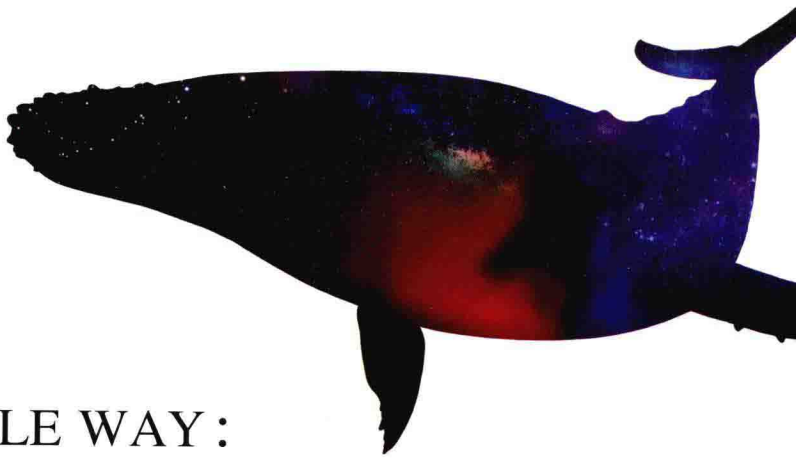


BLUE WHALE WAY:

HOW TO MAXIMIZE THE BIG DATA VALUE?

蓝鲸法则

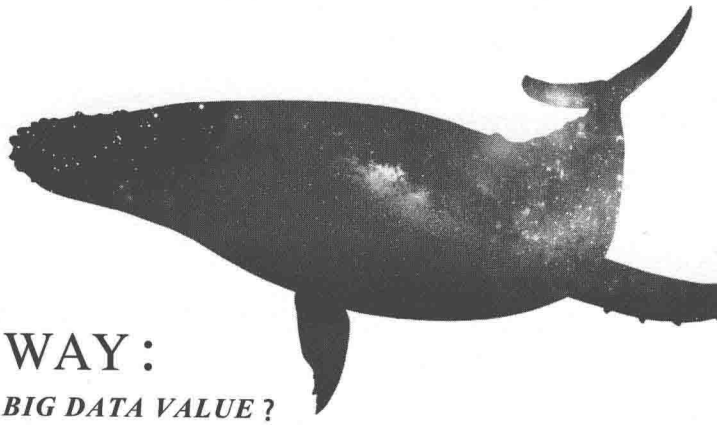
——大数据之道

洪磊　李静　刘先泽　著

迄今为止最全面的互联网大数据指南

为你开启红利的时代

人民出版社



BLUE WHALE WAY:

HOW TO MAXIMIZE THE BIG DATA VALUE?

蓝鲸法则

——大数据之道

洪磊 李静 刘先泽 著

人民出版社

责任编辑:李椒元
装帧设计:肖　辉　孙文君
责任校对:吕　飞

图书在版编目(CIP)数据

蓝鲸法则:大数据之道/洪磊,李静,刘先泽 著.
-北京:人民出版社,2015.12(2016.4)
ISBN 978-7-01-015473-2

Ⅰ.①蓝… Ⅱ.①洪… ②李… ③刘… Ⅲ.①数据管理-研究
Ⅳ.①TP274

中国版本图书馆 CIP 数据核字(2015)第 264224 号

蓝鲸法则
LANJING FAZE
——大数据之道

洪　磊　李　静　刘先泽　著

人民出版社 出版发行
(100706　北京市东城区隆福寺街 99 号)

北京明恒达印务有限公司印刷　新华书店经销

2015 年 12 月第 1 版　2016 年 4 月北京第 2 次印刷
开本:710 毫米×1000 毫米 1/16
印张:14　字数:210 千字

ISBN 978-7-01-015473-2　定价:39.80 元

邮购地址 100706　北京市东城区隆福寺街 99 号
人民东方图书销售中心　电话 (010)65250042　65289539

PREFACE 前言

蓝鲸法则——大数据之道

大数据时代的来临让世界更加的丰满，很多毫不相关的事情也有了一丝一缕的关联，许多事情都正因为大数据的到来而变得更有措施和解决之道。如何在大数据的世界，正能量地引导、影响这个繁杂的世界，并让越来越多难以处理的事情在大数据面前迎刃而解；寻找大数据时代力所能及的方法也是我们所倡导的简约之道，就如同通过 404 一个崩溃的页面而去寻找那些失联在人间的天使，通过大数据的展现让孩子们有一个让亲人看到的场景。

失落的“404 Not Found”

“今天，我想告诉大家一个 404 的故事，但开始之前我们最好先来了解一下到底什么是 404 页面，”2012 年 5 月在 TED 的讲台上社交策略专家雷尼·格里森（Renny Gleeson）正在妙趣横生地讲述着“这就是 404 页面一个让人在浏览页面时让人很崩溃的一种体验，当你点击某网页却无法找到，所显示的默认页面就是 404。”

