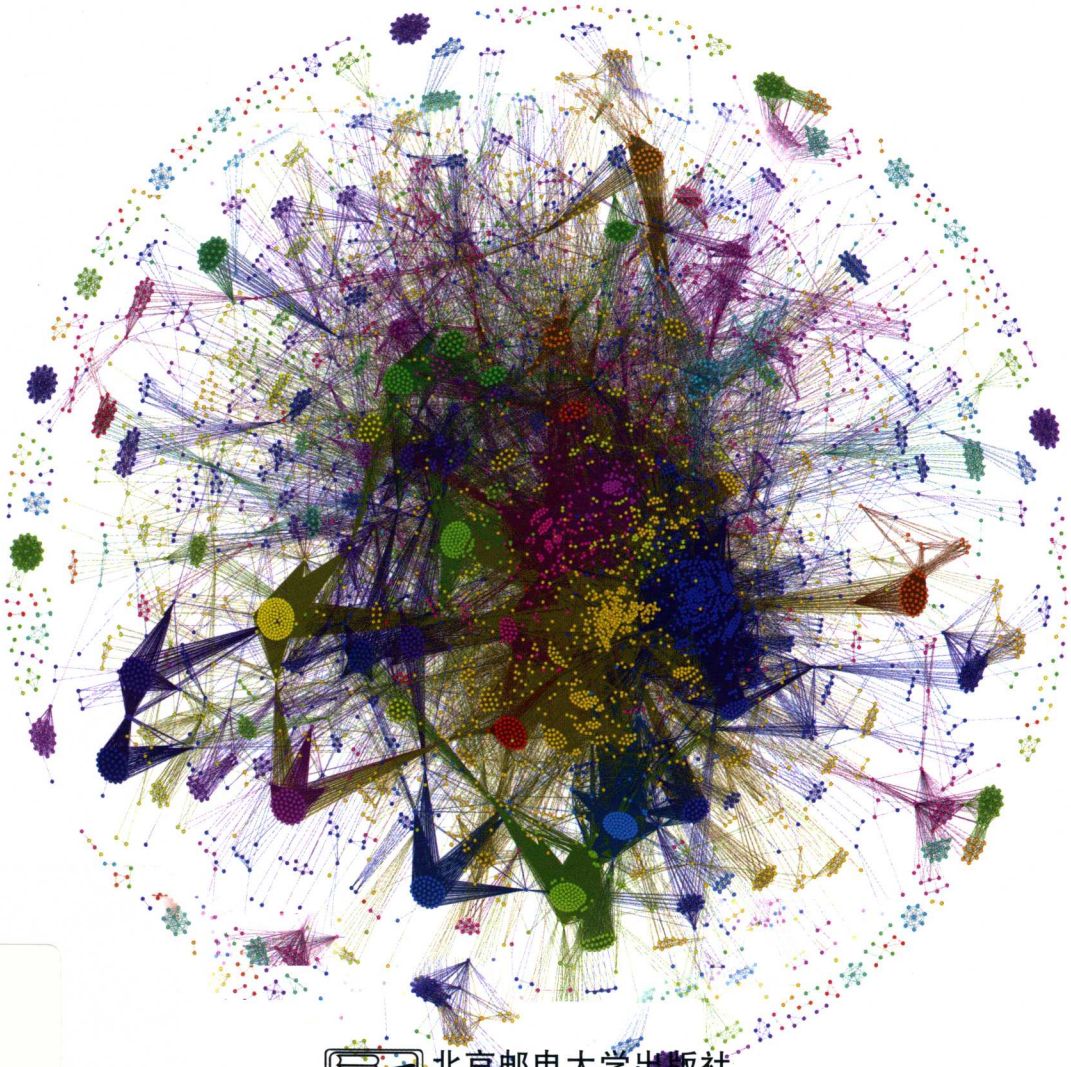
本书由清华大学、上海大学、河北科技大学、国网河北电科院等科研院所的一线教师和高级工程师合著，阿里巴巴、腾讯、百度、51talk、昆仑万维、央视网等企业多位工业界同行联合推荐。

