



大数据管理与数据思维

Big Data and Data Thinking



孟小峰

中国人民大学信息学院

• <http://idke.ruc.edu.cn>





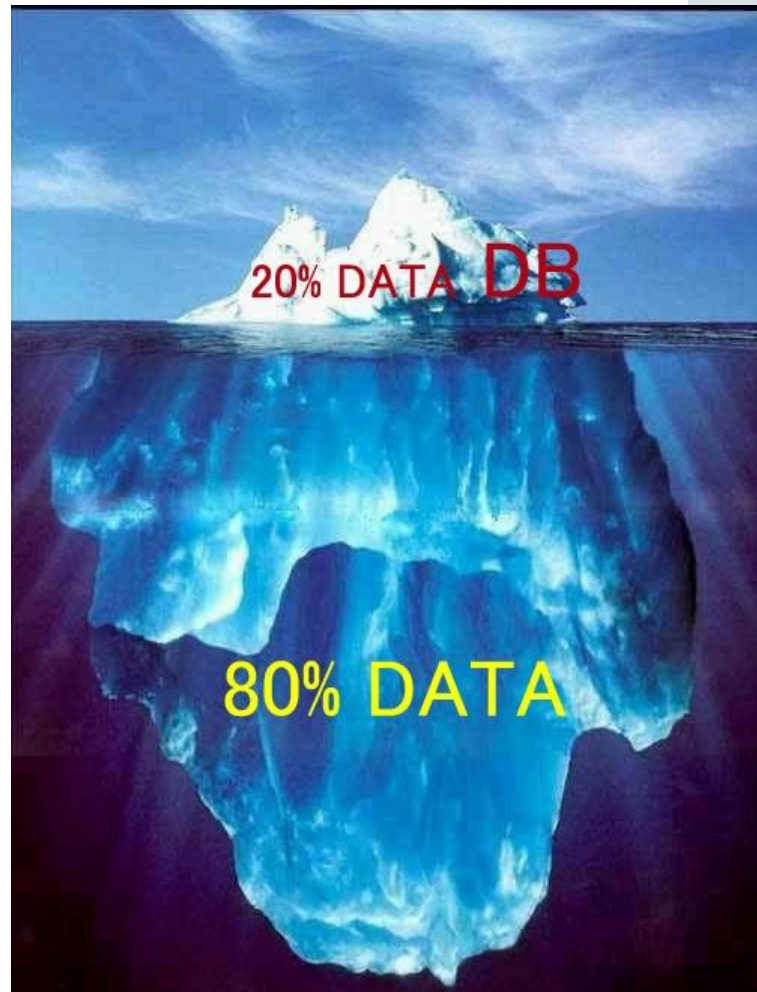
❖ 海量异构

- **281EB in 2009**
- 仍以指数级增长

❖ 内容多样

- 半结构
- 非结构

❖ 动态增长





Data Explosion



MB = 10^6 bytes
a typical book in text format



GB = 10^9 bytes
a one hour video is about 1GB;
data produced by a biology
experiment in one day



TB = 10^{12} bytes
data produced in astronomy data in one night;
US Library of Congress has 1000 TB data;
search log of Bing is 20 TB per day

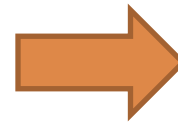


PB = 10^{15} bytes
Facebook uses more than 1 PB for its use;
Google processes 20 PB data per day;
the entire written works of humankind, in



❖ Very Large Database (VLDB)

- MB, 结构数据
- 以数据为对象解决其存储和管理问题

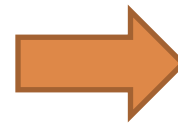


数据工程

Data Engineering

❖ Big Data

- >PB, 非结构数据
- 以数据为资源解决诸领域问题



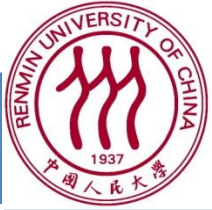
数据思维

Data Thinking





More data vs. better algorithms

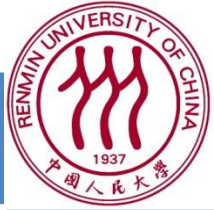


- ❖ **The Netflix challenge (\$1M awarded Sep. 22, 09)**
 - A: sophisticated algorithm
 - B: simple algorithm + additional data (IMDB)
 - *B gets much better results*
- ❖ **Google PageRank:**
 - Rank pages by additional data: link + anchor text
- ❖ **Overture and Google AdWords:**
 - Overture ranks advertisers on their bids
 - Google ranks by bids * CTR
- ❖ **More data beats better algorithms**





The Big Data Makes the Scientific Method Obsolete



❖ by Chris Anderson, **WIRED**



“All models are wrong, but some are useful.” --- George Box

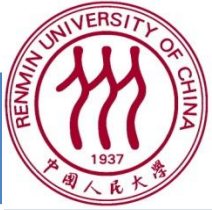


“All models are wrong, and increasingly you can succeed without them.” --- Peter Norvig





Can you succeed without models?



