

"2014"

图书垂直搜索总体设计

系统名称	图书垂直搜索
项目负责人	徐广金
作者	徐广金, 卢俊, 王航, 李晓星, 占元涛
文档提交日期	

百度在线网络技术（北京）有限公司
(版权所有, 翻版必究)

修改记录

No	修改后版本号	修改内容简介	修改日期	修改人
1	1.0	全文初稿	2014.06.12	徐广金
2	1.1	补充部分细节	2014.06.16	卢俊, 王航, 元涛, 晓星

目 录

1 背景.....	1
2 名词解释.....	1
3 设计目标.....	1
3.1 实现的功能.....	1
3.2 设计的性能指标.....	2
4 系统环境（可选）.....	2
4.1 假设及与其它系统联系.....	2
4.2 相关软件及硬件.....	2
4.3 数据规模估计.....	2
5 设计思路及折衷.....	2
5.1 网络数据挖掘/处理的整体方案.....	2
5.2 网页数据抓取方案.....	4
5.3 网页数据时效性.....	4
5.4 数据结构化方案.....	5
5.5 结构化数据的结果检测.....	6
5.6 图片抓取方案.....	6
5.7 结构化结果与聚合的交互方案.....	7
5.8 检索系统选型方案.....	7
5.9 结构化数据存储方案.....	8
5.9.1 评论信息.....	8
5.9.2 图书基本信息.....	8
5.10 中间页数据的来源选择.....	9
5.11 百度阅读数据.....	9
5.12 检索更新.....	10
6 基本介绍.....	11
6.1 系统流程图及说明.....	11
6.1.1 文库/阅读业务架构图.....	11
6.1.2 系统核心流程.....	12
6.1.3 数据上游流程图.....	13
6.1.4 检索模块架构图.....	14
6.2 数据库表设计.....	14
6.2.1 图书评论表（ddbs）.....	14
6.2.2 图书评论排序表.....	15
6.2.3 评论印象表.....	16
6.2.4 图书评论增量表.....	16
) ENGINE=InnoDB DEFAULT CHARSET=gbk COMMENT='评论增量表';.....	17
6.2.5 图书聚合表（中间页使用）.....	17
6.2.6 相关推荐表.....	19
) ENGINE=InnoDB DEFAULT CHARSET=gbk COMMENT='评论印象表';.....	19
6.2.7 图书源信息表.....	19
6.2.8 前端中间页表.....	21

6.2.9 图书抓取整合交互队列表.....	22
7 系统设计.....	22
7.1 检索系统.....	23
7.2 抓取/结构化/聚合系统	23
8 风险评估及对其它系统影响（可选）	23
8.1 已知的或可预知的风险.....	23
8.2 与其它系统可能的影响.....	23
9 技术委员会审核意见.....	23
10 设计评审意见.....	23
11 附件及参考资料.....	23

1 背景

百度阅读的定位是全网的平台化产品，用户可以通过百度阅读查找、阅读、购买包括纸质、电子书在内的所有书籍商品，并且能够在平台中获得商品购买的指导性意见。在这个需求下，全网范围内的图书垂直搜索的重要性就不言而喻。

在当前搜索领域的前沿上，知识图谱的概念方兴未艾，也是我厂的技术趋势。本次项目将要推出的“图书中间页”也是符合该项理念趋势的技术/产品的成果。当用户搜索一本书，他需要的不仅仅是书的标题、作者、价格，也包括了其他人的评价、标签、售卖量等等一系列的相关信息。我们的“图书中间页”正是要满足用户的这一需求，从全网的数据中抓取图书的基本信息和相关属性，并从这些数据中挖掘出更深层次的信息，更好的满足用户在图书阅读、购买等方面的需求。

2 名词解释

spider: 搜索引擎数据流的最上游，负责将互联网上的资源采集到本地，提供给后续建立索引使用

SE: ksarch 提供的一个实时的小数据检索服务，用户通过向检索服务添加数据，建立倒排数据列表，最后即可通过关键字，语句等查询符合条件的内容。

收录服务: 公司内的通用服务，将 spider 的功能服务化，方便各应用方的使用。

PIE: 网页结构化信息抽取平台(Page Information Extraction)，致力于为百度产品便捷、快速地获取全网结构化数据，打造从网页获取、解析、到结构化数据的一体化流程。

cspub: 抓取服务平台。用户只需要申请，提交抓取的 url，即可使用，无需自己搭建抓取环境

GIPS: General Image Process System，是图库(wdm-img) 提供的一款图片处理服务，可将上传的图片做压缩、裁剪等多种处理，然后以 URL 返回在百度图库中的访问地址。接口包括 API、客户端工具、web 浏览器等三种方式提供服务

3 设计目标

3.1 实现的功能

- 1、全网图书（包括纸质、电子）的垂直搜索引擎，包括数据抓取、入库、检索的完整解

决方案。暂定的首期收录网站包括京东、亚马逊、当当、豆瓣等。

- 2、给出网络数据处理的使用规范及整体解决方案，包括数据抓取、归并、更新、存储等各个流程。

3.2 设计的性能指标

This document was truncated here because it was created using Aspose.Words in Evaluation Mode.