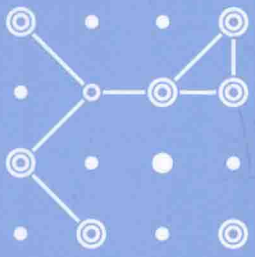
面向实战，基于大数据分析、知识图谱、
人工智能构建现代搜索引擎

# 大数据搜索引擎

## 原理分析及编程实现

刘凡平◎编著



中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

# 大数据搜索引擎原理分析及编程实现

刘凡平　编著

電子工業出版社
Publishing House of Electronics Industry
北京·BEIJING

## 内 容 简 介

本书向读者提供了一套完整的大数据时代背景下的搜索引擎解决方案，详尽地介绍了搜索引擎的技术架构、算法体系及取得的效果，以模块化的方式进行组织。着重介绍了机器学习在搜索引擎中的应用，包括中文分词、聚类、分类等核心的机器学习算法，并结合示例加以介绍和分析，使读者可以更好地理解机器学习在搜索引擎中的价值。还阐述了大数据给搜索引擎带来的新特性，结合目前大数据分析的主流工具，在搜索引擎中构建知识图谱，以及进行日志反馈学习机制，使得搜索引擎更加智能。

本书适合作为互联网行业从业者的技术参考书，也适合作为搜索引擎爱好者的参考读物。



**图书在版编目（CIP）数据**

大数据搜索引擎原理分析及编程实现 / 刘凡平编著. — 北京：电子工业出版社，2016.7

ISBN 978-7-121-29164-7

Ⅰ. ①大… Ⅱ. ①刘… Ⅲ. ①搜索引擎－程序设计Ⅳ. ①TP391.3

中国版本图书馆CIP数据核字（2016）第141781号

策划编辑：李　冰
责任编辑：李　冰
特约编辑：田学清　罗树利
印　　刷：北京季蜂印刷有限公司
装　　订：北京季蜂印刷有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路173信箱　　邮编：100036
开　　本：720×1000　1/16　印张：20.5　字数：525千字
版　　次：2016年7月第1版
印　　次：2016年7月第1次印刷
定　　价：59.00元

# 前　言

搜索引擎本身作为一门综合性的互联网技术，在行业中一直具备较高的关注度。随着最近几年大数据的发展，搜索引擎的关注度越来越高，原因在于搜索引擎技术是大数据应用最前线的领域，也是最容易产生价值的大数据应用。大数据存储、大数据计算都是从搜索引擎中衍生出的新领域。目前搜索引擎技术的发展不仅以大数据为基础，还利用分布式实时计算对数据进行高性能处理，以及利用机器学习将数据变得更具价值。在行业中吸引了包括搜索研发工程师、算法研发工程师、大数据分析工程师、自然语言处理工程师、计算平台架构师、数据挖掘工程师等行业各类精英的关注，这些工程师占据了整个互联网研发体系的 50%～60%，在 BAT 中，甚至超过 60% 的是研发人员。

这类群体薪资水平处于互联网研发人员中较高水平，以猎聘网公布的数据显示，北京地区搜索引擎研发工程师年薪为 35 万～60 万元，大数据工程师年薪为 20 万～25 万元，大数据架构师年薪为 40 万～70 万元，等等。也正是由于薪资水平处于较高位，很多互联网相关从业者也积极关注大数据搜索引擎领域动态。

本书通过介绍大数据下的自然语言处理框架、大数据存储引擎、搜索引擎的分布式实时计算、高性能可扩展爬虫，以及利用大数据构建知识图谱、基于大数据日志的搜索引擎反馈学习等相关信息，不仅使读者对当代搜索引擎研发体系有一定的认识，还可以使读者在搜索引擎领域及大数据领域进行深入思考。

## 本书特色

本书以当前搜索引擎主流技术为基础，密切结合前沿技术发展趋势，行文

通俗易懂，由初步的原理性了解到各模块应用示例，并结合分布式存储、实时计算等，向读者提供了一套完整的大数据时代背景下人工智能搜索引擎的解决方案。

**（1）内容循序渐进、行文有条有序地介绍搜索引擎知识。**

本书充分考虑了不同层次的读者对搜索引擎的理解程度，因此本书由简入深、独特的技术写作视角符合广大读者对于技术类读物的理解需求，使得读者能够在掌握搜索引擎基础的情况下，不断按照搜索引擎的设计深入理解。

