

Web信息处理与应用

第一节 绪论

徐童 2021.9.6



授课教师：徐童 副教授

研究方向：数据挖掘与社交媒体分析

个人主页：<http://staff.ustc.edu.cn/~tongxu/>



课程主页：<http://staff.ustc.edu.cn/~tongxu/webinfo>

课程邮箱：ustcweb2021@163.com

课程QQ群：927315194

- 当我们谈到Web信息处理与应用，我们在谈论什么问题？



案例：《长安十二时辰》中的“**大案牍术**”

- 当我们谈到Web信息处理与应用，我们在谈论什么问题？

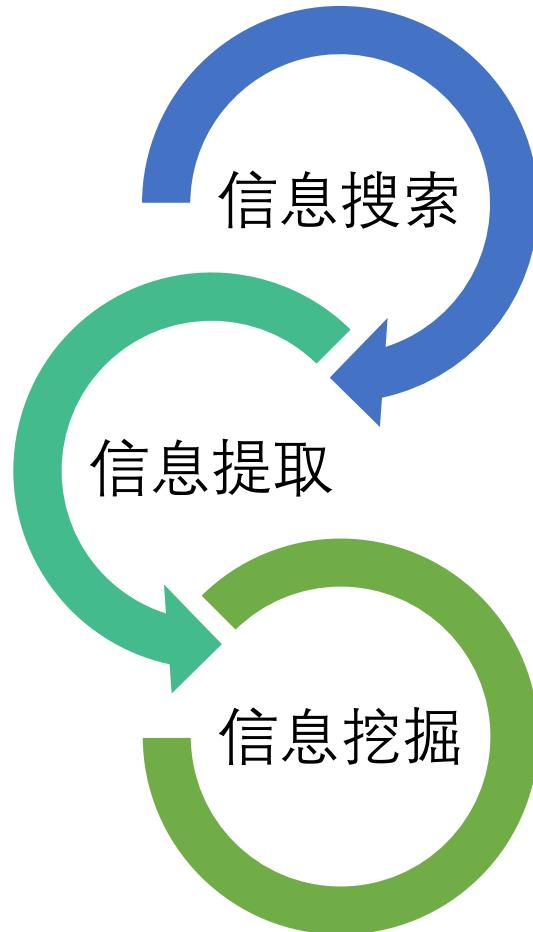


- 问题输入：卷帙浩繁的机密文件
- 问题输出：满足需求的特定人物

解决问题，共分几个步骤？

- 问题输入：卷帙浩繁的机密文件
- 问题输出：满足需求的特定人物

如何解决？



从海量数据中找到可能有用的文档

从目标文档中提取和关联价值信息

通过挖掘被提取信息完成最终决策



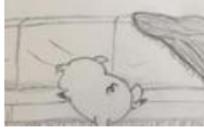
- 从盛唐回到现代，今天的人们在搜索什么？



我们已置身于万物互联的时代，Web浪潮无法回避

- 案例1：文档搜索

The screenshot shows a Baidu search results page. The search query 'Web信息处理' is entered in the search bar. Below the search bar is a navigation menu with tabs: 网页 (selected), 资讯, 视频, 图片, 知道, 文库, 贴吧, 采购, 地图, 更多». The main search results area displays three items:

- Web信息处理与应用复习笔记-GitHub.PDF**
2018年12月2日 - Web信息处理与应用复习笔记-GitHub.PDF, Web信息处理与应用复习笔记
©2017-1熊家靖PB14011026PART1:WebSearch一、Introduction1、web搜索的挑战:数据规...
<https://max.book118.com/html/2...> - 百度快照
- 【专利】WEB信息处理方法及装置_百度学术**
2012
本发明实施例公开了一种WEB信息处理方法及装置,涉及WEB信息处理领域,能够将一个或者多个统一资源定位符对应的预定WEB信息按照预设的规律排行。包括:获取待处理信息,所述...
xueshu.baidu.com - 百度快照
- 基于web的信息处理系统 - 简书**
 2019年5月15日 - 基于web的信息处理系统jenslee 2019.05.15 10:08
字数144 欢迎关注及点赞,后续陆续公布不同源代码实现教学视频。联系微信zlee_com_cn或者扫描头像二维码,备注:...
[简书](#) 简书社区 - 百度快照

