搜索引擎入门知识

1 什么是搜索引擎

搜索引擎是用于查找和排名与用户搜索匹配的 Web 内容的工具

每个搜索引擎都包含两个主要部分：

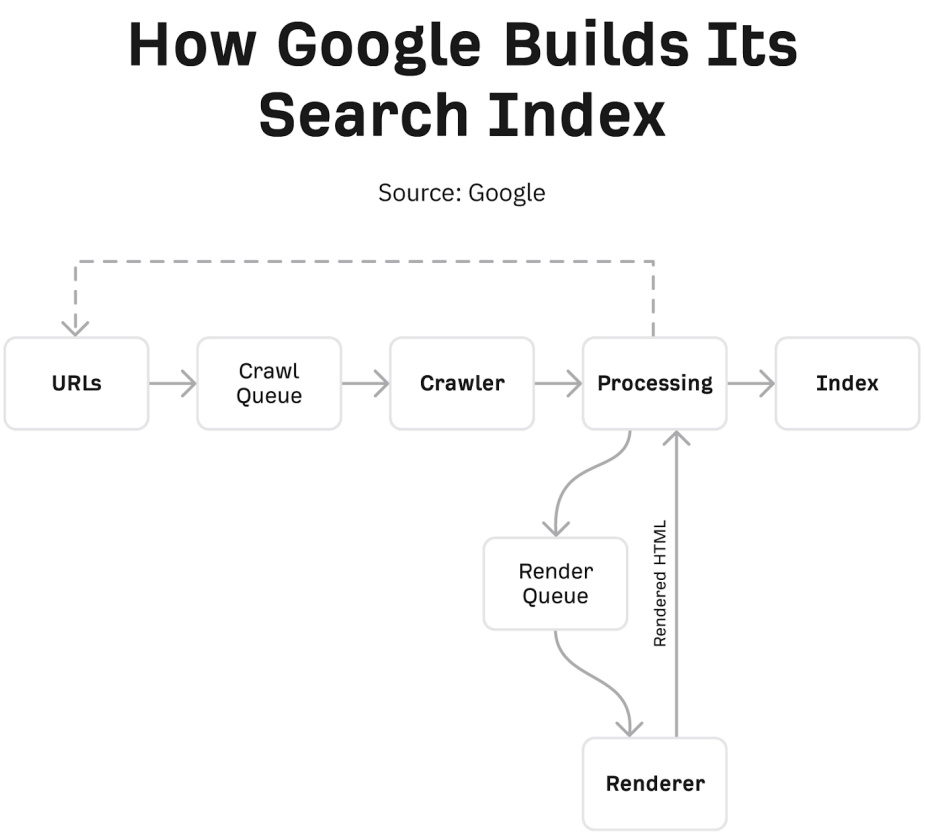
1. **搜索索引**：有关网页信息的数字图书馆
2. **搜索算法**：匹配搜索并进行排名的计算机程序

热门搜索引擎有 Google、Bing、以及 DuckDuckGo

每个搜索引擎都旨在为用户提供最佳、最相关的结果 至少从理论上讲，这就是他们获取或维持市场份额的方式

搜索结果具有两种类型，自然排名结果和付费排名结果
每次有人点击付费搜索结果时，广告客户都会向搜索引擎付费（PPC：*pay per click* 按点击付费广告）

2 搜索索引如何建立



1. **URL**

   Google 通过各种方法发现 URL

   常见有三种方法：**外链**（Google 已经有一个包含数万亿个网页的索引库 如果某人在这些页面中添加了一个链接指向了自己的网站，那么 Google 可以从那些页面中找到链接） **网站地图**（Sitemap 协议列出了你网站上的所有重要页面） **URL 提交**（Google 还允许通过 Google Search Console 提交单个 URL）