大数据
技术及行业应用

如何**定义**大数据？ 如何**应用大数据**？
什么是大数据思维？如何**学习**大数据？
如何**构建**大数据平台？如何在行业中**应用**大数据？

许云峰 徐华 张妍 王杨君 马瑞◎著



北京邮电大学出版社
www.buptpress.com

大数据技术及行业应用

许云峰　徐华　张妍　王杨君　马瑞　著

北京邮电大学出版社
www.buptpress.com

图书在版编目(CIP)数据

大数据技术及行业应用 / 许云峰等著. --北京:北京邮电大学出版社,2016.8
ISBN 978-7-5635-4918-4

Ⅰ.①大… Ⅱ.①许… Ⅲ.①数据处理 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字(2016)第 199031 号

书　　名：大数据技术及行业应用
著作责任者：许云峰　徐华　张妍　王杨君　马瑞　著
责任编辑：王丹丹
出版发行：北京邮电大学出版社
社　　址：北京市海淀区西土城路 10 号(邮编:100876)
发 行 部：电话：010-62282185　传真：010-62283578
E-mail：publish@bupt.edu.cn
经　　销：各地新华书店
印　　刷：保定市中画美凯印刷有限公司
开　　本：787 mm×1 092 mm　1/16
印　　张：16.5
字　　数：469 千字
版　　次：2016 年 8 月第 1 版　2016 年 8 月第 1 次印刷

ISBN 978-7-5635-4918-4　　定　价：35.00 元

序一

自从世界上第一台真正意义上的电子计算机 ENIAC 在宾夕法尼亚大学的诞生开始，数据的存储与组织就成为了计算机技术走向实际应用过程中的一个重要课题。在随后的整整七十年间，以通过高效处理解决实际应用问题为目标的计算机科学得到了长足的发展，但在增长速度的赛跑中，存储器硬件的容量远远地被应用问题的规模甩在后面。进入新世纪后，随着这一矛盾在多样性、价值、速度等方面的日益突出，迫切需要综合已有技术给出系统性的、可扩展的解决方案，而大数据理论与方法也自然地应运而生。

近十年来，这方面的专著、教材层出不穷，但摆在案头的这本《大数据技术及行业应用》却在很多方面令人耳目一新。该书的五位联合作者虽非名声显赫，但都是来自国内著名高校的青年才俊，他们活跃在大数据科学的技术前沿，同时善于汲取与融合来自工业界的先进工程经验和学术界的宝贵学术资源。其中首章简明地介绍了大数据的基本理论，辨析了相关的重要概念和理念。接下来的三章，从私有平台、虚拟化平台、综合性平台等层面系统介绍了现有的成熟技术方法。最后七章是本书的重点，依次剖析了大数据技术在图算法、环境科学、药物数据聚类、电子商务、社交网络、文本挖掘与情感分析、电力系统控制等领域的具体应用。透过这些应用，读者应该能够生动地看到大数据技术也已取得的巨大成效，以及未来发展的广阔前景。

祝贺五位作者的这次成功合作，更祝愿他们能在这一领域继续前行，探索和发展出更多的新方法、新技术。

是为序。

邓俊辉

2016 年 8 月于清华园

序二

2010年，随着中国进入“移动互联网元年”，中国正式进入了移动端设备的全面上网时代。与传统互联网相比，移动互联网解决了任何人、任何时间、任何地点以多种方式上网的问题。伴随着移动互联网的普及，人与人之间、人与物之间、物与物之间实现了全面的互联互通。通过移动互联网，我们可以获得人和人之间互联的社交数据（微博、微信等）、人和物之间互联的行为数据（上网日志信息等）、物与物之间互联的环境数据（智能家居设备采集的环境数据等）。由于通过移动互联网采集的这些数据，具有高容量（Volume）、高生成速度（Velocity）、多样性（Variety），同时具有潜在的应用价值（Value）等特点，所以我们常常称之为“大数据”。为了解决“大数据”的存储管理和计算难题，信息技术领域近些年深入研发了“云”计算技术。“互联网＋”正是在这样一个如何综合利用大数据，实现移动互联网与各个传统行业的深度融合与创新的背景下应运而生的。

互联网技术与应用的发展对于高校和IT行业，特别是信息技术专业领域的学生和工程技术人员，提出了掌握大数据的处理方法与技术的基本要求。本书正是针对这样一个技术与应用发展背景下的要求，系统性的介绍了大数据的相关概念、大数据平台的搭建与综合解决方案，以及国内相关领域的研究学者在环境、医药、电子商务、社交网络、文本挖掘和电力系统等分支领域的应用研究成果。