- ❖ **PageRank: Rank pages without knowing why one page is better than another**
- ❖ **Statistical NLP: translate languages without knowing the languages**
- ❖ **Spell Checker: you can write a reasonably good one in 24 lines of code using big data**





“All models are wrong, and increasingly you can succeed with more of them.” — Peter D. Norvig



“数据思维”

more
s that
kinds of

- In the era of big data, more isn't just more. More is different.





大数据管理面临的机遇与挑战



Science: 数据
处理专题

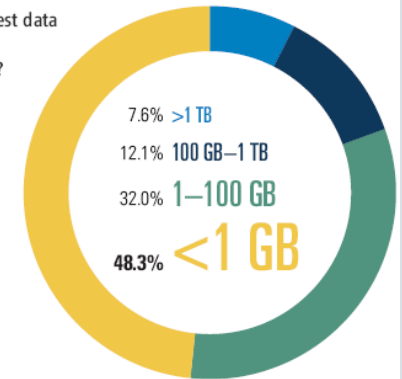
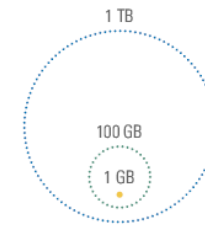
- ❖ 对许多学科而言，海量数据意味着更严峻的挑战；
- ❖ 若能更好地组织和使用这些数据，会有助我们将巨大机遇变为现实。



数据处理中的机遇与挑战

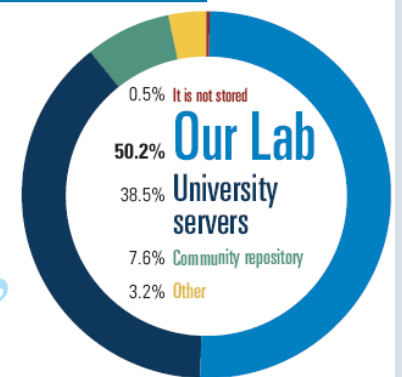
- ❖ 针对海量海量数据管理的挑战性问题，《科学》该刊的**1700**位审稿人和评论家进行了调查。
 - 大约20%人经常使用或分析的数据集超过100GB；
 - 7%的人所用的数据集超过1TB；
 - 约有50%人仅在他们的实验室中存储数据。
- ❖ 这些调查充分说明了人们在抱怨由于缺乏通用海量数据管理方法，而阻碍了数据的存储、使用和有效访问。

What is the size of the largest data set that you have used or generated in your research?



Where do you archive most of the data generated in your lab or for your research?

“Even within a single institution **there are no standards for storing data**, so each lab, or often each fellow, uses ad hoc approaches.”





