# 分布式互联网爬虫及其在垂直领域的应用项目需求分析说明书

# 需求分析说明书

目录

[1.引言 3](#_Toc492476083)

[1.1目的 3](#_Toc492476084)

[1.2背景 4](#_Toc492476085)

[1.3项目目标和要求 7](#_Toc492476086)

[1.4预期读者及阅读建议 12](#_Toc492476087)

[2项目概述 12](#_Toc492476088)

[2.1系统目标 12](#_Toc492476089)

[2.2开发环境 12](#_Toc492476090)

[**3需求** 13](#_Toc492476091)

[3.1一般性需求 13](#_Toc492476092)

[3.2功能需求 13](#_Toc492476093)

[3.3非功能性需求 15](#_Toc492476094)

# 1.引言

## 1.1目的

编写此文档的目的是进一步细化视频网站信息数据系统开发的细节问题，希望能使系统开发工作更具体。为了使用户、软件开发者及分析和测试人员对该系统的集体需求有一个共同的理解，它说明了本系统的各项功能需求、性能需求和数据需求，明确标识各项功能的具体含义，阐述实用背景及范围，提供客户解决问题或达到目标所需要的条件或权能，提供一个度量和遵循的基准。具体而言，编写软件需求说明的目的是为所开发的系统提出：

a) 软件设计总体要求，作为软件开发人员、软件测试人员相互了解的基础。

b) 功能、性能要求，数据结构和采集要求，重要的接口要求。

c) 软件确认测试的依据。

## 1.2背景

**一、行业大背景和潜力——大数据时代的来临**

“大数据”是指以多元形式，自许多来源搜集而来的庞大数据组，往往具有实时性。在企业对企业销售的情况下，这些数据可能得自社交网络、电子商务网站、顾客来访纪录，还有许多其他来源。这些数据，并非公司顾客关系管理数据库的常态数据组。

从技术上看，大数据与云计算的关系就像一枚硬币的正反面一样密不可分。大数据必然无法用单台的计算机进行处理，必须采用分布式计算架构。它的特色在于对海量数据的挖掘，但它必须依托云计算的分布式处理、分布式数据库、云存储和/或虚拟化技术。(在维克托·迈尔-舍恩伯格及肯尼斯·库克耶编写的《大数据时代》中大数据指不用随机分析法（抽样调查）这样的捷径，而采用所有数据的方法）大数据的4V特点：Volume（大量）、Velocity（高速）、Variety（多样）、Value（价值）。

大数据的意义是由人类日益普及的网络行为所伴生的，受到相关部门、企业采集的，蕴含数据生产者真实意图、喜好的，非传统结构和意义的数据 。2013年5月10日，阿里巴巴集团董事局主席马云在淘宝十周年晚会上，将卸任阿里集团CEO的职位，并在晚会上做卸任前的演讲，马云说，大家还没搞清PC时代的时候，移动互联网来了，还没搞清移动互联网的时候，大数据时代来了。

从海量数据中“提纯”出有用的信息，这对网络架构和数据处理能力而言也是巨大的挑战。在经历了几年的批判、质疑、讨论、炒作之后，大数据终于迎来了属于它的时代。2012年3月22日，奥巴马政府宣布投资2亿美元拉动大数据相关产业发展，将“大数据战略”上升为国家战略。奥巴马政府甚至将大数据定义为“未来的新石油”。

大数据时代已经来临，它将在众多领域掀起变革的巨浪。但我们要冷静的看到，大数据的核心在于为客户挖掘数据中蕴藏的价值，而不是软硬件的堆砌。因此，针对不同领域的大数据应用模式、商业模式研究将是大数据产业健康发展的关键。我们相信，在国家的统筹规划与支持下，通过各地方政府因地制宜制定大数据产业发展策略，通过国内外IT龙头企业以及众多创新企业的积极参与，大数据产业未来发展前景十分广阔。

大数据时代，互联网数据被认为规模最大、结果最复杂、潜在价值最高的源数据。虽然在软件领域出现了多款开源的爬虫工具，为用户带来了许多方便，但一般都不能较好的解决数据的两个关键问题：

1、采集质量。数据质量直接决定了数据能否最为可靠数据提供服务。对多样性页面数据进行准确抽取是目前非常有挑战的难点。

2、采集时效性。对于新发布的数据，如何在较短的时间内采集到库中是当前采集系统面临的又一个重要问题。采集频度高则需要有更多的硬件支持，同时要能处理好站点反扒措施。除此之外，采集互联网爬虫的存储、服务、排重等各技术环节都具有转战，行业内也积累了高效的解决方案。

