

1. 大数据采集、挖掘技术

利用爬虫技术对全网信息进行实时收集、挖掘和智能检索，保证了信息的及时性和完整性。

2. 大数据处理技术

- 快速数据去重：消除重复网页、转载文章。
- 网页去噪：去除网页中的噪音连接（互相引用），噪音内容（与应用目的不相符合的内容）。

3. 舆情分类

将采集到的网页信息归类，主要用到自然语言处理中的文本分类技术和文本聚类技术。

4. 舆情挖掘

- 话题检索与跟踪：通过聚类分析，把舆情归入不同的话题类别，通过算法发现感兴趣的话题，跟踪话题发展过程。主要任务是检测新话题，必要时建立新的话题类别。
- 文本情感倾向分析：对体现文本主题偏好、观点、情感的主观性文本进行检测、分析和挖掘。