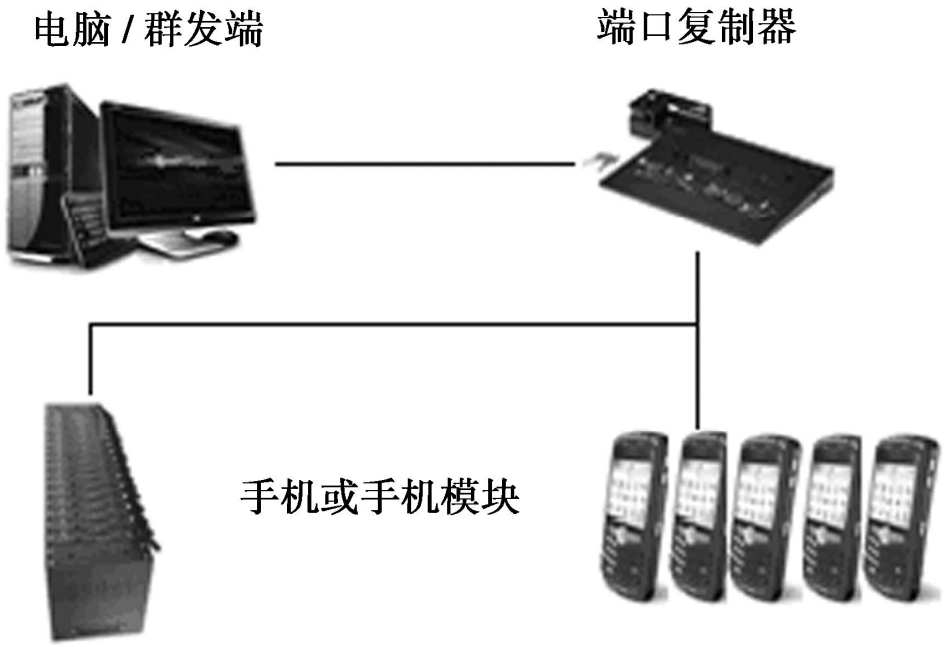
- 1）违反宪法所确定的基本原则的；
- 2）危害国家安全，泄露国家秘密，颠覆国家政权，破坏国家统一的；
- 3）损害国家荣誉和利益的；
- 4）煽动民族仇恨、民族歧视，破坏民族团结的；
- 5）破坏国家宗教政策，宣扬邪教和封建迷信的；
- 6）散布谣言，扰乱社会秩序，破坏社会稳定的；
- 7）散布淫秽、色情、赌博、暴力、凶杀、恐怖或者教唆犯罪的；
- 8）侮辱或者诽谤他人，侵害他人合法权益的；
- 9）含有法律、行政法规禁止的其他内容的。
- 凡是所群发的短信含有上述内容，以及用户认定受到骚扰，或有不良信息的就是不良内容短信。

不良内容短信的特征

- 从网络角度，发送者和接受者之间不存在社会网络关系，不符合正常的短信传播模型；
- 从短信回复率角度，而不良内容短信回复率极低，基本不超过1%；
- 从用户行为角度，不良内容短信发送时，同一时间点上收到不良内容短信的接收号码之间相关性很高，会将同一内容的短信在短时间内大量发送
- 从发送者心理角度，手机号码发送的不良内容短信会尽量用最小的成本实现最大的信息发送量。
- 通过这些特征，可以有效识别和过滤不良内容短信。

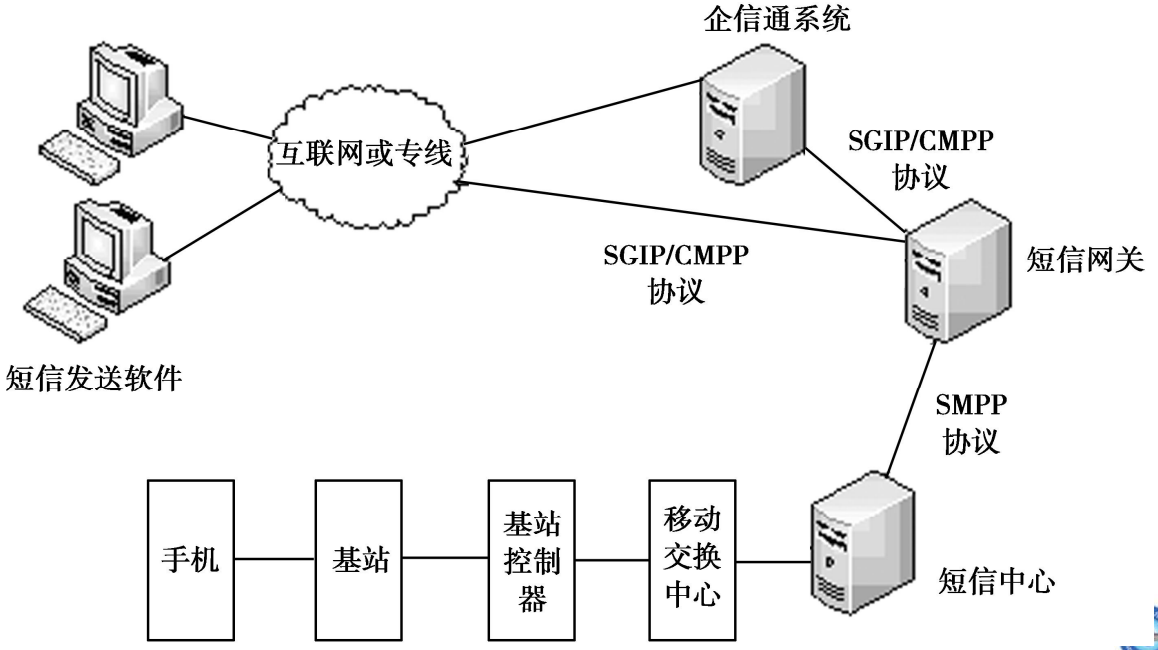
6.1.5 不良内容短信的发送方式

- 手机端口发送不良内容短信多是通过端口复制机将电脑与手机或手机模块链接起来，通过电脑编辑短信内容，经由端口复制机，控制手机或者手机模块进行发送。



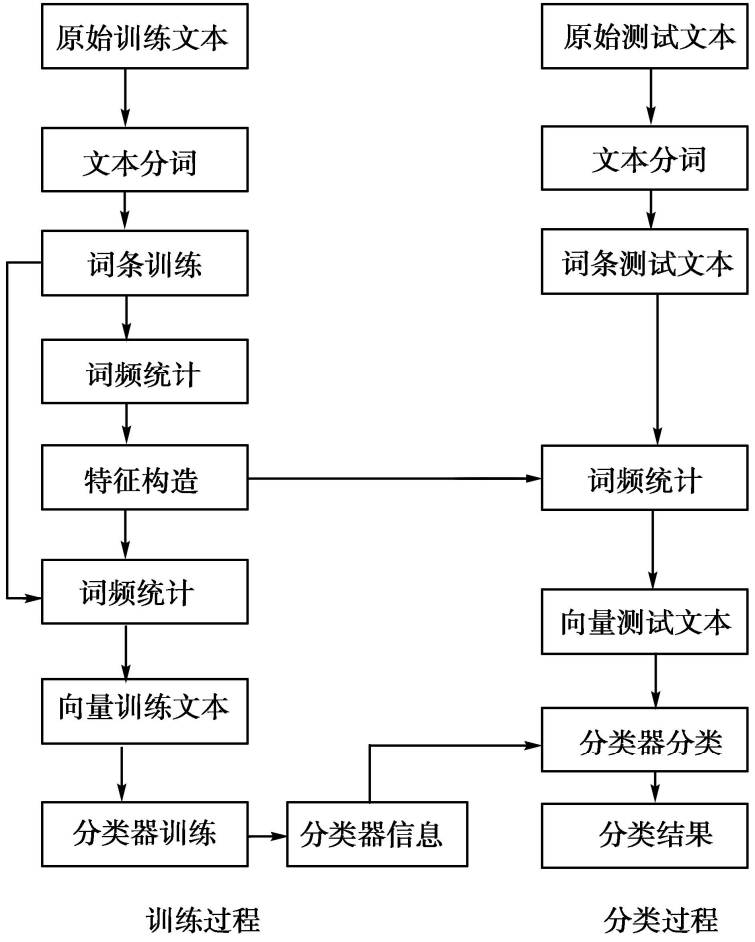
6.1.6 不良内容短信的发送方式

- 网络端口发送不良内容短信采用的群发方法为接入运营商网关，或者直接采用电脑客户端进行发送。该方法具有发送速度快，发送量大的特点。



6.1.7 基于短信内容的识别技术

- 一般而言，文本分类可以分为三个主要的部分
  - 特征构造
  - 分类器训练
  - 分类器在线分类



6.1.8 基于用户的识别

1. 黑白名单技术

- 黑白名单技术优点在于简单高效、处理速度快、系统资源消耗小，易于实施；
- 缺点是需要手动维护黑白名单列表，同时需要及时更新列表的名单，最主要的缺点在于它拦截所有黑名单中发来短信，但是黑名单中用户也可能发送的不是不良内容短信。同时，不加鉴别地接收所有白名单发送来的短信，这其中也可能有不良内容短信。
- 黑白名单技术通常作为一种补充手段。

6.1.8 基于用户的识别

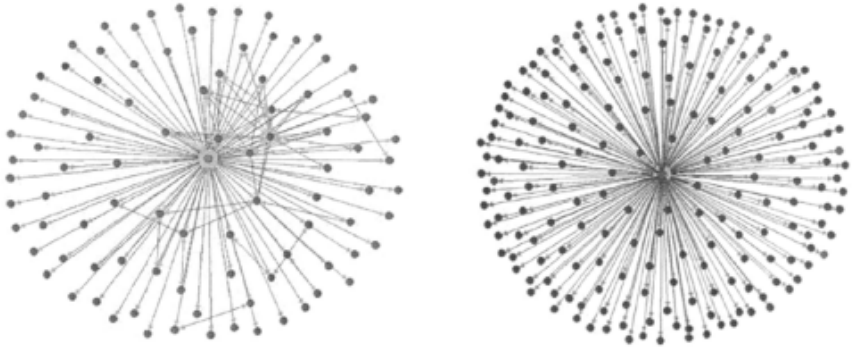
现有系统中，自动生成黑名单的方法主要有如下两种：

- 话单分析机制
  - 机制的优点在于实现简单；
  - 最大的不足是该方案采用了后处理方式
- 在线自动监测机制
  - 优点是对现有系统不产生任何影响,易于实施。
  - 而改造短信中心的方法实施难度较大，一般只能由原SMSC的厂商来实施，而且容易对现有系统构成一定的风险。

6.1.8 基于用户的识别

基于社会网络的用户识别技术

- 从不良内容短信发送的目的性分析，其发送对象多为匿名发送，即发送者与接收者之间不认识。故在网络特性中表现为短信发送者与接收者之间几乎很少存在语音通话记录。



(a)　　(b)

图 6-5　正常短信和不良内容短信发送情况（a）正常用户，（b）不良内容短信用户

6.1.9 不良内容短信识别方法的缺陷

1. 网络端发送短信识别能力不足

（1）短信发送速度快，频率范围宽

（2）无社会性特征，用户一般也不会回复

（3）可供分类识别的特征少

- 基于上述几个原因可知，网络端发送短信的不良内容短信识别难度很大

6.1.9 不良内容短信识别方法的缺陷

- 现有短信中心过滤算法及其不足

（1）内容关键字过滤机制

- 内容关键字过滤机制的优点在于其原理和实现方式都较为简单，应用成本较低。
- 但该机制存在一些重大的局限性：一是关键字选取难度很大，仅通过关键字匹配很难判断出短信的内容合法性，因此很容易造成误判；二是不法分子很容易通过各种方法绕过关键字列表，对此类短信而言，关键字过滤机制形同虚设。

