安徽三联学院

ANHUISANLIANUNIVERSITY

本科毕业论文(设计、创作)

题目： 基 于Web 搜索引擎的设计与实现

Design and Implementation of Web-based search engine

摘 要

网络中的资源非常丰富，但是如何有效的搜索信息却是一件困难的事情。建立搜索引擎就是解决这个问题的最好方法。本文首先详细介绍了基于英特网的搜索引擎的系统结构，然后从网络机器人、索引引擎、Web 服务器三个方面进行详细的说明。在次基础上设计并实现了一种快捷高效的新闻搜索引擎，该搜索引擎是从指定的Web 页面中按照超连接进行解析、搜索，并把搜索到的每条新闻进行索引后加入数据库，然后通过 Web 服务器接受客户端请求后从索引数据库中搜索出所匹配的新闻。

关键词：搜索引擎；网络机器人；索引引擎;Web服务器

Abstact

The network resources are very rich , but how effective search information is a difficult thing . Build a search engine is the best way to solve the problem . This paper first introduced the Internet search engine based on the structure of the system, and then from the network robots , indexing engine , the Web server of the three aspects of the detailed instructions.

Based on the time designed and realized a quick and efficient news search engine, the search engine from the Web page in accordance with specified in the connection for analytical , search , and the search to every news indexing of add to the database . Then through the Web server accept client requests from database search index after the news of the match.

Keywords: search engine ; Network robot ; Indexing engine ; Web server

目　　录

第一章　绪论

1.1搜索引擎出现的背景与意义

网络的出现以与发展对于世界发展的意义是极其重要的，它让地球村的理念变成现实，信息的传输不再受到时间和空间的限制。在没有搜索引擎的时代，用户希望寻找某方面的信息，就必须通过各种途径或者是之间的连接寻找，可以这样说，脱离了搜索引擎的，就像是信息海洋中的一个一个孤岛，用户必将面临巨大的搜索成本，同时必须付出大量的时间和精力。

搜索引擎的出现改变了上述的现象，它通过程序的自动搜寻并建立索引，将这些信息孤岛联系起来，形成了一巨大的信息网，并且运用分布式计算的巨大力量，能够让用户从海量数据中摒除垃圾信息，获取想要的知识。搜索引擎不仅仅是节省了用户的时间，通过挖掉搜寻成本这座墙，它让许许多多的不可能成为可能。

1.2搜索引擎的发展历史与趋势

搜索经历了三代的更新和发展：

第一代搜索引擎出现于1994年。这类搜索引擎一般都索引少于1,000,000个网页，极少重新搜集网页并去刷新索引。而且其检索速度非常慢，　一般都要等待10秒甚至更长的时间。

第二代搜索出现在1996年。这类搜索引擎系统大多采用分布式方案(多个微型计算机协同工作)来提高数据规模、响应速度和用户数量，它们一般都保持一个大约

50,000,000网页的索引数据库，每天能够响应10,000,000次用户检索请求。

第三代搜索引擎年代的划分和主要特性至今没有统一的认识，不过至少可以肯定的是：第三代搜索引擎是对第二代搜索引擎在搜索技术上的改进，主要增加了互动性和个性化等高级的技术，为用户使用搜索引擎获取信息获得更好的体验。至于互动性的评价标准是什么，以与第三代搜索引擎到底比第二代搜索引擎增加了多少价值——尤其是为企业利用搜索引擎开展网络营销增加了哪些价值，目前并没有非常令人信服的研究结论。这也就是目前所谓的第三代搜索引擎并没有表现出太多优势的原因之一。

现在，网络上有很多著名的搜索引擎，百度，google 等等，百度从2005年诞生到现在成为全球最大的中文搜索引擎，可想而知，发展的速度是多么的快，人们对搜索引擎的的需多么的大，百度的日点击率我无法在找到确切的数字，但是我们可以计算一下，截至2008年底，中国网民规模达到2.98亿人，每个网民上网点击百度的次数应该不少于十次吧，像我们要在百度上找资料的网名点击率百次不止，所以百度的日点击率是多么惊人。

搜索引擎经过几年的发展和摸索，越来越贴近人们的需求，搜索引擎的技术也得到了很大的发展。搜索引擎在将来的的发展趋势大概有以下几个方面：

1. 提高对用户输入的理解

为了提高搜索引擎对用户检索提问的理解，就必须有一个好的检索提问语言，为了克服关键词检索和目录查询的缺点，现在已经出现了自然语言智能答询。用户可以输入简单的疑问句，比如“how can kill virus of computer?”。 搜索引擎在对提问进行结构和容的分析之后，或直接给出提问的答案，或引导用户从几个可选择的问题中进行再选择。自然语言的优势在于， 一是使网络交流更加人性化，二是使查询变得更加方便、直接、有效。就以上面的例子来讲，如果用关键词查询，多半人会用“virus”这个词来检索，结果中必然会包括各类病毒的介绍、病毒是怎样产生的等等许多无效信息，而用“how can kill virus of computer?”,搜索引擎会将怎样杀病毒的信息提供给用户，提高了检索效率。