2. **Crawling**

Googlebot 等爬虫抓取（访问并下载）已知的URL的页面

注意 Google 并不总是按照发现页面的顺序对其进行抓取
Google 会根据以下因素对要抓取的 URL 进行排序，其中包括：

- URL 的 PageRank[见下]
- URL 多久更改一次
- URL 是否是新的

3. **Processing**

   Google 会在处理过程中从抓取的页面中提取关键信息 Google 以外的人都不知道有关此过程的细节

   Google 必须渲染页面以理解和提取信息，它将会运行页面的代码以了解外观对用户的影响

4. **Indexing**

   索引是将抓取页面中的信息添加到叫做搜索索引的大型数据库中
   本质上，这是一个由数万亿个网页组成的数字图书馆，Google 的搜索结果都来自于此
   当你在搜索引擎中搜索时，你并不是直接匹配互联网上的结果，而是在搜索搜索引中进行匹配的

3 搜索引擎如何对网页排名

每个搜索引擎都有用于对网页进行排名的独特算法

Google 有 200 多个排名因素 没有人知道所有的这些排名因素，但是关键因素却是已知的：

- **外链 Backlinks PageRank** 是一种根据指向外链数量和质量来判断网页价值的公式：

  其想法来自科学家衡量科学论文“重要性”的方式，即通过查看引用它们的其他科学论文的数量
  Sergey 和 *Larry* 采纳了这个概念，并通过跟踪网页之间的引用（链接）将其应用于网络
  它是如此有效以至于它成为了 Google 搜索引擎的基础，到现在仍然如此

  > *We assume page A has pages T1...Tn which point to it (i.e., are citations). The parameter d is a damping factor which can be set between 0 and 1. We usually set d to 0.85. There are more details about d in the next section. Also C(A) is defined as the number of links going out of page A. The PageRank of a page A is given as follows:*
  >
  > **PR(A) = (1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))**
  >
  > *Note that the PageRanks form a probability distribution over web pages, so the sum of all web pages' PageRanks will be one.*

  d 类似于一个"阻尼系数"，它随机模拟用户在浏览网页时持续点击链接的概率，通常这会随着每次链接点击而减少

  要计算全网的 Pagerank ，首先应该知道至少一个页面的 Pagerank
  不过原始算法是

  > *PageRank or PR(A) can be calculated using a simple iterative algorithm and corresponds to the principal eigenvector of the normalized link matrix of the web.*

  并且这个分数是一个相对衡量标准而非绝对值

  需要指出的是， Google 已经放弃了公开的 Pagerank 指标：

  > *"As the Internet and our understanding of the Internet have grown in complexity, the Toolbar PageRank score has become less useful to users as a single isolated metric. Retiring the PageRank display from Toolbar helps avoid confusing users and webmasters about the significance of the metric."*

由于对于其他排名因素都不存在这样可见的量化标准，这使 PageRank 成为似乎唯一重要的因素
结果人们很快就开始买卖“高 PR”的链接 卖家最初建立这些有不少策略，之一是发表博客评论，即**垃圾链接**（例如“访问我的折扣药品网站……”）

于是 Google 推出了 **nofollow** 属性，允许网站管理员停止通过特定链接传输 PageRank：

> *From now on, when Google sees the attribute (rel=“nofollow”) on hyperlinks, those links won’t get any credit when we rank websites in our search results.*

现在几乎所有的 CMS（*content management system*）系统都默认 “nofollow” 博客评论链接

由于原始公式指出，PageRank 平均分配在网页上的导出链接
网站管理员很快开始有选择地将 “nofollow” 属性添加到他们认为不太重要的页面（如导出链接等），这使他们能够有效地“塑造”他们网站上 PageRank 的流动 例如想要提升一个网站的权力，他们只会从高 PR 页面链接到该页面，并且 "nofollow" 页面上其它所有链接，这样可以发送最大的 PR 到指定的页面