6.1.9 不良内容短信识别方法的缺陷

（2）号码黑白名单过滤机制

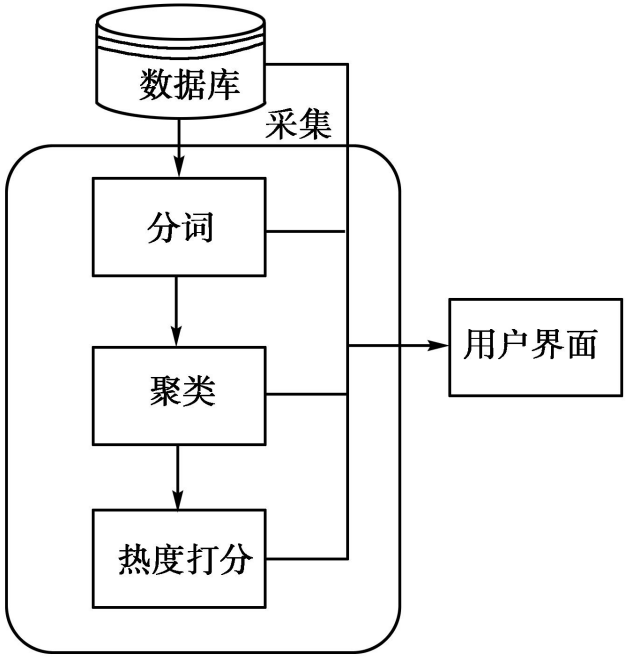
- 这种短信的特点是正常情况下速度不应该很快，数量不应该很多
- 但是针对网络端发送的短信应用系统的不良内容短信过滤，这种短信的特点是用户通过短信接口发送的短信本来就速度快，数量多，所以只能是监控某条同样内容的短信总共发送多少条的机制来监控群发行为。

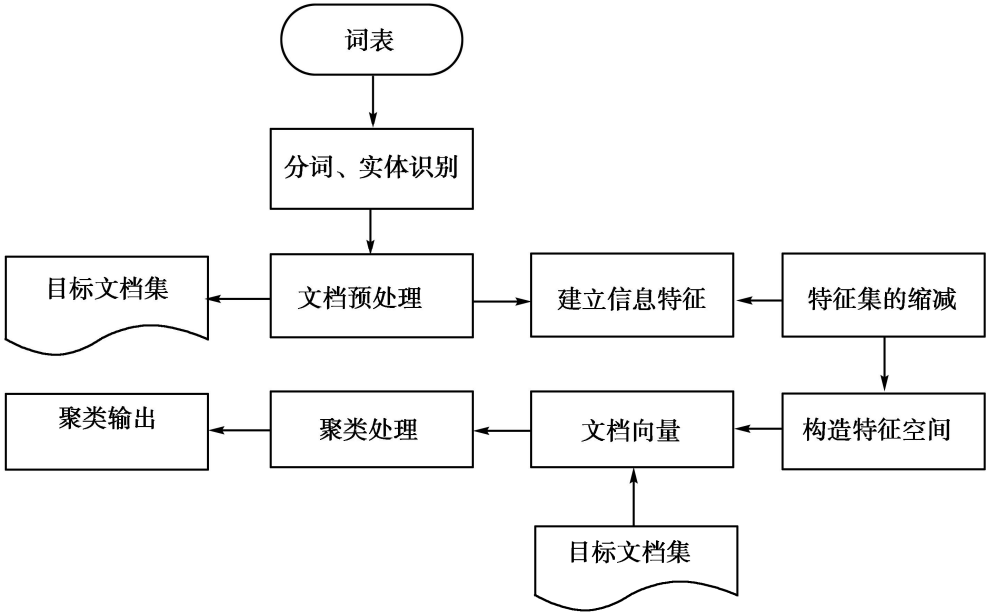
6.1.10 短信热点及其分析必要性

- 随着新的传播技术的快速发展，现在舆情的表达方式越来越多样化、便捷化。
- 网络、短信等传播媒介凭借其特性及其强大的功能改变了以往的新闻和信息传播格局，为公众提供了一个前所未有的自由讨论公共事务、参与政治的活动空间。
- 现在的短信不仅在人们的生活中发挥着重要作用，在政治舞台中发挥的作用也引人注目。

6.1.11 短信话题发现

1. 短信话题发现框架 2. 文本聚类常用算法





6.1.11 短信话题发现

- 常用的聚类算法可以分为以下几类：
  - （1）基于划分的方法——K-means算法
  - （2）基于层次的方法——分裂式层次聚类法
  - （3）基于密度的方法
  - （4）基于网格的方法
  - （5）基于模型的方法

6.1.12 短信话题热度评析

- 热点话题可以从三个方面进行判断：

（1）当谈论一个事件的短信数目越多，在一定程度上可以认为该事件越热；

（2）当谈论一个事件的若干个短信的主题越集中，也从另一个方面说明该事件越热；

（3）当谈论一个事件的短信的平均长度越长，在很大程度上，也能够说明其所谈论的事件越具体也越热。

6.1.12 短信话题热度评析

3个基本参数：类平均长度、类平均相似度和类中文本的数量。

- 聚类过程
  - 计算出类的平均长度
  - 计算类的平均相似度
  - 得到了类平均相似度
  - 为每个类设置类标题

6.2 网络新媒体内容安全

6.2.1 开源情报分析

价值体现：开源情报之概念价值

1.开源情报分析概念

公开的信息或其它资源，包括报纸/刊物、电视、互联网等进行分析后所得到的情报

开源情报概念定义

开源情报数据占比

西方发达国家的国家情报之40%到95%都是以开源情报的形式获取的

西方发达国家较早意识到了开源情报的重要性，成立机构加强获取开源数据

价值体现：开源情报之概念价值

开源情报价值体现

- 情报收集成本小，风险低
- 开源情报内容更加丰富
- 开源情报工作具有隐蔽性

开源情报较于传统数据具有巨大优势

发展历程：开源情报之发展历史

各国发展历史

05年美国成立开放源中心

06年立法启动国家开放源事业计划

欧洲各国定期举办开源情报论坛

瑞士联邦政府建立了跨部门的开源情报工作组

英国建立了专门的开源情报工作

01年澳大利亚就建立了国家开源情报中心

发展历程：开源情报之发展历史

国内研究成果

- 化柏林教授等提出如何把繁杂的大数据进行合理的分析，认为“大数据更需要清洗”
- 2012年，王飞跃提出了面向大数据和开源信息的科技态势解析与决策服务提供了集快速获取文献数据并支持半自动化的从多维角度进行文献解析的框架
- 上海科技情报所建立了以开源情报为基础、面向行业情报服务的第一情报网

国内开源情报的价值未得到充分挖掘

依据指标：开源情报之可靠度

信息源可靠度

评价指标信息源是指传播信息的机构，如报社、出版社、电视台、广播台、政府宣传机构等。第一手信息源能直接接触和完整传递信息，可靠性较高

依据指标：开源情报之可靠度

信息源的可靠程度可依据信息特征来推断

- 形式特征：包括信息源网站、纸质出版物、电子出版物内外包装等产品或媒介的排版美工水平等外在指标
- 组织特征：被评价的信息源是否由一个合法组织来管理运营等相关资格特征
- 链接特征：考察它的链接是否为死链，是否指向可靠性较低的信息源等
- 价值特征：可靠性较高的信息源会围绕某领域、某主题展开报道和论述

依据指标：开源情报之可靠度

信息内容可靠度

1、明确公开源数据、公开源信息和公开源情报的区别

2、考察信息所表述的内容足否合情合理

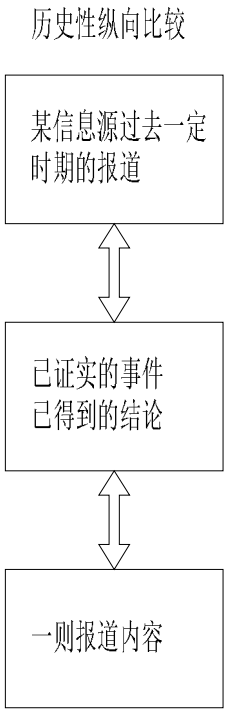
3、高可靠性的内容一般行文直截了当、清晰准确

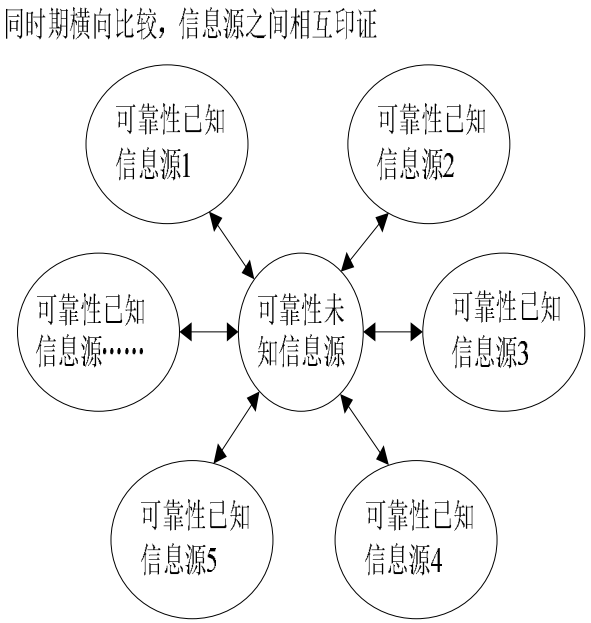
4、从参考引用文献角度考察，高可靠性的内容会为数字、主要观点标引出处

依据指标：开源情报之可靠度

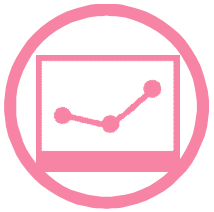
最终情报产品的质量在很大程度上取决于评价者的水平。一旦评价阶段有所偏颇或谬误，很可能导致决策失误

- 不同信息源类别的转化问题
- 中文信息的自动过滤技术
- 不同信息源对某一事件或观点的评判相互矛盾、不易区分





分析方法：开源情报之数据分析



数据定量分析

情报分析越来越多地依赖以计算机为代表的信息技术，利用数据挖掘、机器学习、统计分析等方法，运用关键词词频、词汇共现、文献计量等定量化手段，通过计算或者在计算的基础上辅以人工判断形成分析结论

多源数据融合

把通过不同渠道、利用多种采集方式获取的具有不同数据结构的信息汇聚到一起，形成具有统一格式、面向多种应用的数据集合

相关性分析

两个或者两个以上变量的取值之间存在某种规律性，当一个或几个相互联系的变量取一定的数值时，与之相对应的另一变量的值按某种规律在一定范围内变化

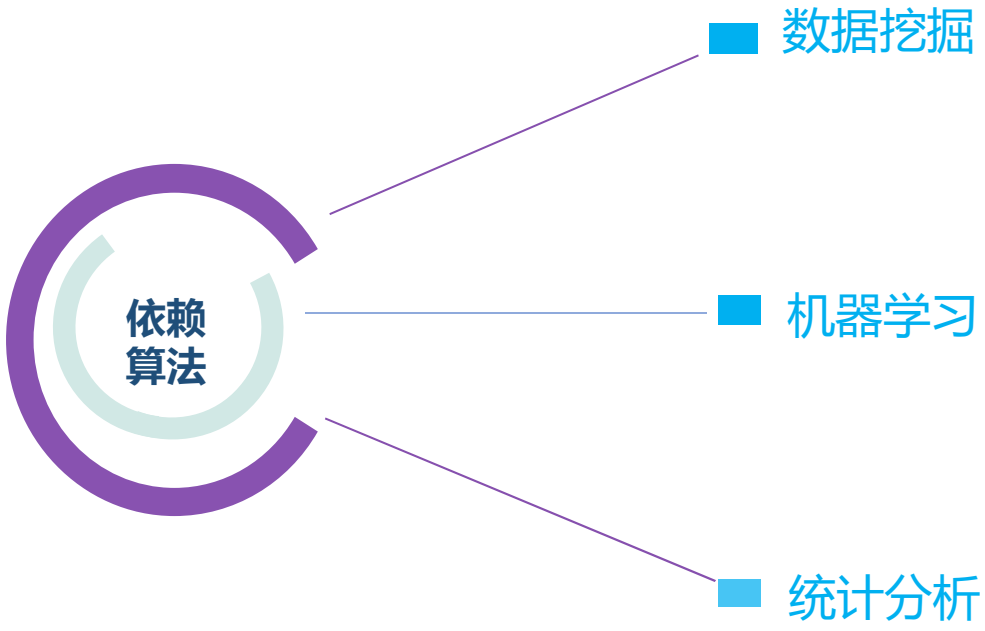
分析方法：开源情报之数据分析

数据定量分析

- 聚类分析
- 关联规则挖掘
- 时间序列分析
- 社会网络分析
- 路径分析
- 预测分析

分析方法：开源情报之数据分析

早期的情报分析更多地依靠人的智力去解读特定的、少量的数据对象，通过人的分析、归纳和推理得出情报研究的结论。随着科学技术的迅猛发展，仅靠人力本身已经无法胜任情报分析工作了，情报分析越来越多地依赖以计算机为代表的信息技术