特定领域中的大数据



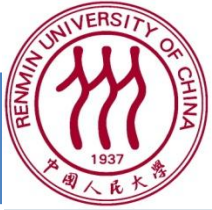
❖ 这些特定领域主要包括：

- 基因组学中的海量数据；
- 神经科学领域中的海量数据；
- 海量数据的可视化技术；
- 社会科学领域中的数据；
- 气候数据；
- 海量生态数据；

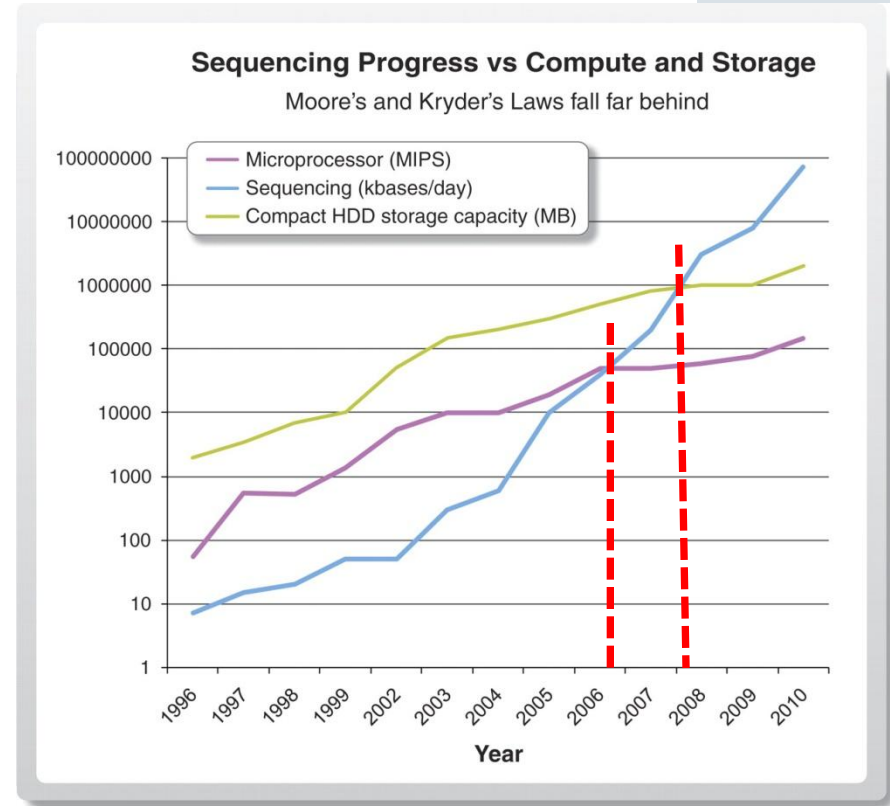




基因组学中的海量数



- 仅在一个基因序列定序器的数据增长速度：
– 10M/day-->40G/day
- 全球有**20**多个基因序列研究中心，每个中心设置多于**10**个基因序列定序器。
- 从右图可知，数据增长速度早已超越摩尔定律





神经科学领域中的海量数据



- 人的大脑大约有**800亿**个神经元，**150万亿**个突触。
- 神经科学信息框架涵盖了所有神经科学的资源，从而促进了已有知识和各种类型数据库的集成。
- 如何对这些海量数据进行组织和时空多维挖掘，存在很大的挑战。

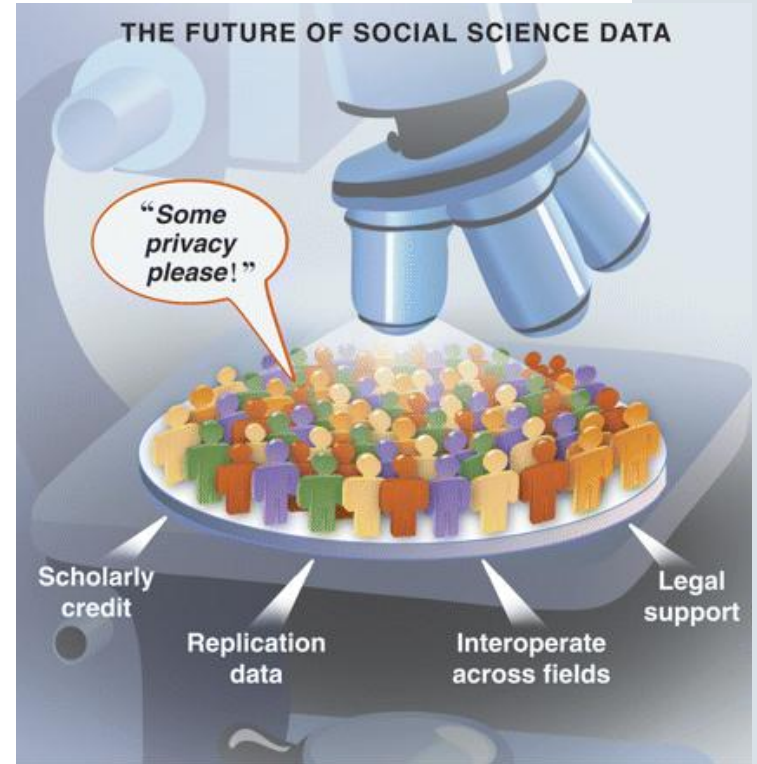




社会科学领域中的数据



- 在社会科学领域中，同样存在着海量的信息化社会科学数据。
- 未来的社会科学数据包括信用卡交易、拷贝数据、电子医疗数据等。
- 这类数据对统计方法、隐私保护方法以及伦理学的研究提出了严峻的挑战。





Some Applications



google.org Flu Trends

[Google.org home](#)

Flu Trends

[Home](#)

United States

Cities (Experimental)

Major cities

[Download data](#)

[How does this work?](#)

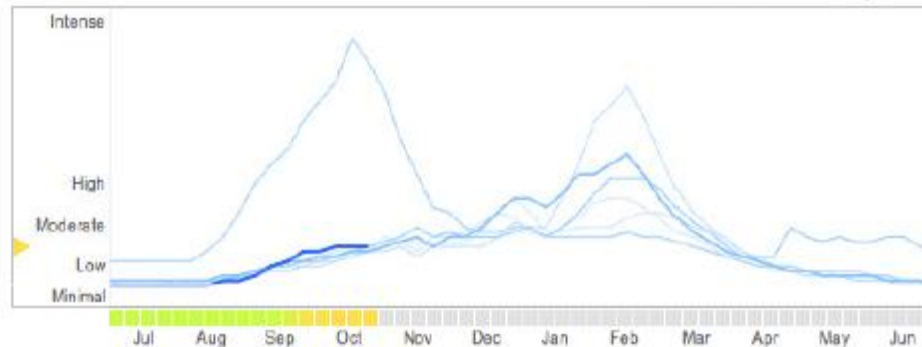
[FAQ](#)

Explore flu trends - United States

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. [Learn more »](#)

