1. **系统需求概述**

**为了实现高效的网络信息采集和分类管理，系统达到如下要求：**

1. **独立运营：**

系统包含整套包括信息采集、信息分类、信息处理、统计报表、后台管理功能；

可以独立完成整个网络信息采集工作，包含添加任意添加关键词、添加账号和用户、添加采集目标网站等，数量无限制。

1. **全面覆盖：**

7\*24小时不间断监测；  
采集覆盖面广，可以独立添加任意网站做信息采集；

能够对微博信息采集（无需调用微博网站提供的API）、手机新闻客户端信息采集、HTTPS类网站信息采集；

境外媒体采集：可对英文类新闻，财经、航运等境外信息媒体进行采集整理。

1. **重点跟踪：**可自定义对某些重大项目、事件及课题进行跟踪采集和分析。
2. **速度快：**

搜集和识别信息速度快，响应时间要求在分钟级；

自由控制每个网站的调度频率，监测速度可以达到分钟级。

1. **信息要素：**需要获取信息标题及摘要、URL， URL解析IP、域名地域、发帖人、发布时间、网页快照、跟帖内容等。
2. **界面友好：**

各种相关设置需在GUI界面中较轻易完成；

支持移动设备的访问；

能为其它办公系统提供各类数据的RSS接口。

1. **后台监控：**能够对后台机器的运行状态、采集情况、数据库等做监控，可以在不用进机房解决所有非硬件故障。

**二、详细技术需求和功能描述**

**系统分为3大块：采集部分，分析部分，用户界面部分。 各部分详细功能如下：**

**其中★为表示为必须满足**

**（一）采集部分详细功能需求**

**内置采集范围**

重点监测目标网站（列表）3000个以上

**★**网站域名300，000个以上

**★**自己加入任意网站（行业性网站、新网站、地方性网站）

**能采集监测的目标类型**

**★**微博网站、新闻网站、资讯网站、论坛、博客、

搜索引擎、查询型网站

图像网站、视频网站

商品目录网站、成员目录网站、在线数据库网站等

**支持各种协议网站**

需要支持HTTP网站、HTTPS通信加密网站、FTP文件下载网站、SOAP API网站、REST API 网站、3G网站

**其他非常规采集**

**★**支持各种Windows程序，比如QQ群，微信群；支持采集格式未知文件中数据、PDF文件中数据、Office文件中数据、图片中的数据

**需支持各种网站登录**

支持多种网站类型、支持多种网站特点、支持多种网页流程、网页编码转换、多种数据处理方式、多种网页配合动作

**能够智能识别**

**★**智能内容识别、格式识别、关键词识别、时间判别

智能下载、智能过滤、智能遍历

**用户接口**

具备图形化向导接口

具备用于快速编辑的简单脚本接口

**支持防采集网站**

**★**可以解决IP访问频率有限制的问题，可以提供防盗链措施，数据加密措施，数据乱码措施，数据隐藏措施，数据图片化措施

**支持二次开发**

可以利用PHP，C#调用采集组件，进行二次开发

**（二）分析部分详细功能需求**

**自动分析功能**

热点：自动显示全网实时热点

**★**自动分类：自动根据设置的关键词进行分类

**★**自动过滤：利用舆情分类信息，自动或根据用户自定义把不关注的热点信息过滤掉，提高信息的准确性

自动聚类：自动识别出相似信息

自动摘要：自动对正文摘要，以方便列表查看

自动人名识别：自动人名识别，以方便自动分类

自动地名识别：自动地名识别，以方便自动分类

负面预判：自动对疑似负面信息做出标记，以便用户优先处理

**内容分析**

主题分析、选择分析、核心词分析、热词分析、专题跟踪分析、网站传播分析、微博传播路径分析

**网站分析**

网站构成分析、网站采集分析、网站单独分析

**操作分析**

工作量分析

**关系数据库要求**

数据库类型：MySQL

**★**支持记录数量：亿级

**★**千万级表查询当天采集单条记录时间：<0.1秒

亿级表查询当天采集单条记录时间：<0.5秒

历史单条记录时间：<1秒

**文件本地保存要求**

支持保存网页HTML快照

支持保存网页图片快照

支持保存图片文件

支持保存视频文件

**三、其他**

**质保期：不少于3年；**