编码：项目级-算法设计说明书-Version\_1.0.0-signed-20200828

文档等级：□发布级 ■项目级 □团队级

采纳标准：团队文档质量标准

网络文学大数据分析系统

**Big Data Analysis System of Network Literature**

算法设计说明书

**Version 1.0.0-signed**

编写人：焦晨阳、全芷玥、阳玉洁

西工大-华迪实训第四小组

**All Rights Reserved**

**目 录**

**[1引言 3](#_Toc22485)**

[1.1 标识 3](#_Toc5098)

[1.2 系统概述 3](#_Toc1775)

[1.3 文档概述 4](#_Toc19138)

[1.4 基线 4](#_Toc12439)

**[2 引用文件 4](#_Toc4963)**

**[3 各需求功能点算法流程描述 5](#_Toc30093)**

[3.1题材收录占比 5](#_Toc6418)

[3.2月份排行 5](#_Toc21351)

[3.3季度排行 6](#_Toc27883)

[3.4年度排行 6](#_Toc24899)

[3.5用户阅读稳定性分析 6](#_Toc12862)

[3.6积分排行 7](#_Toc9012)

[3.7各题材top10 7](#_Toc6860)

[3.8长篇排行 7](#_Toc12811)

[3.9作者积分排行 7](#_Toc6634)

[3.10新晋作者排行 8](#_Toc8804)

[3.11稳定更新排行 8](#_Toc11744)

[3.12 本月完结收益排行 8](#_Toc11554)

[3.13 打赏收益排行 8](#_Toc22767)

[3.14 读者消费排行 9](#_Toc27836)

[3.15 读者订阅偏好 9](#_Toc28651)

[3.16 同类网站排行 9](#_Toc24076)

# 1引言

## 标识

适用系统：网络文学大数据分析系统

中文名称：算法设计说明书(ADS)

英文名称：Algorithm Design Specification (ADS)

文档版本：1.0.0

文档编号：项目级-算法设计说明书-Version\_1.0.0-signed-20200828

## 系统概述

本文档适用于“网络文学大数据分析系统”项目的开发过程。网络文学大数据分析系统项目由西北工业大学软件学院提出，由西工大-华迪实训第四小组成员负责实施，该项目编号为项目级-算法设计说明-Version\_1.0.0-signed-20200828，其软件产品版本号为“1.0”。

该项目基于B/S（Brower/Server）浏览器/服务器结构，采用传统的三层结构方式进行解耦开发。对当下最热门的网络文学阅读机构——“晋江文学”提供的各项数据进行可视化展示和分析。数据分析采用Hadoop和Spark。

## 文档概述

本文档依据《GB/T 8567-2006 计算机软件文档编制规范》制定，属于技术文档，仅限于项目相关人员阅读。

本文档的主要内容共分3部分:引言、引用文件、各需求功能点算法流程描述。对16个需求功能点的算法实现方法和思路进行了详细设计和描述。

本文档的保密范围如下：

①开发人员

②测试阶段人员

③对本文档进行评审的人员或机构

④项目组以及其他有权需要调用本文档的人员

## 基线

基线：GB/T 8567-2006

# 2 引用文件

《GB/T 8567-2006 计算机软件文档编制规范》，国家标准

《项目级-文档编写格式规范-Version\_1.0.0-release-20200812》，工作室编写

《项目级-软件开发计划-Version\_1.0.0-release-20200813》，工作室编写

《项目级-软件配置管理计划-Version\_1.0.0-release-20200813》，工作室编写

《项目级-软件需求规格说明书-Version\_2.0.0-release-20200813》，工作室编写

《项目级-类图设计-Version\_1.0.0-release-20200815》，工作室编写

《项目级-系统架构设计-Version\_1.0.0-release-20200815》，工作室编写

《项目级-数据库设计说明-Version\_1.0.0-release-20200815》，工作室编写

《项目级-界面原型设计-Version\_1.0.1-release-20200815》，工作室编写

《项目级-配置状态报告-Version\_1.0.0-release-20200815》，工作室编写

《需求分析与系统设计》（机械工业出版社）作者：麦沙塞克；翻译：金芝

# 3 各需求功能点算法流程描述

## 3.1题材收录占比

1. 爬取“晋江文学”网站作品库的所有书籍的属性（书籍id + 书籍名称 + 书籍题材 + 书籍积分 + 书籍起始更新时间 + 书籍最后更新时间 + 状态（完结/连载） + 书籍打赏数额 + 一个月内更新天数 + 作者id）；
2. 将数据存入HDFS；
3. 用Spark读取HDFS；
4. 数据处理，提取各个书籍的书籍id及对应书籍题材（书籍有爱情、武侠、奇幻、仙侠、游戏、传奇、科幻、童话、惊悚、悬疑、剧情、轻小说、古典衍生、东方衍生、西方衍生、其他衍生、儿歌、散文、寓言、童谣、历史故事20种题材类型）；
5. 按照题材进行分类计数；
6. 统计各分类书籍数量；
7. 计算每个分类所占比例：A分类占比 = ；
8. 用饼状图展示题材收录占比情况；

## 3.2月份排行

1. 使用前面爬取和存入到Spark中的数据，进行数据处理。提取各个书籍id、书籍积分及对应的书籍起始更新时间，筛选发文时间在第11-40天内的书籍；
2. 将筛选后的结果按照书籍积分由大到小进行排名；
3. 用柱状图展示当月积分排名前十位的书籍作品；

