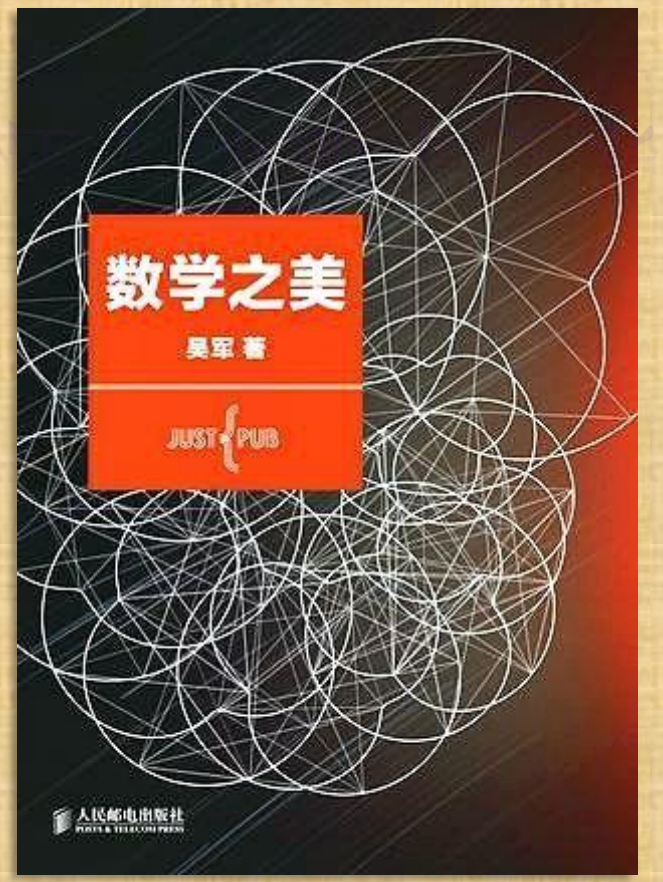
数学的精彩就在于简单的模型可以干大事，
数学的魅力就在于即那个复杂的问题简单化

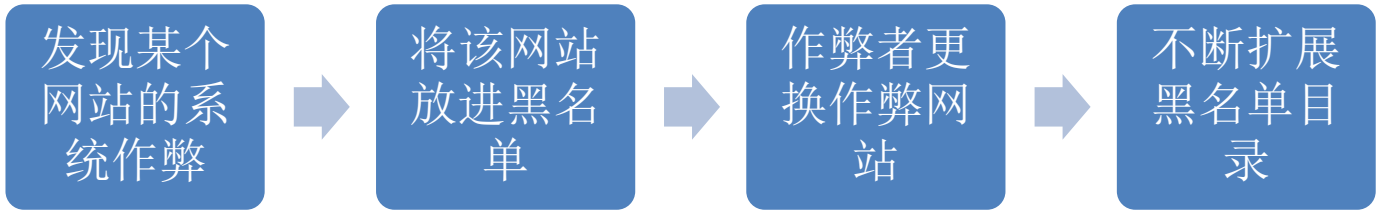


物联网141 朱得元

整体思维

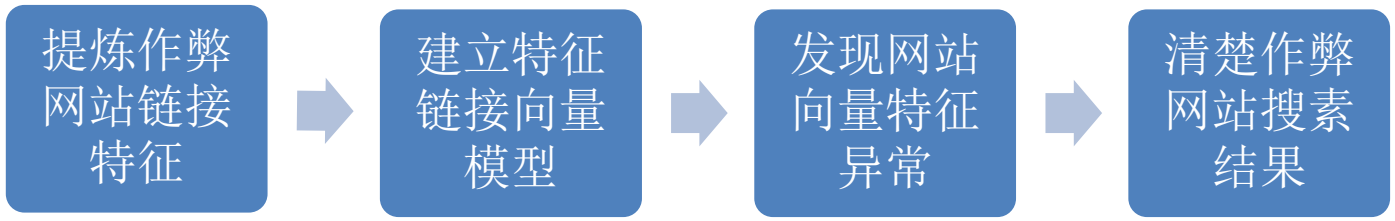
案例：如何抓住搜索引擎排名的作弊者

大部分人：



这其实是一种凑结果的方法，能快速解决问题，但一旦出现新情况就需要不断调整适应，最后导致解决问题的方法越来越复杂而失去效率。

顶尖高手：



分析业务建立数学模型，然后用实践数据验证模型的可靠性，再用经过实践的模型去解决所有相关的问题，这样建立了普遍适应能力的，抗干扰能力强的系统。

简化思维

案例：如何建立一个可用的搜索引擎

5000亿个网页
如何在**最短时间内用最少服务器遍历一遍网页**？

5000亿个网页
如何用**最少空间建立网页内容的索引用于比对**？

5000亿个网页
如何计算出那些**网页的质量度高可优先推荐**？

5000亿个网页
如何计算出哪个网页**最可能是客户查找的网页**？

问题本质

如何在有限时间内最多地爬下最重要的网页?

数学方法

图论

BFS（广度优先算法）
找到一个网站就顺链接下载其上全部下级页面

DFS（深度优先算法）
先找到重要的网站下载重要的页面

这个问题也可以等价于
从北京出发到走遍全国每个城市，怎样走最好？

下载网页

制作索引

网页排名

查询相关

问题本质

如何用最少空间建立网页内容的索引用于比对?

数学方法

布尔代数

1 建立一个关键字词汇表

2 每个关键词建立一个长长的二进制数，每一位代表一篇文献

3 每一位数如果是1则代表一篇文献是否含有某关键词，
100010010010...表示第1篇，第5篇，第8篇，第12篇含有某关键词

