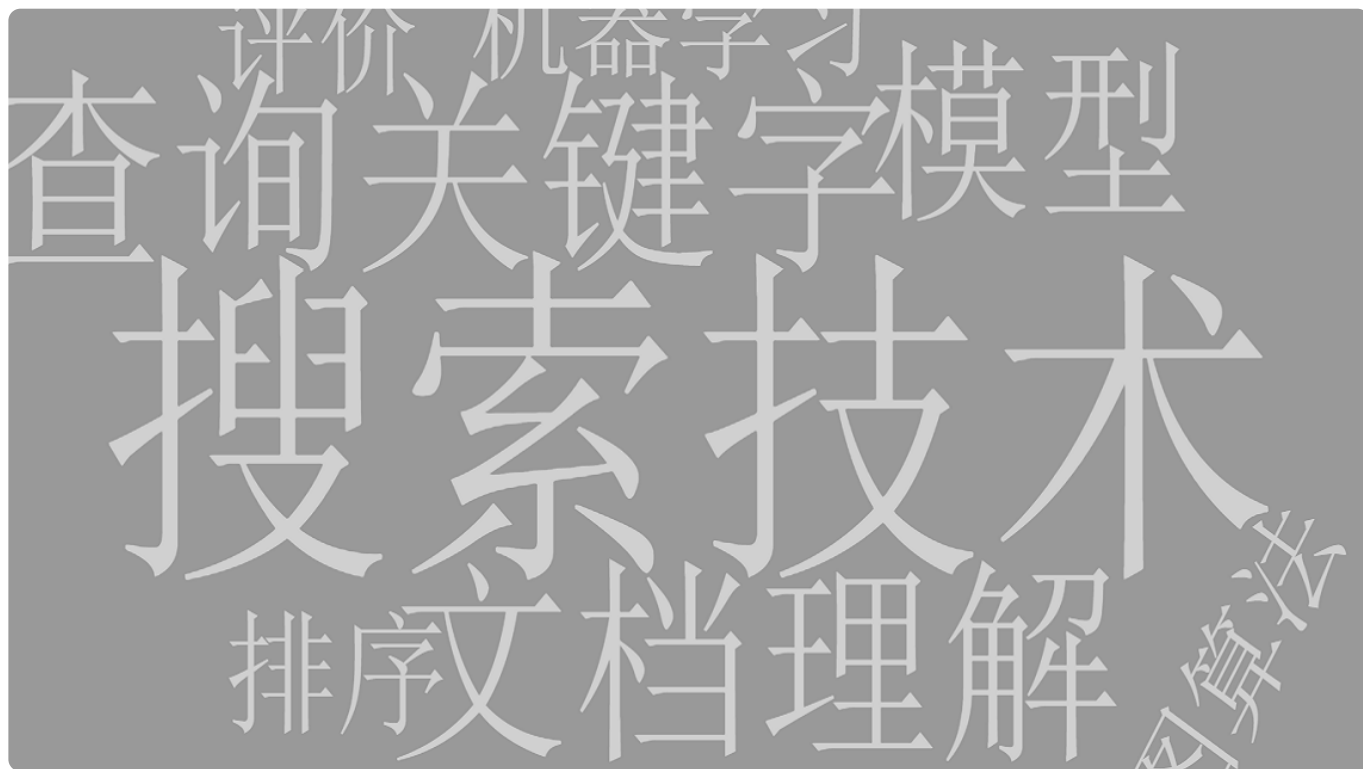


## 复盘 1 | 搜索核心技术模块

2018-02-25 洪亮劼

AI技术内参

[进入课程 >](#)



到目前为止，我们讲完了人工智能核心技术的第一个模块——**搜索**。我们从搜索的核心算法入手，进而讨论了搜索的两个关键组件，分别是查询关键字理解和文档理解，并落实到对搜索系统的评价，然后从宏观视角介绍了搜索框架的历史和发展，最后又从深度学习技术在搜索领域的应用角度，对分享做了一个延伸。

整个模块**共 27 期**，**9 大主题**，希望通过这些内容，能让你对搜索技术有一个系统的认识和理解，为自己进一步学习和提升打下基础。今天我们就来对这一模块的内容做一个复盘。

提示：点击知识卡跳转到你最想看的那篇文章，温故而知新。如不能正常跳转，请先将 App 更新到最新版本。

### 1. 现代搜索架构剖析

从 20 世纪 50 年代有信息检索系统开始，搜索系统大致经历了三个发展阶段。从最开始的“基于文本匹配的信息检索系统”到“基于机器学习的信息检索系统”，再到近几年受深