分析方法：开源情报之数据分析

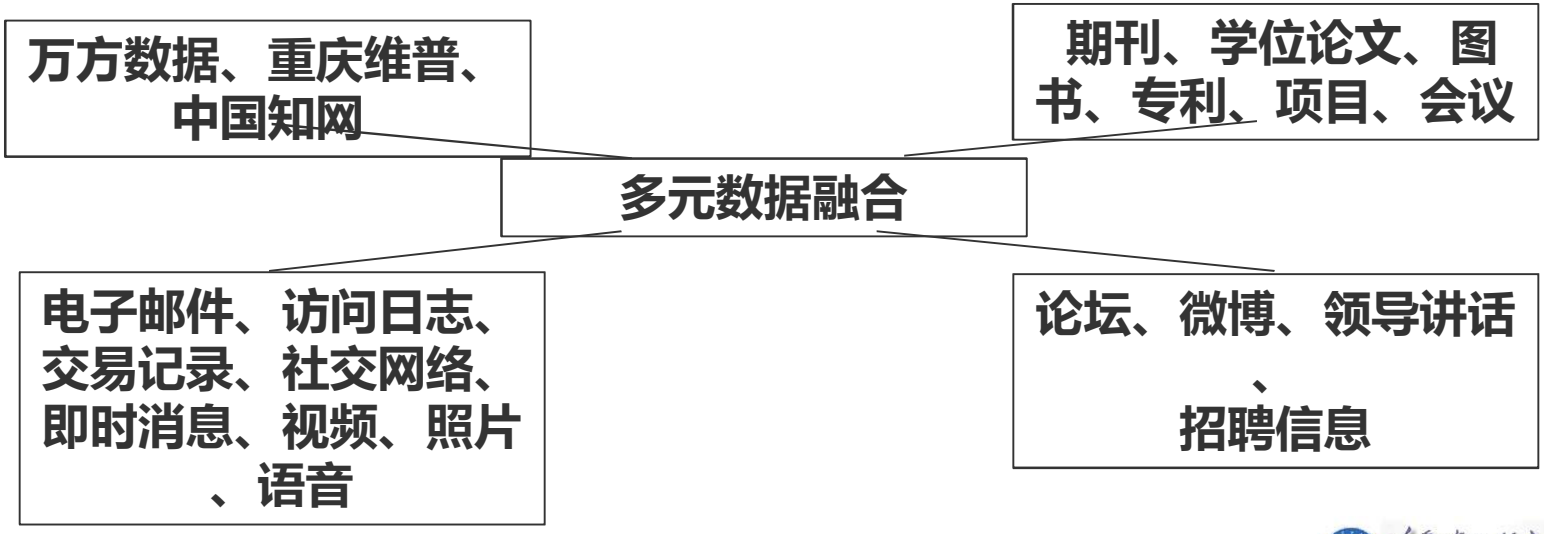
多元数据融合

数据来源渠道

- 不同用户、不同网站、不同来源渠道

数据呈现形式

- 音频、视频、图片、文本等



分析方法：开源情报之数据分析

相关性分析

大数据时代在数据处理理念上有三大转变

- 要全体不要抽样
- 要效率不要绝对精确
- 要相关不要因果

大数据环境下了解数据之间是什么关系，而没必要弄清楚为什么

系统框架：开源情报之具体处理

系统框架

情报采编报子系统

信息采集层依托开源情报数据采集体系，根据采集策略，实时准确采集来自不同数据源的数据，并对数据进行抽取结构化等清洗预处理

01

02

建立并更新原始素材库，为系统提供基础数据。实现数据的归类存储与数据更新。能够按数据来源分类存储原始数据，形成原始资源库，并对其做索引，供系统对原始信息的查找

情报感知分析子系统

03

主要实现提供各种动态快讯、智能简报、热点分析报告、专题深度报告、统计分析报告、季度/年度研究报告、多功能检索、分类导航浏览等功能

大数据服务提供子系统

系统框架：开源情报之具体处理

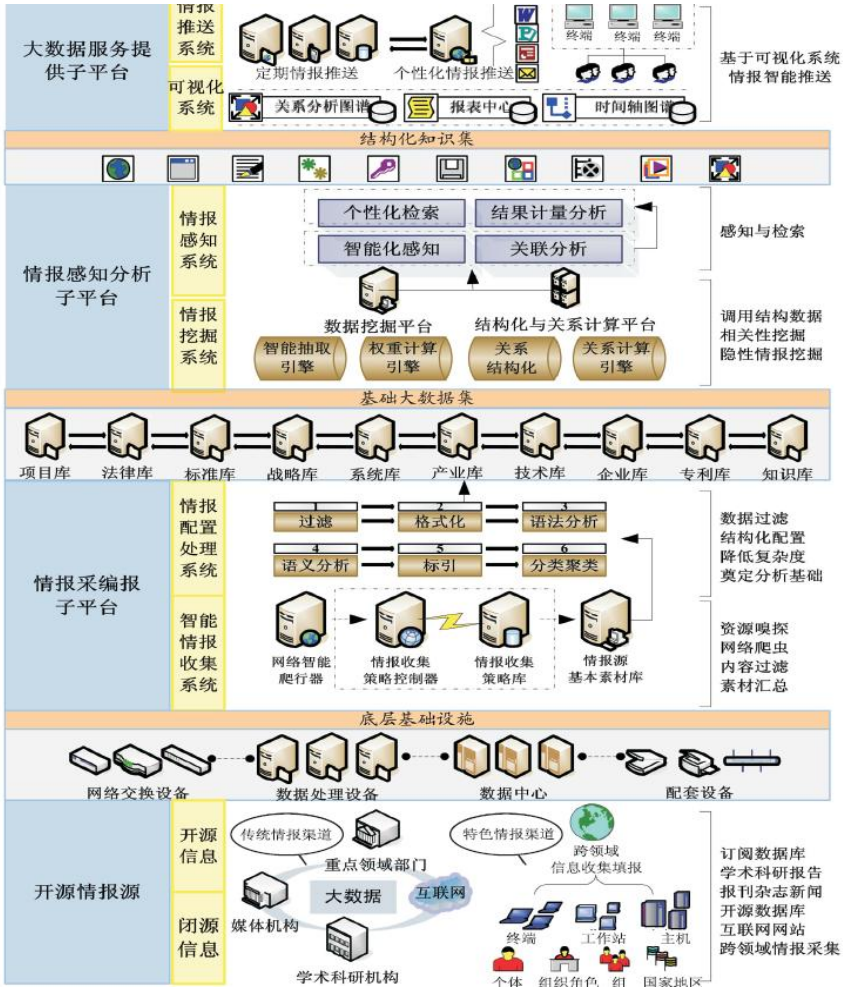
情报采编报子系统

- 实现对网络爬虫获取的原始网页信息做结构化数据抽取
- 支持流数据及动态网页信息的抽取
- 支持网页中内嵌各种文档格式的下载与解析
- 对通过各接口获取的数据，有些需要识别其应用层协议、数据解密之后再抽取其结构化的数据

系统框架：开源情报之具体处理

情报感知分析子系统

- 底层挖掘，即实现文本挖掘的预处理和通用挖掘流程，形成挖掘资料库
- 实时存储，以数据库和文件两种形式存储并索引，按策略做更新，实现多维度检索库
- 定向跟踪，对特定关注对象的定向跟踪分析
- 热点挖掘，热点信息自动聚类，通过机器学习　自动发现热点
- 统计分析，支持对入库信息的智能统计报表
- 演变分析，关注对象的发展、扩散、分布等分析



系统框架：开源情报之具体处理

处理流程

信息采集业务

01

信息采集的主要任务是将互联网、标准资源库、企业资源库、现有工程数据、内部资料和其它来源的数据收集起来，形成原始数据。对采集到的原始数据，做一定的预处理、进行粗分类并存储，形成原始素材库，存储客观的基础素材，并对原始素材库做索引以支持原始信息的定位

02

实现对开源情报做深度挖掘加工，自炼信息关键词、摘要，针对结构化后的数据做索引动提

开源情报加工与分析业务

03

存储情报服务和产品的历史数据。将平台的服务和产品采用多种方式发布、推送给不同的用户,包括订阅、热点周报、专题报告及年度汇总报告等

情报展示与服务业务

系统框架：开源情报之具体处理

信息采集业务

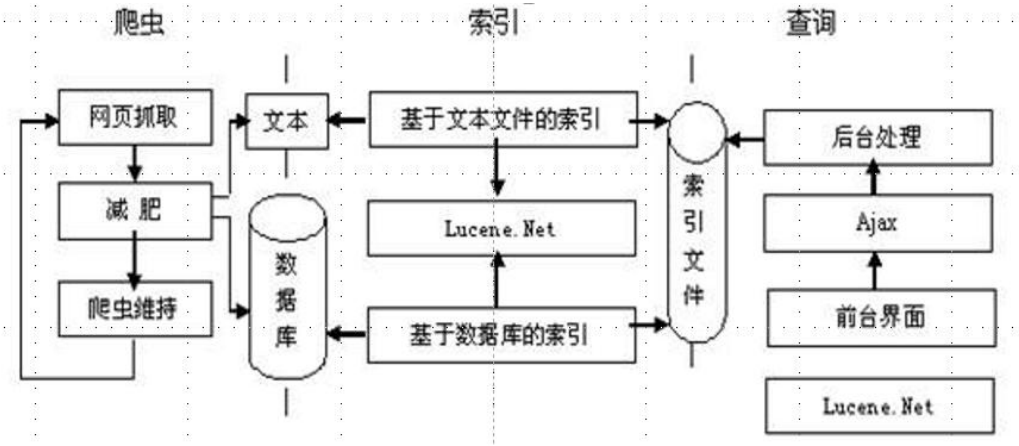
- 爬虫策略设置。首先根据用户提供的主题关键词、相关文档，训练主题向量，并形成训练库，将训练好的主题向量存储在主题向量库中。其次根据用户需求配置爬虫的采集规则和更新频率

- 数据采集。在每一轮数据爬取过程中，爬虫根据设定的采集规则和URL 得分选择一定数量的URL来抓取，接着解析原始网页,提取网页正文和外链。针对每一个外链，根据其对应锚文本与主题向量的相关度赋予分值，各个待抓取链接按照得分高低排序，使得那些主题相关的网页得到优先抓取。同时根据用户设定的更新频率对网页库中已经过期的网页重新采集

系统框架：开源情报之具体处理

开源情报加工与分析业务

- 底层挖掘层：将获取初始数据进行清理并得到规范后每条记录的元数据，之后对其中的文本信息进行分类与聚类
- 上层信息检索部分：全文检索、摘要检索、主题检索、关键词检索高级检索五大功能
- 上层智能分析部分：发现、演变分析、预测三个关联度比较大的功能



未来趋势：开源情报之应用趋势

引入大数据	• 在信息采集、整序、组织、检索、分析和可视化等方面成熟的理论方法和技术应用到大数据的工作。
应用大数据	• 要寻求情报研究的客观性，摒除过多的主观意愿，也需要多种技术来支撑，这一发展趋势是大数据时代下的必然。
探索大数据	• 以开源信息为主，汇集海量数据，通过定量的方式来描述、分析、评判科技发展的态势，服务于科技决策。
利用大数据	• 借助数据挖掘技术，建立与闭源知识对象的索引和相互关系，组建以情报领域知识库，构建情报分析人员专用的情报池。
研发大数据	• 预测未来，不如创造未来。

