

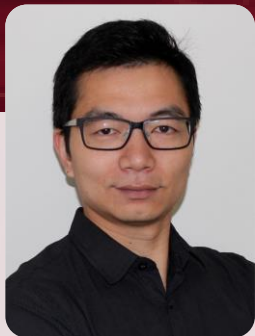
多媒体大数据管理与分析

多媒体大数据的发展与相关研究问题

第一讲

刘家俊
中国人民大学信息学院
2017年7月

个人简介



刘家俊 副教授

❖ 学习

- 本科 (2002-2006) : 南京大学
- 博士 (2009-2012) : 昆士兰大学

❖ 工作

- IBM中国研究院/开发院 (2006-2008) : 研究员
- 澳大利亚联邦科学与工业研究组织 (2012-2015) : Postdoctoral Research Fellow
- 中国人民大学 (2015至今) : 副教授

❖ 研究

- 多媒体与时空大数据的管理、搜索、挖掘和分析
- 30多篇国际会议、期刊论文, 包括多篇顶级IEEE TKDE、ACM TOIS、SIGMOD、ICDE论文
- 多个国际会议/期刊担任程序委员/评审, 如ACM Multimedia 15, VLDB J

❖ 产品

- 5项中/美/澳专利
- 互联网应用: 澳洲留学生社交服务网站 www.buding.com.au

研究课题



多媒体大数据的发展与相关研究问题

- ❖ 多媒体大数据时代的挑战
- ❖ 多媒体大数据关键技术与应用
- ❖ 多媒体大数据若干相关研究领域



多媒体大数据时代的到来

❖ 多媒体：多种媒体的综合，形式包含文本、图片、音频、视频



图片来源：www.nipic.com

多媒体大数据时代的到来

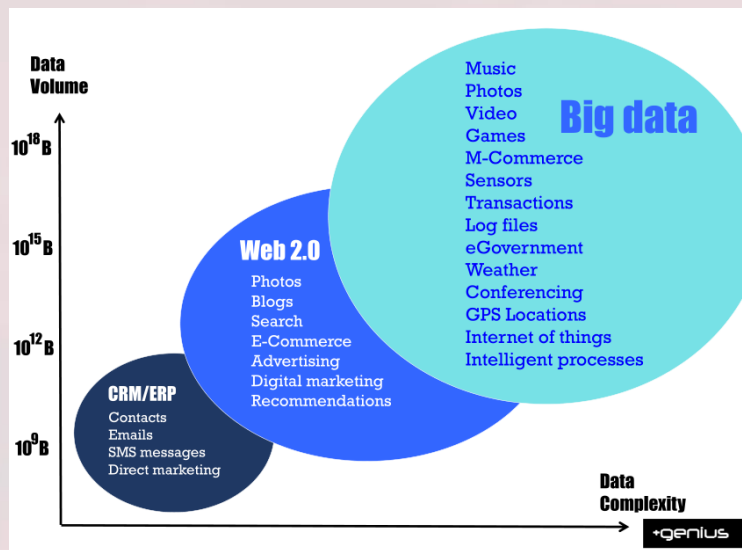
❖ 多媒体数据：

- 多种媒体的综合，形式包含文本、**图片、音频、视频**
- 信息密集，数据量巨大
 - 一本书 $\approx 100\text{KB} \sim 1\text{MB}$
 - 一首MP3 $\approx 3\text{MB} \sim 10\text{MB}$
 - 一张照片 $\approx 2\text{M} \sim 10\text{MB}$
 - 一部电影 $\approx 500\text{MB} \sim 20\text{GB}$



多媒体大数据时代的到来

❖ 发展经历



= ? EB

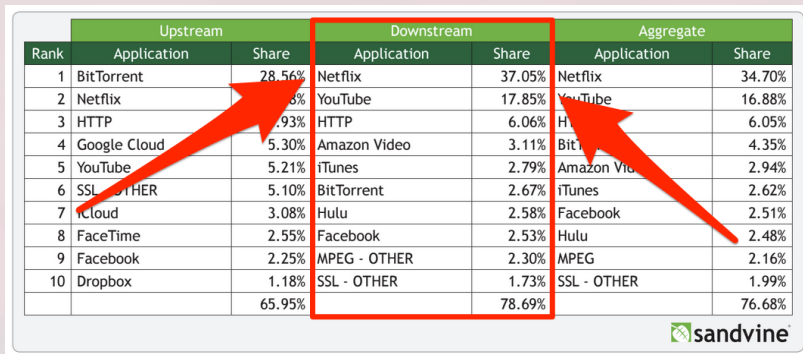


图片来源：geniusworks, jureviciusstudios

多媒体大数据时代的到来

❖ 多媒体愈加成为互联网信息传播的主要载体

- 到2015年底，约70%的互联网流量被用于音视频



| Upstream | | | Downstream | | Aggregate | |
|----------|--------------|--------|--------------|--------|--------------|--------|
| Rank | Application | Share | Application | Share | Application | Share |
| 1 | BitTorrent | 28.56% | Netflix | 37.05% | Netflix | 34.70% |
| 2 | Netflix | 17.85% | YouTube | 17.85% | YouTube | 16.88% |
| 3 | HTTP | 6.06% | HTTP | 6.06% | HTTP | 6.05% |
| 4 | Google Cloud | 5.30% | Amazon Video | 3.11% | BitTorrent | 4.35% |
| 5 | YouTube | 5.21% | iTunes | 2.79% | Amazon Video | 2.94% |
| 6 | SSL - OTHER | 5.10% | BitTorrent | 2.67% | iTunes | 2.62% |
| 7 | iCloud | 3.08% | Hulu | 2.58% | Facebook | 2.51% |
| 8 | FaceTime | 2.55% | Facebook | 2.53% | Hulu | 2.48% |
| 9 | Facebook | 2.25% | MPEG - OTHER | 2.30% | MPEG | 2.16% |
| 10 | Dropbox | 1.18% | SSL - OTHER | 1.73% | SSL - OTHER | 1.99% |
| | | 65.95% | | 78.69% | | 76.68% |