基于搜索关键词，寻找最相关的文档，是信息检索的基本任务。

• 案例2：多模态搜索



从单一的文本信息到更为复杂的多模态搜索，任务与方法都在拓展

- 案例3：面向知识的搜索



人们已不再满足于单纯呈现原始的文档，
而需要更加精炼的知识表达与更加直观的需求解决

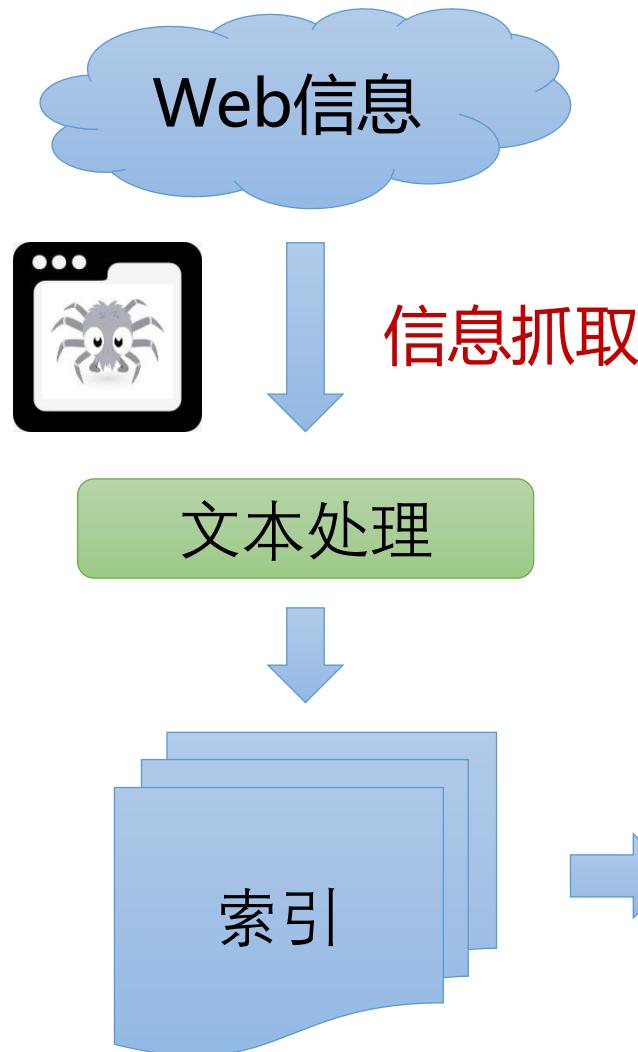
- 案例4：基于分析的搜索

The screenshot shows the TianYanCha.com website interface. At the top, there's a navigation bar with tabs for '查公司' (Search Company), '查老板' (Search Owner), and '查关系' (Search Relations). The search bar contains the text '珠海格力电器股份有限公司'. Below the search bar, there are several data summary sections: '上市信息 999+' (Listed Information), '公司背景 303' (Company Background), '司法风险 924' (Judicial Risk), '经营风险 85' (Operational Risk), '公司发展 82' (Company Development), '经营状况 999+' (Operational Status), '知识产权 999+' (Intellectual Property), and '历史信息 330' (Historical Information). A prominent blue sidebar on the left is labeled '官方信息' (Official Information) and '自主信息' (Self-owned Information), with the number '29' indicating the count of self-owned information items. The main content area displays a table titled '竞品信息 50' (Competitor Information) with 5 rows of data. The columns are: 序号 (Rank), 产品名称 (Product Name), 当前融资轮次 (Current Financing Round), 估值 (Valuation), 成立日期 (Establishment Date), 产品标签 (Product Tag), 所属地 (Location), and 简介 (Introduction). The data rows are as follows:

