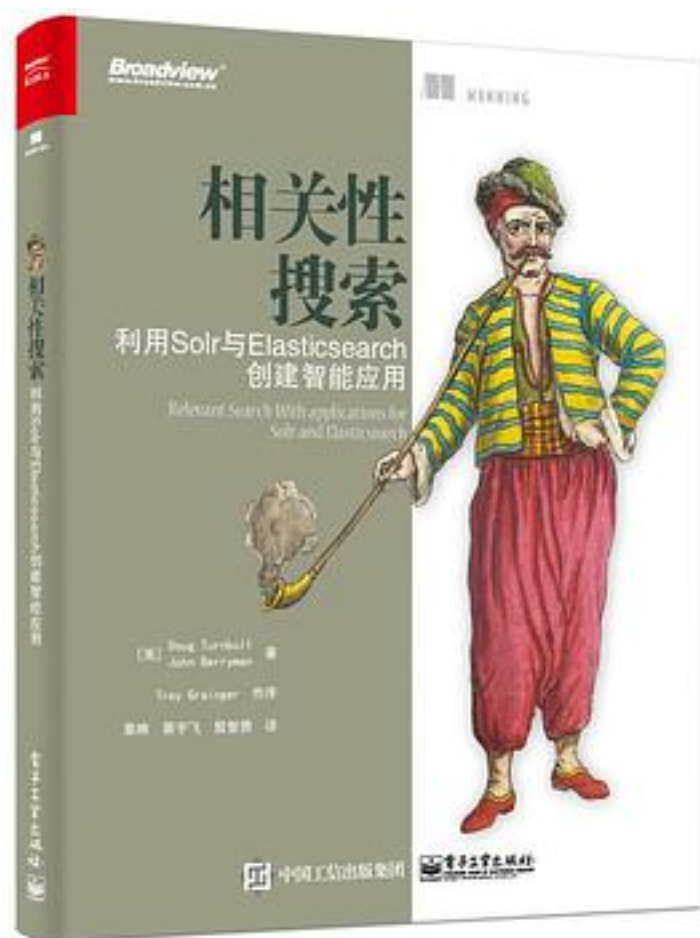


相关性搜索：利用Solr与Elasticsearch创建智能应用



[相关性搜索：利用Solr与Elasticsearch创建智能应用_下载链接1](#)

著者:[美]Doug Turnbull（道格·特恩布尔）

出版者:电子工业出版社

出版时间:2017-10-1

装帧:平装

isbn:9787121327216

《相关性搜索：利用Solr与Elasticsearch创建智能应用》揭开了相关性搜索的神秘面纱

，告诉大家如何将 Elasticsearch 与 Solr 这样的搜索引擎作为可编程的相关性框架，从而表达业务排名规则。从这《相关性搜索：利用 Solr 与 Elasticsearch 创建智能应用》中你可学会如何结合各种外部数据源、分类方法以及文本分析手段对相关性进行编程，以满足用户的个性化需求，将令人满意的搜索结果呈现给用户。此外，相关性搜索也需要一定的软性技能《相关性搜索：利用 Solr 与 Elasticsearch 创建智能应用》还将告诉读者怎样与业务人员协作，为业务找到正确的相关性需求，从而在搜索产品的整个研发生命周期内，实现相关性改进的良性循环。

本书介绍了搜索引擎的基本原理，及相关性搜索的调试技术，用大量实例的方式详述了搜索引擎的诸多特性，以形成一整套针对相关性搜索的系统化方法，并倡导致力于提高搜索质量的企业文化。《相关性搜索：利用 Solr 与 Elasticsearch 创建智能应用》适用于想利用 Elasticsearch 或 Solr 尝试构建智能搜索应用的开发人员。

作者介绍:

Doug Turnbull 在 OpenSource Connections 上领导着一项搜索相关性的咨询业务，在那里他经常发表观点和更新博客。Doug 利用各种搜索和自然语言处理技术（NLP）为多个领域的客户构建语义丰富的相关性搜索体验。

John Berryman 的第一份职业是航空工程师，但在航空领域工作了几年之后，他发现编写程序或解决数学难题才是他喜欢的工作。后来，John 撇下了飞机和卫星，开始全职工作于软件开发、基础架构，以及搜索技术领域。目前，John 供职于 Eventbrite，帮助利用 Elasticsearch 构建事件活动的发现、搜索及推荐。