作为教育部支持的高级访问学者，许云峰老师曾在我所在的清华大学智能技术与系统国家重点实验室从事了为期一年的高级访问学者研究工作，重点研究了基于社交网络数据的社群发现算法。作为近年来在国内互联网社交数据挖掘与分析方面较为活跃的研究学者，他陆续在KBS和ESWA等国际期刊上发表了系列化社群发现的研究成果。本书既是对许云峰老师过去多年来在大数据应用方面的技术总结，也是将他对大数据技术的应用经验分享给相关同学和技术人员的一个很好的形式。相信各位读者能够从中熟悉大数据技术的深刻内涵。

徐华

2016年8月于清华园

前　言

如何定义大数据？如何应用大数据？什么是大数据思维？如何学习大数据？如何构建大数据平台？如何在行业中应用大数据？这一系列的问题，是当前在大数据热的时代背景里，让人感到非常迷茫的问题。本书直面这些问题，在从业者角度解答以上问题，希望能给大数据行业的初学者提供一些帮助。当然我们的观点并不是放之四海而皆准的唯一真理，随着大数据行业的发展，会有更全面的答案出现。希望这本书能起到抛砖引玉的作用。

本书第1章阐述了大数据的相关概念。第2章讲解了基于Hadoop的私有云平台搭建案例。第3章讲解了基于OpenStack和Docker的大数据平台的基础虚拟化平台搭建。第4章讲解了基于CDH和HDP的大数据平台搭建。第5章讲解了基于Spark平台的大数据处理应用，并展示了图挖掘中的经典案例。第6章讲解了大数据技术在环境科学中的应用案例。第7章讲解了一个大数据在DrugBank药物数据库聚类方面的应用案例。第8章讲解了一个大数据在电子商务数据分析中应用的案例。第9章讲解了一个大数据思维在社交网络数据分析中应用的案例。第10章讲解了大数据技术在情感分类中应用的案例。第11章讲解了大数据技术在电力数据分析中应用的案例。本书的案例虽然不能囊括大数据应用的所有领域，但是在不同的角度回答和解决了大部分人在当前大数据环境下面临的问题和挑战。

本书由清华大学、上海大学、河北科技大学、国网河北电科院等科研院所的一线教师和高级工程师合著。其中，许云峰完成了第1章、第7章和第9章。徐华完成了第10章，张妍完成了第3章、第4章、第5章。王杨君完成了第6章。马瑞、陈二松和范辉共同完成了第11章。许云峰和陈书旺共同完成了第2章。许云峰和李媚共同完成了第8章。

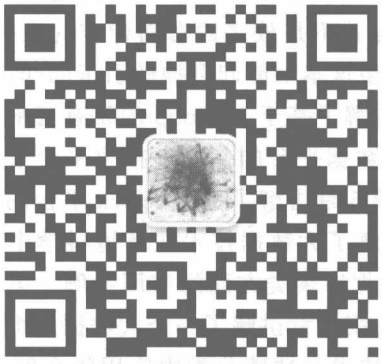
本书由清华大学MOOC著名教育专家邓俊辉老师和清华大学智能技术国家重点实验室徐华老师倾情作序，邓老师多次入选清华大学我印象最深的十大教师(毕业生评选)，徐华老师是我在清华访学期间的指导老师。两位老师治学严谨，是我学习的典范，在此表示深深的感谢。

本书得到来自工业界同行的宝贵意见和建议，他们是：阿里数据经济研究中心常务副主任兼秘书长潘永花，腾讯企业云架构师 潘晓东，百度 云服务架构师 董月照，51talk大数据专家资深架构师郝伟瑞、刘会山，央视网高级数据分析师孙泉，昆仑万维苗雨顺，北京卓新思创CEO庄严，北京星立方科技发展股份有限公司周长亮，北京图森王路，在此表示诚挚的感谢。感谢参加本书的编辑和实验工作的同学，他们是：周莹莹、白云、么学媛、赖杰、俞孝帅、胡江涛、肖明美、刘芳彤、温亚东、李书航、刘利平。同时感谢石家庄京华电子实业有限公司远松灵总经理提供部分应用场景支持，感谢石家庄吾搜网络科技有限公司提供域名和空间支持。

全书共46.9万字，许云峰完成了16.9万字，徐华完成了8.2万字，张妍完成了12.1万字，王杨君完成了3.7万字，马瑞完成1万字，陈书旺完成1万字，李媚完成1万字，陈二松完成1万字，范辉完成1万字，许云岭完成1万字。