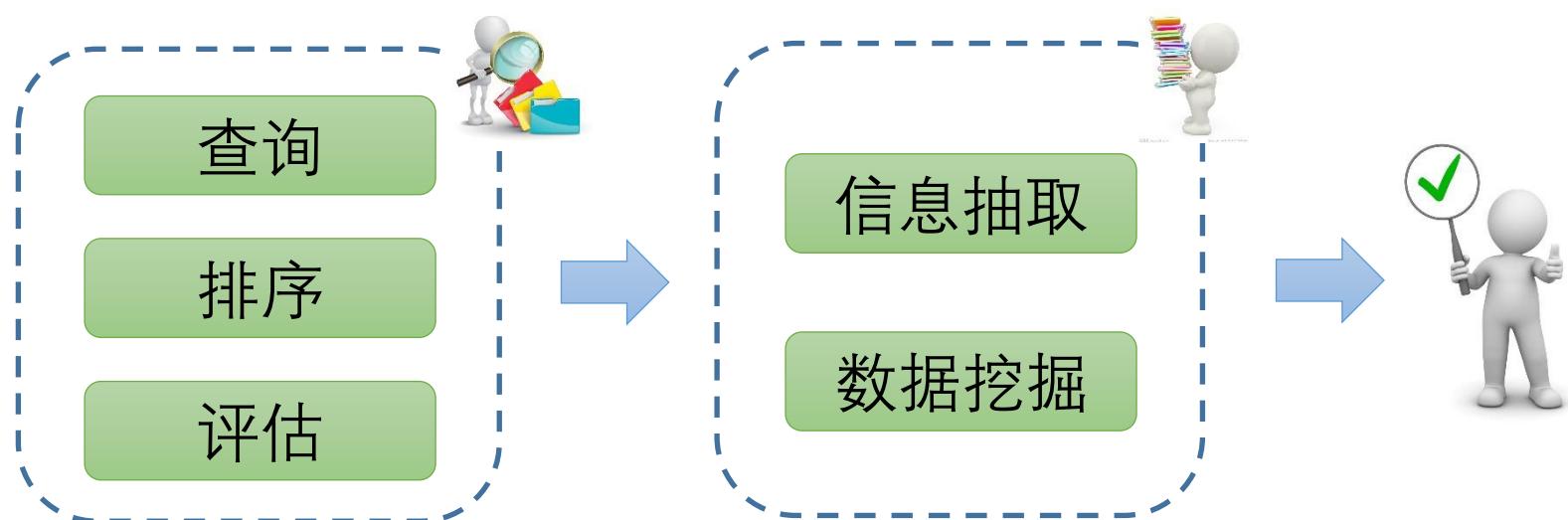
序号	产品名称	当前融资轮次	估值	成立日期	产品标签	所属地	简介
1	老板电器	定向增发	-	2010-12-18	生产制造	浙江	厨房电器生产制造商
2	华自科技	定向增发	-	2009-09-25	生产制造	湖南	水电控制设备系统研发商
3	中国西电	定向增发	-	2008-04-30	生产制造	陕西	输配电及控制设备研发制造商
4	飞科电器	定向增发	-	2006-06-10	生产制造	上海	个人护理电器产品研发生产商
5	*ST圣莱	定向增发	-	2004-03-11	生产制造	浙江	温控器及电热水壶研发、生产和销售

