信息传递：算法-引擎-AI

摘要:

本文讨论了信息传递中的搜索算法及其发展历程。搜索算法在信息传递中起着关键作用，可以理解为在解空间中寻找确定最优解的过程。从原始宗教和科学的求知精神到现代搜索引擎的出现，从最早的FTP 服务到现代的搜索巨头 Google 等，信息传递过程中不断的技术创新、盈利模式探索和服务完善，逐渐形成了技术、盈利和服务的三角闭环。最近，AI 的发展使得搜索引擎的体验有了新的里程碑，AI 模型如 ChatGPT 直接生成内容，提供更高效的信息传递方式，但仍需关注数据驱动 AI 的可解释性问题。

关键词:

信息传递，搜索算法，搜索引擎，盈利模式，AI 模型，可解释性

上一篇我们讨论了信息过滤的过程，研究了推荐算法的来龙去脉，其本质是在 B 端对 C 端的不确定认知下如何寻找信息流的传播途径——信息过滤。如果 B 端对 C 端有着明确的认知，那么在面对 C 端的需求下是有确定结果的。因此，搜索算法应运而生，其本质是寻找在 B 端对 C 端的确定认知下信息流的传播途径——信息传递。

首先，我们需要理解三个名词——搜索、推荐、思考。从解答问题的方面理解：

（1）**搜索**是在一定解空间中寻找确定最优解的过程；

（2）**推荐**是在一定解空间中寻找近似最优解的过程；

（3）**思考**是在未知解空间中证明解存在性的过程。

以上三种过程无疑不是因果认知世界的三大途径，对于信息流的传播途径也是如此。我们通过对于未知信息的思考，形成自己的理解，再去验证我们的理论是否正确；我们面对复杂信息流，如果不了解内部运作模式，那么会去寻找与自己认知、信仰匹配的信息；如果了解，那么会去寻找这种模式下的正确答案。（3）回答的是有没有的问题，是一种存在性证明；（1）（2）回答的是怎么办的问题，是一种思维性证明。OK，明白之后，让我们带着这种世界观，一起去寻找搜索发展的方法论。

最开始的搜索算法可以理解为原始宗教、原始科学。我们人类对于理解不了的事物本能有一种求知问底的态度，坚信一定是有一种因果力量促使事物发生，于是有根据的提出基本教义和基本原理，去寻找、构建这个广阔的未知世界，与现实碰撞匹配。进入信息社会，我们依然采用同样的逻辑，引入了图论中树的方法，在探索过程中，一旦发现原来的选择不符合要求，就回溯至父节点重新寻找，这就是深度优先搜索；在探索过程中，优先寻找相邻节点，知道全部寻找完毕，这就是广度优先搜索，两种方法的实质就是遍历穷举、无监督搜索。如果我们可以加上一些规则来有监督搜索，是不是可以提高效率，于是有了 A^*算法，制定一些启发规则和一个代价函数，估算起始节点经过该节点到达目标节点的代价，节点扩展时总是寻找具有最小代价的节点。

综上，我们可以认为信息传递的过程就是一个匹配更新的过程，

信息过滤也是如此。

- Search and Recommendation are two sides of the same coin
  Search -> *Information Pull* with *explicit* info request (query)
  Recommendation -> *Information Push* with *implicit* info request (user profile, contexts)
- Technically, they can be unified under the same matching view
  - Though they are studied by different communities: SIGIR vs. RecSys
- Deep learning-based matching methods
  - Representation learning-focused
  - Matching function learning-focused
- Matching is a generic problem for a wide range of applications
  E.g., online advertising, question answering, image annotation, drug design