sandvine

图片来源：Sandvine

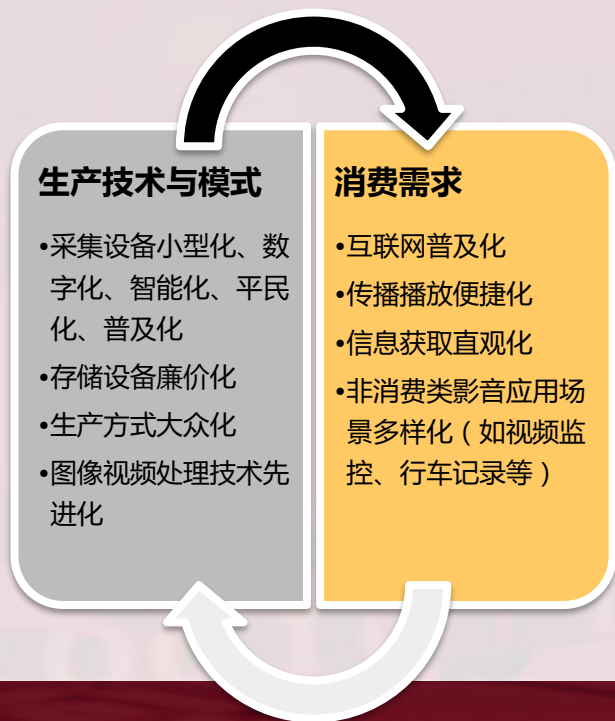
- 到2019年，80%的互联网流量将被用于音视频

数据来源：Cisco VNI Forecast and Methodology, 2015-2020



多媒体大数据时代的到来

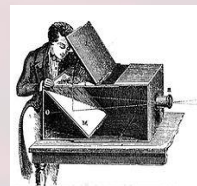
❖ 多媒体数据飞速增长的原因



图片来源：互联网

多媒体大数据时代的到来

❖ 采集设备小型化、数字化、智能化、平民化



图片来源：互联网

多媒体大数据时代的到来

❖ 存储设备廉价化



图片来源：互联网

多媒体大数据时代的到来

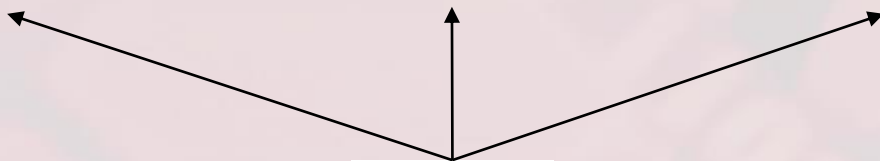
❖ 生产方式大众化



YY LIVE



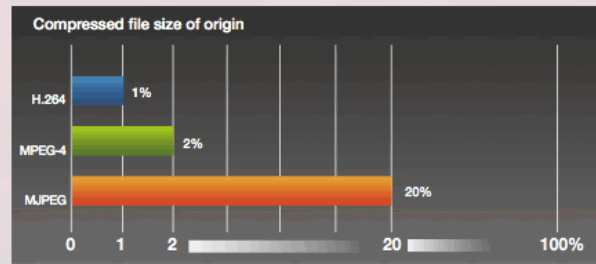
YOUKU 优酷



图片来源：互联网

多媒体大数据时代的到来

❖ 图像视频处理技术先进化



Compressed file size of MJPEG, MPEG-4 and H.264

| | MJPEG | MPEG-4 | H.264 |
|----------------------------|--|---|---|
| Compressed file size | 20% | 2% | 1% |
| Bandwidth comparison ratio | 20 | 2 | 1 |
| Encoding CPU loading ratio | 1 | 4 | 10 |
| Application | <ul style="list-style-type: none">Local storageSnapshot viewing | <ul style="list-style-type: none">Moving picture viewingReal-time transmission | <ul style="list-style-type: none">Moving picture viewingReal-time transmission |

Comparison of MJPEG, MPEG-4 and H.264

多媒体大数据时代的到来

❖ 应用需求不断多样化、普及化

- 照片存储分享
- 新闻视频
- 影视作品点播
- 用户自制内容分享
- 视频监控
- 草根直播



多媒体大数据带来的挑战

01

存不下

多媒体数据的高密度特性决定了其数据量极大，在当前存储设备成本已极低的情况下仍存在不能完全保留数据的困扰。

02

搜不准

即使神经网络的发展已经日新月异，由于往往存在从文字到图像、从语义到视觉特征的鸿沟，现在图像搜索的质量仍然存在较大改进空间

03

查不快

多媒体数据是高信息密度的数据，一般情况下存储的维度极高，难于索引。如何提高查询的效果和效率也是一大挑战。

04

传不动

多媒体压缩技术的发展远远慢于多媒体数据增长的速度，如何高速高效传输多媒体数据仍是难题。



谢谢！

