

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>

骏杰印刷厂印刷

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学位论文奖。

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瀚的多媒体信息资源,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杲、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

序
前言

入 门 篇

第 1 章 互联网、多媒体与大数据	3
1.1 绪论	3
1.2 本书内容结构	6
第 2 章 海量多媒体处理技术综述	8
2.1 多媒体检索技术	8
2.2 高维索引技术	12
2.2.1 集中式高维索引	12
2.2.2 分布式高维索引	13
2.3 降维与聚类技术	14
2.3.1 降维	15
2.3.2 聚类	15
2.4 并行检索技术	15
2.4.1 基于数据分片的负载均衡技术	16
2.4.2 云计算、网格计算及点对点计算	16
2.5 有代表性的海量多媒体系统	20
2.6 本章小结	23

检 索 篇

第 3 章 基于语义特征的多媒体检索	27
3.1 引言	27
3.2 文本检索模型	28
3.2.1 布尔模型	28
3.2.2 向量空间模型	29

3.2.3 聚类检索模型	32
3.2.4 概率模型	32
3.3 TF×IDF 权值	33
3.4 现有支持语义的多媒体检索系统	34
3.5 本章小结	36
第4章 基于内容特征的多媒体检索	37
4.1 基于内容的图像检索	37
4.1.1 图像特征提取	37
4.1.2 图像相似度模型	46
4.1.3 图像检索中的相关反馈	48
4.1.4 现有基于内容的图像检索系统	48
4.2 基于内容的音频检索	50
4.2.1 音频特征提取	50
4.2.2 音频例子检索	56
4.2.3 现有基于内容的音频检索系统	61
4.3 基于内容的视频检索	61
4.3.1 视频预处理技术	61
4.3.2 系统体系结构	64
4.3.3 视频检索技术	64
4.3.4 现有基于内容的视频检索系统	67
4.4 本章小结	70
第5章 基于多特征的多媒体检索	71
5.1 基于多特征的图片检索	71
5.1.1 基于语义和内容的图片检索	71
5.1.2 基于内容和主观性特征的图片检索	78
5.1.3 基于多内容特征的书法字图片检索	89
5.2 基于多特征的音频检索	97
5.3 基于多特征的视频检索	97
5.4 本章小结	99
第6章 跨媒体检索	100
6.1 引言	100
6.2 交叉参照图模型	101
6.3 异构媒体对象相关性挖掘	103

6.3.1 基于语义标注的方法	103
6.3.2 基于链接分析的方法	104
6.3.3 基于异构特征分析的方法	105
6.3.4 其他方法	108
6.4 本章小结	109
第7章 社交媒体检索与推荐	110
7.1 引言	110
7.2 国内外研究现状分析	111
7.3 社交(媒体)对象概率建模	114
7.4 基于多特征融合的社交图片对象查询与推荐	115
7.5 结合视觉特征和标签语义不确定性的社交图片概率查询	116
7.5.1 语义特征概率建模	117
7.5.2 查询算法	118
7.6 结合视觉特征的社交图片主观性概率查询	121
7.6.1 主观性特征概率分布模型	121
7.6.2 查询算法	121
7.7 结合地理标注信息和视觉特征的社交图片复合查询	122
7.7.1 基于代价模型的查询策略选择	123
7.7.2 查询算法	125
7.8 社交对象的相关性概率查询	126
7.8.1 交叉关联概率图模型	126
7.8.2 查询算法	127
7.9 基于用户偏好概率模型的社交图片个性化推荐	128
7.9.1 用户偏好概率分布表	128
7.9.2 个性化推荐算法	131
7.10 本章小结	132
第8章 语义网数据检索	133
8.1 语义网和 RDF 数据	133
8.2 RDF 数据管理研究现状	136
8.2.1 SPARQL 查询语言	136
8.2.2 基于关系数据模型	137
8.2.3 基于图数据模型	143
8.3 面向 RDF 的智能检索方法	146