## 大型搜索框架剖析及历史发展

### 基于文本匹配的信息检索系统的特点：

- 文本匹配系统的基础是一个倒排索引；
- 依赖传统的检索方法，比如TF-IDF或BM25；
- 很难比较自然地处理多模数据；
- 其最大优势也是其最大劣势，即不依靠机器学习。

### 基于机器学习的信息检索系统的特点：

- 开始有了一整套的理论支持；
- 能够很容易地利用多模数据；
- 基于机器学习的信息检索系统也有其局限性。

[查看更多»](#)

## 多轮打分系统概述

---

### 多轮打分系统的特点：

- 每一轮都比上一轮使用的文档数目要少；
- 每一轮都比上一轮使用的特性数目更加复杂，模型也更加复杂。

### 多轮打分系统核心思路：

- 第一轮打分，常常被称作“顶部K”（Top-K）提取。
- 第二轮或以后轮数，即“重排”。

[查看更多»](#)

## 搜索索引及其相关技术概述

经典的索引结构由“字段”（Field）和对应的列表组成。两种可以让索引更加高效的技术：压缩和“略过”（Skipping）。

两种查询关键字处理的策略：“文档优先”（Document-at-a-Time）计算策略和“词优先”（Term-at-a-Time）计算策略。

[查看更多»](#)

### 2. 经典搜索核心算法

## TF-IDF算法及其变种

---

TF，是指“单词频率”。

IDF，是指“逆文档频率”。

TF-IDF背后隐含的是基于向量空间模型的假设。

TF-IDF几个经典的变种：

- 用对数函数对TF进行变换；
- 对TF进行“标准化”；
- 对数函数对IDF进行变换；
- 对查询关键字向量、文档向量进行标准化。

[查看更多»](#)

## BM25算法及其变种

---

### BM25算法的三个主要组成部分：

- 单词和目标文档的相关性；
- 单词和查询关键词的相关性；
- 单词的权重部分。

### BM25的代表性变种：

- BM25F，多个“域”文档上的计算；
- 把BM25和其他文档信息结合起来。

[查看更多»](#)

## 语言模型及其变种

一个语言模型就是一个针对词汇表的概率分布。

查询关键字似然检索模型：

- 确定语言模型的形态；
- 利用最大似然估计算法进行参数估计；
- 平滑策略。

语言模型的两个变种方向：

- 不同类型的平滑策略；
- 在语言模型本身的的定义上做文章。

[查看更多»](#)

### 3. 基于机器学习的排序算法

问题设置：把一个排序问题转换成一个机器学习的问题设置，特别是监督学习的设置。

## 单点法排序学习 (Pointwise)

每一个训练样本都仅仅是某一个查询关键字和某一个文档的配对。

如何评估？精度、召回、F1值、NDCG。

[查看更多»](#)



## 配对法排序学习 (Pairwise)

配对法的基本思路是对样本进行两两比较。

热门的配对法排序算法：RankSVM、GBDT和Rank-Net。

[查看更多»](#)



## 列表法排序学习 (Listwise)

列表法排序学习有两种基本思路。第一种，就是直接针对NDCG这样的指标进行优化。目的简单明了，用什么做衡量标准，就优化什么目标。第二种，则是根据一个已经知道的最优排序，尝试重建这个顺序，然后来衡量这中间的差异。

[查看更多»](#)

### 4. 基于机器学习的高级排序算法



## RankSVM算法

RankSVM算法的核心思想是应用支持向量机到序列数据中，试图对数据间的顺序直接进行建模。

[查看更多»](#)

## GBDT算法

GBDT, Gradient Boosted Decision Tree, 梯度增强决策树。

梯度增强首先还是增强算法的一个扩展,也是希望能用一系列的弱学习器来达到一个强学习器的效果,从而逼近目标变量的值,也就是我们常说的标签值。

梯度增强决策树就是利用决策树,这种最基本的学习器来当作弱学习器,去拟合梯度增强过程中的梯度。

[查看更多»](#)



## LambdaMART算法

LambdaMART“三步曲”:

- RankNet
- LambdaRank
- LambdaMART

[查看更多»](#)



## 查询关键字分类

---

查询关键字从大类上分为信息意图、交易意图以及导航意图三类。

把查询关键字进行分类是对用户行为进行建模的必要步骤。

[查看更多»](#)

## 查询关键字解析

---

### 查询关键字分割技术：

- N元语法
- 互信息的方法
- 条件随机场

### 查询关键字标注方法：

- PRF方法
- 条件随机场

[查看更多»](#)

## 查询关键字扩展

### 查询关键字扩展的两个思路：

- 根据查询关键字和查询结果之间自然结合产生的同义效果
- 从海量的文本信息中分析出词语之间的相关度

[查看更多»](#)

## 6. 文档理解



### 文档分类

文档分类的主要类型有二元分类、多类分类和层次分类。  
如何评估？精度、召回、F1值、NDCG。

[查看更多»](#)