6.2 网络新媒体内容安全

6.2.2 社会网络分析

概述

社会网络指的是社会行动者（social actor）及其间关系的集合。

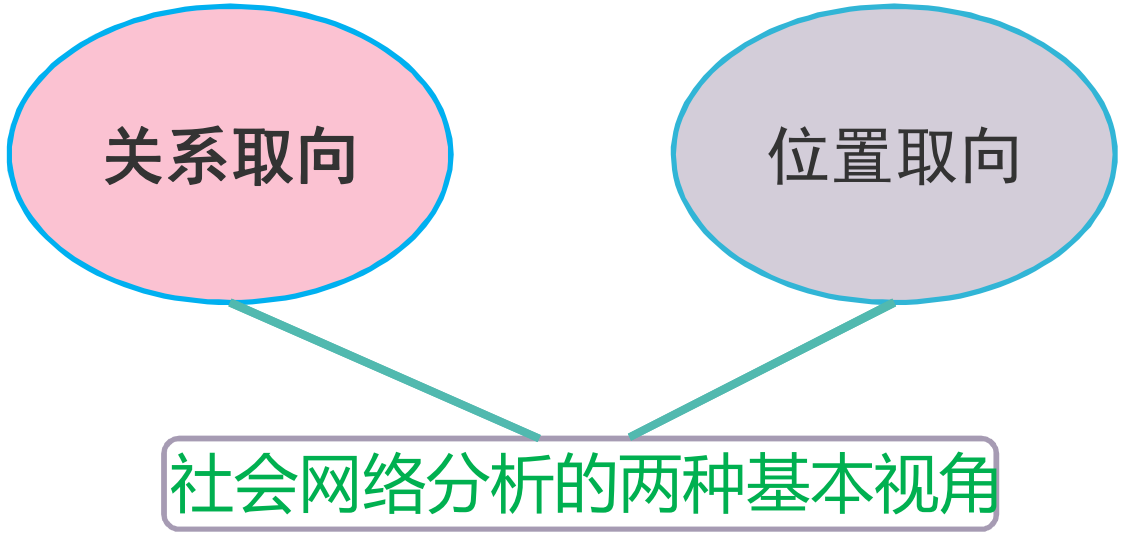
概述

社会网络

- **点：个体、单位、教研室、系、学校，村落、组织、社区、超市、国家等，也包括网上虚拟社群的成员或社群本身。**
- **关系**
  - ①个人之间的评价关系：喜欢、尊重等
  - ②物质资本的传递：商业往来、物资交流
  - ③非物质资源的转换关系：行动者之间的交往、信息的交换
  - ④隶属关系：属于某一个组织
  - ⑤行为上的互动关系：行动者之间的自然交往，如谈话、拜访等
  - ⑥正式关系（权威关系）：正式角色也是有关系性的，如教师/学生、医生/病人、老板/职员关系等
  - ⑦生物意义上的关系：遗传关系、亲属关系以及继承关系等

概述

社会网络分析主要是研究社会实体的关系连结以及这些连结关系的模式、结构和功能。



概述

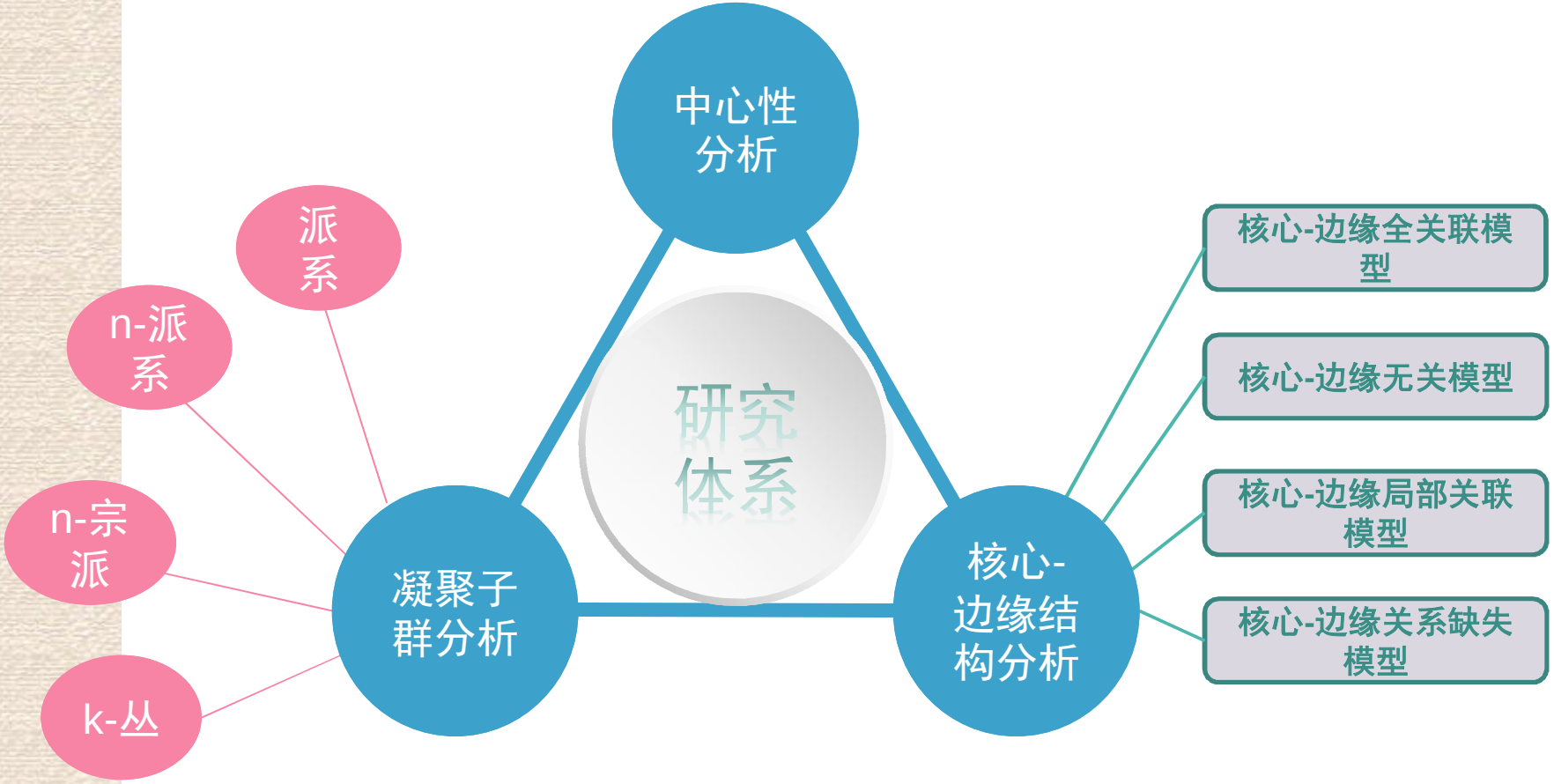
关系取向分析内容

- 规模（range）
- 强度（strength）
- 密度（density）
- 内容（content）
- 不对称关系（asymmetric relation）
  与对称关系（symmetric relation）
- 直接性（direct）与间接性（indirect）

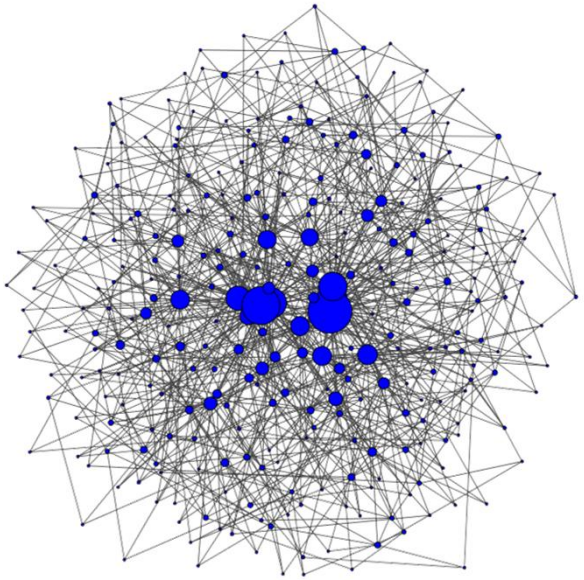
位置取向分析内容

- 结构等效（structural equivalence）
- 位置（position）
- 角色（role）

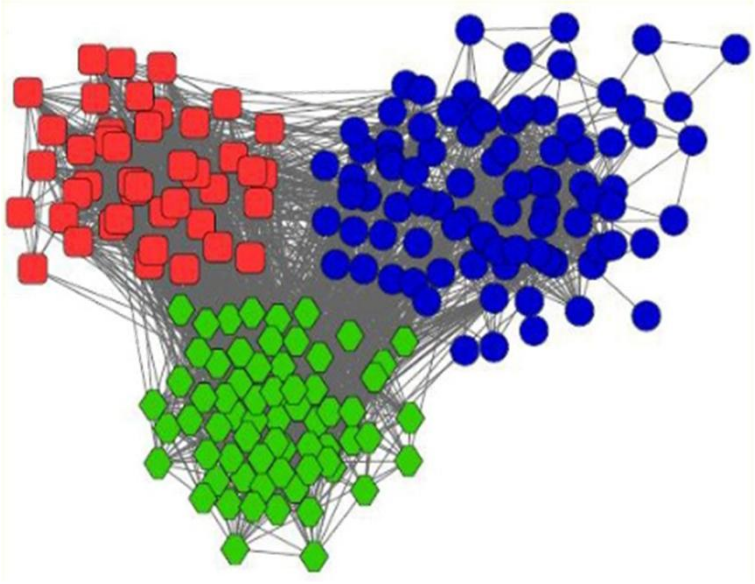
研究体系


中心性分析
派系
n-派系
n-宗派
k-丛
凝聚子群分析
研究体系
核心-边缘结构分析
核心-边缘全关联模型
核心-边缘无关模型
核心-边缘局部关联模型
核心-边缘关系缺失模型

分析模型

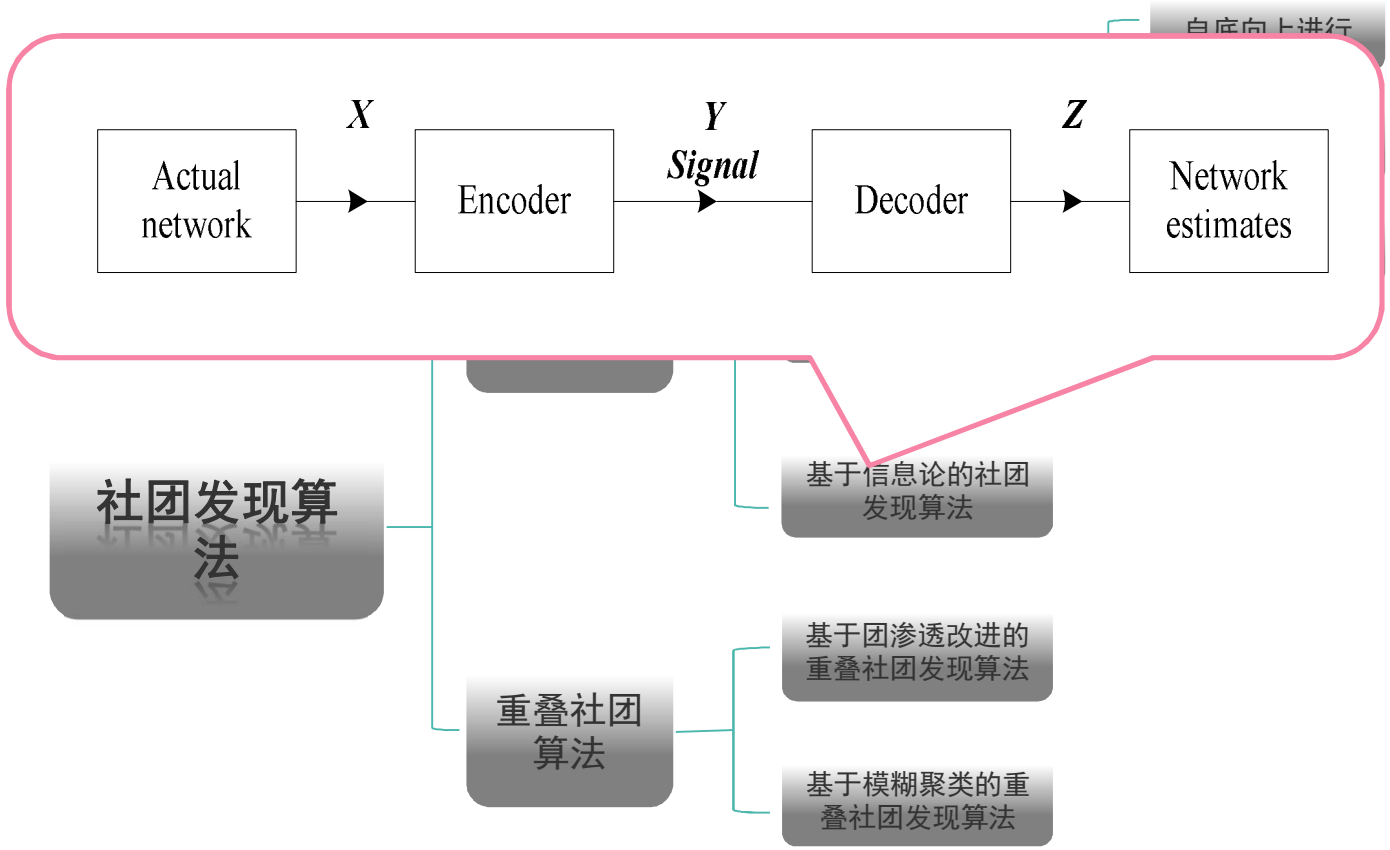


BA模型

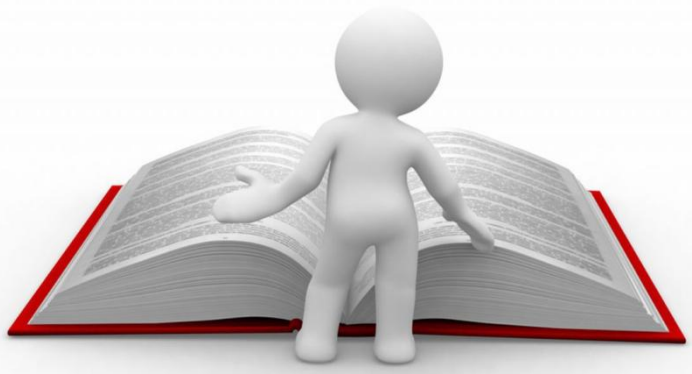


社团结构模型

分析模型


Actual network
X
Encoder
Y
Signal
Decoder
Z
Network estimates
社团发现算法
基于信息论的社团发现算法
重叠社团算法
基于团渗透改进的重叠社团发现算法
基于模糊聚类的重叠社团发现算法