人们需要直接从文档中获得的信息，
更需要从信息中分析和总结出的规律与结论

- 本课程所要解决的问题

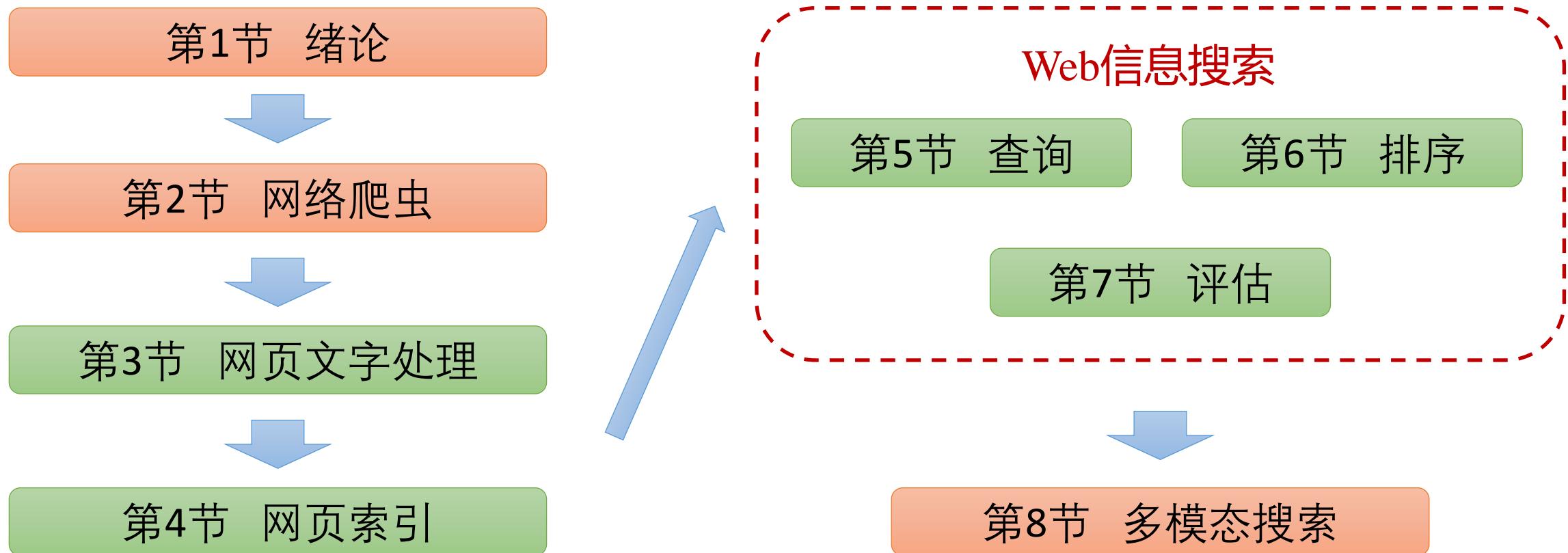


- Web信息如何获取? 【网络爬虫】
- Web信息如何整理与存储? 【文本处理、索引】
- Web信息如何搜索? 【查询/排序/评估】
- 如何提炼价值信息与知识? 【信息抽取】
- 如何分析信息并支撑应用? 【网络数据挖掘】