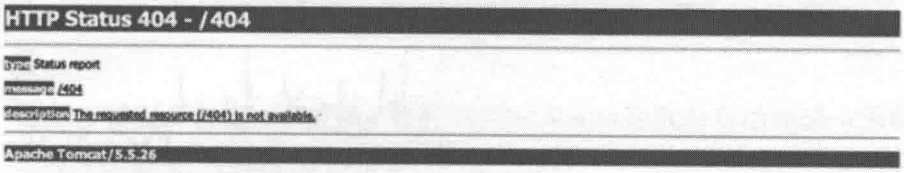


图 0-1 404 页面

404 是 HTTP 其中一种的标准回应信息，通俗地说，当用户浏览网页时，服务器无法正常提供信息，例如用户输入了错误链接；或者无法相应并且找不到原因，这时候页面会出现这个 404 页面。404 是一个信息码，反馈给用户的信息便是：” Not Found。网页不存在”。 据说在第三次科技革命之前，互联网的形态就是一个大型的中央数据库，这个数据库就设置在 404 房间里面。那时候所有的请求都是由人工手动完成的，如果在数据库中没有找到请求者所需要的文件，或者由于请求者写错了文件编号，用户就会得到一个返回信息：room 404 : file not found。

这种遭遇比比皆是，大的网站有，小的网站也有。404 想告诉我们的是“对不起，你想要的这里我们没有”。这的确不是什么好的体验，用雷尼·格里森的话来说好比当你欣赏风景如何的美景和美人的时候，你也想置身其中看个究竟，而这时突然跳出一个“404”门卫告诉你走错地方了这里没有你想要的，感觉就像被人扇了一记耳光。

4XX 系列错误是网页客户端错误的状态码，每串数字对应着一个错误状态。常见的也不仅仅是 404 网页错误，还有 403 错误（服务器禁止回应）和 405 错误（回应资源不被支持）。但人们唯对 404“情有

独钟”。看来是大家对于这种突然之间到来的”Not Found”失落感受太是强烈了。

如何解决“404 Not Found”问题，互联网上有着大量的信息分享，但你会发现大家对它的热度并没有消退的势头，通过百度指数基于“404 Not Found”作为关键词的热点趋势来看持续 6 年保持上涨态势。可以说这种失落感还在持续地上升中。

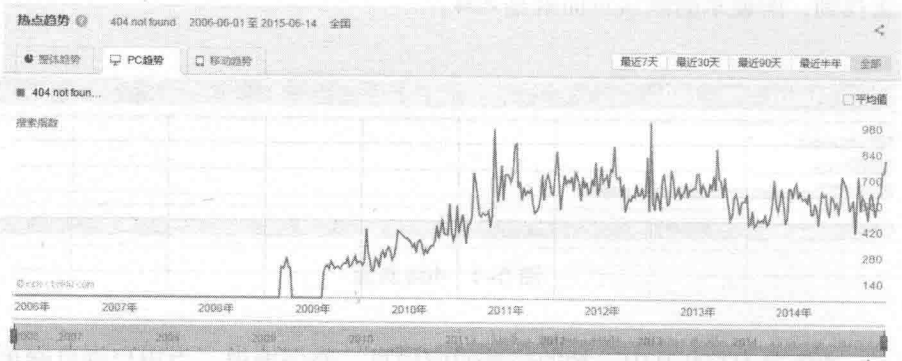


图 0–2 “404 not found”百度指数趋势

寻找“失落”的孩子

2012 年初欧洲的一家公益组织 Child Focus 推出了一项名为“Not Found”的计划，利用网站的 404 页面来帮助找回那些走丢的孩子们。该项目的主要内容就是利用人们上网时经常碰见的 404 错误页面来向公众展示那些被拐卖或是失踪的儿童信息，以便利用广阔的网络力量来让这些失去孩子的家庭重获幸福。

来自欧洲失踪儿童联合会的代表 Valeria Setti 在接受采访时表示，这一项目将允许该组织在任何一个 404 错误页面上刊登并显示失踪儿童的姓名和照片等粗略信息，在这些信息还会提供一个 NotFound.org 官方网站的详细资料页面，供知情人查阅。

Child Focus（关心孩子）在旗下的 Notfound.org 网站上，有这样一段话：“欧盟每年都会有上千孩子走失。你可以借 NotFound 共享自己的一份

