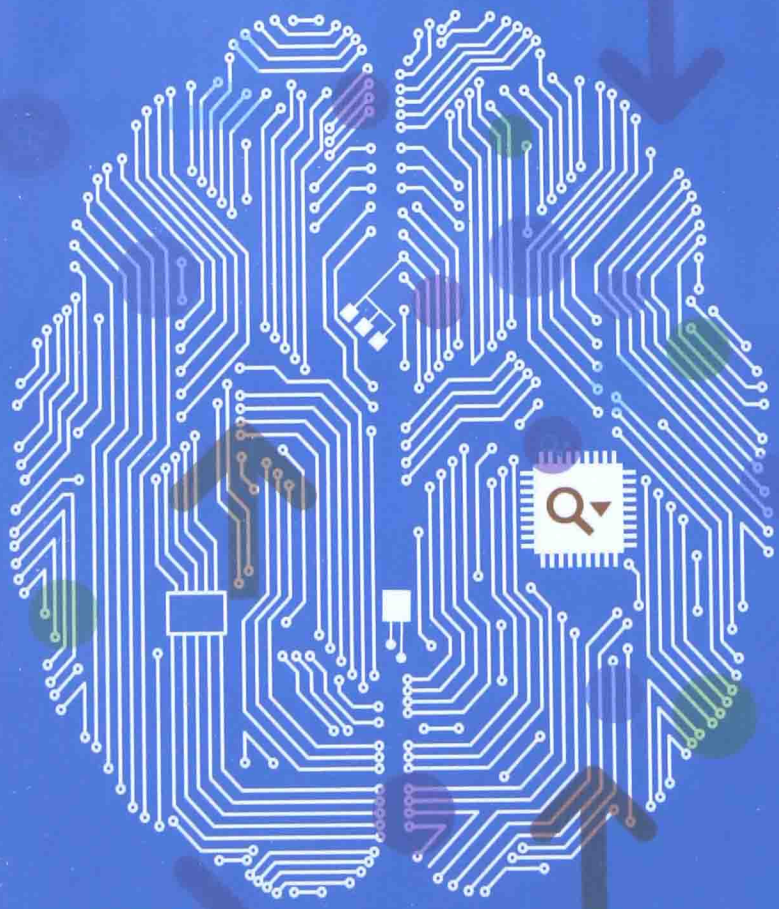
que

Google
Semantic Search

谷歌语义搜索

[英] David Amerland 著
程龚 译



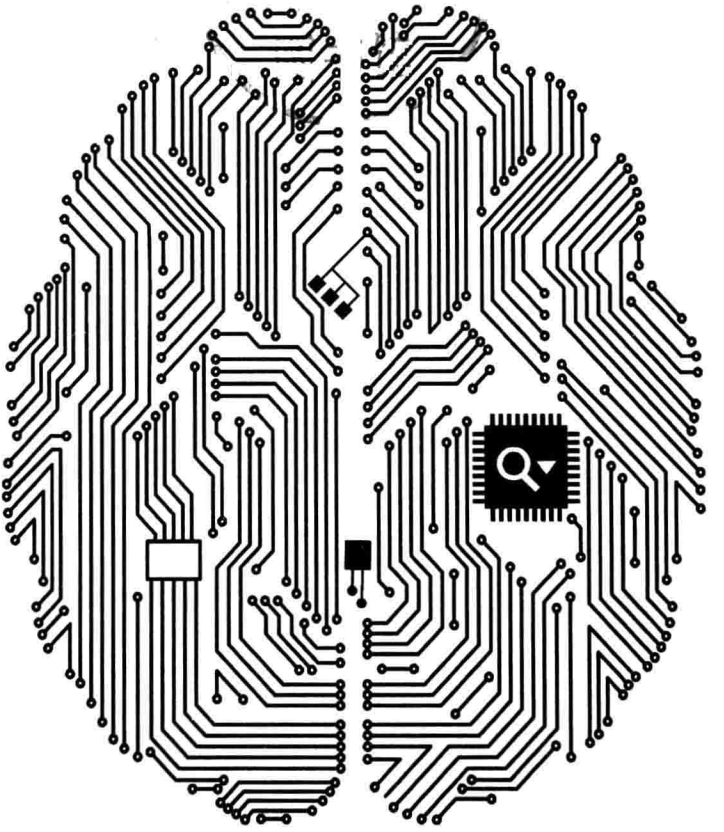
增加公司的站点流量
提升品牌影响力　升华线上形象

人民邮电出版社
POSTS & TELECOM PRESS

Google
Semantic Search

谷歌语义搜索

[英] David Amerland 著
程龚 译



人 民 邮 电 出 版 社
北 京

图书在版编目（CIP）数据

谷歌语义搜索 / （英）阿默兰德（Amerland, D.）著；程龚译. -- 北京：人民邮电出版社，2015.3
ISBN 978-7-115-37626-8

Ⅰ. ①谷… Ⅱ. ①阿… ②程… Ⅲ. ①语义结构－网络检索 Ⅳ. ①G354.4

中国版本图书馆CIP数据核字(2015)第013320号

◆ 著 [英] David Amerland
译 程 龚
责任编辑 傅道坤
责任印制 张佳莹 焦志炜

◆ 人民邮电出版社出版发行 北京市丰台区成寿寺路 11 号
邮编 100164 电子邮件 315@ptpress.com.cn
网址 http://www.ptpress.com.cn
大厂聚鑫印刷有限责任公司印刷

◆ 开本：700×1000 1/16
印张：13.75
字数：241 千字 2015 年 3 月第 1 版
印数：1－3 000 册 2015 年 3 月河北第 1 次印刷

著作权合同登记号 图字：01-2013-9201 号

定价：45.00 元

读者服务热线：(010)81055410 印装质量热线：(010)81055316
反盗版热线：(010)81055315

内 容 提 要

语义搜索是指搜索引擎的工作不再拘泥于用户所输入请求语句的字面本身，而是透过现象看本质，准确地捕捉到用户所输入语句后面的真正意图，并以此来进行搜索，从而更准确地向用户返回最符合其需求的搜索结果。