## 文档聚类

文档聚类可以分为扁平聚类和层次聚类，也可以分为硬聚类和软聚类。

最基础的文档“扁平聚类”方法当属“K均值算法”。

[查看更多»](#)



## 多模文档分类

多模数据，其实就是说数据有多种模式的表达途径。

多模数据建模的核心思路就是数据表征。

[查看更多»](#)

### 7. 经典图算法

## PageRank算法

---

如何定义一个页面的PageRank? 当前页面I的PageRank值, 是I的所有输入链接PageRank值的加权和。

[查看更多»](#)



## HITS算法

---

一组概念: “权威”结点和“枢纽”结点。

HITS算法的好处是为用户提供了一种全新的视角, 对于同一个查询关键字, HITS提供的权威排序和枢纽排序能够帮助用户理解自己的需求。

[查看更多»](#)

## 社区检测算法之“模块最大化”

社区检测算法的核心就是要根据给定的一组结点和它们之间的关系，在无监督的情况下找到这些社区，并分配哪些结点属于哪个社区。

我们希望社区里结点之间的联系紧密，在模块化目标函数里，就表达为两个结点的连接数目减去这两个结点之间的“期望连接数”。模块化最大化就是指，对于同一个社区中的所有结点，我们希望这个差值的和最大化。

[查看更多»](#)

### 8. 基于深度学习的搜索算法



## 深度结构化语义模型

---

深度结构化语义模型，是利用深度学习技术对搜索算法进行改进的一个经典尝试。

利用深度学习技术来进行搜索建模，可以从三个方面入手：查询关键字的表达、文档的表达和匹配函数。深度学习的主要应用，就是成为查询关键字和文档表达的提取器。

[查看更多»](#)

## 卷积结构下的隐含语义模型

这个模型是利用卷积神经网络技术来表征查询关键字和文档。

整个模型就是希望先从原始的文字信息中，利用保留顺序的一个移动窗口提取最基本的特征；然后利用卷积神经网络的标配，卷积层加池化层，来提取空间位置信息；最后利用一个全部的展开层来学习下一步的系数。

[查看更多»](#)



## 局部和分布表征下的搜索模型

这是一个结合了学习完全匹配的局部表征和模糊匹配的分布表征的统一的搜索模型。

从整个模型来看，局部表征和分布表征的主要区别在于如何处理查询关键字和文档的匹配信息。

[查看更多»](#)



## 搜索系统评测基础指标

线下评测

基于二元相关度的评测指标：精度、召回、F值

[查看更多»](#)



## 搜索系统评测高级指标

DCG, Discounted Cumulative Gain, 折扣化的累积获得。

nDCG, Normalized Discounted Cumulative Gain, 把DCG加以“归一化”的指标。

[查看更多»](#)

## 如何评测搜索系统的在线表现？

在线可控实验是建立因果联系的重要工具，也可以说是唯一完全可靠的工具，其基础是统计的假设检验。利用因果推论对实验结果进行分析，是越来越受关注的机器学习前沿知识。

[查看更多»](#)

### 积跬步以至千里

最后，恭喜你在这个模块中已经阅读了70047 字，听了220 分钟的音频，这是一个不小的成就。在人工智能领域的千里之行，我们已经迈出了扎实的第一步。

恭喜你！获得一张内部通关卡

## 你已学习

27 期 | 70047 字 | 220 分钟

## 8个关卡

第一关：搜索



第二关：推荐系统



第三关：广告系统



第四关：自然语言处理及文本处理



第五关：计算机视觉



第六关：人工智能国际顶级会议



第七关：数据科学家养成



第八关：数据科学团队养成



AI技术内幕

感谢你在专栏里的每一个留言，给了我很多思考和启发。期待能够听到你更多的声音，我们一起交流讨论。


# AI 技术内参

你的360度人工智能信息助理

洪亮劼

Etsy 数据科学主管  
前雅虎研究院资深科学家



新版升级：点击「 请朋友读」，10位好友免费读，邀请订阅更有**现金**奖励。

© 版权归极客邦科技所有，未经许可不得传播售卖。页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 062 | WSDM 2018论文精读：深度学习模型中如何使用上下文信息？

下一篇 063 | 简单推荐模型之一：基于流行度的推荐模型

## 精选留言 (2)

 写留言



学无止境

2018-03-08

请问视频垂直领域中文查询关键字解析通常怎么做？

展开 ∨

 1



白杨

2018-05-17

文档优先的策略，我想到一个场景是，时间优先的新闻文档。

有一个实际的问题是，现在的开源搜索引擎比如sphinx或lucene底层好像没有开放这样的接口出来，也可能是开放了我没查到，老师能解答一下吗？

展开 ∨





下载APP