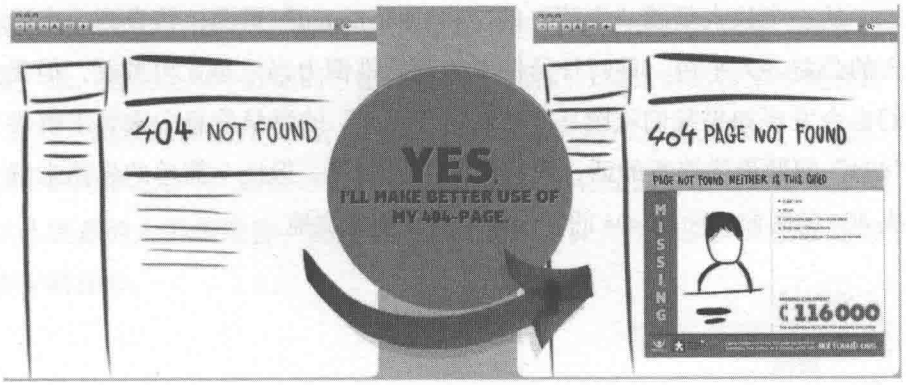


图 0-3 404 页面找回走丢的孩子

力量。只需安装我们的应用，在你的网站 404 页面就会自动加载一张走失孩子的照片。让我们一起携手送他们回家吧。”

此外，该类页面还设计有一个即时分享按钮，该按钮的作用是让用户能够很方便地把页面信息发送给他们所熟悉的企业或个人并以此让那些被刊登孩子的信息受到更多的人所关注。

该项目首先在欧盟国家推行起来，2013 年 1 月，益云公益、CSDN、百度寻人和中国计算机学会等也联合推出了中国版“404 页面儿童打拐”项目。腾讯 QQ 空间等多个产品也上线了附带儿童走失信息的 404 页面。这些儿童走失信息来自中华社会救助基金会、宝贝回家寻子网等公益机构。目前，已有包括微软 Bing、58 同城、淘宝、蘑菇街等 1100 多家大中小型网站加入，总展现次数达数十万，而在微博上，“404 公益”也获得了大量网友的转发、关注。

“百度正能量之寻亲行动”的活动，将网页搜索的 404 页面创新为公益广告，用以展示 24 个已经被解救孩子的信息，帮助他们寻找父母。百度的 404 页面是当用户请求的页面不存在、链接错误或具有攻击性时产生的提示页面，其中由于用户访问带有攻击性的风险页面所导致的 404 提示页面每天的展现量就达到了千万。百度将网页搜索的 404 页面创新为公益广告，不仅提示网友访问出错，还展示 24 个已经被解救孩子的信息，相当于推出了一个巨大的公益寻人平台。

从一个让人倍感“失落”的“404 Not Found”页面，转身到一个巨大的公益寻人平台。日新月异技术的背后推倒力当然是无可置疑，但我们也会发现如果我们试图从“404 Not Found”的海量信息中来寻求解决“404”问题最优答案的话，我们可能不得而终。但换个简单的思路来解决404的时候，而让404的“失落”有了用武之地。

蓝鲸法则

蓝鲸被认为是已知的地球上生存过的体积最大的动物，长可达33米，重达181吨。尽管体型巨大，平时行动缓慢，常常静止不动，却能在水中沉浮自如，尾巴灵活地摆动，既是前进的动力，也起着舵的作用，前进的时速高达28公里。1000万年后，鲸的后肢已经退化，它抛弃了陆地生活，而完全适应了这浩瀚的海洋，可以说归功于两点：

（1）懂得很好利用海水的浮力，而不需要像陆地动物祖先那样支撑身体，所以鲸通过让身体变大，就达到了在水中自由沉浮的目的；

（2）用占身体不到5%的“鳍”来而产生推进、平衡及导向作用，来畅游海洋。如：背鳍其长度不及体长的1.5%，控制在海洋中的身体平衡；而用尾鳍可作上下摆动，是游泳的主要器官。有些种类还具有背鳍，用来平衡身体。

正如我们迷失在大数据的海洋中一样，现阶段我们不可能处理所有的数据，但需要了解这些数据的“能力”而让我们可以在大数据中“浮游”起来，而不止于一味地潜入数据的深处甚至完全把它煮沸，因为，我们已经早已告别了数据库只存放在一间房间的时代。

仅在身体5%的“鳍”的玄妙之处在于诠释了用最小作用力来创造世界的真谛，这也告诉了我们从认识数据“能力”的目的是等到简单的答案，而起到对事物的推动、平衡或是导向作用，而不是在大数据中获得更多繁杂的东西，以阻碍你的前进。

套用到“404”来寻找失踪的孩子这个案例上我们知道，如果我们一味地去处理404数据来发现404数据的可用之处的时候我们很有可能是

找不到答案的，而我们能理解到大量的404数据带给人们的是一种“失落”感，这便是404数据的“能力”，当然我们以集纳的态度来审视这种“能力”能带给我们什么呢？“对的，是倍感失落之后激发起的‘走失’孩子的同情之心”！这是一个很简单的信息，也正是因为信息的简单才可以让更多的人都能理解，都能感同身受；而由此产生的洞察的确也更加的深刻有力。