National

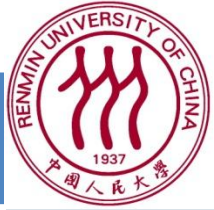
2011-2012 Past years



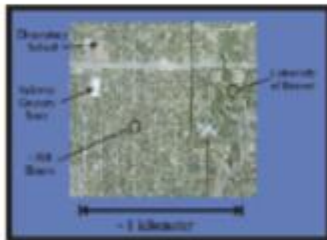
<http://www.google.org/flutrends/>



Some Applications



A Sampling of Other Use Cases



- [Disease Early Warning](#): Remote surveillance of disease and prediction of epidemics.
- [Population Census](#): Supplements traditional census and mapping in developing regions.
- [Humanitarian Crisis Mapping](#): Can detect and monitor a growing range of crisis types.
- **Water Resources**: Monitor water quality and availability and alleviate water shortage problems.
- [Food Security](#): Famine early warning, rainfall and water requirements estimations, agr production estimates and irrigation and fertilizer supply & demand.
- **Global Education Programs**



❖ 大数据特征

- 多源异构：存在较大的异质性
- 分布广泛：分布在各个区域
- 动态增长：增长快，更新快
- 数据-模式：先有数据后有模式



如何高效管理海量数据？





例：多源异构



- Syntax
 - (format)
- Schema
 - (model)
- Semantics
 - (meaning)

Study A

METADATA (from EML)		Study A: White Mountains Area col. units: sq. meter PIRU = <i>Picea rubens</i> BEPA = <i>Betula papyifera</i>		
date	site	species	area	count
10/1/1993	N654	PIRU	2	26
10/3/1994	N654	PIRU	2	29
10/1/1993	N654	BEPA	1	3

Study B

METADATA (from EML)		Study B: Green Mountains Area sampled: 1 sq. meter picrub = <i>Picea rubens</i> betpap = <i>Betula papyifera</i>	
date	site	picrub	betpap
31 Oct 1993	1	13.5	1.6
14 Nov 1994	1	8.4	1.8

Integrated Data

study	date	site	species	density
A	10/1/1993	N654	<i>Picea Rubens</i>	13.0
A	10/3/1994	N654	<i>Picea Rubens</i>	14.5
A	10/1/1993	N654	<i>Betula papyifera</i>	3.0
B	10/31/1993	1	<i>Picea Rubens</i>	13.5
B	10/31/1993	1	<i>Betula papyifera</i>	1.6
B	11/14/1994	1	<i>Picea Rubens</i>	8.4
B	11/14/1994	1	<i>Betula papyifera</i>	1.8

metadata
'promoted'
to become
data

format
normalized
using
metadata

species metadata
from study B
is now data
(picrub/betpap
column headings)