2. 对检索的结果进行处理

对检索的结果处理，有以下几个方向：其一，使用评价，就是将网页的数量算作网页评分因素之一，这样搜索的结果就更加的能够满足用户的要求，在这个方面google(.google.cn) 的“评价体系”已经做出了相当出色的成绩。其二，使用大众

访问性，就是将访问数量(也可以叫做点击数量)算作网页评分的因素之一，这样想，sina,,cn　这样的的分数会很高，而这样的很多时候都是用户想找的，这样能够提高搜索引擎的准确率。其三，去掉结果中的附加信息。有调查指出，过多的附加信息**加重了用户的信息负担，为了去掉这些过多的附加信息，可以采用用户定制、容过滤等检索技术。**

3. 确定搜集返回，提高针对性

在这个方面现在的发展的方向是：其一，垂直主题搜索。垂直主题的搜索引擎以其高度的目标化和专业化在各类搜索引擎中占据了一系席之地，比如象股票、天气、新闻等类的搜索引擎，具有很高的针对性，用户对查询结果的满意度较高。其二，非Www信息的搜索。搜索引擎提供了例如 ftp 等非Www信息的搜索。其三，多媒体搜索。搜索引擎还提供了例如包括声音、图像等等多媒体信息的检索。

4. 提供更优化的检索结果

在这个方面有两个主要的发展方向：其一，纯净搜索引擎。这类搜索引擎没有自己的信息采集系统，利用别人现有的索引数据库，主要关注检索的理念、技术和机制等。其二，元搜索引擎。元搜索引擎(metasearch　enging)是将用户提交的检索请求到多个独立的搜索引擎上去搜索，并将检索结果集中统一处理，以统一的格式提供给用户，因此有搜索引擎之上的搜索引擎之称。它的主要精力放在提高搜索速度、智能化处理搜索结果、个性搜索功能的设置和用户检索界面的友好性上，查全率和查准率都比较高。

第二章搜索引擎的结构

2.1系统概述

搜索引擎是根据用户的查询请求，按照一定算法从索引数据中查找信息返回给用户。为了保证用户查找信息的精度和新鲜度，搜索引擎需要建立并维护一个庞大的索引数据库。 一般的搜索引擎由网络机器人程序、索引与搜索程序、索引数据库等部分组成。