目录: 第1章 搜索的相关性问题

1.1 我们的目标：掌握相关性技术研发的技能

1.2 为什么搜索的相关性如此之难

1.2.1 什么是具备“相关性”的搜索结果

1.2.2 搜索：没有银弹

1.3 来自相关性研究的启示

1.3.1 信息检索

1.3.2 能否利用信息检索解决相关问题

1.4 如何解决相关性

1.5 不只是技术：管理、协作与反馈

1.6 本章小结

第 2 章 搜索—幕后揭秘

2.1 搜索101

2.1.1 什么是搜索文档

2.1.2 对内容进行搜索

2.1.3 通过搜索来探索内容

2.1.4 获取进入搜索引擎的内容

2.2 搜索引擎的数据结构

2.2.1 倒排索引

2.2.2 倒排索引的其他内容

2.3 对内容进行索引：提取、充实、分析和索引

2.3.1 将内容提取为文档

2.3.2 充实文档以清理、强化与合并数据

2.3.3 执行分析

2.3.4 索引

2.4 文档的搜索和获取

- 2.4.1 布尔搜索：AND/OR/NOT
- 2.4.2 基于 Lucene搜索的布尔查询（MUST/MUST_NOT/SHOULD）
- 2.4.3 位置和短语匹配
- 2.4.4 助力用户浏览：过滤、切面和聚合
- 2.4.5 排序、结果排名，以及相关性
- 2.5 本章小结
- 第3章 调试我们的第一个相关性问题
- 3.1 Solr和Elasticsearch的应用：基于Elasticsearch的例子
- 3.2 最了不起的数据集：TMDB
- 3.3 用Python语言编写的例子
- 3.4 第一个搜索应用
- 3.4.1 针对 TMDB Elasticsearch索引的第一次搜索
- 3.5 调试查询匹配
- 3.5.1 检查底层查询策略
- 3.5.2 剖析查询解析
- 3.5.3 调试分析，解决匹配问题
- 3.5.4 比较查询条件和倒排索引
- 3.5.5 通过修改分析器来修正我们的匹配
- 3.6 调试排名
- 3.6.1 利用 Lucene的解释功能来剖析相关性评价
- 3.6.2 向量空间模型、相关性解释信息和我们
- 3.6.3 向量空间模型在实践中的注意事项
- 3.6.4 通过对匹配的评价来度量相关性
- 3.6.5 用 $TF \times IDF$ 计算权重
- 3.6.6 谎言、该死的谎言和相似度
- 3.6.7 决定搜索词重要性的因素
- 3.6.8 解决 Space Jam和 alien的排名问题
- 3.7 问题解决了？工作永远做不完！
- 3.8 本章小结
- 第4章 驾驭token
- 4.1 将token作为文档特征
- 4.1.1 匹配的流程
- 4.1.2 token，不只是单词
- 4.2 控制查准率和查全率
- 4.2.1 查准率和查全率的例子
- 4.2.2 查准率或查全率的分析
- 4.2.3 一味提高查全率
- 4.3 查准率和查全率—让鱼和熊掌兼得
- 4.3.1 评价单一字段中特征的强度
- 4.3.2 超越 $TF \times IDF$ 的评价：多搜索词与多字段
- 4.4 分析策略
- 4.4.1 处理分隔符
- 4.4.2 捕获同义词的语义
- 4.4.3 在搜索中为专指性建模
- 4.4.4 利用同义词为专指性建模
- 4.4.5 利用路径为专指性建模
- 4.4.6 对整个世界分词
- 4.4.7 对整数分词
- 4.4.8 对地理数据分词
- 4.4.9 对歌曲分词
- 4.5 本章小结
- 第5章 多字段搜索基础
- 5.1 信号及信号建模
- 5.1.1 什么是信号