图 1 Deep Learning for Matching in Search and Recommendation Summary

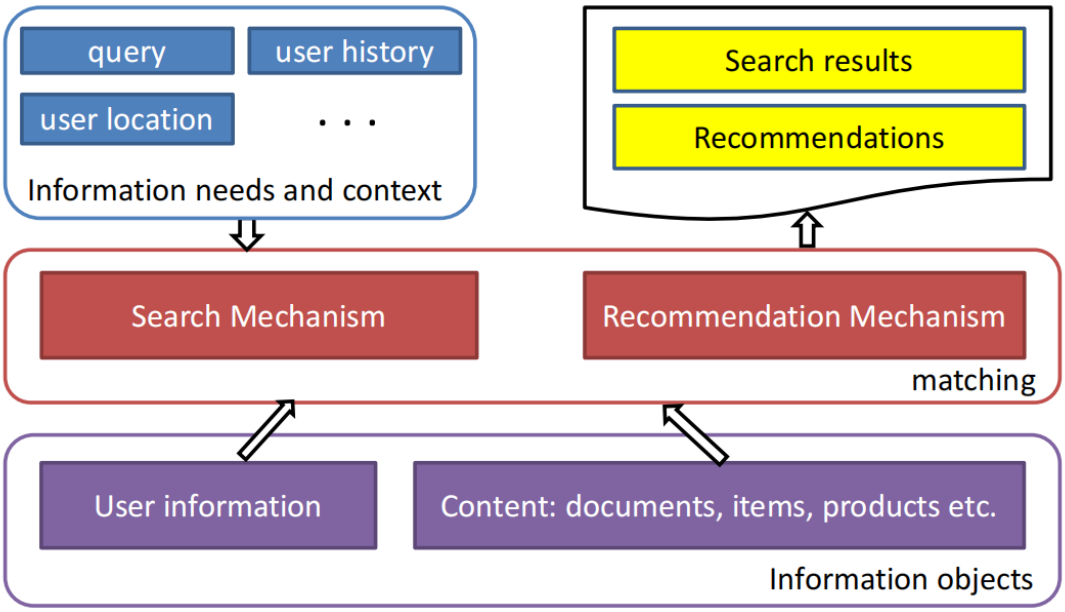

图 2 Unified View on Matching in Search and Recommendation (Hector et al, CACM'11)

搜索算法真正深入千家万户归功于搜索引擎的出现。现代第一个搜索引擎 Archie(1990)，服务于 FTP，通过正则表达式来匹配用户查询与文件名称来实现查询。随后，World Wide Web Wander（1993）世界上第一个网络爬虫出现了，自动代理收集 URL，同年，ALIWED（1993）检索标题标签，但文章内容无法索引。

InfoSeek（1994）正式推出搜索服务，李彦宏是核心工程师之一，同年，杨致远与 David Filo 创立 Yahoo！（1994），稍后，第一个可以索引全文内容的搜索引擎 Web Crawler（1994）推出，而后 Lycos（1994）创立，根据搜索频率排序。接着，Excite（1995）搜索引擎正式上线，但存在信息丢失、太多或者无关的问题，稍后，Alta Vista（1995）诞生，并在搜索引擎做了很多开创工作——第一个允许用户使用自然语言搜索的搜索引擎，第一个尝试使用自己的数据创建完整网页索引的搜索引擎，第一个扩展了布尔操作符在搜索中的使用，允许搜索者限制从一个域得到结果的数量，第一个允许多语言搜索，第一个允许人们在搜索文本内容的同时搜索图像、视频和音频的网站。

1998 年，GoTo(后改称 Overture)正式开始竞价排名业务，成为 PPC 点击付费广告形式的鼻祖。随后，Direct Hit（1998）创办，主要采用用户点击率来列出搜索结果排名，被作弊者利用。下半年，Google（1998）公司创立。2001 年，百度成立。2002 年，Google AdWords 推出 PPC 形式，随后推出 Adsense 内容广告系统。

2009 年，Google 以创始人之一的 Larry Page 命名的 Page Rank（谷歌 PR 值）正式被大众知晓。2010 年，Google 在旗下 Chrome 浏览器中推出 Google Instant，即用户在输入关键词并未按下 Search 键时即可看到搜索结果。之后，各大 APP 软件均有搜索模块。

统览这个搜索引擎的发展历程，我们可以发现一条清晰的路线——从技术到盈利再到服务的三角闭环，并伴有资本兼并和重组统

一。在这个过程中 Google 公司的贡献和地位尤为突出，这也来源于它从意义和意图、相关性、品质、用户体验、个性化对网页进行排名的策略。

时间来到 2022 年底，ChatGPT 的出现打开了搜索引擎的缺口，搜索之后看到依旧是各类信息的集合，还需要大脑思考整理，而 AI 通过 Transformer、Diffusion 直接生成我们所需要的内容，是信息传递过程中新的里程碑，不过基于数据驱动的 AI 还需要 knowledge-driven 来弥补可解释性的问题。

By building learning systems, we don't have to write these rules anymore. Increasingly, we're discovering that if we can learn things rather than writing code, we can scale these things much better.

——*John Giannadrea,Google,2015*

参考文献：

[1] Jun Xu, Xiangnan He, and Hang Li. 2018. Deep Learning for Matching in Search and Recommendation. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18). Association for Computing Machinery, New York, NY, USA, 1365–1368.

[2] https://baike.baidu.com/item/搜索算法/2988274

[3]https://zhuanlan.zhihu.com/p/208846943

[4]https://baijiahao.baidu.com/s?id=1626167337781038888

[5]http://www.hepou.com/site/excite.html

[6]https://www.semrush.com/blog/google-search-algorithm/?kw=&cmp=AA_SRCH_DSA_Blog_EN&label=dsa_pagefeed&Network=g&Device=c&utm_content=665538834698&kwid=dsa-2147915049507&cmpid=18361936995&agpid=154786738681&BU=Core&extid=91684392298&adpos=&gclid=CjwKCAjwq4imBhBQEiwA9Nx1Bp2wUYi-Y3orTuwrgCaVSXZUDrIS90NJzHWhb_KLYRGTrLNoIbcgthoCDR0QAvD_BwE

[7]https://www.google.com/search/howsearchworks/how-search-works/ranking-results/

[8]https://www.wired.com/2016/02/ai-is-changing-the-technology-behind-google-searches/