【项 目 名 称】

需求说明书

目录

1 引言

1.1 编写目的

该文档首先给出了整个系统的整体网络结构和功能结构的概貌，反映出搜索引擎系统的结构，试图从总体架构上给出整个系统的轮廓，然后又对功能需求、性能需求和其它非功能性需求进行了详细的描述。为开发人员、维护人员、需求人员间提供共同的协议而创立基础，对软件功能的实现作使命描述，作为软件人员进行设计和编码的基础；作为需求人员和开发人员之间的共同文档，为双方相互了解提供基础；确定系统测试及验收内容。该文档详尽说明了这一软件产品的需求和规格，这些规格说明是进行设计的基础，也是编写测试用例和进行系统测试的主要依据。同时，该文档也是用户确定软件功能需求的主要依据。

1.2 范围

本文档的适用范围为项目的开发人员、业务或需求分析人员、测试人员、用户文档编写者、项目管理人员，也适用于客户。

该产品是在积累了丰富业务经验的基础上进行开发的，在需求上，充分考虑了具体用户的实际情况。

1.3 定义

搜索引擎是指一种 web 上应用的软件系统，他以一定的策略在 web 上搜集和发现信息，在对信息进行处理后和组织后，为用户提供 web 信息查询服务。从使用者的角度来看，这种软件系统提供一个网页界面，让他通过浏览器提交一个词语或者短语，然后很快返回一个可能和用户输入内容相关的信息表。

1.4 参考资料

搜索引擎——原理、技术于系统

Java how to program

Java 程序设计教程

2 项目概述

2.1 目标

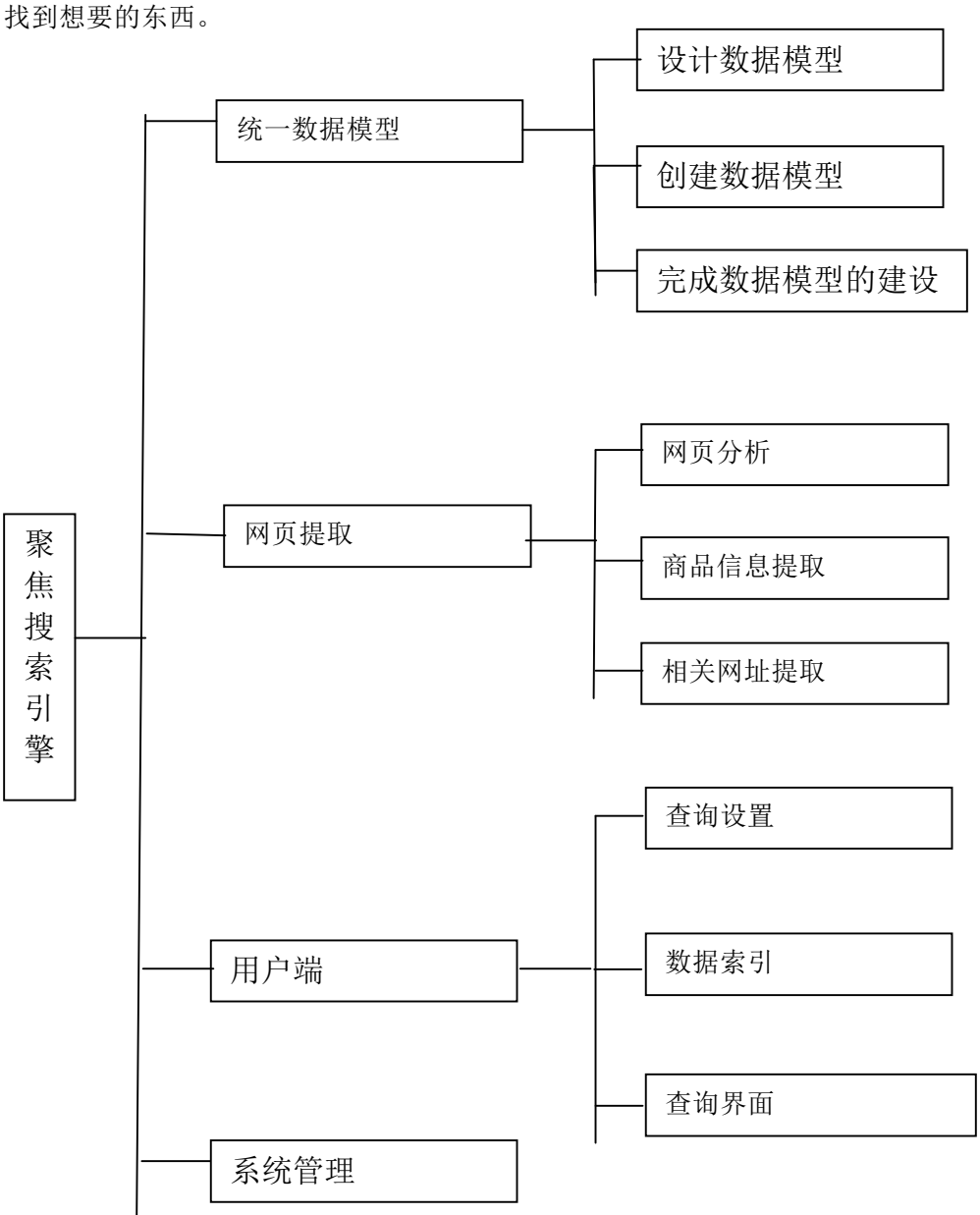
本系统的目标是为了使普通用户能够在互联网上方便的共享资源，为用户提供一个统一的资源平台，用户通过使用本系统提供的客户端应用程序，可以方便的搜索和下载互联网上各种不同访问