分析方法



社会网络分析常用方法

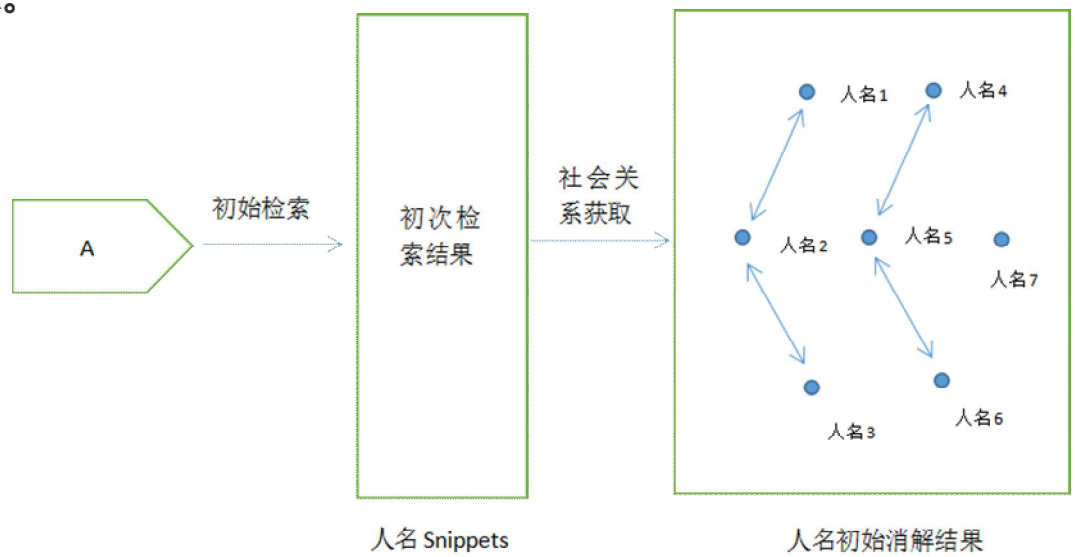
基于命名实体检索结果的社会网络构建

基于内容分析的社会网络构建

分析方法

一、基于命名实体检索结果的社会网络构建

例：以检索人名A为例，初始检索返回一组Snippet，抽取每个Snippet中的人名。假设任何两个人名共同出现在某个Snippet中就认为两人具有社会关系，共现的次数作为这种关系的度量。从而可以对出现在所有Snippet中的人名构建关系矩阵，矩阵元素，表示人名i和人名j的共现次数。由于是利用人名A的社会网络来对人名A检索得到的有效Snippet进行重名消解，关系矩阵中不包含人名A。



人名A初始关系

分析方法

二、基于内容分析的社会网络构建

在对输入文章进行分词标注、共指消解等预处理之后，通过名词合并及主动词识别，得到存在关系的实体之间的关系指向和关系描述，最后通过有向图把存在关系的实体进行链接，最终形成有向关系网络。这样不仅能够通过对一个新闻事件的分析得到对事件中实体之间的关系的指向，更能根据关系图中每个点的出度、入度确定各个实体在事件中的重要程度，而且可以确定点与点之间的相对关系紧密程度，并给出比较合理的点与点之间关系的描述。

优点：

- 基于文本内容分析，结果更加可靠
- 对所有的不同实体之间的关系进行抽取
- 采用有向图对社会网络进行可视化表现

分析方法

基于浅层分析与机器学习的汉语零指代消解

具体步骤：

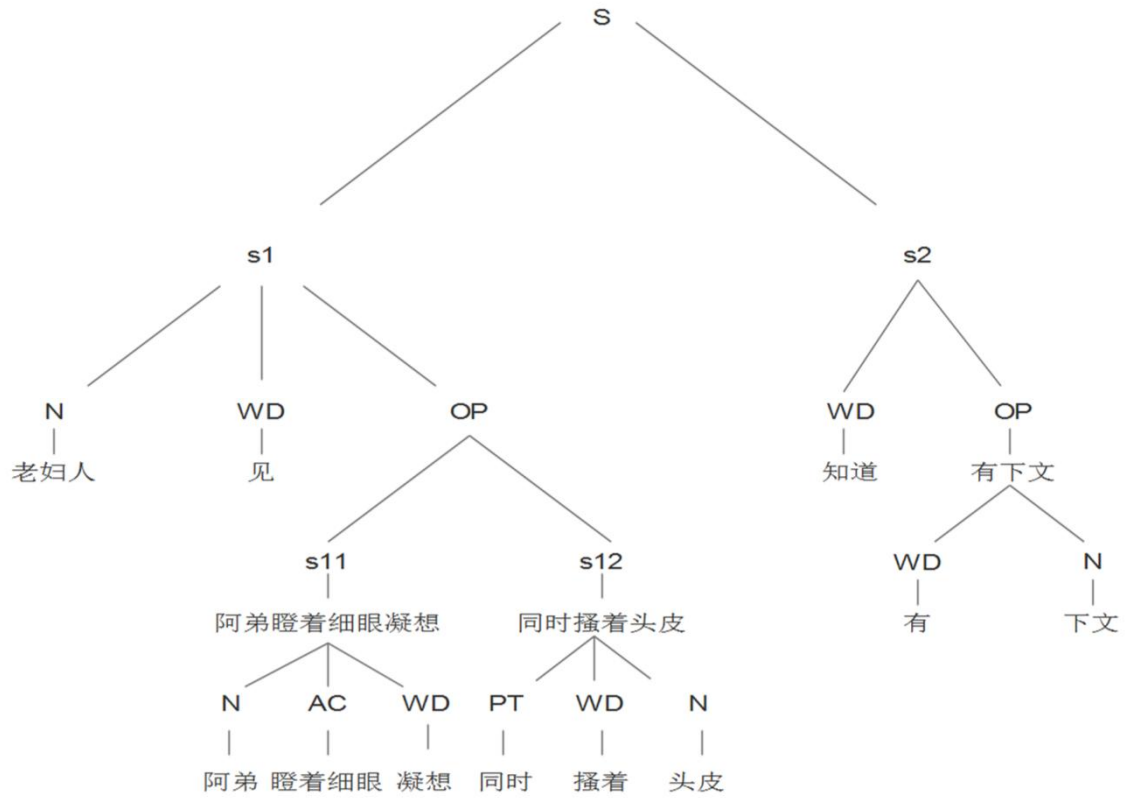
①名词合并

规则序号	规则描述
1	名词+和+名词（N+HE+N）
2	名词+以及+名词（N+以及+N）
3	名词+的+名词（N+DE+N）
4	名词+和+名词+的+名词（N+HE+N+DE+N）
5	名词+和+名词+和+名词+的（N+HE+N+HE+N+DE）
6	名词+的+名词+和+名词（N+DE+N+HE+N）
7	名词+的+名词+和+名词+的（N+DE+N+HE+N+DE）
8	动词+名词+的+名词（V+N+DE+N）
9	动词+名词+名词（V+N+N）
10	动词+的+名词+名词（V+DE+N+N）
11	名词+名词（N+N）
12	介词+名词+名词（P+N+N）
13	介词+名词+的+名词（P+N+DE+N）
14	介词+动词+的+名词+名词（P+V+DE+N+N）
15	介词+动词+名词+名词（P+V+N+N）

分析方法

分析方法

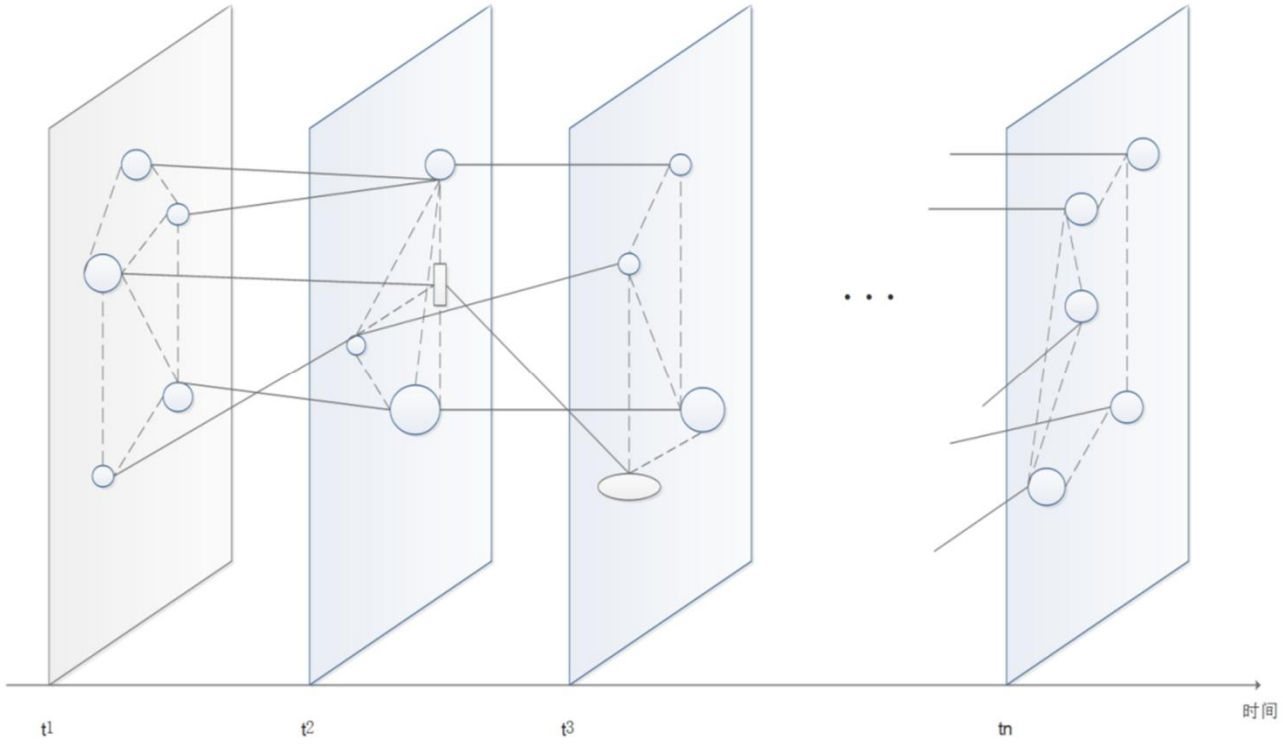
例：“老妇人见[阿弟瞪着细眼凝想，同时搔着头皮]，知道有下文……”



