PPCblog.com呈现给我们一幅由Jess Bachman（在WallStats.com工作）精心描绘的示意图，这张流程图展示了每天拥有3亿次点击量的Google搜索按钮背后搜索引擎在那不到1秒的响应时间内所进行的处理。

　　这是我刚付印的最新示意图，这张流程图演示了在你点击Google搜索按钮后，在Google返回查询结果前那一眨眼的功夫里，Google是如何处理你的搜索请求的？这可是搜索巨人Google年赢利额高达200亿美元的杀手级应用，也是Internet首屈一指的商业和技术神话，大家肯定都想知道Google这棵摇钱树背后的秘密。

**Google官方对其搜索技术的叙述**

　　我们搜索技术的后端软件会在服务器侧触发一系列执行时间不到1秒的并行计算，Google问世前的传统搜索引擎的搜索结果严重依赖于关键词在页面上出现的频度，我们使用了200多个指标信号（其中包括我们拥有专利的PageRank页面等级加权算法）用来检查万维网的链接结构（佩奇和布林最初的想法是把万维网的链接结构用图论的有向无环图来建模）并决定网页的重要程度，我们假定一个网页的重要程度取决于别的页面对它的引用，就像学术论文中的引用指数一样，重要的论文总是会被很多其他论文引用。然后我们再根据搜索条件进行超文本匹配分析（对bot抓取的页面内容进行关键词倒排索引检索）确定跟搜索请求最相关的网页。综合最重要的网页和跟搜索请求最相关的网页两个方面，我们就能按重要程度和用户搜索请求相关程度把查询结果排序后呈现给我们的用户。

**数据中心：**Google用来索引世界的塔

Google的数据中心高度机密，我们能了解到的不多：

1． 在美国本土有19个以上的数据中心，其余17个数据中心分布在美国以外的世界各地。

2． 每个数据中心有50万平方英尺那么大，建造一个数据中心要花费约6亿美元。

3． Google数据中心是世界上最高效的设施之一，而且也非常环保，几乎没有碳排放。

4． 数据中心使用50到100兆瓦的电力，由于需要冷却，通常建在便于用水的地方。

5． Google服务器安置在一个一组容得下1160台服务器的有房子那么大的标准集装箱容器中。

**处理流程**

1．你写博客、或在Twitter上推微博、更新站点等诸如此类往Web上添加内容的操作

2．Google bots程序（一种作为搜索引擎构件的智能代理程序）抓取你网页的title和description、keyword等内容

（1）Google爬虫沿着链接路径周游万维网，如果没有超文本路径到你的站点，你的站点将不会被索引

（2）如果你在robots.txt中设置不许索引，Google爬虫程序将不会抓取你的网页

（3）如果链接到你站点的超文本链接上有nofollow标签，Google爬虫将不会从这些链接路径周游到你的站点。

（4）Google也能通过blog软件或xml站点地图找到你的网站

（5）从PageRank越高的网站链接到你的网站的链接越多，你的网站的PageRank就越高。

（6）Google爬虫将周游所有未标注为nofollow的链接

3．一旦被Google爬虫访问到，网页几秒内就被索引了

（1）网页内容被存储在一个倒排索引中

  ① 网页标题和链接数据被保存在一个索引中，用于广度优先搜索

  ② 网页内容保存在另一个索引中，以用于检索频率不高的长尾、个性化、深度优先搜索

（2）当你用Google搜索时，你并没有在检索时时更新的万维网，而是在检索Google的缓存，Google定期更新其索引库，在Twitter实时搜索等的竞争下，Google的索引库更新周期趋短。

4．Google基于链接评估域名和网页的总体PageRank值。

5．检查网页以防止作弊行为

(1) Google的搜索质量和反垃圾信息审查和优化算法

(2) 1万多远程测试用户评价搜索结果的质量

(3) Google征请用户对有PageRank讹诈嫌疑的垃圾信息进行举报

(4) Google接到 (美国)数字千年版权法案的通知，要求Google从搜索结果中剔除涉嫌盗版的内容

6．在对页面做了损害分析后，现在每个页面都有很多用于辅助用户搜索的数据片（比如检索关键词）反向引用着它

7．用户发出搜索请求

（1）Google搜索质量工程师Patrick Riley：在大多数Google搜索中，你的搜索处于许多并行的控制过程或Google实验室的创新项目组过程中，可以说每一个查询请求都会参与一些Google的创意实验。

8．Google会用同义词匹配与你的搜索关键词语义相近的查询结果

9．生成初步的查询结果

（1）Google当然能返回成千上万数量无限的查询结果，但一般只显示不到1000条的查询结果，出于“少则得，多则惑”的考虑。

（2）对查询结果做本地化处理，本土站点在查询结果中优先出现

10．对查询结果集按权威性和PageRank进行排序，重复的查询结果被剔除。

(1) Google根据关键词、广告类型、用户所处位置找出相关的被竞价拍卖的关键词广告

(2) 关键词广告必须遵守当地法律条文

  ① 广告业主的非法广告将被取缔

  ② 如果关键词的搜索流量过低或关键词广告点击量偏低，则会被自动禁用

  ③ 出于商业策略，像亚马逊这样的客户会给予优惠折扣。

(3) 关键词相关广告按收益潜力（对关键词进行竞价拍卖后的广告质量不断进行评估）排序

(4) 对广告业主来说广告内容一般都是固定的，但有时使用动态关键词使关键词广告与搜索关键词相关度更高

  ① 一些广告本身允许增加易变的附属信息，比如网站链接、电话号码、产品链接、地址等

(5) 当广告拥有了相当高的点击率，则会显示在搜索结果列表的上方，以使其更显眼。

(6) 其余的广告依序显示在相应的位置

11．对查询结果进行过滤处理

(1) 对通常的查询（比如在Google首页上发出的搜索请求），Google会把相关的专题性垂直搜索结果（比如新闻、购物、视频、书籍、地图等）也加到返回的查询结果中

(2) 个性化方面：用户访问过的网站在查询结果列表中会更靠上

(3) 大量使用锚点的网站有可能被从查询结果中删除

(4) 搜索结果集的聚簇性：如果网页被其他高PageRank的网站引用，则网页的重要性会大大提高。

(5) 趋势分析：对搜索流量爆增或有大量新闻的搜索关键词，Google会在新的查询结果中增加额外的PageRank权值。（Google有反映关键词搜索流量的Google趋势专题页面）

(6) 同一个域名下的多个网页如果具有相同的PageRank会被归为一组。

12. 最终返回给浏览器端的用户一个人性化的、布局良好的、查询结果和广告泾渭分明的有机查询结果页面。

　　所有这些步骤在总共不到1秒的响应时间内完成，每天3亿次的点击量给Google带来了超过200亿美元的年收入。