本书在初稿完成后，经多位工业界和学术界同行审阅，获得大家普遍认可。同时各位专家学者也提出了很多宝贵的意见和建议，我们做了对应修改，但是由于时间仓促，难免会有疏漏。在本书付梓之际，本人诚惶诚恐，失眠多日，但是在探寻真知的过程中偶得的宝贵知识和经验不敢私藏，所谓“愚者千虑，必有一得”，希望本书中的工程经验和学术观点能够对广大读者有所启发，并起到抛砖引玉的效果。同时欢迎读者与我们交流，并提出宝贵意见，联系信息如下。

本书官方公众号：大数据技术及行业应用



本书官方网站：http://dxmall.com.cn:2016

许云峰

2016年8月于河北科技大学

目　录

第1章　大数据相关概念

大数据就是互联网发展到现今阶段的一种表象或特征而已，我们无须神化它或对它保持敬畏之心，在以云计算为代表的技术创新大幕的衬托下，这些原本很难收集和使用的数据开始容易被利用起来了，通过各行各业的不断创新，大数据会逐步为人类创造更多的价值。

1.1　什么是大数据？

大数据并不是一个新的概念，大数据其实是随着计算机技术、通信技术、物联网技术的发展而必须面对的一个普遍的问题。类似计算机发展史上的软件危机，信息技术每发展到一定阶段就会遇到数量太大处理不了的问题，可以说大数据是一种信息技术发展的现象。20世纪60年代，随着商业软件的发展，原来的文件已经不能满足商业数据存储的要求，于是产生了关系型数据库。相对于当时的传统文件，关系型数据库是一种处理大数据的经典解决方案。戏剧性的是，随着互联网技术的发展，传统数据库已不能解决对互联网文档数据的存储，于是产生了基于文档应用的NoSQL技术。可见在信息技术发展的过程中，人类在不停面对来自大数据的挑战。

麦肯锡咨询公司最早提出大数据时代的到来："数据，已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素。人们对于海量数据的挖掘和运用，预示着新一波生产率增长和消费者盈余浪潮的到来。"IBM将大数据的特征归纳为4个"V"：量(Volume)，多样(Variety)，价值(Value)，速度(Velocity)。特点有四个层面：第一，数据体量巨大。大数据的起始计量单位至少是P(1 000个T)、E(100万个T)或Z(10亿个T)；第二，数据类型繁多。比如，网络日志、视频、图片、地理位置信息等；第三，价值密度低，商业价值高；第四，处理速度快。

麦肯锡咨询公司最早给出了大数据的定义：大数据是超过传统数据库工具的获取、存储、分析能力的数据集，并不是超过TB的才叫大数据。维基百科对大数据的定义：大数据是指无法在可承受的时间范围内用常规软件工具进行捕捉、管理和处理的数据集。作为多年从事数据处理工作的IT从业者，我们给出的大数据定义是：大数据是超过传统数据库工具、传统数据结构、传统程序设计语言、传统编程思想的获取、存储、分析能力的数据集。本书对大数据定义的补充是基于我们在长期的大数据处理中遇到的一些具体挑战。

(1) 关于传统数据库的局限。我们都知道Oracle数据库是大家常用的企业应用数据

库。但是对于现在的一个小的物流公司来说，使用这个 Oracle 已经不能满足一个小公司获取、存储、分析能力的需求。对于一个二线城市中的一个小物流中心来说，每天保守地说如果是 1 万件的吞吐量，那么在数据库中会至少产生 1 万条记录，一年就会产生 365 万条记录，3 年就会是 1 000 万多条数据。对于千万级数据的调优，需要一个工作 5 年以上的数据库工程师才能胜任，据了解这样一个工程师的年薪应该是 20 万元以上，那么对于一个小的物流公司负担这样的薪水就有些困难了，所以大部分小物流公司对这些物流数据的处理措施是定期删除数据，但是这些数据其实是蕴含着非常宝贵的商业价值的。如果采用大数据时代的数据库管理工具，比如 MongoDB 等，那么成本就会降低很多，当然会使用基于大数据需求的数据库管理工具的人才也是非常难找，目前只有一些重点大学成立了大数据专业，比如清华大学的大数据学院等。

