TURING
图灵程序设计丛书

自制搜索引擎

How to Develop a Search Engine

[日] 山田浩之　末永匡 / 著　胡屹 / 译

2600行代码

真实体验搜索引擎的开发过程

开源搜索引擎Senna/Groonga的开发者亲自执笔

Google、百度的工作机制

中国工信出版集团

人民邮电出版社
POSTS & TELECOM PRESS

TURING
图灵程序
设计丛书

自制搜索引擎

How to Develop a Search Engine

[日] 山田浩之　末永匡 / 著　胡屹 / 译

人民邮电出版社
北　京

图书在版编目（CIP）数据

自制搜索引擎/（日）山田浩之，（日）末永匡著；
胡屹译. -- 北京：人民邮电出版社，2016.1
（图灵程序设计丛书）
ISBN 978-7-115-41170-9

Ⅰ. ①自… Ⅱ. ①山… ②末… ③胡… Ⅲ. ①互联网络－情报检索 Ⅳ. ①G354.4

中国版本图书馆CIP数据核字（2015）第282984号

内 容 提 要

本书聚焦于Google和Yahoo!等Web搜索服务幕后的搜索引擎系统，首先讲解了搜索引擎的基础知识和原理，接着以现实中的开源搜索引擎Senna/Groonga为示例，使用该引擎的源代码引导读者亲自体验搜索引擎的开发过程。这部分讲解涉及了倒排索引的制作和压缩、检索的处理流程以及搜索引擎的优化等内容。最后又简单介绍了一些更加专业的搜索引擎的知识和要点，为读者今后进一步学习打下了基础。本书适合所有对搜索引擎感兴趣的技术人员阅读。

◆ 著　[日]山田浩之　末永匡
译　胡　屹
责任编辑　乐　馨
执行编辑　高宇涵
责任印制　杨林杰
◆ 人民邮电出版社出版发行　北京市丰台区成寿寺路11号
邮编　100164　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京隆昌伟业印刷有限公司印刷
◆ 开本：880×1230　1/32
印张：6.5
字数：187千字　2016年1月第1版
印数：1-4 000册　2016年1月北京第1次印刷
著作权合同登记号　图字：01-2015-1264号

定价：39.00元

读者服务热线：(010)51095186转600　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京崇工商广字第0021号

版权声明

●**免责声明**

本书所述内容仅以提供信息为目的。因此，请读者务必根据自身的义务和判断运用书中的信息，由此产生的任何后果，技术评论社、原书作者、人民邮电出版社以及译者概不负责。

由于本书所提供的是截至2014年8月8日的信息，所以在使用时亦会出现所述内容已发生变更的情况。

另外，软件版本的升级会导致其功能和界面等与本书所述不符。因此，在购买本书前，请您务必确认软件的版本号。

请在阅读并同意了上述注意事项之后再使用本书，否则，技术评论社、原书作者、人民邮电出版社以及译者恐怕难以回复您的询问。这一点还望诸位读者事先知晓。

●**关于商标和注册商标**

本书所述的产品名称多为各相关公司的商标或注册商标。另外，本书已省略了™、® 等符号。

译者序

《自制搜索引擎》一书终于和读者们见面了，“自制”系列图书的家族中又多了一名新成员。近几年，图灵先后出版了几本“自制”系列图书，如《30天自制操作系统》《自制编程语言》《两周自制脚本语言》等。在这些书中，我们不用去读枯燥乏味的原理和晦涩难懂的算法，只需跟随作者的脚步，即可从零开始，一步步地创造出操作系统或编程语言的雏形。

《自制搜索引擎》一书也不例外。在这本不到200页的书中，作者先用简明扼要、通俗易懂的语言为我们讲解了搜索引擎的结构及核心概念，紧接着又带领我们剖析了一个名为wiser的原创搜索引擎的源代码。理论与大量源代码的结合帮助我们迈入了搜索引擎的大门，只要用心阅读并实际操作，就能制作出一个可以在计算机上运行的简易搜索引擎。然而与其他计算机技术一样，虽然搜索引擎的入门很简单，但要成为这个领域的技术专家却并不容易，离不开大量的知识积累和实践。所以在分析完源代码以后，作者又带领我们优化了现有的wiser搜索引擎，并简单地介绍了一些更加专业的知识，以启发我们深入思考，为进一步学习铺平了道路。