8.4 本章小结	148
----------------	-----

索 引 篇

第 9 章 文本索引	151
9.1 倒排文件索引	151
9.2 签名文件索引	152
9.3 本章小结	153
第 10 章 高维索引	154
10.1 集中式高维索引	154
10.1.1 基于数据和空间分片的索引方法	154
10.1.2 基于向量近似表达的索引方法	156
10.1.3 基于空间填充曲线的索引方法	156
10.1.4 基于尺度空间的索引方法	157
10.1.5 基于距离的索引方法	158
10.1.6 基于数据分布的索引方法	162
10.1.7 基于 LSH 函数的索引方法	163
10.1.8 子空间索引方法	163
10.2 分布式高维索引	164
10.3 不确定性高维索引	166
10.3.1 相关工作	166
10.3.2 预备工作	167
10.3.3 ISU-Tree 索引	168
10.3.4 CU-Tree 索引	174
10.4 实例：基于局部距离图的交互式书法字索引	178
10.4.1 问题定义及动机	178
10.4.2 局部距离图索引	179
10.4.3 超球心重定位	182
10.4.4 索引更新算法	184
10.4.5 伪 k 近邻查询算法	185
10.4.6 实验	186
10.5 本章小结	189

第 11 章 多特征索引	190
11.1 通用多特征索引	190
11.2 图片多特征索引	191
11.2.1 结合语义和内容的多特征索引	191
11.2.2 基于视觉和主观性特征的商品图片多特征索引	191
11.2.3 书法字图片多特征索引	196
11.2.4 社交图片的多特征索引	203
11.3 音频多特征索引	208
11.3.1 基于内容的音频多特征索引	208
11.3.2 基于内容及语义的音频多特征索引	209
11.4 视频多特征索引	210
11.4.1 基于多特征哈希的视频索引	210
11.4.2 基于多特征索引树的视频索引	210
11.5 跨媒体索引	211
11.5.1 预备知识	211
11.5.2 索引生成算法及其可扩展性	211
11.5.3 查询算法	216
11.5.4 实验	217
11.6 社交（媒体）对象的相关性索引	220
11.7 本章小结	222

降 维 篇

第 12 章 降维技术	225
12.1 引言	225
12.2 无监督降维	227
12.2.1 主成分分析	227
12.2.2 多维尺度分析	227
12.2.3 局部保留映射	228
12.2.4 Isomap 降维	229
12.2.5 其他降维方法	230
12.3 半监督降维	230
12.3.1 基于类别标记的方法	230

12.3.2 基于成对约束的方法	233
12.3.3 基于其他监督信息的方法	235
12.4 监督降维	235
12.4.1 线性判别式分析降维	235
12.4.2 其他降维方法	236
12.5 本章小结	237

聚 类 篇

第 13 章 聚类技术	241
13.1 引言	241
13.2 基于划分的聚类算法	244
13.2.1 k-Means 算法	244
13.2.2 k-Medoids 算法	244
13.2.3 k-Modes 算法	245
13.3 基于层次的聚类算法	246
13.3.1 BIRCH 算法	246
13.3.2 CURE 算法	246
13.3.3 CHAMALEON 算法	247
13.3.4 其他层次聚合算法	247
13.4 基于密度的聚类算法	247
13.4.1 DBSCAN 算法	248
13.4.2 OPTICS 算法	248
13.4.3 其他密度聚类算法	248
13.5 基于网格的聚类算法	249
13.5.1 STING 算法	249
13.5.2 CLIQUE 算法	249
13.5.3 其他网格聚类算法	250
13.6 基于模型的聚类算法	250
13.6.1 MRKD-Tree 算法	250
13.6.2 SOON 算法	251
13.6.3 粒子筛选算法	251
13.7 其他聚类算法	251

13.7.1 模糊聚类算法	251
13.7.2 基于图论的聚类算法	252
13.7.3 AP 聚类算法	252
13.8 本章小结	252
第 14 章 文本聚类	253
14.1 k 平均文本聚类算法	253
14.2 层次式文本聚类算法	254
14.3 基于后缀树的 Web 文本聚类算法	254
14.4 基于密度的 Web 文本聚类算法	255
14.5 本章小结	256
第 15 章 图片聚类	257
15.1 引言	257
15.2 基于文本特征的 Web 图片聚类	257
15.2.1 候选图片聚类名的学习	257
15.2.2 合并和裁剪聚类名	258
15.3 基于多特征的 Web 图片聚类	258
15.3.1 Web 图片的三种表达	258
15.3.2 使用文本和链接信息聚类	262
15.4 基于相关性挖掘的 Web 图片聚类	263
15.4.1 图片-文本相关性挖掘	265
15.4.2 图聚类算法	266
15.5 基于多例学习的 Web 图片聚类	266
15.5.1 基于 EM 的多例聚类算法	266
15.5.2 启发式迭代优化算法	267
15.6 基于概率模型的个性化社交图片聚类	267
15.6.1 问题定义	267
15.6.2 上下文信息相似度量	269
15.6.3 用户偏好概率模型	270
15.6.4 聚类算法	270
15.7 本章小结	272
第 16 章 音频聚类与分类	273
16.1 引言	273
16.2 基于拟声词标注的音频聚类	274