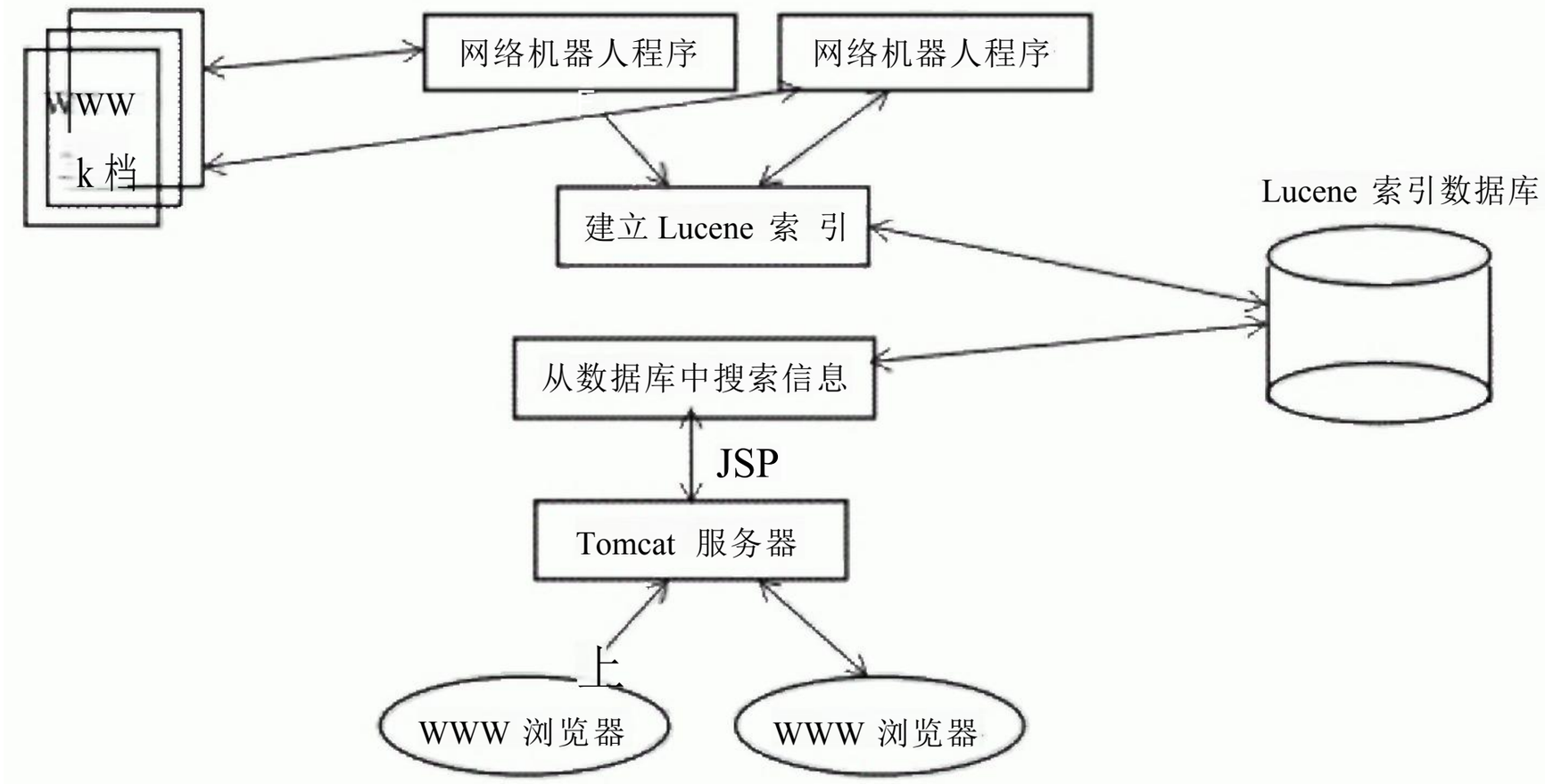

图 1 搜索引擎的系统结构

2.2搜索引擎的构成

22.1 网格机居人也称为“网格缺妹”(Sidar),是一个功能报强的删**原制九文档**

可以在扫描WEB 页面的同时检索其的超并加入扫描队列等待以后扫描@因为WEB 虫六载高清无水印泛使用超，所以一个 Spider 程序理论上可以访问整个WEB 页面。

为了保证网络机器人遍历信息的广度和深度需要设定一些重要的并制定相关的扫描策略。

2.2.2索引与搜索

网络机器人将遍历得到的页面存放在临时数据库中，如果通过SQL 直接查询信息速度将会难以忍受。为了提高检索效率，需要建立索引，按照倒排文件的格式存放。如果索引不与时更新的话，这样用户用搜索引擎也不能检索到。

用户输入搜索条件后搜索程序将通过索引数据库进行检索然后把符合查询要求的数据库按照一定的策略进行分级排列并且返回给用户。

2.2.3 Web服务器

客户一般通过浏览器进行查询，这就需要系统提供Web服务器并且与索引数据库进行连接。客户在浏览器中输入查询条件，Web 服务器接收到客户的查询条件后在索引数据库中进行查询、排列然后返回给客户端。

2.3搜索引擎的主要指标与分析

搜索引擎的主要指标有响应时间、召回率、准确率、相关度等。这些指标决定了搜索引擎的技术指标。搜索引擎的技术指标决定了搜索引擎的评价指标。好的搜索引擎应该是具有较快的反应速度和高召回率、准确率，当然这些都需要搜索引擎技术指标来保障。

召回率： 一次搜索结果中符合用户要求的数目与用户查询相关信息的总数之比

准确率： 一次搜索结果中符合用户要求的数目与该次搜索结果总数之比

相关度：用户查询与搜索结果之间相似度的一种度量

精确度：对搜索结果的排序分级能力和对垃圾网页的抗干扰能力

2.4 小 结

以上是对于基于因特网的搜索引擎的结构和性能指标进行了分析，本人在这些研究的基础上利用 JavaTM技术和一 些Open Source工具实现了一个简单的搜索引擎——新闻搜索引擎。在接下来的几章里将会就本人的设计进行详细的分析。

第三章网络机器人

3.1什么是网络机器人

网络机器人又称为 Spider 程序，是一种专业的 Bot 程序。用于查找大量的 Web 页面。它从一个简单的Web 页面上开始执行，然后通过其超访问其他页面，如此反复，理论上可以扫描互联网上的所有页面。

基于因特网的搜索引擎是Spider 的最早应用。例如搜索巨头 Google 公司，就利用网络机器人程序来遍历 Web 站点，以创建并维护这些大型数据库。

网络机器人还可以通过扫描Web 站点的主页来得到这个站点的文件清单和层次机构。还可以扫描出中断的超和拼写错误等。

3.2 网络机器人的结构分析

Internet是建立在很多相关协议基础上的而更复杂的协议又建立在系统层协议之上。Web 就是建立在 (Hypertext Transfer Protocol ）协议基础上，而 又是建立在 TCP/IP(Transmission Control Protocol /Internet Protocol) 协议之上，它同时也是一种 Socket 协议。所以网络机器人本质上是一种基于 Socket 的网络程序。