话语片段层次分析结果

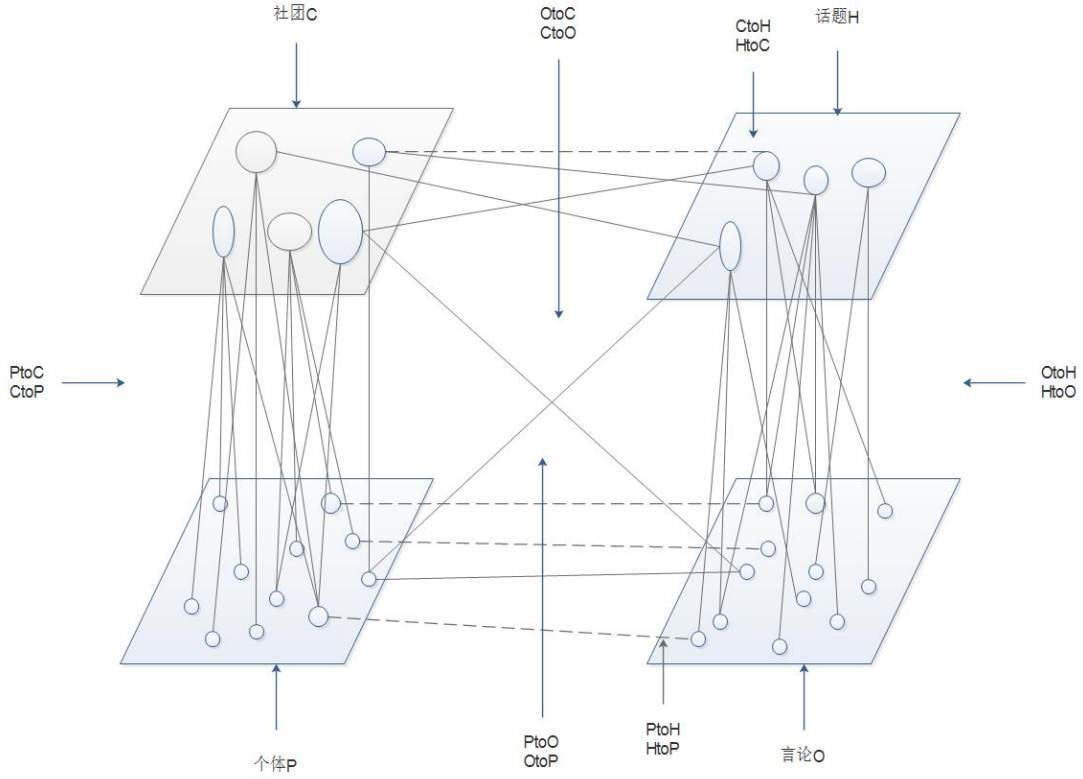
安全应用

社团和话题之问具有密切的关系



安全应用

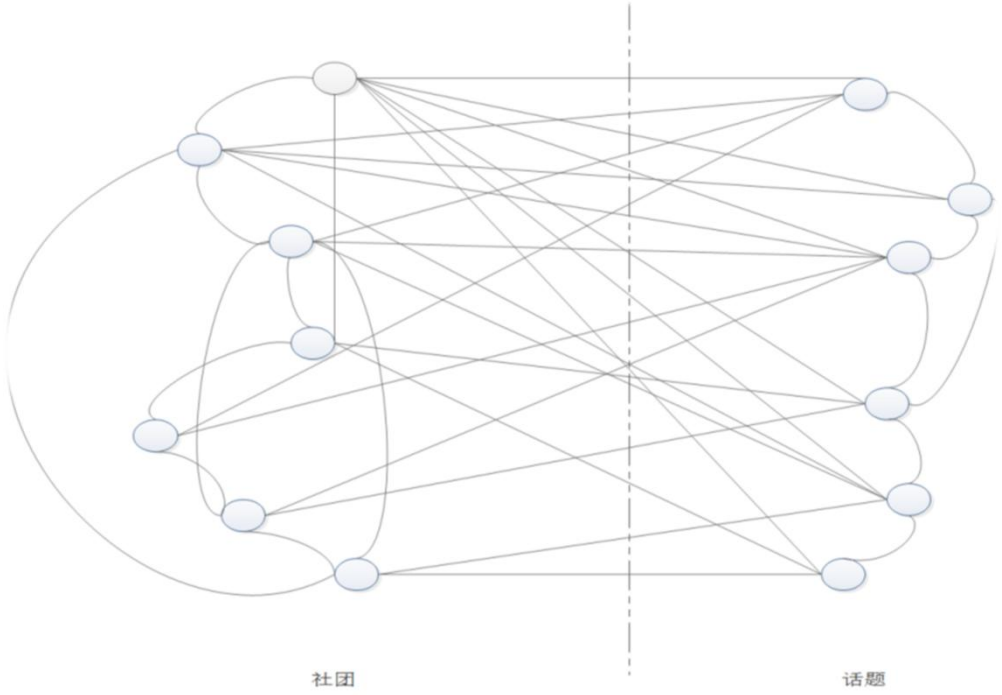
社团挖掘和话题监控的互动模型



个体概念和函数的示意图

安全应用

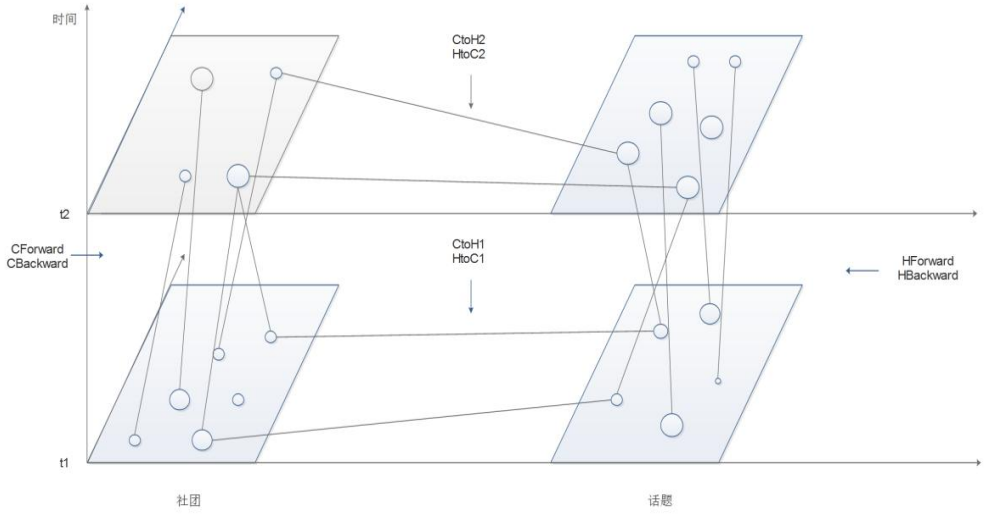
$\forall p_1, p_2,$如果$p_1 \neq p_2$，那么$PtoO(p_1) \neq PtoO(p_2)$



社团挖掘和话题监控的二分图模型

安全应用

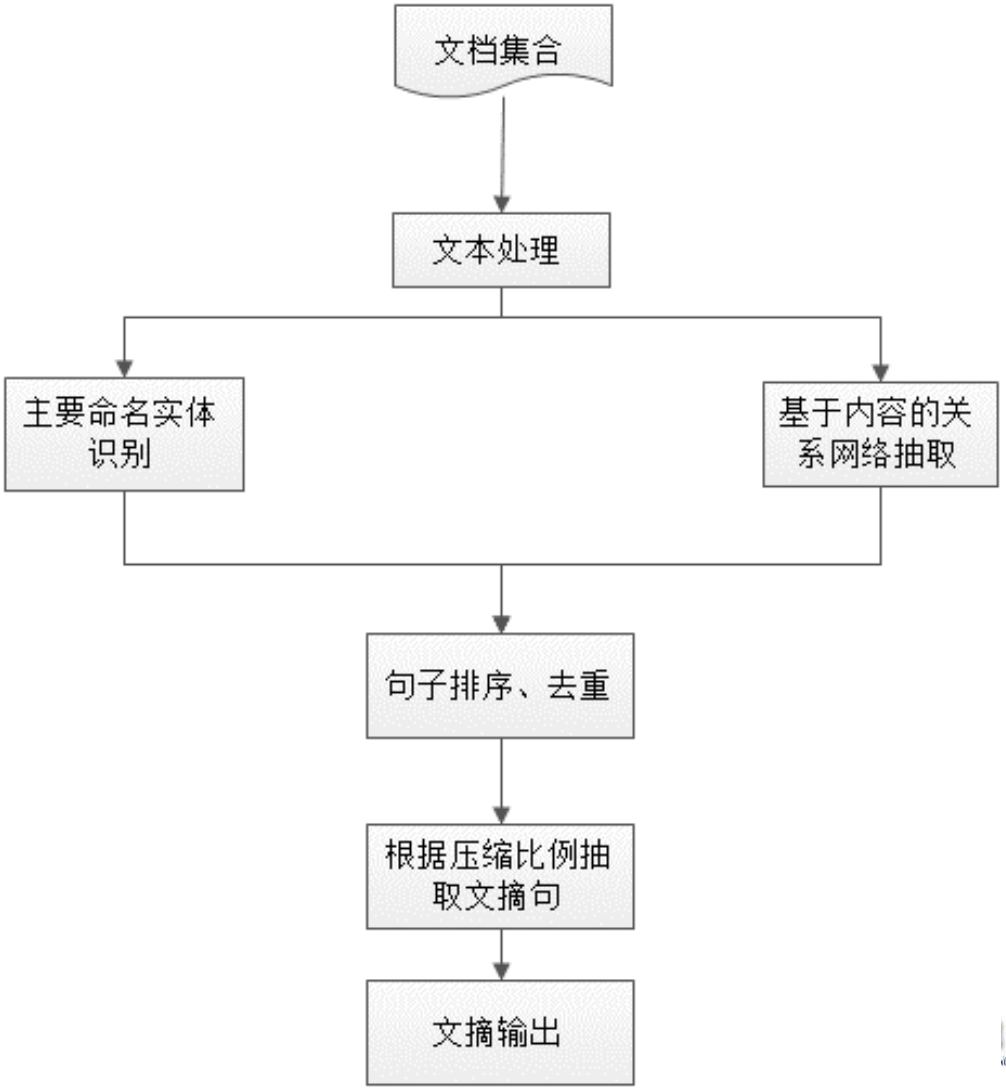
在静态模型中增加时间维就可以得到社团演变和话题演变的动态互动模型，即把上面讨论的各个概念，比如P、O、C和H都放入到一个事件空间来考虑，那么它们都是动态变化的。特别地，社团跟踪和话题跟踪的任务就是找出不同时刻的社团、话题之间的关系，模型见右图：



社团演变和话题演变动态互动模型图

安全应用

进行文摘方法

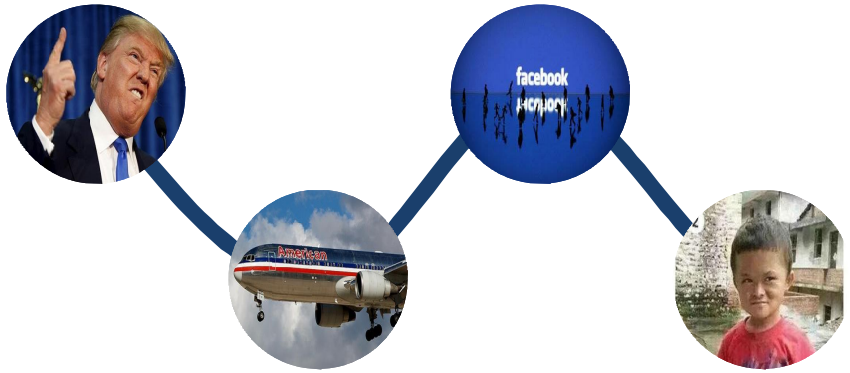


发展趋势

发 展 趋 势

①节点重要性的定义

③网络结构和网络行为是
如何影响节点重要性评价



②各种指标间的内在联系

④如何在这种具有大数据特征的
时变网络中对节点重要性排名

6.2 网络新媒体内容安全

6.2.3 网络舆情分析

网络舆情分析概述

人大选举…
总统换届…
…
萨德部署…

两会召开…
…
经贸合作…

网络舆情

舆情指在一定的社会空间内，围绕中介性社会事项的发生、发展和变化，作为主体的民众对作为客体的国家管理者产生和持有的社会政治态度。如果把中间的一些定语省略掉，**舆情就是民众的社会政治态度**。