- 围绕“检索”、“抽取”与“挖掘”三条主线

第一部分：Web信息处理与检索



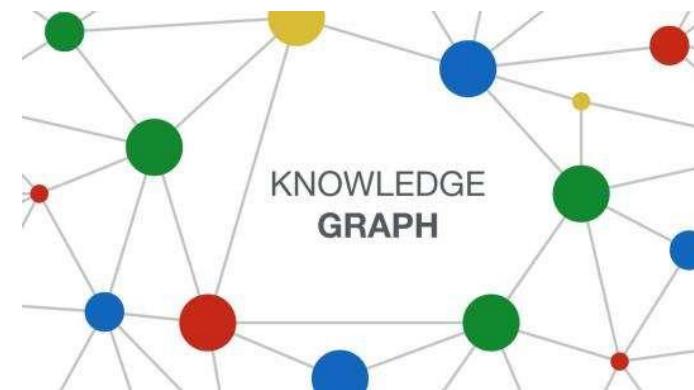
- 围绕“检索”、“抽取”与“挖掘”三条主线

第二部分：Web信息抽取与知识图谱

第9节 实体识别



第10节 关系抽取



- 围绕“检索”、“抽取”与“挖掘”三条主线

第三部分：面向Web信息的数据挖掘

第11节 数据准备

基本数据挖掘方法

第12节 分类算法

第13节 聚类算法

Web信息应用

第14节 推荐系统

社会网络分析方法

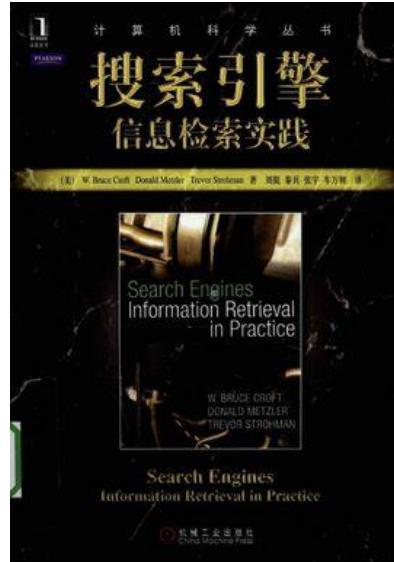
第15节 社会网络

第16节 社会传播

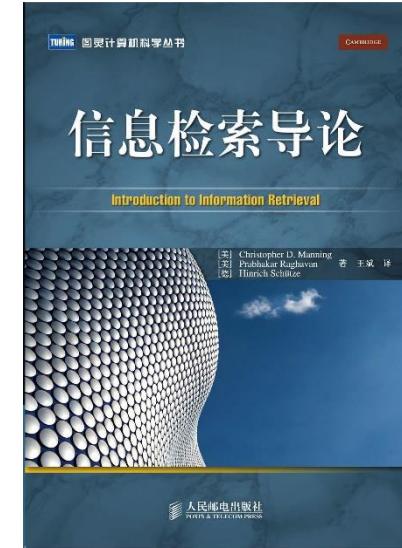
- 课程形式：讲课+实验
 - 理论学时60学时 + 实验学时30学时（无单独实验课安排）
- 成绩组成
 - 60%考试成绩+30%实验成绩+10%作业成绩（部分章节）
- 实验安排
 - 信息检索、信息抽取、信息挖掘各一次实验
 - 实验大约将在9、10、11每个月的下旬布置，每次为期4周
 - 鼓励两人一组，可以单人进行（无优惠政策）
- 考试形式：半开卷（一张A4纸）

参考书目

- 参考书目



搜索引擎：信息检索实践，
W. Bruce Croft 等著，刘挺 等
译，机械工业出版社



信息检索导论，Christopher
D. Manning 等著，王斌 译，
人民邮电出版社

可参考部分领域内重要会议论文：SIGIR、KDD、CIKM、WWW, etc.

- **Web 信息基础**

- Web信息起源

- Web搜索发展史

- Web搜索的挑战

- 信息检索概述

- 数据挖掘概述

起初阿帕创造阿帕网络。

阿帕网络是空虚混沌。渊面黑暗。

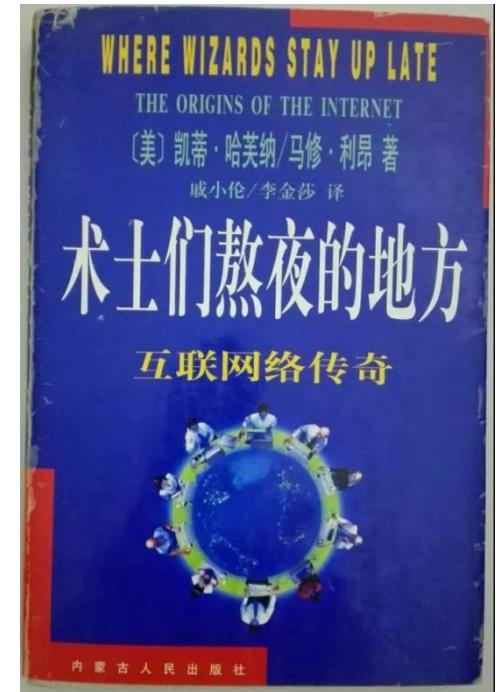
阿帕的灵运行在网络里面。阿帕说：‘要有一个协议。’ 就有了一个协议。阿帕看它是好的。