3.2.1如何解析 HTML

因为Web 中的信息都是建立在HTNL 协议之上的，所以网络机器人在检索网页时的第一个问题就是如何解析 HTML。在解决如何解析之前，先来介绍下HTML 中的几种数据。

我们在进行解析的时候不用关心所有的标签只需要对其中几种重要的进行解析即可。

文本：除了脚本和标签之外的所有数据

注释：程序员留下的说明文字，对用户是不可见的

简单标签：由单个表示的　　L　标签

开始标签和结束标签：用来控制所包含的 HTML 代码

我们在具体解析这些HTNI 标签有两种方法：通过 JavaTM 中的 S者通过 Bot 包中的ETEPage　类来解析，水入在实际编程中采用后者

Bot 包中的ILPage　　类用来从指定URL中读取数据并检索出有用的信息）　　给18．com

出该类几种重要的方法。

NLPage 构造函数：构造时象并指定用于通讯的　　对象

Public HMPage(　　)

GetForns 方法：获取最后一次调用Open 方法检索到的表单清单

Public Vector getForms()

Get　　方法：获取发送给构造函数的　　对 象

Public　　get　　()

Getimage 方法：获取指定页面的图片清单

Public Vector getlmage()

GetLinks 方法：获取指定页而的连接清单

?ublic Vector getlinks()

Open 方法：打开一个页面并读入该页而，若指定了回调对象则给出所有该对象数据

Public void open(String url,HTMLEditorKit.ParserCallback a)

3.2.2 Spider 程序结构

网绝机器人必须从一个网页迁移到另一一个网页所以必须找到该页面上的超连接。程序首先解析网页的TL　代码，查找该页面的超连接然后通过道归和非递归两种结构来实现 Spider 程序。

虽然这望只描述了一个队列·但在实际编程中用到了四个队列，他们每个队列都保存着同一一处理状态的R。

等待队列：在这个队列中， URL 等待被Spider 程序怒理。新发现的URL 也被加入到这个队列市“

处理队列：当Spider 程序开始处理时；他们被送到这个队列中

错误队列：如果在解析网页时出错， URL 将被送到这里。该队列中的 URL 不能被移入其他队列中

完成队列：如果解析网页没有出错， URL 将被送到这里。该队列中的 URL 不能被移入其它队列中

在同一时间 URL 只能在一个队列中，我们把它称为URL 的状态。

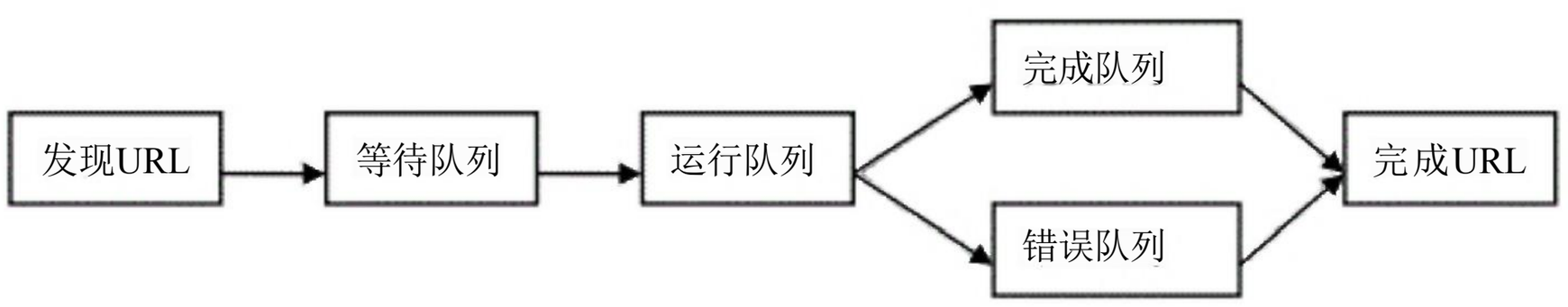

图 1 队列的变化过程

以上的图表示了队列的变化过程，在这个过程中，当一个URL 被加入到等待队列中时 Spider 程序就会开始运行。只要等待队列中有一个网页或 Spider 程序正在处理一个网页·程序就会继续他的工作。当等待队列为空并且当前没有任何网页时·Spider 程序就会停止它的工作。

3. 2. 3如何构造 Spider 程序

在构造Spider 程序之前我们先了解下程序的各个部分是如何共同工作的。以与**如何对这个程序进行扩展。**

流程图如下所示