本书是最易读的语义搜索入门图书，共12章，涵盖了什么是语义搜索、什么是知识图谱、SEO的新发展、信任和作者排序、什么是TrustRank、内容如何成为营销、社交媒体营销和语义搜索、不再有谷歌的“第一页”、影响的传播和语义搜索、实体抽取和语义网、语义搜索的四个V、搜索如何变为无形等知识。本书除了剖析谷歌的内部工作模式和新专利之外，还讲解了社交网络对SEO性能的影响。

本书适合搜索领域从业人员、SEO从业人员以及网络营销人员阅读。

关于译者

程龚，现就职于南京大学计算机科学与技术系、计算机软件新技术国家重点实验室，担任助理研究员。2006 年毕业于东南大学计算机科学与技术专业，获取工学学士学位，2010 年毕业于东南大学计算机软件与理论专业，获取工学博士学位，师从瞿裕忠教授。他的研究兴趣包括：语义搜索、语义 Web、Web 科学和数据集成。近年来主持一项国家自然科学基金项目和一项江苏省基础研究计划项目，并参与过包括国家973计划课题在内的多个科研项目。他在International Semantic Web Conference（ISWC）等国际会议和 Journal of Web Semantics 等国际期刊上发表过多篇论文，并曾两次获得 ISWC 的最佳研究论文提名奖。他的主要业余爱好是旅行、摄影和篮球。

关于作者

早在那个只需要一本 80 页纸的目录就可以列全所有网站，搜索引擎优化（Search Engine Optimization，SEO）的技巧还停留在堆砌关键词和夹带不可见文本的时代，David Amerland 就与 Web 结识了。

从那个搜索引擎优化和社交媒体营销尚不繁荣的时代起，他就已经开始与大型跨国企业和个人企业家合作，帮助他们制定与其内在文化相符的优化和营销策略，以向他们的目标受众传递价值。

他为福布斯（Forbes）、英国惠普（HP UK）和今日社交媒体（Social Media Today）撰稿，也在自己的网站 HelpmySEO.com 上写博客。写作和上网冲浪之余，他也会花一些时间做一些关于“社交媒体是如何改变一切”的演讲。

献辞

和其他每本书一样，这本书也献给 N。她对我意味着一切。但我也想把这本书献给另一个 N，也是一位女性，有着尖尖的耳朵和一条尾巴。在深夜时分，当我奋笔疾书时，她的陪伴让世界更多了一些暖意。我以不同的方式向你们俩致以谢意。