4 计算机要找出哪些文字含用户搜索关键词只需要做一次布尔运算

5 布尔运算的效率最便宜的微机一秒钟可以进行数十亿次

6 海量网页就构成了一个海量索引

7 索引还需要记录每个词的位置和次数

8 巨大的索引超出计算机内存，需要设计计算机的分布式运算能力

问题本质

如何计算出那些网页的质量度高优先推荐?

数学方法

PageRank算法

1 PageRank算法核心思想就是一个网页被很多其它网页所链接，特别是高质量的网页所链接，那么它的网页质量就高，相应排名也高。

2 为了计算网页的质量排名，就需要知道其关联的网页质量排名，这就产生了一个是先有鸡还是先有蛋的怪圈。

3 利用二维矩阵相乘迭代算法解决这个问题，假定所有网页排名都是一个相同初始值，通过这种迭代算法一定可收敛到网页真实排名。

4 计算海量网页排名计算量非常大，利用稀疏矩阵计算技巧可简化计算，最后谷歌发展出MapReduce并行计算工具减少服务器负担。

5 佩奇和布林成功关键是把整个互联网当做一个整体对待，以往的算法只注意了网页内容和查询语句的相关性，忽略了网页之间的关系。

问题本质

如何计算出最可能是客户要查找的网页?

数学方法

关键词权重的概率论计算（TF-IDF）

1 包含关键词多的网页应该比少的网页相关度高，但是长网页岂不是占了便宜？所以需要计算“关键词的频率”，也就是关键词次数除以网页的总字数。

2 如果一个搜索包括N个关键词，那么需要计算每个关键词在网页中出现的总词频（TF）。

3 你得删除掉很多无用的虚词或副词，也就是不同的关键词应该有不同的权重，使用最多的权重是“逆文本频率指数（IDF）”，也就是取关键词在网页中出现的次数除以网页总数的对数。

4 把每个关键词的词频和权重做加权求和，就可以得到搜索结果的相关性。

5 最后的搜索排名主要由相关性和网页排名综合决定。

世界上最好的读书分享者，不是讲演时获得多少掌声，而是讲演后，听众会不会好奇地去买一本回去细细品读。

谢谢

THANK YOU