**二、前景分析——爬虫与大数据**

爬虫即网络爬虫，是一种自动获取网页内容的程序。是搜索引擎的重要组成部分，因此搜索引擎优化很大程度上就是针对爬虫而做出的优化。

在大数据时代下，人类社会产生的数据每天都在爆炸性的增长，在依托数据进行一些分析时，传统的采集数据的方式，像问卷，访谈，人工的定向统计等方法因为采集的样本少，可信度低等缺点已经不能再满足需求。在这种情况下，爬虫技术正在大数据领域悄然崛起！相比之下，爬虫可以采集互联网上海量的信息提供给需要者进行下一步的分析，预测，改进等工作，由于爬虫可以实时的，全面的而又有针对的采集互联网上信息，是大数据的基础，因此爬虫技术是当下一门正在走红或者说已经走红的技术，前景广阔。互联网数据定向采集可做特定领域的问题发现、问题跟踪等。本项目将以视频分享方面的采集数据为基础，采集建库并建设通用的数据关键词服务，结合新媒体文化对垂直领域的关注事件，进行媒体传播排名以及媒体传播趋势。

综上所述，开发一款具有较高数据质量、对新数据探测周期短、可灵活扩展的分布式互联网爬虫具有技术理论的挑战，同时具有较高的应用价值。对于提升学生的实践能力、设计能力、独立解决问题的能力有极大的帮助。

## 1.3项目目标和要求

#### 1.3.1项目目标

视频分享方面的采集数据为基础，采集建库并建设通用的数据关键词服务，结合媒体文化对垂直领域的关注事件，进行媒体传播排名以及媒体传播趋势。

#### 1.3.2项目要求

(1)基本要求

软件操作简便快捷，数据采集部分可以通过简单易懂的方式配置新网站数据的采集。数据分析模块通过基关键词过滤以及统计学思想计算媒体传播排名及趋势。

界面设计风格采用Web形式，用户界面简单明了，符合业务办理流程，具有人性化特点。软件设计结构模块化、程序结构化、数据格式标准化、代码统一化，各种文档资料规范化。

（2）设计要求

数据流程：

数据采集模板

d

模板抽取

下载内容

去重

抽取列表

下载频道页

互斥获得种子

用户内容

Web服务器

结果库

排名计算

关键词过滤内容

数据采集

数据分析

传播趋势

数据库

模板库

模板服务器

模板管理员

模块功能：

1)数据采集模板配置工程

数据采集模板配置工程主要完成采集站点的频道模板配置，配置重要信息有：

`website name` varchar(64) DEFAULT '0' COMMENT '站点名称',

`channel\_name` varchar(256) DEFAULT '0' COMMENT '频道名称',

`channel\_url` varchar(256) DEFAULT '0' COMMENT '频道url',

`channel\_url\_nextpage` varchar(256) DEFAULT '0' COMMENT '频道下一页url',

`start\_time` timestamp NULL DEFAULT NULL COMMENT '开始时间',

`stop\_time` timestamp NULL DEFAULT NULL COMMENT '结束时间',

`status` varchar(64) DEFAULT '0' COMMENT '抽取正常、抽取异常',

`title` varchar(256) DEFAULT '0' COMMENT '文章标题的xpath',

`author` varchar(64) DEFAULT '0' COMMENT '文章作者的xpath',

`pubtime` varchar(32) DEFAULT '0' COMMENT '发布时间的xpath',

`content` longtext COMMENT '正文的xpath',

`source` varchar(50) DEFAULT '0' COMMENT '转发来源的xpath'

模板工程需要保证系统的可用性，该工程基于Tomcat结合Web工程实现，只要实现对页面的表格填写信息写入到模板库中。

配置模块在写库之前需要对配置进行测试验证，如果测试通过则写入到模板库，如果失败则提示，并要求用户重新配置模板。

模板配置内容包括两类：一、频道中视频列表的抽取，保证能把首页上的视频url抽出来，如果想更深一步可以考虑进行翻页抽取url。二、具体页面的内容抽取，如标题、作者、发布时间、播放量等。

该子工程是整个基于爬虫做数据分析应用系统的基础。同时因为网站模板存在改版问题，模板管理同学需要能及时发现失效模板并进行修复。

对失效模板的精确展示并提供方便的修改入口，能极大的提升模板管理的成本。模板失效需要配备报警机制，对模板失效时间要及时进行邮件等形式的报警。

2) 数据采集

获取互斥种子：该模块从模板库读取种子信息，互斥特性则是为了实现分布式时的并行模式。在模板库中的数据量变的很大，则采集工程需要进行分布式部署，同时能够避免重复采集同一模板数据。

下载频道页：根据模板中频道url直接下载列表，读取所有的视频url，配置上其它必须信息，打包成一个下载元数据，放到缓存队列中。

去重：去重的目的就是为了避免对同一url文章的重复下载，浪费资源，更有可能触发该网站的禁止访问策略。该模块从下载任务队列中取出下载的元数据，在开始真正下载之前先进行去重处理。