16.2.1 动机	274
16.2.2 实现	275
16.3 基于隐马尔可夫模型的音频分类	279
16.4 其他聚类与分类方法	280
16.5 本章小结	280
第 17 章 视频聚类	281
17.1 引言	281
17.2 基于多特征的视频聚类算法	281
17.2.1 视频信息获取	282
17.2.2 视频片段相似度度量	282
17.2.3 上下文信息相似度度量	283
17.2.4 聚类处理	283
17.3 其他视频聚类算法	285
17.4 本章小结	285

并行处理篇

第 18 章 海量多媒体分布式并行相似查询处理	289
18.1 基于数据网格的 k 近邻相似查询	289
18.1.1 预备工作	290
18.1.2 支撑技术	291
18.1.3 GkNN 查询算法	297
18.1.4 理论分析	300
18.1.5 实验	304
18.1.6 具体应用：基于数据网格的书法字检索	308
18.2 移动云计算环境下的医学图像查询处理	310
18.2.1 预备工作	311
18.2.2 支撑技术	314
18.2.3 两种索引结构	324
18.2.4 MiMiC 查询算法	328
18.2.5 实验	330
18.3 本章小结	336

第 19 章 分布式并行环境下的多重相似查询优化	337
19.1 引言	337
19.2 预备工作	338
19.3 动态查询层次聚类	341
19.4 pGMSQ 算法	342
19.5 实验	345
19.6 本章小结与展望	347

应 用 篇

第 20 章 多媒体技术在数字图书馆中的应用	351
20.1 引言	351
20.2 国内外数字图书馆的发展	353
20.3 数字图书馆的优势	355
20.4 多媒体检索在数字图书馆中的重要性	355
20.5 代表性的数字图书馆系统	356
20.6 本章小结	359
第 21 章 网络舆情分析与监控	360
21.1 背景和意义	360
21.2 网络舆情概述	361
21.3 国内外研究现状	363
21.4 总体框架及体系结构	364
21.5 关键技术	366
21.5.1 基于 Mashup 的舆情信息采集与整合	366
21.5.2 舆情信息预处理	367
21.5.3 舆情信息动态挖掘	375
21.5.4 舆情服务	391
21.6 本章小结	392
第 22 章 基于视觉和感性计算的网络购物——淘淘搜	393
22.1 背景和意义	393
22.2 国内外技术现状	393
22.3 搜索引擎框架	395
22.4 系统体系结构	396

22.5	关键技术	397
22.5.1	数据采集、过滤及建库	397
22.5.2	提取主、客观特征	400
22.5.3	搜索引擎设计与实现	401
22.6	原型系统——淘淘搜	404
22.7	本章小结	406
第 23 章	移动商品视频搜索——酷搜	407
23.1	引言	407
23.2	国内外技术现状	408
23.3	关键技术	409
23.4	系统分析	410
23.4.1	功能性需求分析	410
23.4.2	非功能性需求分析	410
23.5	系统设计	411
23.5.1	总体结构设计	411
23.5.2	功能模块设计	412
23.5.3	数据库设计	413
23.6	系统实现	415
23.6.1	数据采集模块	415
23.6.2	数据检索模块	416
23.6.3	数据显示模块	417
23.6.4	数据推送模块	418
23.6.5	后台管理模块	419
23.7	本章小结	420
总 结 篇		
第 24 章	挑战及发展趋势	423
24.1	面临的挑战	423
24.2	发展趋势	425
24.3	本章小结	427
参考文献		428

入门篇