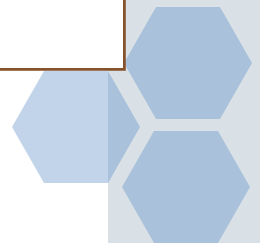
density
calculated
using
metadata





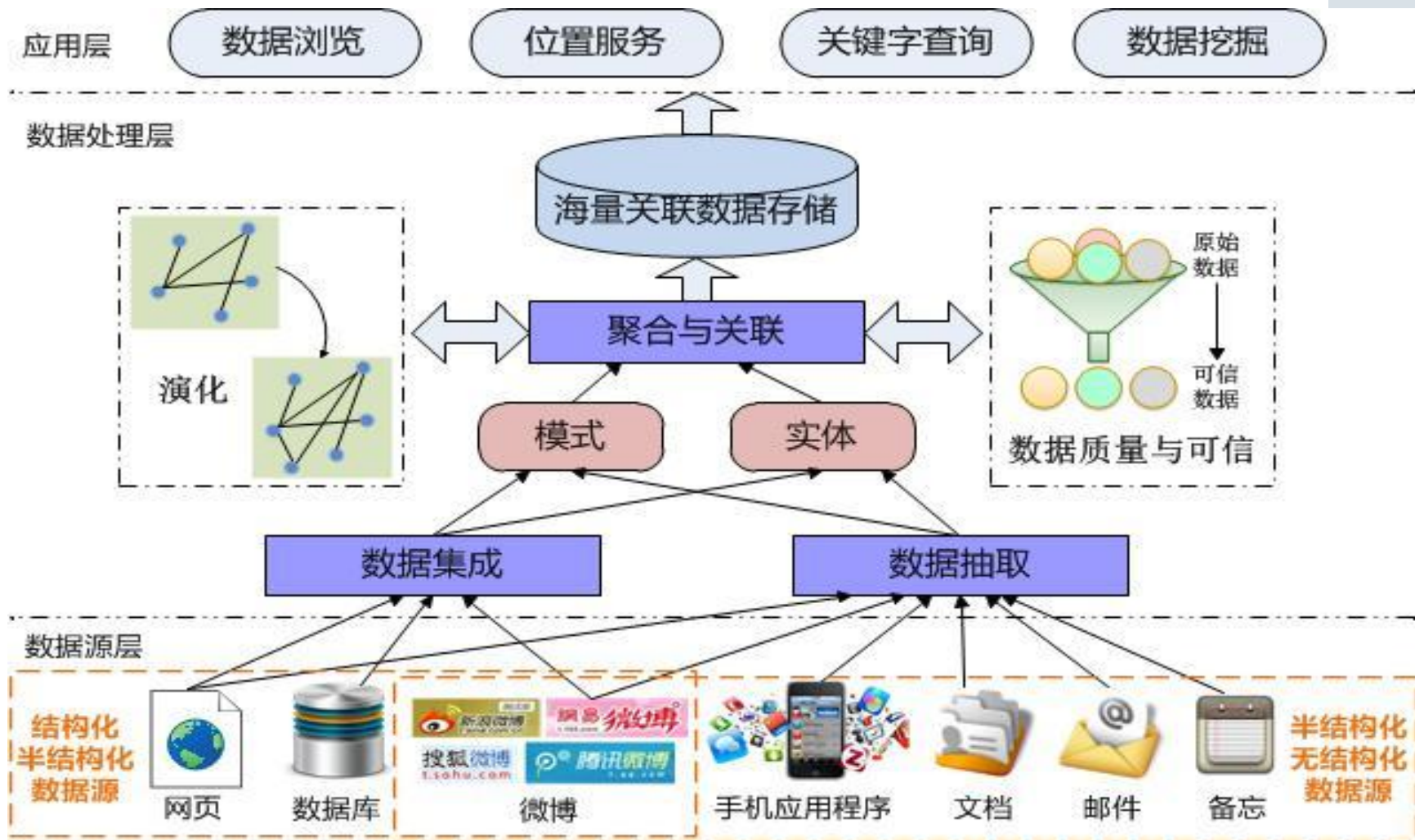
Google Scale in Hardware and Storage

- Giga 10^9 , Tera 10^{12} , Peta 10^{15} , Exa 10^{18} , Zetta 10^{21}
- Publicized: Bigtable of 70 petabytes, 10M ops/sec.
- Some representative numbers:
 - Storage: 10^{18} -> 10^{20-21}
 - Users: 10^9 -> 10^{10}
 - Network: 10^{20} , now, -> $10^{21}/\text{yr}$ (32 KB/sec. for 1B people)
- Warehouse computing possibilities





大数据管理框架

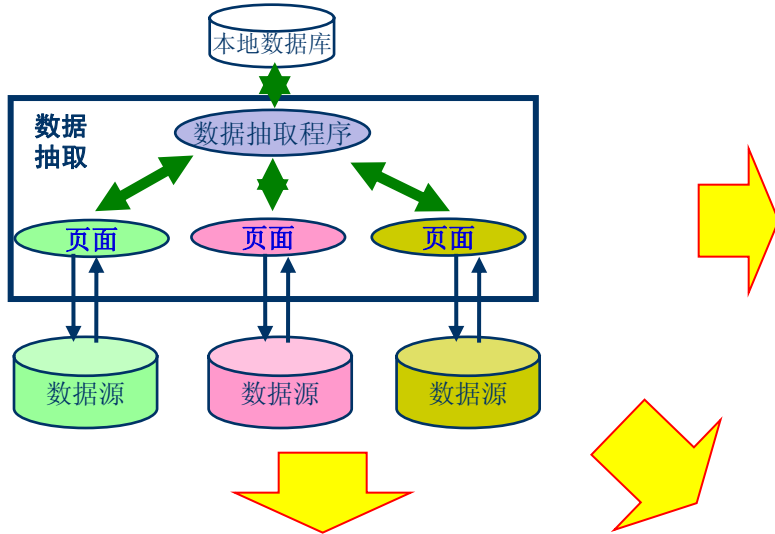




系统成果



❖ 基于数据抽取的数据集成方法



❖ 求职领域：工作通数据集成系统

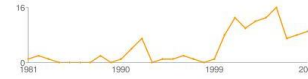


集成近**100**个数据源，数据量超过**300**万条

❖ 学术领域：计算机中文文献集成系统C-DBLP

王珊 Wang Shan

中国人民大学信息学院



共检索到结果120条 查看王珊的合作作者(共114个) 在DBLP中查找该作者文献 查看该作者的单位历史 查看项目历史

论文列表

No.	Paper Information
2008	
120	朱青 王珊 常利军: 基于概率学习导航的分布式信息查询, 第二十五届中国数据库学术会议, 2008, 计算机科学, 35 (增刊) (10A): 14-18
119	谢州市 周文静 李娜 王珊 张华: 基于数据库的视频数据随机访问方法, 第二十五届中国数据库学术会议, 2008, 计算机科学, 35 (增刊) (10A): 226-229
117	图数据库学术会议, 2008, 计算机科学, 35 (增刊) (10A): 220-225
114	王珊 肖旭芹 张斌松 陈虹: 支持What-if分析的OLAP系统研究, 计算机学报, 2008, 31 (09): 1573-1587