阅读本书几乎不需要任何有关搜索引擎的知识储备，但由于wiser是用C语言编写的，所以您最好还是能有些C语言的编程经验。“啊，用C写的啊？”也许您也和我当初一样，一听是C语言就泄气了。的确，C语言不是那么好用。指针是个难点不说，有些语句的写法也显得很诡异，而且还缺乏丰富的内置函数和数据结构。但如果您坚信某某语言才是世界上最好的语言，并要因此放弃本书的话，那么我建议您先下载wiser的源代码读一读再做决定。wiser的源代码仅有大约2600行。即使

只瞥一眼，也应该能够发现这些源代码不但具有详细的注释、清晰的结构，而且遵循了良好的命名规范。仔细地阅读后，甚至还能看到有些地方应用了回调函数、设计模式等所谓的“现代”编程技巧。不仅如此，作者还通过引入了名为 uthash 的代码库简化了对字符串、列表和哈希表的操作。例如要向列表中添加元素时，只需使用形如“LL_APPEND(*list, element);”的一行代码，这就大大增加了代码的可读性。相信您读到最后也会由衷地感叹：原来 C 语言也能这么好用啊。

对于想要开发搜索引擎的读者来说，本书的作用自不必说。而对于专注于其他领域的开发者，甚至对于那些只是想学门新技术来娱乐一下的程序员来说，读读本书也是大有裨益的。例如，我们可以从中学到如何高效地求得多个大集合的交集，如何压缩存储大量的整数，如何运用 sar 命令查看并分析系统的性能等。即使我们不从事搜索引擎的开发工作，这些算法和技术也会对日常的工作有所启发和帮助。所以，读过了本书，就算您并不打算做一个搜索引擎出来，也能得到一些收获。

值得一提的是，在本书中很多叙述得较为简练甚至一笔带过的段落中，其实隐藏着大量的知识。在掌握了搜索引擎的核心技术后，不妨查查资料、写写代码，试着去掌握这些更高级的知识，搞清楚里面专业术语的含义。例如，书中提到了字典树（Tier）、Suffix Array 等国内教材中罕见的数据结构，那么我们能不能用自己熟悉的编程语言实现它们？作者开发的开源搜索引擎 Groonga 采用了内存映射文件技术，那么内存映射文件的机制是什么……在不断探索这些问题的过程中，我们不但能把这本不算厚的书读得越来越厚，也能使自己的知识量不断增长。

最后，在这里衷心感谢在翻译过程中给予我支持与鼓励的各位。欢迎诸位读者批评指正，提出宝贵的建议。希望所有对搜索引擎感兴趣的读者都能从本书中获益。

胡屹

2015 年 10 月于北京

前言

本书聚焦于 Google 和 Yahoo! 等 Web 检索服务幕后的搜索引擎，旨在阐明这种系统内部的工作机制。诸位读者通过第 1 章的学习，掌握了搜索引擎的基础知识和原理之后，就可以从第 2 章开始，对照着示例搜索引擎的源代码体验搜索引擎的开发过程了。这种原理和实践的有机结合，有助于大家更加深入地理解搜索引擎的构造。

一直在企业和大学从事搜索引擎研发工作的山田负责搜索引擎原理的写作，并完成了整体构思和统稿的工作。开源搜索引擎 Senna/Groonga 的开发者、拥有多个检索服务实战经验的末永在书中介绍了实践和运用搜索引擎时的要点。这种内容上的相互补充使得原理和实践有机地结合在了一起。

若从本书获得的知识和经验能有助于诸位读者创造出划时代的软件和服务，我们将感到不胜荣幸。

山田浩之

于 2014 年 8 月

Contents 目录

第1章

搜索引擎是如何工作的

在体验搜索引擎的开发过程之前，我们先在第 1 章介绍一下搜索引擎的基本概念。搜索引擎的基础是应用于信息检索、数据库等领域的信息技术，要想开发搜索引擎，横跨多个领域的广泛知识是不可或缺的。在本章我们尽可能通俗易懂、简明扼要地总结归纳了这些知识。由于本章讲解的是后续章节的背景知识，所以恳请诸位认真地读下去。