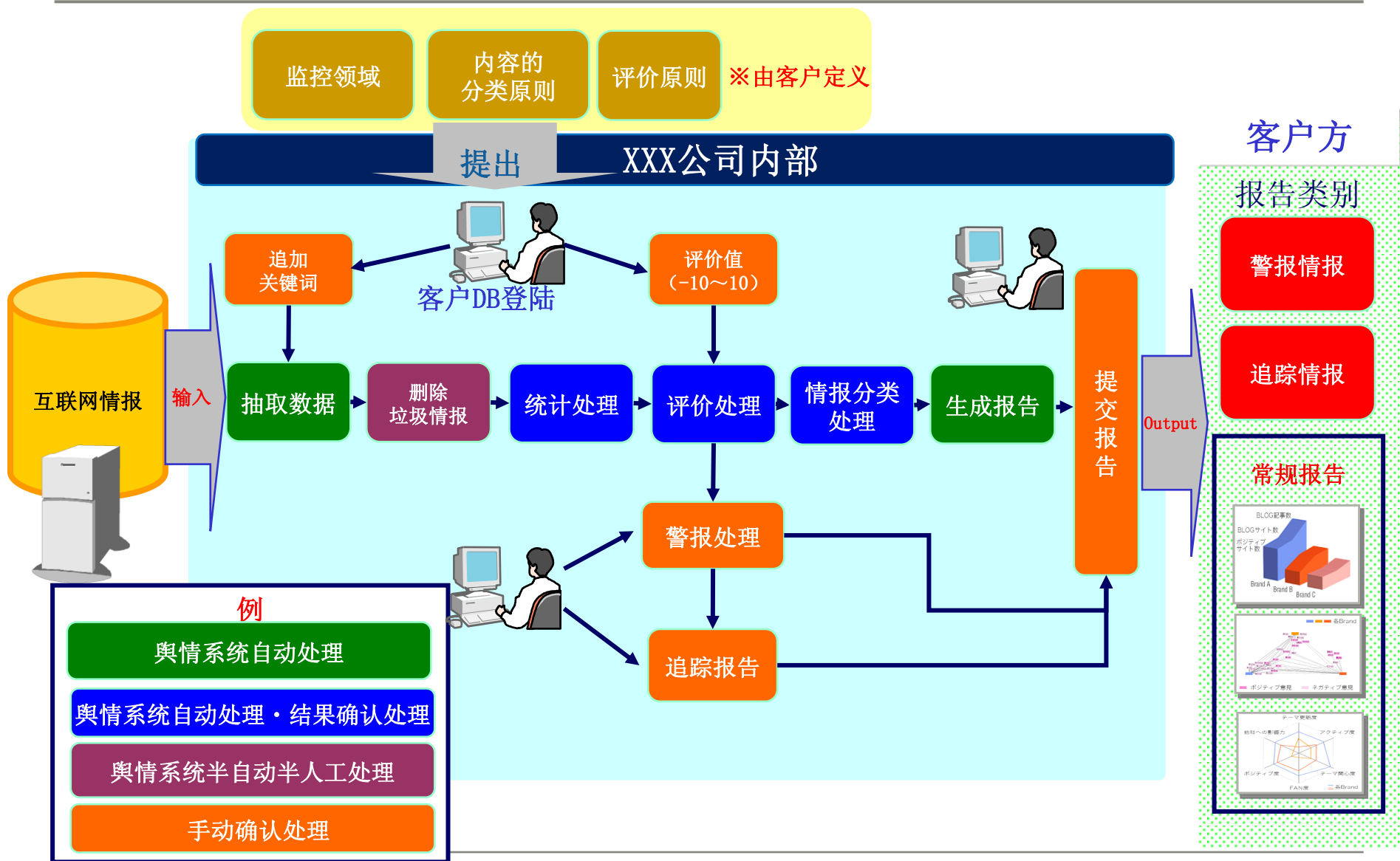


北京图灵开物技术有限公司



輿情系统机能概要



輿情系统概要说明

【机能・处理概要】

	机能・处理名称	实施者	内容
自动处理	抽取数据	系统	根据设定的Keyword抽取必要数据
	删除垃圾数据		从收集到的信息中删除垃圾信息
	统计处理		删除重复数据
	评价		对数据内容进行 -10~10的评分
	分类		根据设定标准对内容进行分类（按部门进行分类）
	生成报告		数据可视化，生成报告
	警报处理		遇到比设定值低的负面评价进行警报处理
手动处理	数据确认	操作者	对抽取出的数据进行确认
	删除垃圾数据		对自动无法删除的垃圾信息进行删除
	各行程确认		对各行程结果进行确认
	发送报表		生成可视化报告
	追踪服务		根据警报对紧急事件进行报告服务