集成近**6**万作者的信息，日访问量**6000**次，半年累计访问超过**100**万次

❖ 新闻领域：舆情监控系统



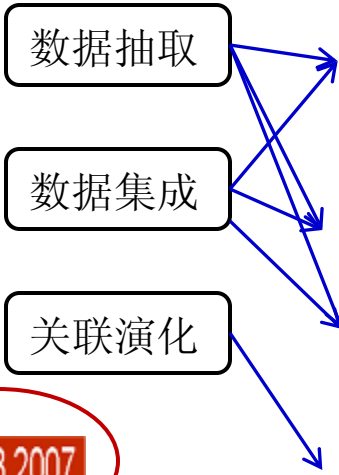
集成**5**个代表性论坛，**1000**多新闻媒体，十万个网上博客



计算机中文文献集成系统C-DBLP(1)



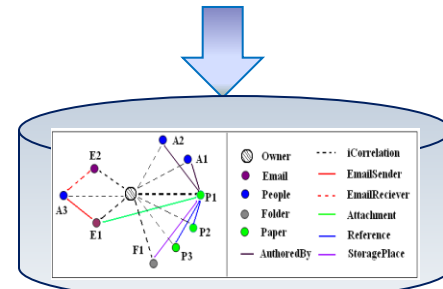
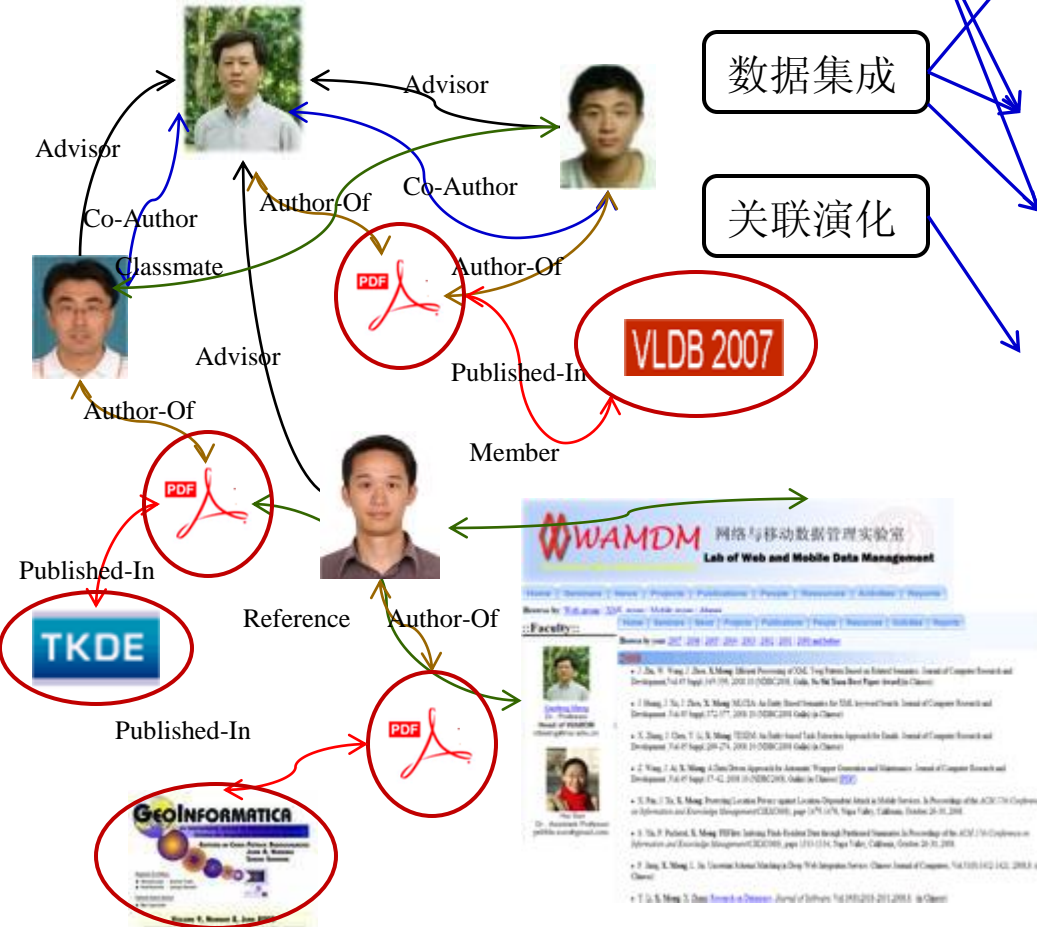
❖ 文献集成系统SearchScholar



实体:
作者, 论文, 期刊, 会议, 研究机构, ...