致谢

没有任何一本书是单靠一个人就可以完成的，这本书也不例外。第 6 章中对内容质量、内容管理及其对语义搜索的影响的评论，承蒙 www.asmartsolutions.ca 的 CEO Sergey Adrianov 的指点，他勇敢地让我把这些写进书里。这章中使用的关于谷歌搜索及其与谷歌产品和服务之间关系的图表得到了 Frontiercoaching（www.frontiercoaching.com）的 Bob Barker 以及 Thomas Power 的授权，他们关于数字领域发展的某些想法经常让我激动得彻夜难眠，一直思考。

自从 Google+成为我生活的一部分以后，这还是我的第一本写在纸上的书。因此，我非常感谢所有那些与我分享思想火花，以及就此向我提供反馈或者补充他们自己观点的人们。在越来越多以致难以一一致谢的人们当中，我需要特别提及如下几位：Jeff Jockisch，在他的帮助下，我对身份和信任有了更深的认识；Gideon Rosenblatt，他的 Google+社区为关于作者身份对搜索中的排序的影响的一些有趣的观点提供了讨论场所；Mark Traphagen，他孜孜不倦的工作才使得每个人都没有偏离作者身份这一主题；Bill Slawski，是一位搜索引擎优化专家，他是我认识的最专注于谷歌专利及其意义的人；Aaron Bradley，他对语义搜索的理解堪称典范；J. C. Kendall，用实例证明了工业界可以更多遵循的那些搜索引擎优化伦理；Dan Petrovic，他频繁的搜索引擎优化实验帮助我论证了我个人的观点；NOD3X 的 Lee Smallwood，慷慨地对数据进行了可视化，使我得以将观点表述得更透彻；还有那些只是与我在网上结识，但他们的慷慨、帮助、智慧和细心让我深深折服的人们。在此向你们深表谢意。

要感谢杰出的高级策划编辑 Katherine Bull，她从不催促我，即便我知道有时候她已经非这样做不可了；组稿编辑 Amber Avines，她对这本稿件的初审是非常宝贵的；项目编辑 Andy Beaster，我敢肯定要是没有他，一些文件早就不见了；感谢 Pearson 的编辑团队，他们让我的文字更容易理解。最后，我要感谢搜索引擎优化的思想家和实践者 AJ Kohn，他使我在写作本书时将天马行空般的思考落在了实处。我要对他致以特别的谢意。

上述所有人的努力都为这本书创造了价值，如果书中还发现有一些错误，责任完全在我。

前　言

搜索正发生着变化。当然这不是一个新的话题。从某个角度来说，它从第一天起就在一条不断变化的轨迹上运动。然而如今，这条轨迹的弧线已经以更快的速度和更陡的角度来匹配 Web。

事实上，不能认为搜索与 Web 是各自独立的，没有了搜索的 Web 无法正常运转。这种共生关系带来了各种各样的问题，因为它成为了一种推拉效应的一部分，其中，Web 表示那些在其中积极工作的人们，他们想要将所有错误的东西都推出去，而搜索则想要把所有东西都拉进来。

当所有事物都进入 Web 之后，这场索引信息的争斗转变为将其正确分类的争斗。因为 Web 的增速是如此惊人，任何分类都必须是机器驱动的并具有可伸缩性（scalable），这只能以两种方式发生：A，有人的辅助；B，没有人的辅助。

我们从马尔可夫链（Markov chain）和布尔算法（Boolean algorithm）非常突然地切换到了不断变化的伦理领域和对做“正确”或“错误”的事的意愿上来。这里的假设是，一旦某个事物可以被解构并且其工作方式可以被理解，人们就可以试着借势（gaming it）来优化它的效率。

这正是在搜索上发生的事情。当搜索因部分借助人力而有助于理解那些被编入索引的数据应该被如何分类之后，它便基于有关借势的必要技术创造了一个完整的产业。

当搜索引擎用更复杂的方式来反击我们借势其算法在 Web 上获得更大曝光度

（visibility）的尝试之后，“我们”和“他们”之间的推拉效应就被强化了。搜索引擎更新的每个周期都会带来“阵痛”，并造成曝光度的损失，因此，这不得不通过寻找借势搜索的新方法来应对，如此往复。