形式的资源的同类资源。这里所指的资源是指包括文档，视频，音频，图片等在内的各种类型的文件。该系统具有下载统一性，搜索功能强大和平台无关等优点。

该系统主要是针对目前的搜索引擎的非实时性和通用性和模糊搜索等，即不知道资源的可访问性的缺点而提出来的，同时进行了其他扩展，比如，对于目前可使用的获取资源的方式太多，但各有利弊，从而无从选择的缺点，提出了综合各种访问协议的资源形成统一的资源平台的想法；对于可访问资源太多，无从选择的缺点提出了搜索最热门的资源等功能。

2.2 产品功能

本软件为用户提供一个搜索平台，让用户能搜到想要的一类的一系列东西，使用户更能快捷的找到想要的东西。



2.3 用户特点

本系统最终用户是所有想在互联网上搜索和下载资源的普通用户，系统的操作人员即是普通用

户，系统的维护人员是具有一定的 java 语言编程能力，拥有一定的网络信息知识的技术人员。本系统预期的使用频度将会很高。

2.4 假定和依赖

本项目是否能够成功实施，主要取决于以下几点：

（1）为了项目的开发下的条件和实施，在必要时对现有的业务流程进行合理的调整。

（2）学校为调研、开发和实施过程提供必要的工作环境和系统运行环境，这些环境有助于软件开发开展工作。

（3）学校为软件系统的运行提供必要的且能够满足系统运行条件的硬件环境和通讯环境，不合适的硬件环境和通讯环境将会影响系统的性能。

3 具体需求

3.1 功能需求

3.1.1 功能描述

随着各种 B2C、C2C、B2B 的网站的成长和各种测评网站的建设。足不出户的网络购物已然成为大家生活中的一部分。但是随之而来想要从各种网站中找到最好、最划算的商品也非常困难。所以提供一个高质量的在线商品搜索平台无疑会为网络购物带来更好的未来。虽然用户可以通过 Google、百度、Yahoo 等通用搜索引擎，在海量的网络数据中取得一些信息。但是通用性搜索引擎也存在着一定的局限性

3.1.2 输入

能够在指定的网页入口中通过爬虫按照建立的模型分析商品的有效信息（基本属性），并保存相应数据信息。如：淘宝网针对某个商品的描述，除了基本属性外还包括了一些详细描述、商家信息、成交记录、付款方式等。我们需要剔除除了商品信息以外的部分，做到准确分析。可以自行设置网页入口，准确找到网页中的属性信息。