(2) 关于传统数据结构的局限。大数据时代的数据具有 4 V 的特性，数据类型繁多，数据量大，要求处理速度快，基于内存操作，因而传统的数据结构必须经过优化创新才能适应大数据时代的应用需求。例如 Google 的 BigTable 技术，Apache 软件基金会的 HDFS 文件结构，Mahout 中的数据结构，HIVE 数据库，学术界的热点知识图谱技术，这些都是顺应大数据时代而生的大数据时代的数据结构。这些数据结构相比与传统编程语言中的数组、结构体、ArrayList、HashMap 等数据结构具有更好地适应大数据 4V 特点的特性。现在已经广泛应用的大数据的数据结构都是根据大数据的具体应用场景进行优化和创新的，例如 Mahout 中的 FastByIDMap、FastIDSet 和 GenericItemPreferenceArray 等，相对于单机环境的 Java Collections 框架，它们降低了对内存的占用(欧文，2014)。

(3) 关于传统的程序设计语言的局限。大数据应用环境下，产生了许多新兴的编程语言，如 R 语言、Python、Scala 等，这些语言天生具有操作分布式计算环境和进行机器学习运算的基因。以 R 语言为例，其完成一个逻辑回归分析并生成图像，大概只需要 5 行代码，而 Java 和 C# 却需要调用 N 多类库，写上数十行程序。Python 和 Scala 可以轻松操作Hadoop 和 Spark 平台，而这些在 C# 和 Java 环境中，需要 N 多的类库和配置，再加上数十行代码

(4) 关于传统编程工具中的报表、datagrid 等控件的数据展示能力是有限制的。如需要展现上万个节点之间的关系，使用水晶报表或者 JFreeChart 是无法完成用户需求的，需要 EChart、Three. js 等大数据时代的报表工具。

(5) 关于传统编程思想的局限。Jeffrey Scott 等提出在大数据环境下，传统的编程思想和框架产生了瓶颈(Vitter，2008)，因为传统的编程思想和框架将寄存器、缓存、内存、磁盘统一编址，是基于所有的存储器具有相同访问时间的假设(Vitter，2008)。因而用这样的思想来处理大数据，会造成效率低下，并且在这种框架和编程思想下开发的应用程序不能适应大数据时代的应用。例如对 600 万条数据做一次查询需要十几秒，如果采用大数据时代的编程思想，可以将查询控制在毫秒级别，当然还有服务器配置高低的问题，这里我们讨论的是在相同的运算环境下。

综上所述，我们概括了传统数据库、数据结构、程序设计语言、编程思想的局限，提出了大数据的定义。

1.2　大数据有多大?

大数据到底有多大?这个问题在大数据的定义里是可以界定的,大数据的大是相对于传统的数据库、编程语言、数据结构、编程思想和框架的处理能力的大,但是相对于现在的数据增长来说,以前的工具、方法思想都是传统的。当遭遇新的智能硬件革命、新技术浪潮的时候,现在的技术工具方法等又会在未来成为传统的且阻碍潮流发展的。因而大数据的大是一个相对的大,人类在未来世界里会不断地遭遇大数据的问题,从而会有新的解决方案提出。因而大数据不是一个新问题,会是一个一直存在的问题。

1.3　大数据是一种思维方式

在大数据环境下,程序员、决策者、领导层等,应该具有大数据的思维方式,为什么要这么说?我们接触过教师、官员、企业主、金融从业人员、互联网从业者等很多人,发现大家总有这么几个思维定式:(1)认为大数据就是将传统的经验知识应用到海量数据里去,当然这是应用大数据的一种方式,但是大数据其实是可以展示很多现象的,这些现象中有人类在传统数据环境下不能发现的客观规律。如果一味地坚持已有知识的应用,而错过从全局角度去发现、理解大数据现象下新的客观规律,就如同一叶障目不见泰山。(2)习惯将传统的思维方式带入到大数据应用环境中。举个例子,一个编程能力很强的研究生在数据采集过程中总是很苦恼,工作进度很慢,我很奇怪,就问他为什么这么苦恼?进度这么慢,这明显不是他的风格。他告诉我说,在大数据采集过程中,总有一些网站出现莫名其妙的异常,而且这些异常超出了正常的逻辑。他一直试图把这些异常解决掉,但总是解决不了,所以耽误了进度。其实这个学生陷入一个传统程序设计思维在大数据应用场景下的误区。因为大数据采集过程中的异常来源是非常广泛的,如网络异常、网站改版、各种 Web 服务器的访问策略异常等,有的是人为的,有的是硬件的,还有的是软件造成的。传统程序设计思维因处理的应用场景单一,异常是可控的,程序员可以将程序写得非常完美,但是在大数据环境下,一天可能有上亿条数据的吞吐量,并且应用场景复杂,因而个别数据的不完整是可以忽略不计的,因为大数据观察的是大趋势、大方向,个例的差异是可以忽略的,如果一味地追求程序的完美,要处理所有异常,相对于大数据价值密度低、整体价值高的特点,在时间成本和用人成本上是得不偿失的。