语义搜索有着终结这一轮回的前提和前景。理解语义搜索的最好方式是将其比作一个探照灯，打探了 Web 上所有不同的数据结点，并跟随它们绘制了一幅画面来刻画它们是如何连接的，它们属于谁，谁创建了它们，他们还创建了其他什么，他们是谁，他们曾经是谁以及他们是干什么的。

语义搜索的最基本层面是将含义（meaning）用于 Web 上不同数据结点之间的连接，让我们得以对它们建立起前所未有的清晰理解。这是颠覆性的。Web 由数据组成，数据则被大量性（volume）、高速性（velocity）、多样性（variety）和真实性（veracity）这些概念所支配，一旦我们找到一种方式来完满地应对这四个概念，我们就已经解决了搜索问题。

目前我们还没有做到。语义搜索应用的新动态在这四个概念之间徘徊，它们中任何两者之间的平衡都远没有被解决，更不要提全部四个了。如果解决了如何索引每分钟都在生成的大体量数据这一问题，那么如何用一种满足时间需求的方式来对其分类就变得至关重要了。在分类和分优先级的速度问题（即速率）被解决后，内容的多样性就成为了一个问题。

在所有这三个方面以及搜索结果中的质量问题都最终得以解决之后，起源和信任（即真实性）问题就抬起了它丑陋的头。接下来，非常突然地，其他三个概念中的每一个都再次成为问题：你该如何来验证以如此惊人的速度涌入的数据、快速地评价它并成功地应对它的所有变种呢？

答案在于增量地解决。语义搜索不同于过去我们拥有过的任何技术，它可以缓存它所计算出的所有特征，因而它所照亮的数据结点在它离开之后不会再回到黑暗之中以等待再次照亮。这样它就在进行一种学习，并变得愈发聪明、愈发敏捷、愈少犯错误和愈可靠，也变得愈难被借势利用。

本书谈论的是语义搜索，即关于它是什么、它怎样运转以及你现在可以做些什么来从中获益。写这本书时我主要关注的是谷歌，有三个原因。首先，谷歌在搜索的语义索引方面有显著的进展。其次，Google+社交网络在帮助网站提高在线曝光度方面扮演了一个重要角色，如果低估了它，就会在搜索中丧失一个巨大的机遇。再次，谷歌是世界上最主要的搜索引擎，占据了 95%的全球移动搜索市场和超过 80%

的全球桌面搜索市场。如果不去关注它，就没有任何商业意义了。

语义搜索用很多方式将我们带回到了 Web 的那个黄金年代——就在线的工作而言，只要你有工作的热情、自信和精力，一切皆有可能。

我们又面临这样的景况，对此我很兴奋。我希望这本书成为你数字旅途的指南，但我更希望它成为你亟需的一块跳板，使你得以让自己的工作在世界上以数码或其他形式留下一丝痕迹。

David Amerland

2013 年，曼彻斯特

目　录

1

第1章
什么是语义搜索

搜索是我们在 Web 上浏览的途径。如果你的业务在搜索中不可见的话，它就很难被你的客户发现。搜索首先是一种营销，并正经历着一场巨大的变革。

在本章中，我们讨论谷歌搜索中出现的新元素、为什么会发生这种变化以及它将如何以你能想到的几乎每种方式来影响你的业务。本章提供一份清单，罗列了你为了能利用上即将到来的变化而需要去做的每件事情，并且本章的每一节帮助你理解你需要做些什么才能充分利用谷歌的语义搜索。

1.1　向语义搜索迁移

如今，当我在谷歌的搜索框中敲入一条搜索查询之后，我会异乎寻常地感觉到屏幕的另一侧有一种智能给了我答案。以前可不是这样的，而如今我所感觉到的这种智能是搜索技术领域从未取得过的、最具突破性的进展之一。当然，就像手机和3D 电视一样，其概念并不是全新的，并且差一点就没能成为现实。

一本关于语义搜索的书不可避免地会以“语义搜索究竟是什么”这样平淡无奇的问题作为开篇语。答案可能极具技术性并且复杂难懂，它可以包含数学甚至一些哲学概念（当它们适用于数学时），但本书并不打算从技术的角度来简单地满足你的好奇心。我在本书中给出的关于语义搜索的一些解释在某种程度上是有所局限的，但它们为帮助你更好地理解语义搜索提供了一切所需。