【客户自定义内容】

	项目	内容
客户 自定义情报	关键词情报	根据所要收集信息的内容设定Keyword
	分类	设定分类标准
	．．．．	

輿情系统的性能

■ 硬件和人员配置

- 4台独立IP机器, 2名校对人员+1名后备

■ 采集范围

- 收集到的网页已覆盖18万网站(*3月19日~30日两周统计结果)
- 单日采集50000相关网页(平均)

■ 处理时间

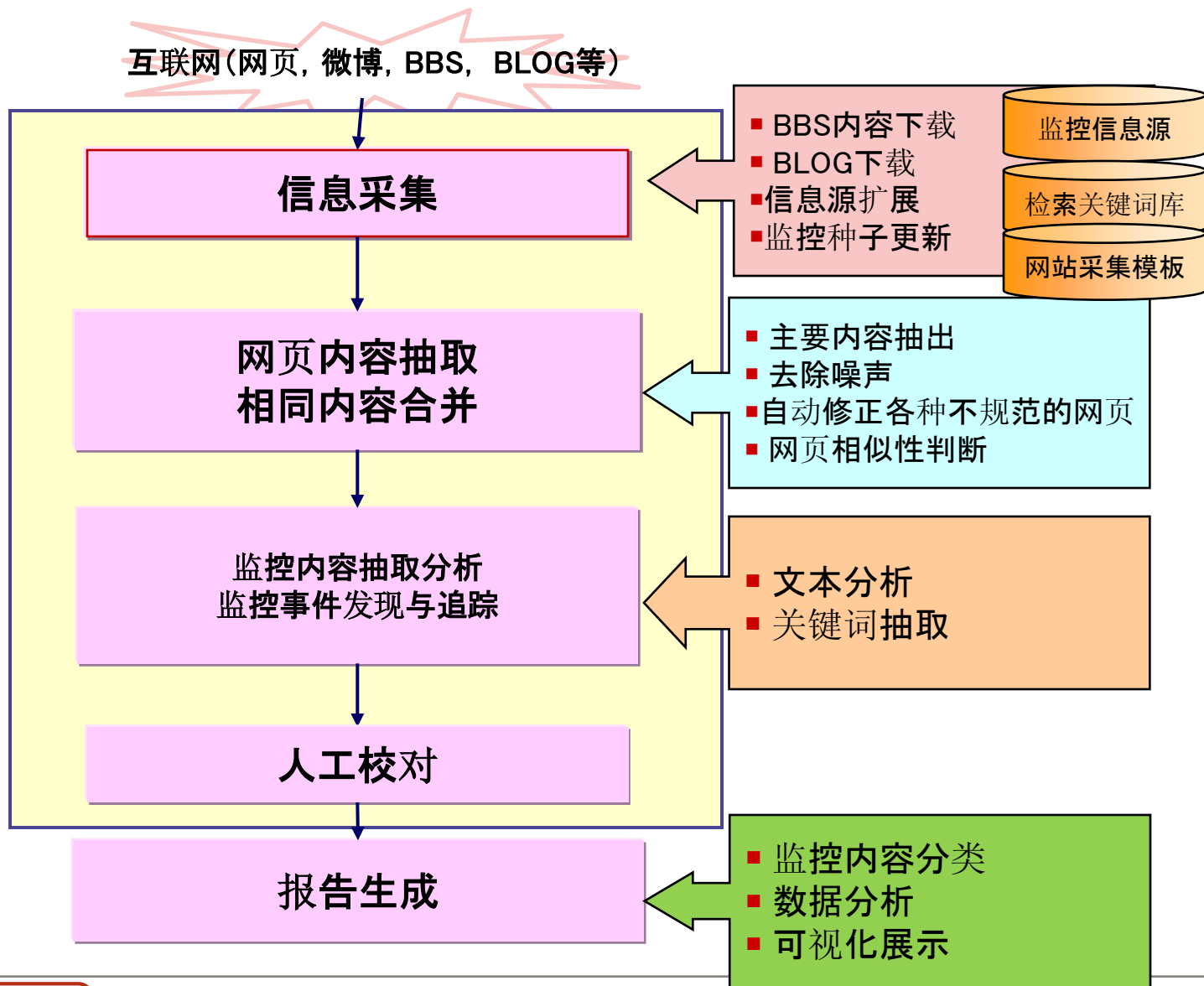
- 24内生成监控报告

■ 目前监控上限

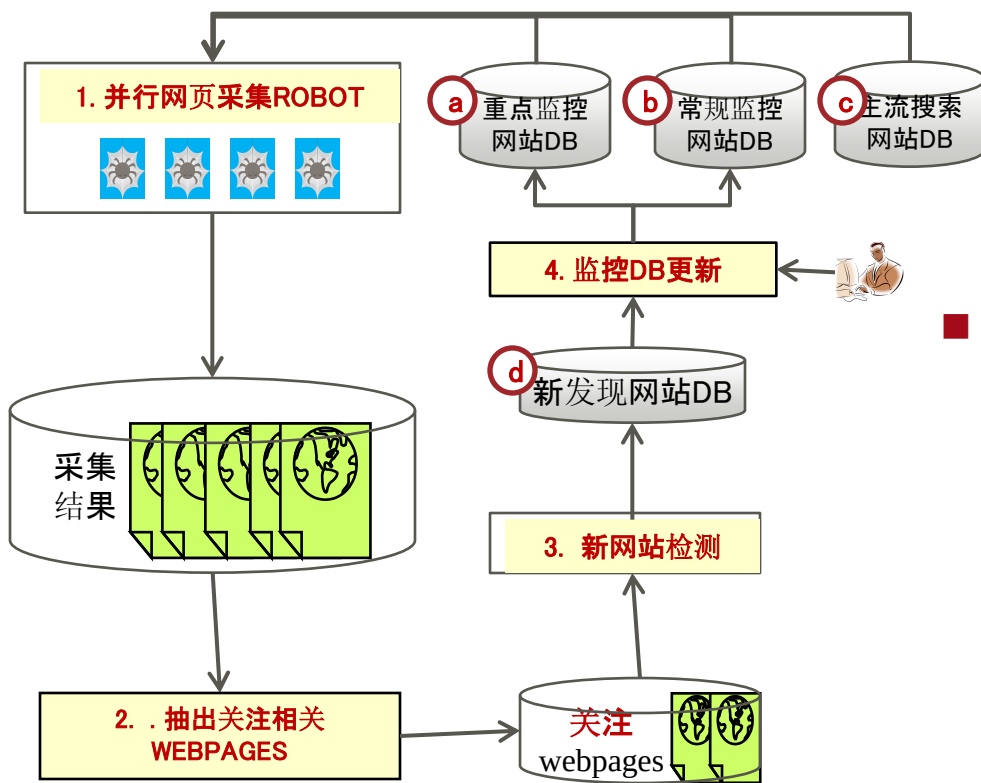
- 每日监控网页<300,000 网页
- 信息激增时, 可通过增加机器, 提升每日监控网页
- 参考: 2009年~至今, 微软中文搜索引擎bing共收集到约380万包含“该汽车企业”关键字的网页, 由此估计单日有关“该企业”的报道可能超过30万

监控流程图

■ **网络信息监控**: 即从互联网中获取用户感兴趣的内容, 并对内容进行分析、追踪



互联网信息自动采集过程



■ 监控信息源

- 重要的新闻门户和信息最可能出现的网站，每3小时监控一遍
- 常规相关网站，24小时内采集一遍
- 搜索门户及BBS、BLOG、微博搜索引擎，24小时内采集一遍
- 监控信息源扩充

■ 采集过程

- 并行采集ROBOT，依次采集a,b,c三个DB中指定的网站内容，生成采集结果。其中，ROBOT可多个并行采集
- 分析网页，抽出与监控内容相关的网页
- 检测是否有新的信息源，包含用户监控相关信息，生成新发现信息源
- 监控信息源更新，人工确认加入新网站及删除无效的监控网站