- 5.1.2 从源数据模型开始
- 5.1.3 实现信号
- 5.1.4 信号建模：为数据的相关性建模
- 5.2 TMDb—搜索，人类最后的边疆
 - 5.2.1 违反基本法则
 - 5.2.2 让嵌套文档扁平化
- 5.3 在以字段为中心的搜索中给信号建模
 - 5.3.1 从 best_elds 开始
 - 5.3.2 控制搜索结果中的字段偏好
 - 5.3.3 可以使用信号更精准的 best_elds 吗
 - 5.3.4 让失败者分享荣耀：为 best_elds 校准
 - 5.3.5 利用 most_elds 统计多个信号
 - 5.3.6 在 most_elds 中缩放信号
 - 5.3.7 什么时候其他匹配才无关紧要
 - 5.3.8 有关 most_elds 的结论是什么
- 5.4 本章小结
- 第6章 以词为中心的搜索
 - 6.1 什么是词为中心的搜索
 - 6.2 我们为什么需要以词为中心的搜索
 - 6.2.1 猎寻“白化象”
 - 6.2.2 在“星际迷航”的例子中寻找白化象问题
 - 6.2.3 避免信号冲突
 - 6.2.4 理解信号冲突的机理
 - 6.3 完成第一个以词为中心的搜索
 - 6.3.1 使用以词为中心的排名函数
 - 6.3.2 运行以词为中心的查询解析器（深入底层）
 - 6.3.3 理解字段同步
 - 6.3.4 字段同步和信号建模
 - 6.3.5 查询解析器和信号冲突
 - 6.3.6 对以词为中心的搜索进行调优
 - 6.4 在以词为中心的搜索中解决信号冲突
 - 6.4.1 将字段合并成自定义全字段
 - 6.4.2 利用 cross_elds 解决信号冲突
 - 6.5 结合以字段为中心和以词为中心的策略：鱼与熊掌兼得
 - 6.5.1 将“相似字段”分到一组
 - 6.5.2 理解相似字段的局限
 - 6.5.3 将贪婪的简单搜索和保守的放大器结合起来
 - 6.5.4 以词为中心与以字段为中心，查准率与查全率
 - 6.5.5 考虑过滤、放大，以及重新排名
 - 6.6 本章小结
- 第7章 调整相关性函数
 - 7.1 何谓评价调整
 - 7.2 放大：通过突出结果来实现调整
 - 7.2.1 放大：最后的边疆
 - 7.2.2 放大时—选择加法运算还是乘法运算，布尔查询还是函数查询？
 - 7.2.3 选择第一扇门：利用布尔查询进行加法放大
 - 7.2.4 选择第二扇门：利用数学运算进行排名的函数查询
 - 7.2.5 函数查询实践：简单的乘法放大
 - 7.2.6 放大处理的基础：信号，处处是信号
 - 7.3 过滤：通过排除的方法对结果进行调整
 - 7.4 满足业务需求的评价调整策略
 - 7.4.1 搜索所有影片
 - 7.4.2 对放大信号进行建模
 - 7.4.3 构造排名函数：增加具有较高价值的层级

- 7.4.4 利用函数查询对具有较高价值的层级进行评价
- 7.4.5 忽略 $TF \times IDF$
- 7.4.6 捕捉综合质量指标
- 7.4.7 达成用户的时效性目标
- 7.4.8 结合函数查询
- 7.4.9 把一切联系起来
- 7.5 本章小结
- 第8章 提供相关性反馈
 - 8.1 搜索框中的相关性反馈
 - 8.1.1 利用“即输即搜”提供即时结果
 - 8.1.2 利用“搜索补全”帮助用户找到最佳查询
 - 8.1.3 利用搜索建议来修正输入和拼写错误
 - 8.2 浏览期间的相关性反馈
 - 8.2.1 构建基于切面的浏览
 - 8.2.2 提供面包线导航
 - 8.2.3 选择其他的结果排序方式
 - 8.3 搜索结果清单中的相关性反馈
 - 8.3.1 什么信息应该出现在搜索结果中
 - 8.3.2 通过文本片段与高亮提供相关性反馈
 - 8.3.3 对相似文档分组
 - 8.3.4 在用户搜不到结果时给予帮助
 - 8.4 本章小结
- 第9章 设计以相关性为核心的搜索应用
 - 9.1 Yowl! 一个绝佳的新起点
 - 9.2 信息和需求的收集
 - 9.2.1 理解用户及其信息需求
 - 9.2.2 理解业务需求
 - 9.2.3 找出必要及可用的信息
 - 9.3 搜索应用的设计
 - 9.3.1 将用户体验可视化
 - 9.3.2 定义字段和模型的信号
 - 9.3.3 信号的组合与平衡
 - 9.4 部署、监控和改进
 - 9.4.1 监控
 - 9.4.2 找出问题并解决它们
 - 9.5 知道什么是恰到好处
 - 9.6 本章小结
- 第10章 以相关性为核心的企业
 - 10.1 反馈：以相关性为核心的企业所依赖的基石
 - 10.2 为什么以用户为中心的文化比数据驱动的文化更重要
 - 10.3 无视相关性的天马行空
 - 10.4 相关性反馈的觉醒：领域专家和专业用户
 - 10.5 相关性反馈的成长：内容管理
 - 10.5.1 内容管理员的角色
 - 10.5.2 与内容管理员缺乏交流的风险
 - 10.6 让相关性更加流畅：工程师/内容管理员的结对
 - 10.7 让相关性加速：测试驱动的相关性
 - 10.7.1 理解测试驱动的相关性
 - 10.7.2 使用带用户行为数据的测试驱动相关性
 - 10.8 超越测试驱动的相关性：学习排序
 - 10.9 本章小结
- 第11章 语义和个性化搜索
 - 11.1 基于用户概况的个性化搜索
 - 11.1.1 收集用户的概况信息