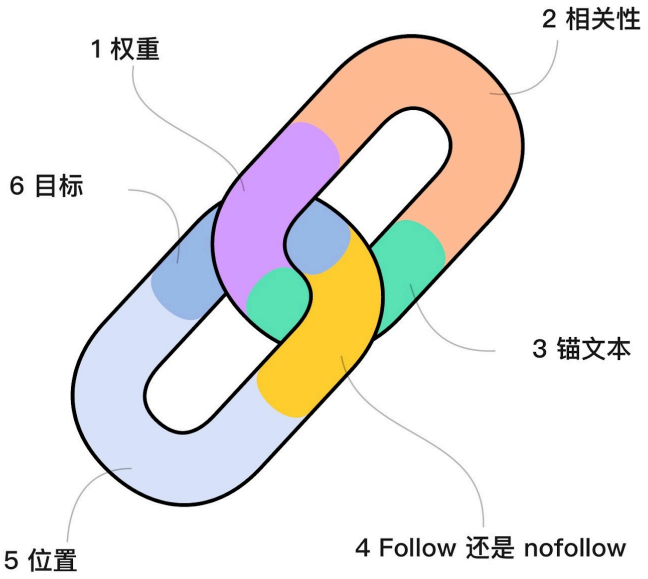
Google 做出了调整，使 PageRank 雕刻（*sculpting*）不再有效：

> *So what happens when you have a page with “ten PageRank points” and ten outgoing links, and five of those links are nofollowed? [...] Originally, the five links without nofollow would have flowed two points of PageRank each [...] More than a year ago, **Google changed how the PageRank flows so that the five links without nofollow would flow one point of PageRank each.***

我们不知道这是否仍然是 “nofollow” 数学的工作方式，但可以肯定向某些链接添加 “nofollow” 标签，对于将更多“链接权重”汇集到页面上的其它链接上是没有帮助的

并且其他因素可能也会影响给定链接传输的价值
2016年，Toolbar PageRank 被正式取消

综合判断一个良好的链接（数量和质量）通常具有几个关键的属性：



- **相关性 Relevance** 相关性不仅仅只有关键词匹配，Google 还使用交互数据来评估搜索结果是否与搜索词相关（如用户觉得页面对自己有帮助吗？）

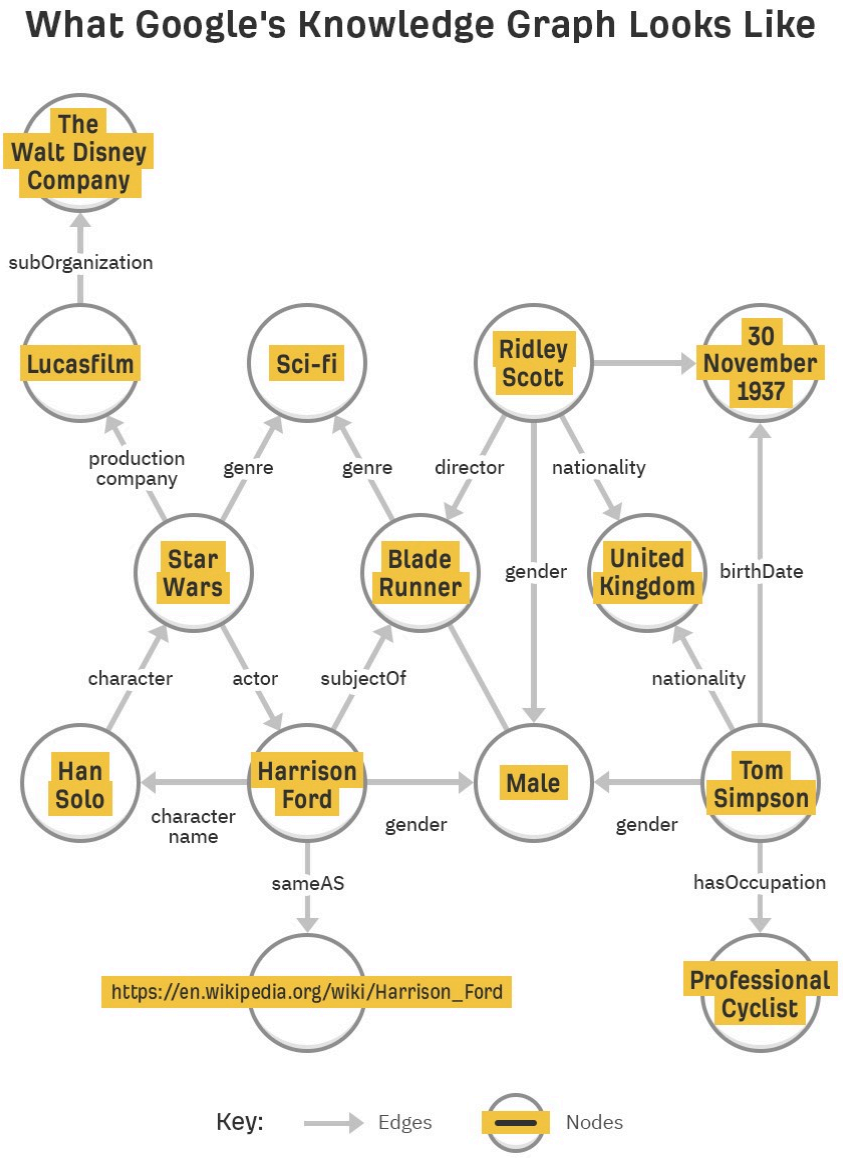
  例如 “苹果” 的所有排名靠前的结果都与技术公司有关，而不是水果 Google 从互动数据中知道大多数用户正在寻找前者而不是后者的信息