3.1.3 输出

针对网页分析结果和数据的展示，通过用户输入搜索关键字，可以展示用户所搜索的商品的相关信息。根据关键字，搜索商品信息，提供商品展示界面。展示商品价格、评价等信息，如：近期商品的价格趋势图、购买数量趋势图等。提供商品的横向、纵向关联信息展示，可以将相似属性（与该商品相同的价格区间、品牌、配置等信息）进行扩展展示。

3.2 性能需求

（1）查询页面一般响应时间不能过长，影响用户的使用。

（2）要保持数据库中的信息是最新的。

（3）支持多用户并发使用，并保证性能不受影响。

3.3 外部接口需求

3.3.1用户接口

无特殊需求。

3.3.2硬件接口

服务器端建议使用专用服务器。

3.3.3软件接口

无特殊需求。

3.3.4通信接口

无特殊需求。

3.4 属性

在软件的需求之中有若干个属性，下面指出其中的几个（注意：对这些决不应理解为是一个完整的清单）。

3.4.1可用性

（1）方便操作，操作流程合理

尽量从用户角度出发，以方便使用本产品。

（2）支持没有计算机使用经验、计算机使用经验较少及有较多计算机使用经验的用户均能方便地使用本系统。

（3）容错能力

系统具有一定的容错和抗干扰能力，在非硬件故障或非通讯故障时，系统能够保证正常运行，并有足够的提示信息帮助用户有效正确地完成任务。

（4）用户可自定义

为了满足业务的不断变化，一些重要的参数应该可以灵活设置。

（5）联机帮助与操作指南。

3.4.2安全性

（1）权限控制

根据不同用户角色，设置相应权限，用户的重要操作都做相应的日志记录以备查看，没有权限的用户禁止使用系统。普通用户只可查询商品，系统管理员可以维护系统。

（2）记录日志

本系统应该能够记录系统运行时所发生的所有错误，包括本机错误和网络错误。这些错误记录便于查找错误的原因。日志同时记录用户的关键性操作信息。

3.4.3可维护性

3.4.4可转移/转换性

本搜索引擎兼容性强，可在多种环境下运行。

3.5 其他需求

3.5.1 数据库

4 运行环境需求

4.1 设备

该系统为 B/S 三层结构，它的运行环境分客户端、应用服务器端和数据库服务器端三部分。

以下是系统的软件环境。

（1）客户端

操作系统：Windows2000 Professional/XP 或更新版本。

浏览器：IE6 以上，其它常见浏览器。

（2）应用服务器端

操作系统：Windows2000 Server 或更新版本。

应用服务器：Tomcat 5.5 或更新版本。

数据库访问：JDBC。

（3）数据库服务器端

操作系统：Windows2000 Server 或更新版本。

数据库系统：SQLServer 2000 或更新版本。

4.2 支持软件

对具体开发环境和语言不做要求。

5 附录

不同领域、不同背景的用户往往具有不同的检索目的和需求，通用搜索引擎所返回的结果包含大量用户不关心的信息。通用搜索引擎的目标是尽可能大的网络覆盖率，有限的搜索引擎服务器资源与无限的网络数据资源之间的矛盾将进一步加深。万维网数据形式的丰富和网络技术的不断发展，图片、数据库、音频/视频多媒体等不同数据大量出现，通用搜索引擎往往对这些信息含量密集且具有一定结构的数据无能为力，不能很好地发现和获取。通用搜索引擎大多提供基于关键字的检索，难以支持根据语义信息提出的查询。为了解决上述问题，定向抓取相关网页资源的聚焦爬虫应运而生。聚焦爬虫是一个自动下载网页的程序，它根据既定的抓取目标，有选择的访问万维网上的网页与相关的链接，获取所需要的信息。与通用爬虫不同，聚焦爬虫并不追求大的覆盖，而将目标定为抓取与某一特定主题内容相关的网页，为面向主题的用户查询准备数据资源。

聚焦爬虫的特点从“聚焦”两字便可以体现，它的搜索和下载会只针对特定的信息和网站。需要根据一定的网页分析算法过滤与主题无关的链接，保留有用的链接并将其放入等待抓取的 URL 队列。并会针对抓去的页面按照所需的信息进行分析、过滤，提取出有用的信息并建立相关索引。在后续的分析过程中得出的信息，将为后续的抓取过程给出反馈和指导。