**（2）技术前瞻性强，注重最新主流技术在现代搜索引擎中的应用。**

本书充分利用了最新技术发展的应用成果，在自然语言处理的基础上不仅结合大数据分析，还包括分布式计算、机器学习、知识图谱等当前大数据应用与分析处理的主流技术，摒弃了传统过时的研发体系及算法。本书中相关研发成果在当前甚至在未来3～5年，都具有实际意义。

**（3）将技术理论与应用范例结合，具备较高的商业实用价值。**

本书内容紧密结合当前一线工程师工作研究成果，将众多的技术理论以实际工作经验的方式展示应用效果。本书介绍的内容也广泛结合工作中的应用示例，并以搜索引擎工程实践的脉络流程介绍技术要点，使读者在短时间内能够掌握当前搜索引擎研发的技术理论。

## 本书结构

本书按照由浅入深、循序渐进的顺序对现代搜索引擎原理和实现进行介绍。全书分为10章，各章的主要内容如下。

第1章针对搜索引擎发展的过去、现在、未来的相关概要介绍，以及现代搜索引擎与大数据、人工智能的相互关系，使广大读者能够在了解现代搜索引擎的背景之下，去了解本书的后续内容。

第 2 章是对搜索引擎原理与技术的初步分析，从模块方面大致介绍爬虫、索引、缓存等；从技术方面大致介绍自然语言处理、知识图谱技术、海量数据存储、分布式计算等。目的是使得读者对搜索引擎的体系结构、部分技术有一定认识，便于读者深入了解后续章节。

第 3 章从自然语言角度开始深入分析原理和实现，自然语言是搜索引擎进行文本处理的基础，其中包括分词、词性分析、语义分析、关键词抽取、核心句抽取、聚类分类等。读者将会从本章中获得当前主流的自然语言处理技术相关知识。

第 4 章主要是针对大数据存储引擎的介绍。大数据存储是搜索引擎最先遇到的问题，解决数据存储问题可以使搜索引擎在数据分析、索引构建、知识图谱等工作持续进行。读者在本章会了解到大数据存储引擎的架构体系、数据模型、数据压缩、负载均衡等。

第 5 章介绍了分布式实时计算。由于搜索引擎处理的是海量数据，数据分析必须依靠具有较强数据处理能力的计算平台，因此搜索引擎通过分布式实时计算去处理大数据并在尽可能短的时间内返回处理结果。本章中，读者会了解到分布式实时计算设计架构、负载均衡及通信设计等相关知识。

第 6 章对爬虫进行了深入分析。读者在本章中将会深入理解分布式可扩展爬虫的体系架构，以及对网页如何进行解析，并抽取出结构化的数据信息。本章还涉及链接去重、网页去重、广告识别等相关算法原理。

第 7 章详细介绍了知识图谱构建。知识图谱是智能化搜索引擎重要的组成部分，利用大数据分析构建出较为合理的知识图谱信息是当前主流的方式。读者将会从本章中深入了解到知识图谱的详细构建过程，以及利用机器学习原理对知识图谱中的实体抽取、关系抽取等相关技术进行。

第 8 章详细分析了索引构建机制。索引的设计与构造是搜索引擎能够进行

快速检索的核心要件，主要针对文件检索的倒排索引与用于智能提示的字典树索引。本章不仅对倒排索引做了深入分析，对倒排索引的压缩、分布式存储等也做了详细介绍。

第 9 章深入分析了搜索引擎的整个对外服务工作流程。包括大数据分布式缓存、搜索智能提示、个性化搜索、图片搜索、搜索与广告等。读者通过本章可以详细了解到文本纠错算法、动态摘要算法、网页排序算法及搜索引擎的评价体系。

第 10 章探讨和分析了搜索引擎日志与搜索引擎本身的关系。搜索引擎日志记录了用户与搜索系统交互的整个流程。通过日志挖掘，不仅可以发现用户的自有特征和行为规律，还可以有效地帮助搜索引擎提升性能和效果。日志作为搜索引擎的核心数据之一，一直使搜索引擎技术中的各类算法不断向前发展。读者通过本章将学会通过搜索引擎日志分析用户特征、用户的部分搜索意图等相关知识。

## 读者对象

- 适合对自然语言处理及机器学习应用领域有兴趣的读者。
- 适合对现代搜索引擎相关算法有兴趣的读者。
- 适合对大数据分析、数据挖掘应用有兴趣的读者。
- 适合互联网行业的不同层次从业者。
- 适合从事搜索引擎优化的网络营销读者。
- 适合高校中学习计算机、软件工程等相关专业的读者。

# 目 录