  不过互动数据远非 Google 做到这一点的唯一方法
  Google 投资了许多技术来帮助理解诸如人、地点、事物之类的实体之间的关系
  知识图谱是这些技术之一，本质上是一个巨大的实体之间关系的知识库

归功于“知识图谱”，Google 可以超越关键词匹配的限制

以 *“apple paper app”* 的第二个结果为例，该结果在页面上的任何地方都没有提及 “apple” 一词

Google 之所以知道这是一个相关的结果，部分是因为它在知识图谱中提到了与苹果密切相关的实体，例如 iPhone、iPad



- **新鲜度 Freshness** 对于 *“what’s new on amazon prime”* 这样的搜索，新鲜度很重要 因为搜索者想了解最近添加的电影和电视节目 而对于 *“how to tie a tie”* ，因为方法没有任何变化，因此搜索结果是昨天还是 1998 年都无关紧要
- **话题权威性 Authority** 搜索系统是否认为网站具有权威性也取决于搜索词 搜索系统可以将疾病控制中心（*cdc.gov*）的站点视为 *“CDC mosquito stop bites”* 搜索的权威站点，但可能不会将其视为对 *“restaurant recommendations”* 搜索的权威站点
- **页面速度 Page speed** 当然，Google 表示，只有你页面速度过慢时才需要考虑速度问题，它只需要足够快就不会对用户产生负面影响

  你可以在 PageSpeed Insights 中检查任何网页的速度和核心页面指标的信息
  核心页面指标由三个指标组成，用来评估网页的加载性能、交互性、以及视觉稳定性
- **移动端友好 Mobile-friendliness** 65% 的 Google 搜索发生在移动设备上
  Google Search Console 中的移动设备可用性报告可以检查任何网页的移动设备友好性问题

4 搜索引擎如何个性化结果

Google 利用个人信息定制搜索排名，常见有以下三种：

- **位置 Location** 如果你搜索 *“italian restaurant”* 之类的内容，则地图中的所有结果均为本地的餐厅 *“buy a house"* 类似，Google 会返回本地的列表而不是其他国家的销售列表页面
- **语言 Language** Google 知道向西班牙用户显示英语结果毫无意义

  不过 Google 在某种程度上依赖于网站站长去做分割
  如果你有多种语言的网页，除非你告诉 Google，否则 Google 可能不会意识到这种情况
  可以使用称为 *hreflang* 的 HTML 属性来执行此操作
- **搜索历史 Search history** 最明显的例子可能是，当下次运行相同的搜索时，Google 会将先前点击的搜索结果排名更高（尤其是短时间内多次访问）

author: *Laplace* ***Time: 2023.01.10***