- 11.1.2 将概要信息与文档索引紧密关联
- 11.2 基于用户行为的个性化搜索
 - 11.2.1 引入协同过滤
 - 11.2.2 使用共现计数的基本协同过滤算法
 - 11.2.3 将用户行为信息与文档索引紧密关联
- 11.3 构建概念性搜索的基本方法
 - 11.3.1 构建概念性信号
 - 11.3.2 利用同义词对内容进行扩充
- 11.4 利用机器学习来构建概念性搜索
 - 11.4.1 概念性搜索中短语的重要性
- 11.5 连接个性化搜索与概念性搜索
- 11.6 推荐是一种广义的搜索
 - 11.6.1 用推荐代替搜索
- 11.7 祝愿大家有一个美好的相关性搜索之旅
- 11.8 本章小结
- 附录A 直接根据TMDB建立索引
- 附录B Solr读者指南
 - • • • • [\(收起\)](#)

[相关性搜索：利用Solr与Elasticsearch创建智能应用_下载链接1](#)

标签

搜索引擎

搜索

elasticsearch

计算机科学

计算机

软件开发

ElasticSearch

分布式

评论

深入浅出，鞭辟入里。Elasticsearch的书本就不多，更多是从技术层面拆解语法，或者从Elastic Stack的角度讲解日志搜集。本书则聚焦在搜索，从相关性的视角讲解智能搜索，并给出明确的应用场景，并基于此应用场景深入讲解了搜索的内核、原理和优化等。从道的角度，进行建模，让人对Elasticsearch的认识提高一个维度。建议和极客时间《Elasticsearch核心技术与实战》一同服用。道术结合，效果愈佳。

个人感觉比《从Lucene到Elasticsearch:全文检索实战》要好很多。《相关性搜索》更多的是讲搜索的原理，如何实现与用户需求更相关的搜索结果，而《从Lucene到Elasticsearch:全文检索实战》有很多内容是ElasticSearch官方文档API，原理也不够深入，道与术的区别。

不囿于具体的语法，而是从高一层讲述相关性在搜索引擎中的重要性、实现和调优，加深了对搜索引擎的理解。不过建议在读之前需要掌握基本的ES语法，可以先看《ElasticSearch实战》

写的还是很接地气的

非常不错的相关性搜索实践，深入浅出得带我们看到了一个企业该怎么实现一套优秀的搜索系统

深入浅出，收获很大

[相关性搜索：利用Solr与Elasticsearch创建智能应用 下载链接1](#)

书评

摘抄书中一段话作为评论吧：“这不是一本介绍ElasticSearch的书籍。（以ES作为示例

讲解)，在使用搜索引擎的时候，我们关心的是其与相关性有关的那些特性，而全然不会涉及其他特性或知识点，这些特性包括：内容分析、数据提取、特征缩放、和性能表现。有很多不错的关于solr和 e...

[相关性搜索：利用Solr与Elasticsearch创建智能应用_下载链接1](#)