## 3.3季度排行

1. 使用前面爬取和存入到Spark中的数据，进行数据处理。提取各个书籍id、书籍积分及对应的书籍起始更新时间，筛选发文时间在第41-130天内的书籍；
2. 将筛选后的结果按照书籍积分由大到小进行排名；
3. 用柱状图展示当下季度积分排名前十位的书籍作品；

## 3.4年度排行

1. 使用前面爬取和存入到Spark中的数据，进行数据处理。提取各个书籍id、书籍积分及对应的书籍起始更新时间，筛选发文时间在第131-310天内的书籍；
2. 将筛选后的结果按照书籍积分由大到小进行排名；
3. 用柱状图展示当年积分排名前十位的书籍作品；

## 3.5用户阅读稳定性分析

1. 使用3.2、3.3、3.4处理的数据（按月、季度、年书籍积分排行），选取月份（季度、年份）书籍积分排名前200名；
2. 分别将月份（季度、年份）书籍积分排名前200名按照对应的书籍类型进行分类，在某个类型中计算加权平均值；

例：在月份排名中有“爱情”类书籍78本，分别对应积分为，则“爱情”类书籍在该月份的受欢迎程度为；

1. 以横坐标为月、季度、年，纵坐标为对应各类型受欢迎程度，用20条折线（20种书籍类型）展示各类型书籍月、季度、年用户阅读稳定趋势变化情况；

## 3.6积分排行

1. 使用前面爬取和存入到Spark中的数据，进行数据处理。提取各个书籍id和书籍积分；
2. 将筛选后的结果按照书籍积分由大到小进行排名；
3. 用词云显示该图书积分、题材等信息。

## **3.7各题材top10**

1. 使用前面爬取和存入到Spark中的数据，进行数据处理。提取各个书籍名称、书籍题材和书籍积分；
2. 先将筛选结果以书籍题材进行分类，再针对各题材以书籍积分为单位进行由大到小的排序。
3. 用表格展示各题材积分排名前十位的书籍作品。

## **3.8长篇排行**

1. 使用前面爬取和存入到Spark中的数据，进行数据处理。提取各个书籍id、书籍积分及对应的书籍起始更新时间、最近一周更新字数，筛选书籍起始更新时间距今超过一年的书籍；
2. 将筛选后的结果按照最近一周更新字数和书籍积分由大到小进行排名；
3. 用柱状图展示长篇排行中书籍作者、积分等信息。

## **3.9作者积分排行**

1. 使用前面爬取和存入到Spark中的数据，进行数据处理。提取各个作者名称和作者积分；
2. 将筛选后的结果按照作者积分由大到小进行排名；
3. 用词云显示该作者的代表作信息。

## **3.10新晋作者排行**

1. 使用前面爬取和存入到Spark中的数据，进行数据处理。提取各个作者名称、作者入站时间和用户申请成为作者30天内所写书籍积分；
2. 将筛选后的结果按照申请成为作者30天内所写书籍积分由大到小进行排名；
3. 用条形图显示该作者创作的图书及积分信息。

## **3.11稳定更新排行**

1. 爬取“晋江文学”网站作品库的所有书籍倒数三个月内所获得的小红花数；
2. 将数据存入HDFS；
3. 用Spark读取HDFS；
4. 使用前面爬取和存入到Spark中的数据，进行数据处理。提取书籍名称以及对应书籍倒数三个月内所获得的小红花数；
5. 将筛选后的结果按照如下公式进行计算，并将计算结果由大到小进行排名；

公式：



1. 用表格显示该图书更新信息。

## 3.12 本月完结收益排行

1. 使用前面爬取和存入到Spark中的数据，进行数据处理；
2. 筛选状态为完结的书籍，并提取各书籍id、书籍题材、书籍名称及其完结后一个月（30天）内的收益总额；
3. 将筛选后的结果按照收益总额由大到小进行排名；
4. 用气泡图展示本月完结收益排名前十位的书籍作品。

**3.13 打赏收益排行**

1. 使用前面爬取和存入到Spark中的数据，进行数据处理；
2. 提取各个书籍id、书籍题材、书籍名称、书籍作者以及该书籍获得的书籍打赏数额；
3. 将筛选后的结果按照书籍打赏数额由大到小进行排名；
4. 用柱状图展示截至当前时间书籍打赏数额排名前十位的书籍作品。

## 3.14 读者消费排行

1. 使用前面爬取和存入到Spark中的数据，进行数据处理；
2. 提取读者id、读者名称、读者近期所阅读的书籍题材信息以及该读者截至当前时间的打赏数额；
3. 将筛选后的结果按照打赏数额由大到小进行排名；
4. 用条形图展示截至当前时间读者打赏数额排名前十位的读者信息。

## 3.15 读者订阅偏好

1. 使用前面读者消费排行所显示的读者信息，根据读者id在spark中筛选出读者信息；
2. 分析读者近期打赏的书籍信息，提取书籍id、书籍名称以及书籍题材信息；
3. 按照读者id的顺序计算各个读者阅读书籍的题材占比，即；
4. 用饼状图展示截至当前时间在读者消费排行前十的读者在某一题材下的订阅百分比。

## 3.16 同类网站排行

1. 爬取晋江文学城、起点中文网、简书、豆瓣读书、笔趣阁、纵横中文网、潇湘书院和文章阅读网的用户量数据；
2. 将数据存入到Spark中进行数据处理；
3. 按照用户量数据由大到小的顺序进行排名；
4. 用折线图展示各网站用户量信息。