关联:
作者关系, 论文发表关系, 合作者关系,
隶属关系, 导师关系, 参考文献关系...

关联发现、删除、更新



- 浏览**
基于任务
- 查询**
多种形式
- 分析**
丰富多样



计算机中文文献集成系统C-DBLP(2)



王珊 Wang Shan
中国人民大学信息学院



研究兴趣:

数据查询处理 关系数据库模式 数据管理策略 数据仓库优化 移动环境数据

论文列表

承担项目历史

No.	Paper Information	项目起止年份	项目名称
120	王珊, 王珊, 董利军, 基于概率图 (JG) 14-18	2009~2011	DDAB1 调节神经
119	董利军, 周文峰, 王珊, 董利军, 王 (JG) 226-22	2005~2007	网络环境下数据
118	曹博, 王秋月, 董利军, 王珊, (JG) 32-36	2005~2007	标记的mdr1基因小鼠的研究
117	张冠松, 董利军, 孙凡, 美国国立卫生研究院, 2008, 计算机研究与发展, 45(4)	2004~2008	因特网上非规范
116	董利军, 曹博, 王珊, S-CER-2004	2004~2008	因特网上非规范
115	张冠松, 董利军, 王占伟, 奥伟, 国际数据库学会, 2008, 计算机		
114	王珊, 董利军, 张冠松, 陈红, (JG) 14-18		
113	董利军, 孙小舟, 周玉, 王珊, (JG) 14-18		
112	曹博, 王珊, 董利军, 王秋月, (JG) 14-18		
2007			
111	王珊, 董利军, 董利军, 基于安全 (JG) 14-18		
110	王珊, 曹博, 基于计算机网络分析的语音通信信号处理, 计算机工程, 2008		
109	李俊强, 李翠平, 王珊, 杨皓, 一种PAML方法, 第二十四届中国数据库学术 (JG) 349-352		
108	曹博, 董利军, 王珊, DETECTOR: 基于关系数据库通用的在线入侵检测 (JG) 5		

单位历史信息

中国人民大学信息学院
中国人民大学数据与知识
中国人民大学信息系 (1)

合作作者列表

33	周立柱	[10]
34	周胜	[24]
35	周龙繁	[73]
36	唐元晶	[13] [15]
37	姚佳丽	[101] [108] [62] [9]
38	姚娜达	[22]
39	孙冠凡	[56]
40	孟小峰	[23] [51] [47] [20] [9] [37] [53] [41] [35] [69] [73] [92] [103] [36] [33] [28] [30] [50] [46] [57] [64] [79] [54] [42]
41	李晓东	[115]
42	程定春	[22]
43	董利军	[120]
44	董利军	[45]
45	孙冠凡	[21] [19]
46	董俊	[91] [116] [106]



输入学者名字, 可以查询出其发表的文章, 参加的科研项目、研究兴趣、科研成果分析、合作者情况等大量信息。



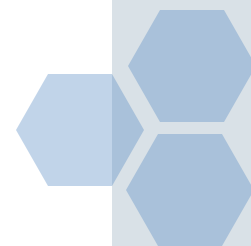
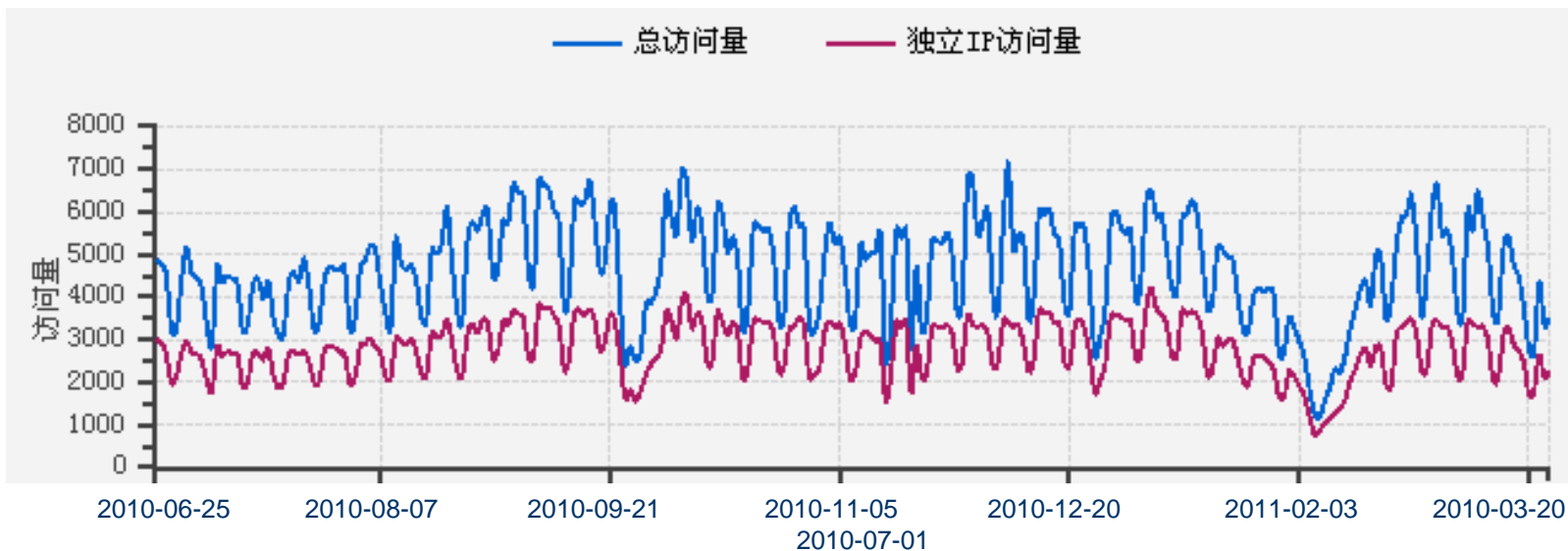