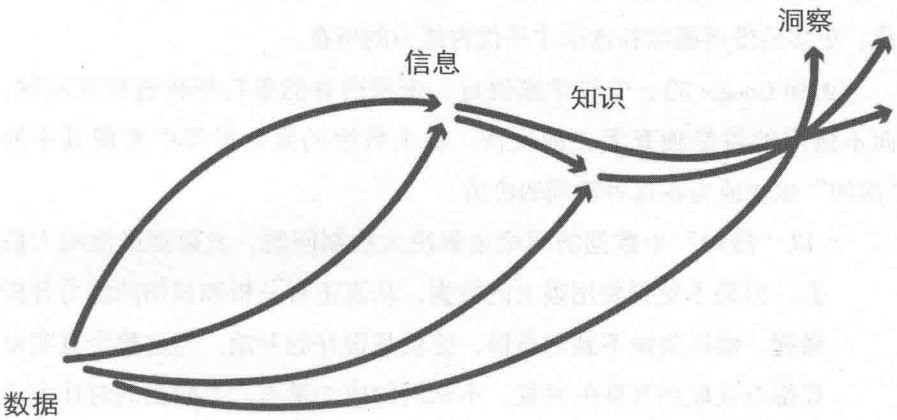


图 0–4 蓝鲸法则——大数据时代的简约之道

蓝鲸法则：大数据之道

1. 不要被大数据的大而沉溺其中，了解数据懂得利用数据的“浮力”才是关键；

2.“以简约为目标”将数据转为为“信息”、“知识”，最终形成洞察，洞察及行动；

3. 洞察可以通过“数据”，“信息”，“知识”流程式，组合式，直通车式的来获得。

“蓝鲸法则”和“大数据”一样是个新鲜的名词，“蓝鲸法则”强调的是“简约”，而大数据给人的印象更多是“繁杂”，也正是用蓝鲸的这种“简约”来化解“繁杂”。造物主用“极值”创造了这个“丰富”的世界，大禹治水用的是疏通而非一味的围堵，一张一弛文武之道，两个看似自

行矛盾体的结合而往往能组成问题的最优方案。

“蓝鲸法则”的大简之道，看上去朴实无华，甚至可以说是当前技术前进过程中的一个“退步之选”，然而这种“退步”也正是当下应对我们无法应对繁杂的创新之道，为什么这么说，我们可以回忆一下在我们一路走来的知识积累过程，如果某人记忆力好唯一是很容易成为一个成绩好的学生，而计算机处理的世界里随着数据存储成本的降低、在线存储成本的为 0，我们已经习惯了数千的记忆思维，从此刻开始已经发生了变化，更多的懂得删除和选取才是优秀能力的所在。

正如 Google 的一位科学家说过，未来所有的事物都将链接互联网，而不链网的将是独有无二的特例。在大数据的繁杂世界中发现其中的“简约”也便成为在这种特例的价值。

- 以“简约”小数据的思维来解决大数据问题。大数据虽然叫大数据，但是不见得要用很大的数据，从真正有分析和预估的能力分析做起，做以前做不到的事情，这就是很好的开始。当这整个真实世界都与互联网互联的时候，不联网的成为稀有。大数据的时代由于诸多原因大数据变的便宜又容易，小数据变的昂贵又困难。
- 让大数据处理方式和大数据发展相适宜。以正确的方式舍弃恰当的非核心元素，往往会做得更好，这也便是小数据处理之道。小数据处理之道可以更清晰地看清楚复杂事物的本质，克服过量的辅助问题。精简大数据发掘小数据，重新考虑正在做的一切。征服大数据不在于控制所有数据，而在于吸收那些有用的部分。
- 通过融入商业思维人工干预的方式来解读数据。处理完的数据是一种直白的信号，而当你要在其中获得怎样的信息、知识和洞察那就需要人工干预来进行解读了。我们要保持紧随大数据的持续演进，今天的大可能是明天小，新的数据源也将出现。

不仅如此我们还必须直面一个看似简单的问题：“我们是否要用大数据来寻找问题的因果关系。”大数据是能看到一个全量的趋势，趋势中混杂着各种类型的巨大数据，这些巨大数据间关联我们知果而不晓因，的确关联信很是重要也给我们以新的视角了来省事问题，如果我们不用大

数据来发现因果关系了，而只看相关性的话那对于事物的发展和问题的研究，好比我们分析大数据对于结果没有一定方向性的预盼，答案的轮廓只能在最后一刻来揭晓，那一切也只在一种偶然的状态下来前行了。

目录 CONTENTS