网络舆情分析概述

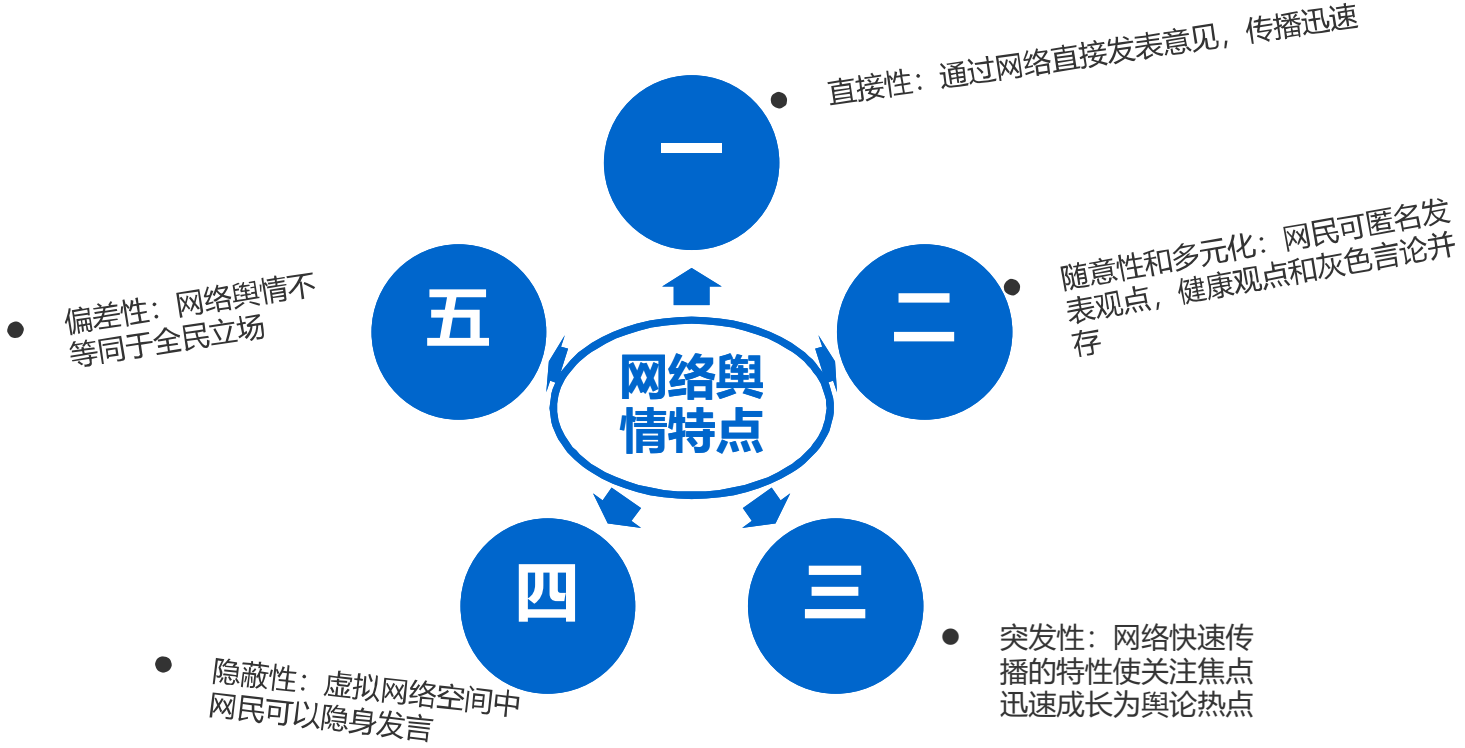
网络舆情主要表现形态

表现形式分类

- 文本
- 图像
- 音频、视频等

网络舆情分析概述

网络舆情分析的特点



网络舆情分析概述

人工舆情监控存在问题

- 舆情收集不全面
- 舆情发现不及时
- 舆情分析不准确
- 舆情利用不便利

网络舆情分析应具备功能

- 舆情分析引擎
- 自动信息采集
- 信息抽取

网络舆情分析关键技术

信息采集技术　热点发现　热点评估　主题跟踪　分析处理

网络舆情分析关键技术

①信息采集技术

网络舆情分析关键技术

②热点发现技术

主要算法：

- Single-pass：动态聚类表现良好
- K-means：基于硬划分的无监督聚类算法
- KNN：基于类比学习的非参数分类技术
- SVM：多热点事件识别
- SOM：人工神经网络