计算机中文文献集成系统C-DBLP(4)



❖ 文献集成系统SearchScholar

- ❖ 12个期刊(1960.01-2011.02), 1 个会议(NDBC 2000-2011), 8万多篇论文, 6.8万多个作者
- ❖ 日访问量超过5000次, 累计访问超过350万次





成果意义

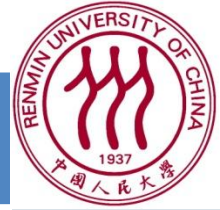


- 建立了一种将数据结构化管理的途径，为解决特定领域的大数据集成问题奠定了基础
- 进而为大数据管理提供一种新的解决思路





日本海啸





数据海洋中的海啸



- ❖ 数据海啸一： **Web**网面海量数据
- ❖ 数据海啸二： 微博海量数据
- ❖ 数据海啸三： 移动**App**海量数据

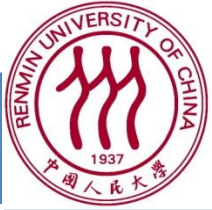


数据海啸：微博海量数据





社会的数字化与数字的社会化



❖ 社会的数字化：数据足迹（**data print**）

- 在数字化时代，各色人等有意无意留下的数据足迹越来越丰富
- 数据足迹是有社会意义（**social meaning**）的，蕴含着社会结构

❖ 数字的社会化：

- 数据足迹及其结构本身就是社会结构和过程的一个环节，不断塑造着新的社会秩序和关系



数据思维：计算社会科学

- ❖ 一切社会解释、监控、预测与规划都离不开对数据足迹的收集、整理和分析
- ❖ 计算社会科学方法：
 - 基于特定社会需要，在特定社会理论指导下，收集、整理和分析数据足迹，以便进行社会解释、监控、预测与规划的过程和活动

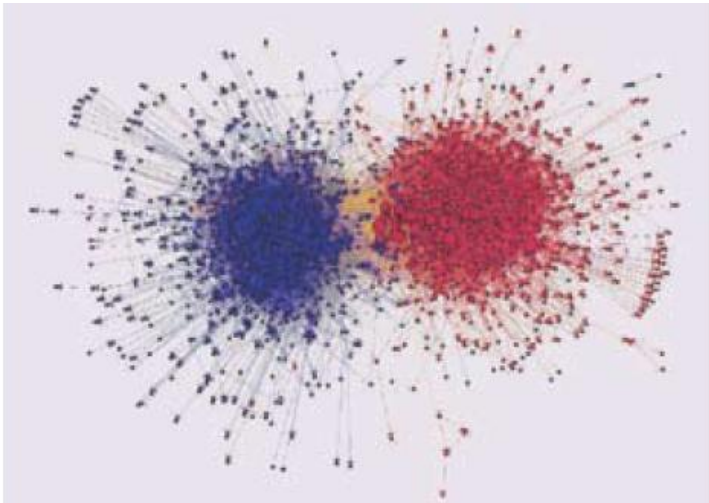


图1 来自博客（blogosphere）空间的数据：上图是一个政治博客群的链接结构（从2004年开始），红色节点代表保守派，蓝色节点代表自由派。橙色链接从自由派博客指向保守派，紫色链接反之。每个博客节点的大小反映了指向该博客的其他博客的数量。复制自文献[8]。



总结和展望

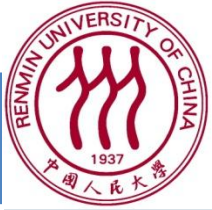


寻找应对数据海啸的方舟.....





总结



- ❖ 积累数据财富应该成为我们的习惯，它或许也会成为国力的标志
- ❖ 研究“数据思维”的方法，或许会有意义，是下一个十年我们面临的机遇期





谢谢!

未来每**18**个月产生的数据量
等于有史以来的数据量之和

--**Jim Gray**
1998图灵奖获奖演说