下载内容页：把url对应的页面现在下来。在碰到禁止访问的IP限制后，能够采用动态IP的方式解决访问频道的限制。一般情况下可以不采用动态IP的方式，因为对于同一个网站的采集频道同时会受到模板数据量、采集程序部署数量、采集种子延迟控制等几个因素印象，不会对网站造成较大的采集压力。

模板抽取：该模板是核心模板，是关系数据质量的关键环节。通过模板配置进行对应信息的抽取，比如标题、作者、发布时间、播放量。如果关键字段抽取失败则通过反馈模块更新模板库的状态字段，标志为：模板失效，模板管理人员可通过报警后web前端进行查看和修改。抽取成功的字段则写入到存储中。小文件存储类型可采用分布式文件系统有：glusterFS、fastDFS等，在数据生命周期不长的情况下，也可以采用传统DB作为存储服务。用传统DB作为采集结果数据存储的优点：利于管理、操作灵活，开发技术门槛低，并且具有很高的稳定性。

3）数据分析

关键词过滤：通过高效的关系过滤技术，将新采集的数据进行过滤打标签，对符合订阅条件的文档写入到服务库中。关键词过滤是具有较高算法设计需求的模块，需要具有较高的编程基础和算法设计能力。

播放量排名计算：近实时从符合条件的数据集中取出最新数据，并进行统计计算。

视频传播趋势：该模块主要是对关键词过滤命中的数据集，在时间维度上的跟踪统计，并以比较明显的方式展示给用户。

## 1.4预期读者及阅读建议

开发小组人员

# 2项目概述

## 2.1系统目标

爬虫通过对国内目前主流的视频分享网站如优酷，哔哩哔哩，Acfun等进行采集并将采集到的信息存储到数据库中

通过采集到的信息进行以播放量为关键字的分析，结合社会坏境，对视频进行排名。

### 2.2开发环境

**模板配置与数据分析应用**：

开发工具：Eclipse、PyCharm

数据库：MySQL

开发技术：Java、Python、sql、hibernate、sql、maven、git

中间件：Tomcat 、jdk1.7

**采集工具**：

开发工具：Eclipse、PyCharm

数据库：MySQL

开发技术：Python、sql、多线程、队列

**环境要求**：

CPU：1GHz 以上

内存：4GB以上

## 3需求

### 3.1一般性需求

| 序号 | 描述 | 优先级 |
| --- | --- | --- |
| GR-1 | 数据存储模板配置以及模板失效处理 | Must |
| GR-2 | 信息采集，信息存储 | Must |
| GR-3 | 根据数据分析并得出结论并展示给用户 | Must |

### 3.2功能需求

| 序号 | 描述 | 优先级 |
| --- | --- | --- |
| FR-1 | 获取互斥种子：该模块从模板库读取种子信息，互斥特性则是为了实现分布式时的并行模式。在模板库中的数据量变的很大，则采集工程需要进行分布式部署，同时能够避免重复采集同一模板数据。 | Must |
| FR-2 | 下载频道页：根据模板中频道url直接下载列表，读取所有的视频url，配置上其它必须信息，打包成一个下载元数据，放到缓存队列中。 | Must |
| FR-3 | 去重：去重的目的就是为了避免对同一url文章的重复下载，浪费资源，更有可能触发该网站的禁止访问策略。该模块从下载任务队列中取出下载的元数据，在开始真正下载之前先进行去重处理。 | Must |
| FR-4 | 下载内容页：把url对应的页面现在下来。在碰到禁止访问的IP限制后，能够采用动态IP的方式解决访问频道的限制。一般情况下可以不采用动态IP的方式，因为对于同一个网站的采集频道同时会受到模板数据量、采集程序部署数量、采集种子延迟控制等几个因素印象，不会对网站造成较大的采集压力。 | Must |
| FR-5 | 模板抽取：该模板是核心模板，是关系数据质量的关键环节。通过模板配置进行对应信息的抽取，比如标题、作者、发布时间、播放量。如果关键字段抽取失败则通过反馈模块更新模板库的状态字段，标志为：模板失效，模板管理人员可通过报警后web前端进行查看和修改。抽取成功的字段则写入到存储中。小文件存储类型可采用分布式文件系统有：glusterFS、fastDFS等，在数据生命周期不长的情况下，也可以采用传统DB作为存储服务。用传统DB作为采集结果数据存储的优点：利于管理、操作灵活，开发技术门槛低，并且具有很高的稳定性。 | Must |
| FR-6 | 模板配置，详见1.3模块功能.1 | Optional |
| FR-7 | 分析以及结果展示详见1.3模块功能.3 | Optional |

### 3.3非功能性需求

| 序号 | 描述 | | | 优先级 |
| --- | --- | --- | --- | --- |
| NFR-1 | 性能需求：优化系统，提高数据爬取速率，实现多线程。 | | Must | |
| NFR-2 | 容错需求：失效模板处理，中断处理，数据实时安全存储。 | Must | | |
| NFR-3 | 易读性需求：展示界面清晰易懂 | | | Optional |

5.进度安排