由此可见,大数据环境下的数据思维方式是需要慢慢建立和适应的。那么大数据思维是什么?大数据思维主要包括两个方面:(1)从什么角度看数据。(2)怎样使用数据。

第一方面,大数据时代的到来,使人类以史无前例的低成本去获取和利用数据,人类的本能是利用已有的知识和经验去理解、分析和利用这些数据,但是所有新事物的出现都会带来新科学规律的发现,例如天文望远镜的出现使人类可以更直观地观测宇宙,而抛弃了原来的猜想方式,发现了木星的四个卫星等客观规律。大数据时代的到来,肯定会帮助人类发现新的客观规律。例如,我们依赖大数据技术和可视化技术对社交网络进行研究,从而更精准

地对社区概念进行定义(Xu, Xu et al. 2015)。可见大数据可以给人类一个更有高度、更全局的观察视角对客观世界进行分析和发现。

第二方面,怎样使用大数据。首先要明确使用大数据的目的是解决问题、发现规律,那么如何根据要解决的问题去使用大数据?我们就需要使用数据的方法,通常我们叫作量化的方法。所谓量化的方法,就是在解决问题的过程要可衡量、可评估、有明确的定义。车品觉等提出了 PIMA 定义(车品觉,2014):要解决的问题是什么,或者说目的是什么?(P);在要达到的这个目的的过程中要有非常明确的定义(I);在解决问题的过程中用的手段必须是可以量化的(M);解决问题的结果是可以评估的(A)。这实际上一个相当严谨的研究问题的体系,是一种严谨的数据思维。这个 PIMA 框架可以提供一个非常易用的大数据流程。

综上所述,大数据思维是一种从全局角度去明确问题、定义过程、量化过程、评估结果的数据思维方式。其不但可以跳出已有知识的界限,从全局角度发现新的规律,而且可以将已有知识、规律应用于大数据的解读过程中,是我们向科学领域进军的新的得力工具和武器。

1.4 大数据思维的应用案例

1.3 节中我们提出了大数据思维的定义,这个定义局限于当前我们对大数据的认识。学界和工业界的学者和技术专家们一定会有不同于我们定义的大数据思维的意见,但是本节还是要把一些我们之所以产生这样定义的案例阐述一下。

随着计算和存储成本的降低,人类可将生产、工作、生活中产生的海量数据进行存储和分析,这些海量数据中蕴含着丰富的关系数据。在已有的计算和可视化条件下,结合云计算和社区发现技术从全局角度分析和挖掘数据中蕴含的丰富信息,可以为自然科学和商业应用等领域提供一种新的观察、分析和挖掘数据的视角,为发现数据中蕴含的结构、功能、规律等信息提供新的强有力的工具和方法。下面从正在研究的两个领域简要介绍下我们提出的大数据思维定义的产生背景。

(1) 药物研究方面。当前国际上对新药物研发都是基于天然产物的活性新化合物的发现和已知结构化合物二次开发两种思路相结合来开展的。例如我们对 Drugbank 数据库中的八千多种药物基于靶标和作用酶进行聚类,图 1-1 中展示的是聚类后的可视化图,通过该图我们可以了解到不同药物种类的分布,药物聚类的内部结构和外部边界,从而指导新药的研发。该图是当前 4K 屏的可视化条件下的极限,上部和下部小的类别没有全部显示出来。聚类算法采用的是我们提出的骨干度算法(Xu, Xu et al. 2015),该算法采用从全局角度去明确问题、定义过程、量化过程、评估结果的思维方式,这部分工作已经发表在期刊《Expert Systems with Applications》上,并得到国际同行的认可。

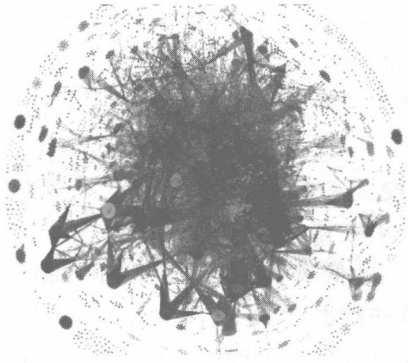


图 1-1 DrugBank 数据库中的八千多种药物基于靶标和作用酶进行聚类后的可视化图