我是“知识就是力量”的坚定信徒——但仅当知识可以被理解的时候。因此，如果有时候我简化了一些技术细节以至于语义搜索听起来有点过于简单了，是因为我渴望达成你阅读这本书的原因：找出你需要做些什么来帮助你的业务在 Web 上取得更大的曝光度。

为了弥补这一简化，我在书的末尾提供了完整的参考资料和学术文献，其中大部分可以在 Web 上免费获取并为你营造许多个晚间阅读的快乐时光。那么不再多说，让我们来看看什么是语义搜索以及为什么它是我们数字世界中的一件大事情。

“语义”是一个希腊语词汇，意指“含义”，语义领域一直忙于研究词语的含义和逻辑语用。在 Web 搜索中，语义搜索标志着一种过渡——从面向以一定概率值包含我们所找信息的单一网页的“笨”搜索，过渡到一种能够提供真正答案或将我们引向一个与我们使用的搜索查询无关并且在过去传统的关键词触发的结果中不会出现的一个网页上的智能搜索。

语义搜索作为一个概念，起源于常被称为互联网之父[①]的 Tim Berners-Lee 在2001 年《科学美国人》（*Scientific American*）上发表的一篇文章。其中，他解释了语义搜索的本质是通过数学来摆脱当今搜索中使用的猜测和近似，并为词语的含义以及它们如何关联到我们在搜索引擎输入框中所找的东西引进一种清晰的理解方式。

---

① 原文误称 Tim Berners-Lee 是互联网（Internet）之父，事实上应该是万维网（Web）之父。互联网之父一般是指 Vint Cerf 和 Bob Kahn。——译者注

从概念上讲，语义搜索最多就是这些。这一改变允许我们做出过渡——从一个链接之网——带给我们一些继而不得不在搜索要找的信息时人工检查的可能答案，过渡到一个答案之网——这些答案是从海量数据的复杂关联和交互中综合得到的，基本上就出现在页面上等待我们立即阅读，或者最多通过点一次鼠标就能获取。

前语义时代的 Web 传送的是一些链接，它们出现在搜索结果中是因为它们表示的页面包含了关键词。语义 Web 传送的是与我们在搜索中键入的问题直接相关的确切答案和页面。

尽管这作为一种过渡可能听起来很简单，但实际并非如此，证据在于当这个概念得以传播了十几年之后，我们才只是刚刚开始直面语义搜索这一现实。语义搜索如此难以实现的原因涉及只有回想起来才觉得是显然的两个因素。第一个是数据。要让一个搜索算法能搞清楚在搜索框中输入的一个词汇的含义并“理解”它，所需要的与之相关的数据量都远超出当前我们已准备好存取的量。更重要的是，不仅仅需要数据，也需要一种有意义的排序和分类，这些处理数据的方式使其从人的角度而言开始变得有意义。

第二个原因是可伸缩性。要让语义搜索在组成 Web 的数以万亿的页面上运转，只能以一种既不需要人的介入又能保证搜索结果质量的方式来扩展到这个数量级。这里的难点在于搜索的质量一直在被人为地一点点微调。当你我执行了一次搜索，并且翻阅了五六页的链接也没能找到我们问题的确切答案之后，我们便很不高兴，并回去重新搜索一次。通过改善自己的搜索查询，我们控制着查找的准确性。搜索中不准确的结果经常是由不够精确的搜索查询导致的。

“关键词”作为搜索返回结果所依赖的途径，对于想要在搜索中排名更高的业务、想要更快地获得更精确结果的个人以及有时候向出价最高者售卖服务的 SEO 专家而言，已经进入了他们的字典里。

我们在搜索中由于所用搜索词含义的误解会得到不精确的结果，这没什么大不了的，因为我们知道无论如何我们总能钻取到要找的信息。人的大脑和眼睛可以快速理解一个网页上呈现的内容，并用一种计算机做不到的方式来决定其是否包含要找的答案。这种低效的方法也能作为应对错误和虚假信息的一种质量保障。

为了更好地说明这一点，以一次传统的搜索为例，例如“肉毒杆菌”（Botulinum），会得到一些页面描述其作为人类已知的最致命的物质之一的功效，同时也描述了它在整容手术中作为肉毒杆菌霉素（Botox）使用。接下来，我作为一个操作员，可