海量网络多媒体信息高效处理：概念与技术

庄毅 著

The concepts and technologies

Efficient Processing of Large-Scale Web-Based Multimedia Information

科学出版社

海量网络多媒体信息高效处理：概念与技术

庄 毅 著

科学出版社

北 京

内 容 简 介

本书较为系统地从数据库层面对海量网络多媒体信息的高效处理进行介绍和讨论。本书分为8篇24章，力求从检索、索引、降维、聚类及并行处理等5个方面在深度和广度上进行阐述，侧重于提高查询效率。同时结合最新的网络多媒体研究现状及发展趋势，进行深入阐述和分析。另外，结合最新应用，如数字图书馆、网络舆情分析与监控及网络购物等进行介绍。

本书可作为高等院校计算机科学、图书情报等专业的研究生或高年级本科生的参考资料或教学用书，对从事海量网络多媒体数据处理研究、应用和开发的广大科技人员也有很大的参考价值。

图书在版编目(CIP)数据

海量网络多媒体信息高效处理：概念与技术/庄毅著. —北京：科学出版社，2013

ISBN 978-7-03-037415-8

Ⅰ.①海… Ⅱ.①庄… Ⅲ.①互联网络-信息管理-研究 Ⅳ.①TP393.4②G203

中国版本图书馆CIP数据核字(2013)第092562号

责任编辑：余 丁 王 苏 / 责任校对：张怡君
责任印制：张 倩 / 封面设计：陈 敬

科学出版社 出版
北京东黄城根北街16号
邮政编码：100717
http://www.sciencep.com

骏走印刷厂印刷

科学出版社发行 各地新华书店经销

*

2013年6月第 一 版 开本：B5(720×1000)
2013年6月第一次印刷 印张：29 1/4
字数：567 000

定价：98.00元

(如有印装质量问题，我社负责调换)

序

海量网络多媒体信息的高效处理一直是国内外学术界研究的热点问题。庄毅博士在网络多媒体数据库研究领域进行了多年深入的研究，特别是在多媒体数据高维索引及查询方法研究上取得了优异的成绩，积累了丰富的经验。目前已在国际国内权威刊物和学术会议上发表了多篇高质量学术论文。由于其成绩出色，获得了2008年中国计算机学会优秀博士论文奖。

该书是作者近年来从事海量网络多媒体数据管理研究工作的总结，大部分研究成果在书中有所反映。同时，还加入了一些最新的相关技术，如社交媒体的概率查询与个性化推荐、网络舆情分析与监控及云计算环境下的并行查询等。

该书较为系统地从数据库层面对网络多媒体信息的检索、索引、降维、聚类及并行处理等概念和技术进行介绍，侧重于提高查询处理效率。该书分8篇24章，从深度和广度上对海量高维多媒体信息的高效处理技术进行阐述。同时作者结合最新的网络多媒体研究现状及发展趋势，进行深入介绍和探讨。最后，将多媒体查询技术应用于数字图书馆、网络舆情分析与监控及在线网络购物等领域，取得不错的效果。

我相信该书的出版将会对国内海量网络多媒体信息处理研究领域起到一定的参考和推动作用。

李　青　教授
香港城市大学电脑科学系
香港网络协会(HK Web Society)主席
2012年10月

前　言

随着多媒体和网络技术的迅猛发展，互联网已经形成了一个巨大而复杂的多媒体信息空间。其包含的海量多媒体信息资源具有以下特点：①数量巨大，增长迅速；②内容丰富，形式多样；③结构复杂，分布广泛；④无序混乱，杂乱无章。面对互联网中这些浩翰的多媒体信息资源，如何对其进行快速准确的检索及高效的处理已经成为一个很重要的研究课题。

本书从数据库层面对海量网络多媒体信息检索、索引、降维、聚类及并行化处理等概念和技术进行较为系统的介绍，侧重于提高查询效率；同时结合最新的网络多媒体研究现状及发展趋势，进行深入阐述和分析；另外，结合最新应用，如数字图书馆、网络舆情分析与监控及网络购物等进行介绍。本书分为 8 篇 24 章，力求从深度和广度上对海量高维多媒体信息的高效处理技术进行阐述。

在入门篇中，简要地对面向互联网的海量多媒体数据查询、索引及管理进行综述。

在检索篇中，针对网络多媒体数据的特点，介绍三种查询方法，如基于语义的多媒体查询、基于内容的多媒体查询和基于多特征的多媒体查询。同时结合最新网络多媒体技术的发展趋势，介绍三种新型查询技术，即跨媒体检索、社交媒体检索和语义网数据检索。

在索引篇中，针对海量高维多媒体数据相似查询存在的计算量大及维度高的问题，分别介绍文本索引、(多)高维索引及多特征索引。

在降维篇中，针对海量高维多媒体数据存在的“维数灾难”的问题，介绍三类降维方法：无监督降维、半监督降维及监督降维。

在聚类篇中，介绍几种主流的聚类方法。同时，结合多媒体数据类型，分别介绍文本、图像、音频及视频方面聚类的最新研究进展。

在并行处理篇中，针对在单机环境下，海量多媒体数据查询性能低下的问题，分别介绍了数据网格环境下及移动云计算环境下的可扩展并行查询技术。该技术包括海量高维数据的分布式优化存储、索引支持下的快速高维数据集的缩减、并行流水线处理及高效的自适应数据传输机制等；同时，针对频繁的用户查询请求，提出基于网格环境的高维相似查询的多重查询优化方法，进一步提高查询密集条件下的海量多媒体检索的并发性。

在应用篇中，结合海量多媒体信息管理中的实际应用展开，如数字图书馆、网络舆情分析与监控及网络购物等。

在总结篇中，对海量多媒体信息处理技术进行总结，并且展望未来发展。

在本书的撰写过程当中，得到了香港科技大学电脑科学与工程系陈雷副教授和加拿大 Simon Fraser 大学裴健教授的支持和鼓励，陈教授审阅了本书的部分章节，并提出了许多宝贵的意见。香港城市大学李青教授在百忙之中为本书作序。在此向他们致以衷心的感谢。本书的第 8 章由北京大学邹磊副教授撰写，在此表示感谢！同时感谢在本书撰写过程中给予无私帮助的同事和朋友，他们是浙江工商大学的凌云教授、姜波教授和张华副教授，南京财经大学的伍之昂副教授，美国 Facebook 公司的杨骏博士和昆士兰大学的邵杰博士等，还有为本书第 23 章开发移动商品视频搜索系统的陈一枭、王晓晴同学等。

本书得到了国家自然科学基金（项目编号：61003074）、浙江省自然科学基金（项目编号：Z1100822，Y1110644，Y1110969，Y1090165）、2012 年度浙江省科协育才工程项目和浙江省科技厅重点创新团队（计划编号：2010R50041 及项目编号：2012R10041-06）的支持，同时得到了 2012 年杭州市青年科技人才培育工程项目资助，在此一并表示感谢！

最后，谨以本书献给我敬爱的父母和我的妻子，没有他们的支持，本书很难顺利出版，在这里特别表示感谢。

由于网络多媒体技术发展日新月异，加上本人学识浅陋，书中必有许多不足之处，希望读者提出意见。

庄　毅

2012 年 12 月

于杭州

目　　录

索 引 篇

降　维　篇

聚 类 篇

并行处理篇

应　用　篇

总 结 篇

入门篇