图灵研发中心信息采集特点

■ 多级监控, 保证监控的时效性

- 没有一个系统能够监控整个互联网, 只能有重点地监控
- 重点保证: 热门网站和重要领域网站
- 覆盖度保证: 通过各大搜索引擎提高监控覆盖率

■ 高质量的监控信息源

- 全部覆盖Alexa Top500 *中文相关网站和重要领域网站

■ 信息源动态更新, 及时发现新的热点

- 对采集结果进行分析, 发现新信息源
- 自动发现与人工检测相结合

■ 高质量的网站内容采集和内容抽出

- 100多种网站采集模板等, 应对不同网站格式, 采集率100%
- 针对不同网站的实时动态采集策略, 根据网站更新速度, 调整采集频率

*Top500 网站排名信息由 ALEXA公司提供

Alexa是一家专门发布网站世界排名的公司, 其拥有的URL数量超过GOOGLE。

检索引擎与舆情系统的比较

■ 共同点

- 监控对象:互联网上的公开信息

■ 舆情系统续悬鱼的优势与价值

- 检索引擎方法:返回大量垃圾和重复信息,信息杂乱,更新不及时;
- 舆情分析:相同信息合并,去除垃圾信息,信息及时更新;

■ 主要搜索引擎与该汽车企业的舆情监控范围比较

- 实验说明:在8个中文检索引擎中检索“**某企业***”关键词,选取网页日期是2012年5月2日的检索结果,作为检索引擎的监控结果

	监控 网页数	企业相关 网页数	覆盖率对比
主要搜索引擎	2817	213	覆盖40.5%トヨタCGM监控结果
汽车企业舆情系统	---	526	覆盖100%搜索引擎监控结果

■ 结论

- 企业舆情监视范围更广
- 企业舆情监视效率更高

*:企业名称不能公开

人工调查与舆情分析系统的比较

■ 实验说明

- 时间:2012年5月2 日, 人员:1 名, 人工调查时间: 8小时
- 人工监控方法: 互联网上自由浏览, 收集与企业相关网页
- 手动调查:(单日累积)113件

该企业舆情系统:(2012年5月2日单日累积) 570件

人工监控网页数	舆情监控网页数	系统中覆盖的人工监控结果
113	570	97.1%

運營事例

- 顧客A公司使用檢索引擎(BAIDU, GOOGLE等)、手工检索与影响该企业形象的负面网页。图灵采用为企业构建的輿情分析系统自动收集该企业负面网页。对比结果如下：

手動調查：(一个月累積) 86件

輿情系統：(一个月累積) 1242件

与国内主力厂商的性能比较

主力厂商	针对客户	信息采集源	每日最大采集量	数据安全	评价分析
中科新天	政府 企业 事业单位	用户指定+搜索引擎	最大12个ROBOT并行采集	承诺, 无具体方案	有
本果舆情	政府 公安 企业	用户指定+搜索引擎	4万	未知	未知
南京擎天	政府 公安	最大2500, 不可扩展	10万	未知	未知
TRS	政府 大众消费	用户指定+搜索引擎	5万	承诺, 无具体方案	有
方正智思	政府 传媒 教育	最大1000	未知	未知	未知
图灵系统	企业 政府	领域网站+自动发现+ 用户指定+搜索引擎 无上限(与服务器数量相关)	5万(4个ROBOT) 可扩展	有方案	强

图灵开物舆情系统独有特点:

1. 中文评价分析, 2008、2009年两年第一
2. 多集监控, 保证时效, 信息源动态更新
3. 汉语分析技术(分词、自动分类)
4. 中日双语系统

輿情系统能做什么，不能做什么

■ 能做

- 收集互联网中的公开的中文文本信息
- 提供监控报告
- 保存监控报告中涉及网页的URL
- 监控报告和URL目前保存1年

■ 不能做

- 互联网信息量庞大(中文有183万网站), 不能100%覆盖
- 互联网中非公开信息, 例如facebook中对个人或朋友开放的信息, 以及论坛中需要特殊权限的信息
- 即时通讯工具信息采集: QQ群, MSN
- 方言, 声音, 图像, 视频信息、手机信息
- 判断信息真伪(但可以甄别可能存在的水军发言人)
- 信息处置(删除信息等)

谢

谢