Single-pass 聚类算法

K -means

KNN 最邻近法

SVM支持向量机

SOM 神经网络聚类

网络舆情分析关键技术

③热点评估与跟踪

热点评估

- 词频统计：TF-IDF与领域词典结合
- 情感分类：基于概率论 基于信息论

对热点舆情进行等级评估与阈值设定

热点跟踪

- K最近邻算法 分类准确性高，速度慢
- 朴素贝叶斯算法 分类效率相对稳定，误差率易受干扰

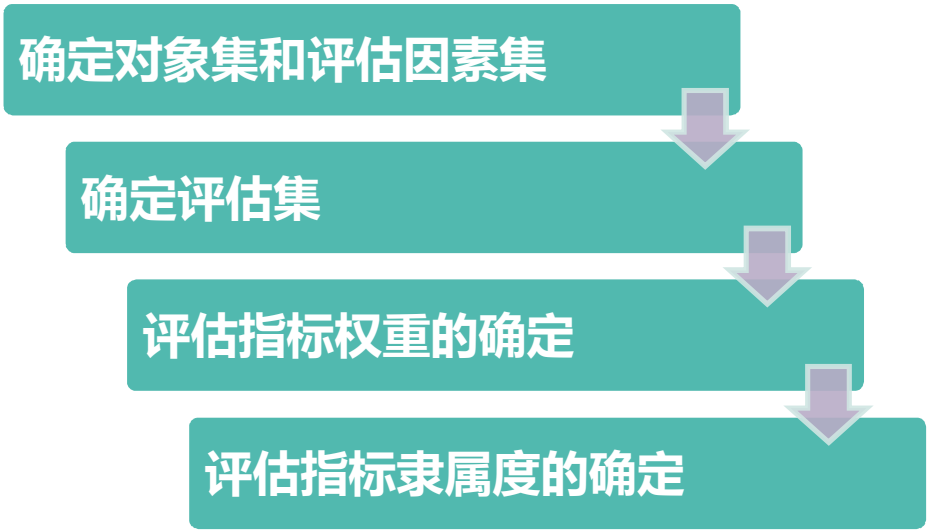
通过对热点的快速分类实现跟踪

网络舆情分析关键技术

④舆情等级评估

- 网络舆情的等级评估通常采用综合评判方法，即对受到多种因素制约的事物或现象做出一个总体评判。

- 我国对网络舆情的等级评估采用多级模糊综合评判模型，模型的确定主要涉及算子的选择。



网络舆情安全评估模型一般构建步骤

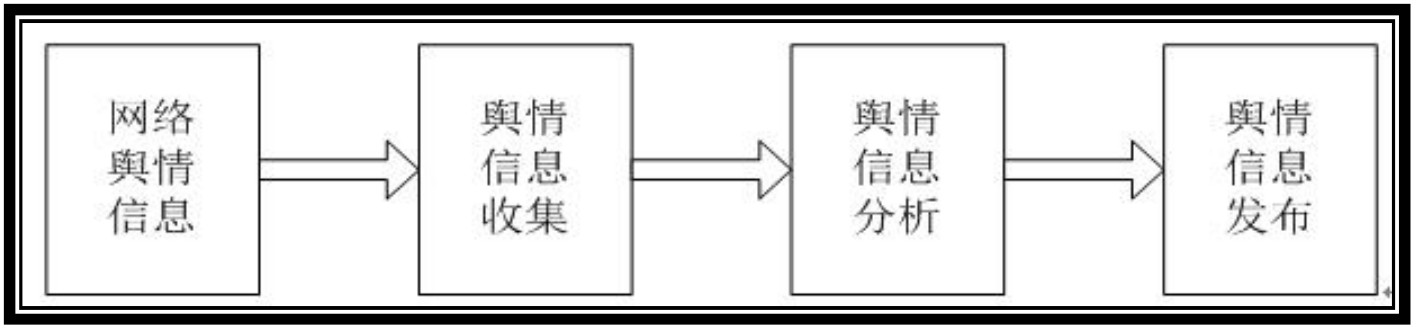
网络舆情分析系统框架

总体设计原则：流程化、标准化、模式化

系统关键：信息采集和舆情分析

信息源的选择：人工设定和机器学习

典型舆情系统业务流程：

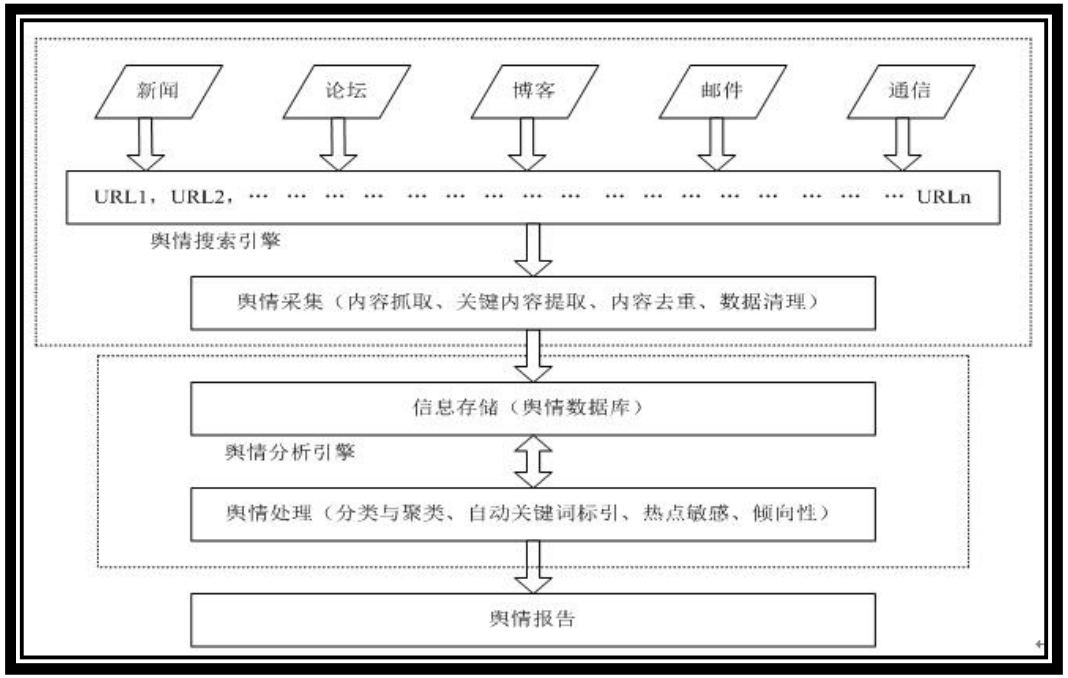


网络舆情分析系统框架

① 系统总体架构

■ 信息采集：从数据源进行网页抓取，经正文提取、内容去重等操作，然后将数据表示为便于处理的形式。

■ 舆情分析：利用分类、聚类算法对信息进行分析处理，形成舆情简报传递给前台。



网络舆情分析系统框架

② 关键技术分析

包括信息源的选择和信息的采集在传统搜索引擎基础上进行拓展搜索的广度和深度影响系统性能

功能：

- 信息的概念化
- 焦点的发现
- 事件的追踪

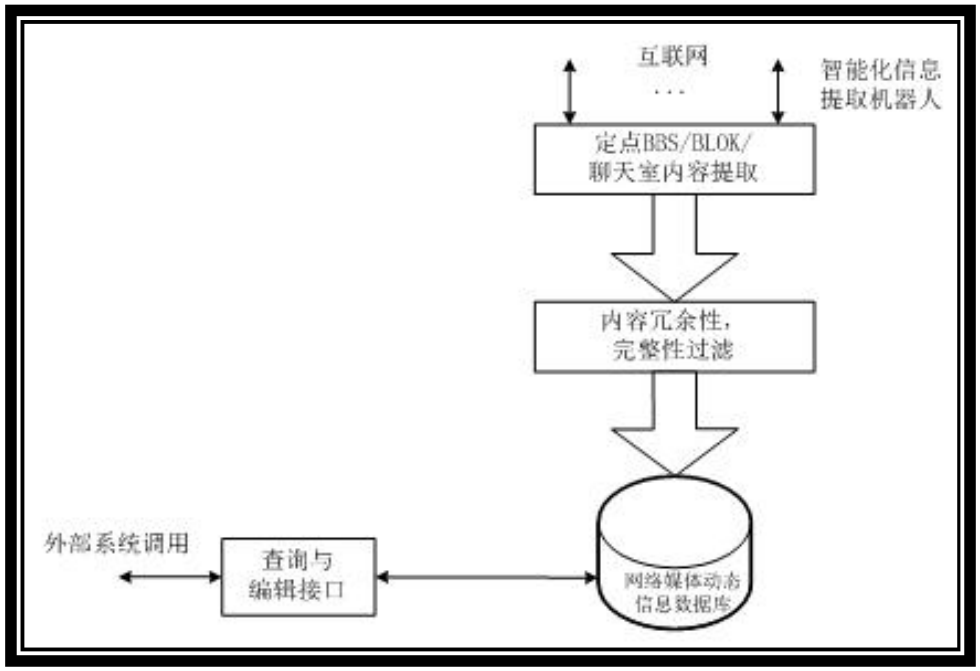
主要技术：

- 文本聚类
- 文本分类
- 文本倾向性分析

网络舆情分析常用方法

①高仿真网络信息深度抽取

- 重点研究原创网络互动式动态信息提取
- 形成高性能动态信息提取系统，组成舆情监控系统的信息获取模块



网络舆情分析常用方法

②信息自动提取机器人技术

1、个性化可配置的信息自动提取技术

2、互动式信息的智能提取技术

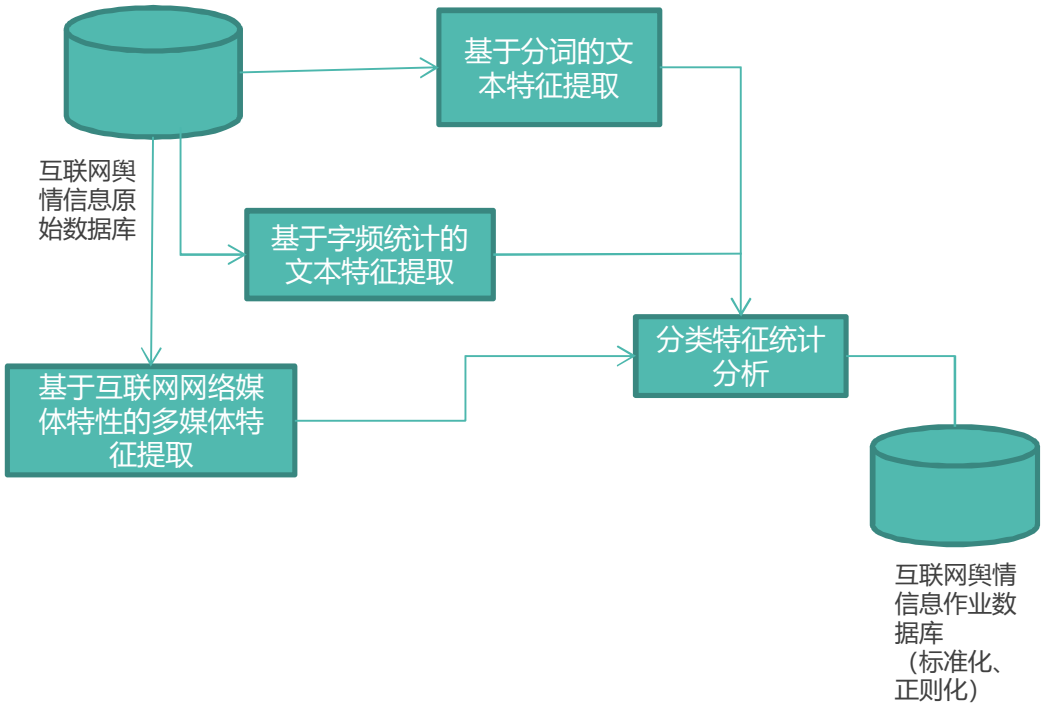
3、网页编写语言的实时语义理解技术

4、多线程内容提取技术

网络舆情分析常用方法

③基于语义的海量文本特征快速提取与分类

- 重点研究网络文本媒体的语义特征提取
- 形成基于语义的文本特征快速提取与分类系统，组成舆情监控系统的信息分析模块



网络舆情分析常用方法

④多媒体群件理解技术

综合字词、标点和模式匹配的文本核心信息快速提取

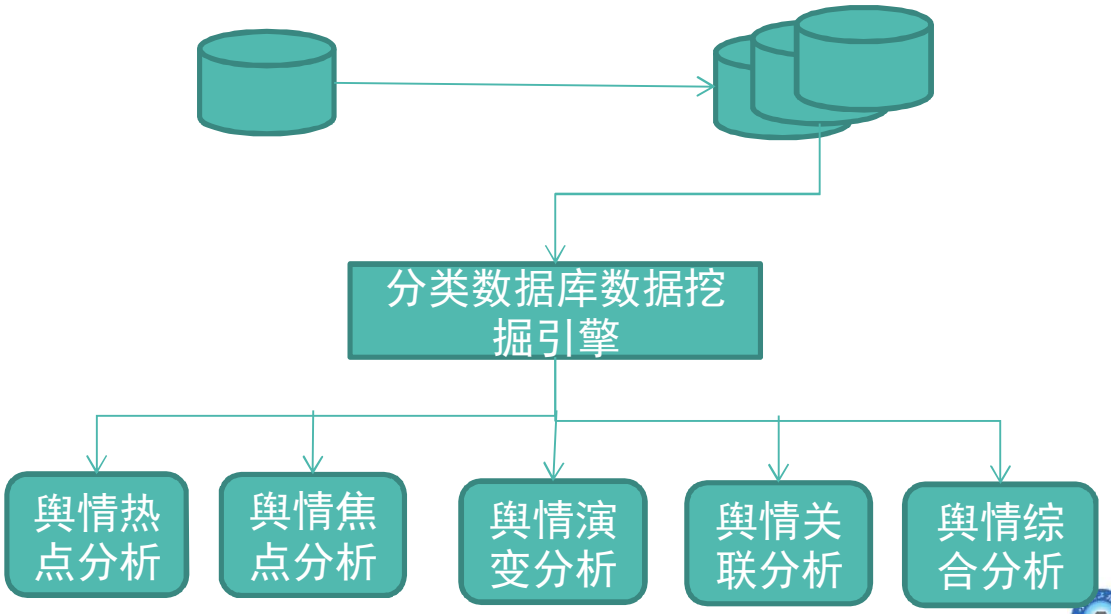
图像核心信息快速提取技术

综合环境信息和相关媒体信息的多媒体群件理解技术

网络舆情分析常用方法

⑤非结构信息自组织聚合表达

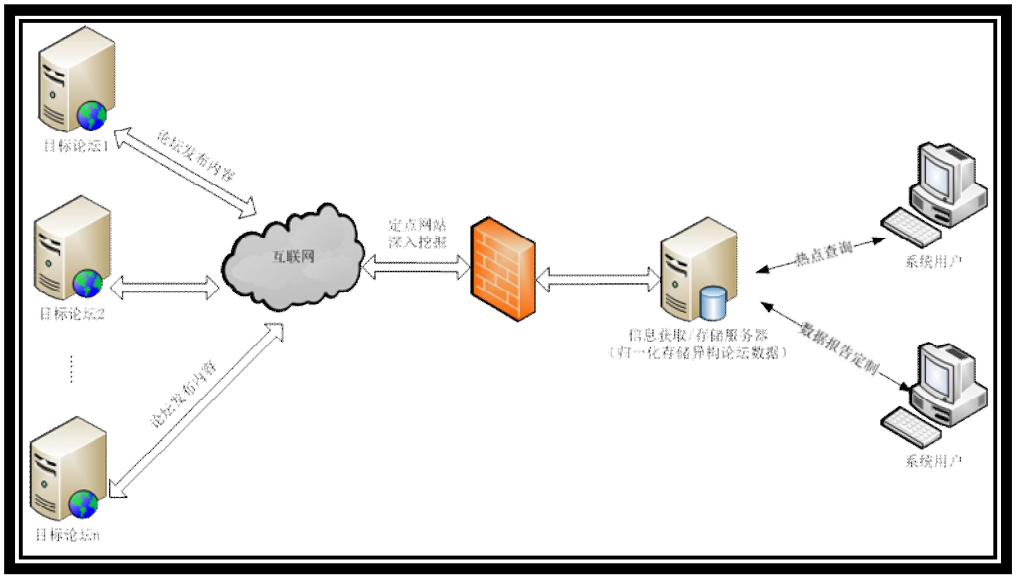
- 数据分析模块
- 数据仓储模块
- 数据类型挖掘引擎模块



网络舆情分析的典型应用

互联网论坛信息监控案例

利用网络协商与人机对话模拟等技术实现目标站点的信息获取，进行归一化存储，最终呈现系统目标站点的讨论分析情况和具体内容。



网络舆情分析的典型应用

定点站点
深入挖掘
机制

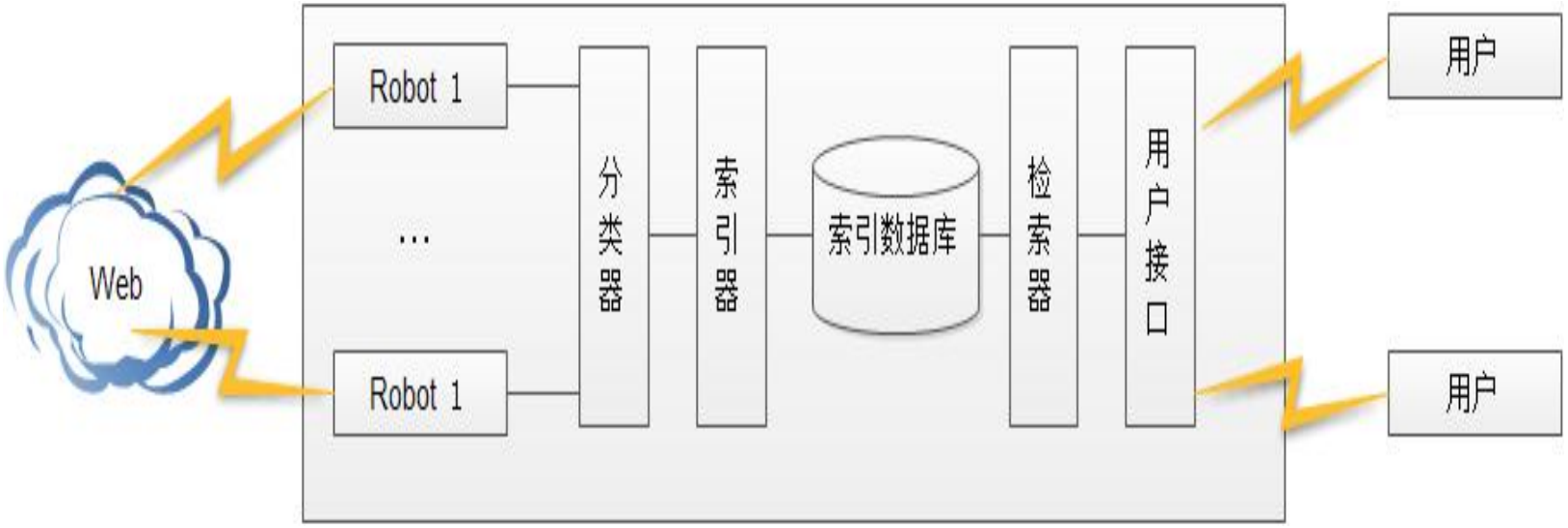
异构数据
归一化存
储与目标
站点热点
查询

监控目标
热点自动
发现功能

网络舆情分析的发展趋势

①针对信息源的深入信息采集

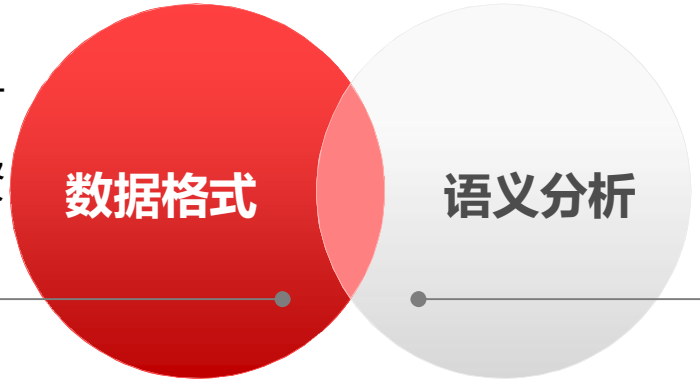
信息采集的深入性和全面性是重点解决问题。



网络舆情分析的发展趋势

②异构信息的融合分析

采取通用的具有高度扩展性的数据格式进行资源整合



采取基于语义等应用层上层信息的抽象融合分析

网络舆情分析的发展趋势

③非结构信息的结构化表达

通常，非机构信息的结构化表达被归结为文本信息提取问题，信息提取技术可分为五个层次。

1	专有名词	如：人名、地名、机构名、日期、时间、邮件地址等
2	模板要素	应用模板的方法搜索和识别别名词性条目的相关信息
3	模板关系	应用模板的方法搜索和专有名词之间的关系
4	同指关系	解决文本中的代词指称问题
5	脚本模板	根据应用目标定义任务框架