阿帕说：‘要有更多的协议。’ 事就这样成了。

阿帕看这是好的。

阿帕说：‘要有更多的网络。’ 事就这样成了。

——丹尼·科恩，《第一行动》



- Web信息起源：1965，超文本概念提出
- Ted Nelson在1965年提出了超文本的概念。
 - HyperText，源自于“非连续性著述”（Non sequential writing）的理念，即分叉的、允许读者作出选择的文本。
 - 以海量数据为基础，使原先的线性文本变成无限延伸、扩展的非线性文本。
- 超文本传输协议（HTTP，HyperText Transfer Protocol）
- 超文本标记语言（HTML，HyperText Markup Language）



- Web信息起源：1969，因特网起源
- 1969年，互联网的原型**ARPANet**由美国国防部研究计划署（DARPA）所制定的协定下诞生，首先用于军事连接。
 - 起初只有4个结点，分布在UCLA等四所大学的4台大型计算机。
 - ARPANet的试验较好地解决了异种机网络互联的一系列理论和技术问题，并推动了TCP/IP协议的诞生（1983）。
- 1986年，美国国家科学基金会（NSF）建立**NSFNet**广域网，逐渐取代了ARPANet。



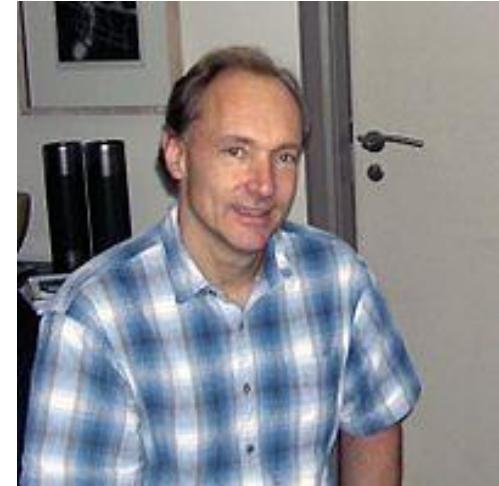
NSFNET T3 Network 1992



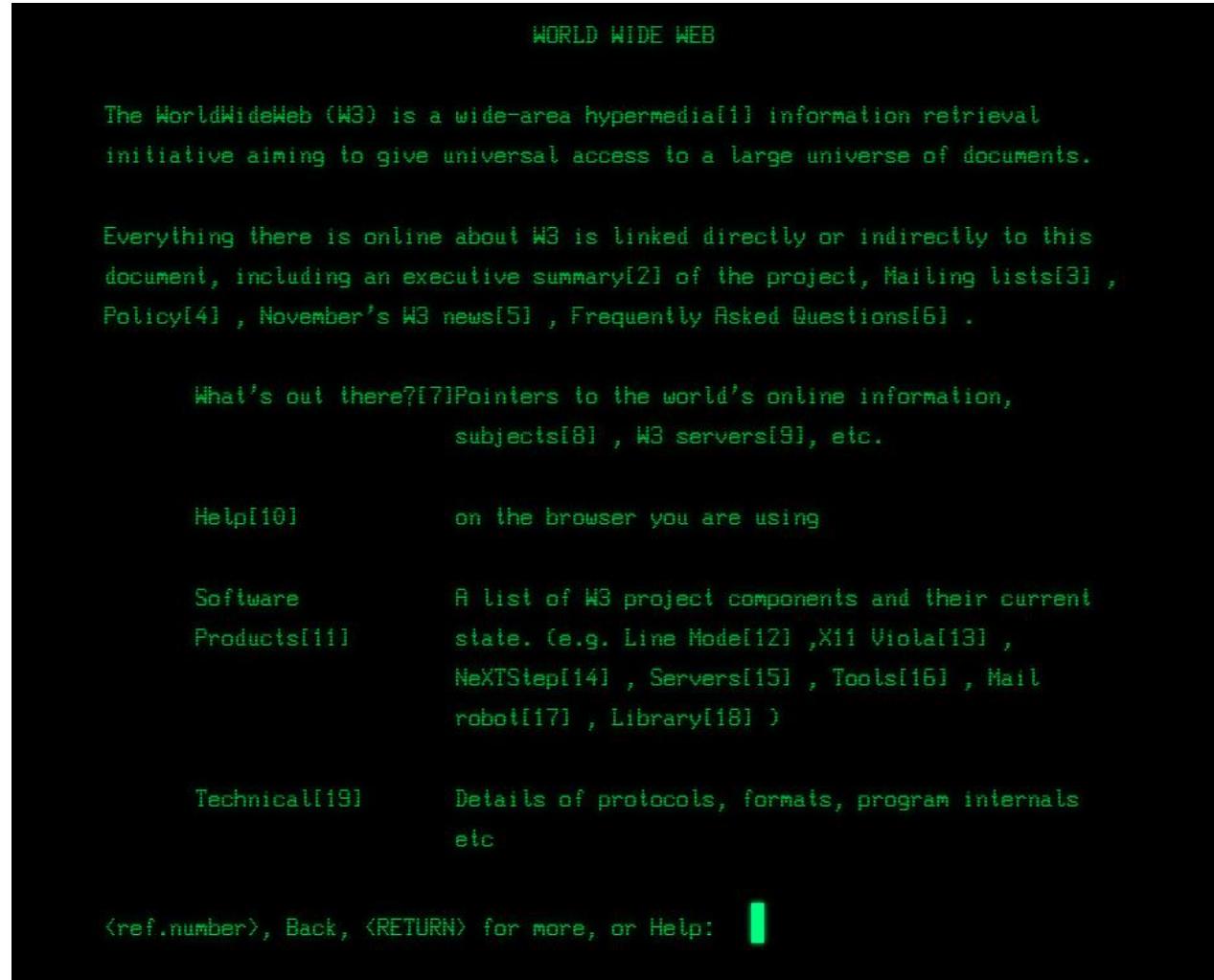
- Web信息起源：小插曲，中国与互联网
- 1987年，在德国和中国间采用CSNET协议建立了email连接。
- 1987年9月20日，北京计算机应用技术研究所王运丰教授从北京向海外发出中国第一封电子邮件。
 - Across the Great Wall, we can reach every corner in the world (越过长城，我们可以到达世界的每一个角落)
 - 另一种说法：1986年8月25日由高能物理所吴为民教授发出第一封电子邮件



- Web信息起源：1989，万维网诞生
- 1989年，欧洲核子物理研究所（CERN）的Tim Berners Lee（万维网之父）等人首次提出了一个分类互联网信息的协议，即World Wide Web
 - 在1990年，他写出了第一个网页：
<http://info.cern.ch>
 - 他定义了URLs、HTML、HTTP等的规范，使网络能够为大众所使用。
 - 他创立了万维网联盟（World Wide Web Consortium, W3C）并担任主席



- Web信息起源：世界上第一个网页



- Web的五个特点

- **图形化**

- 将文字、图形、音频、视频等多模态信息集合于一体

- **平台无关**

- 访问万维网对系统平台没有约束

- **分布式**

- 信息分布于不同的站点，在物理上分开，在逻辑上一体

- **动态性**

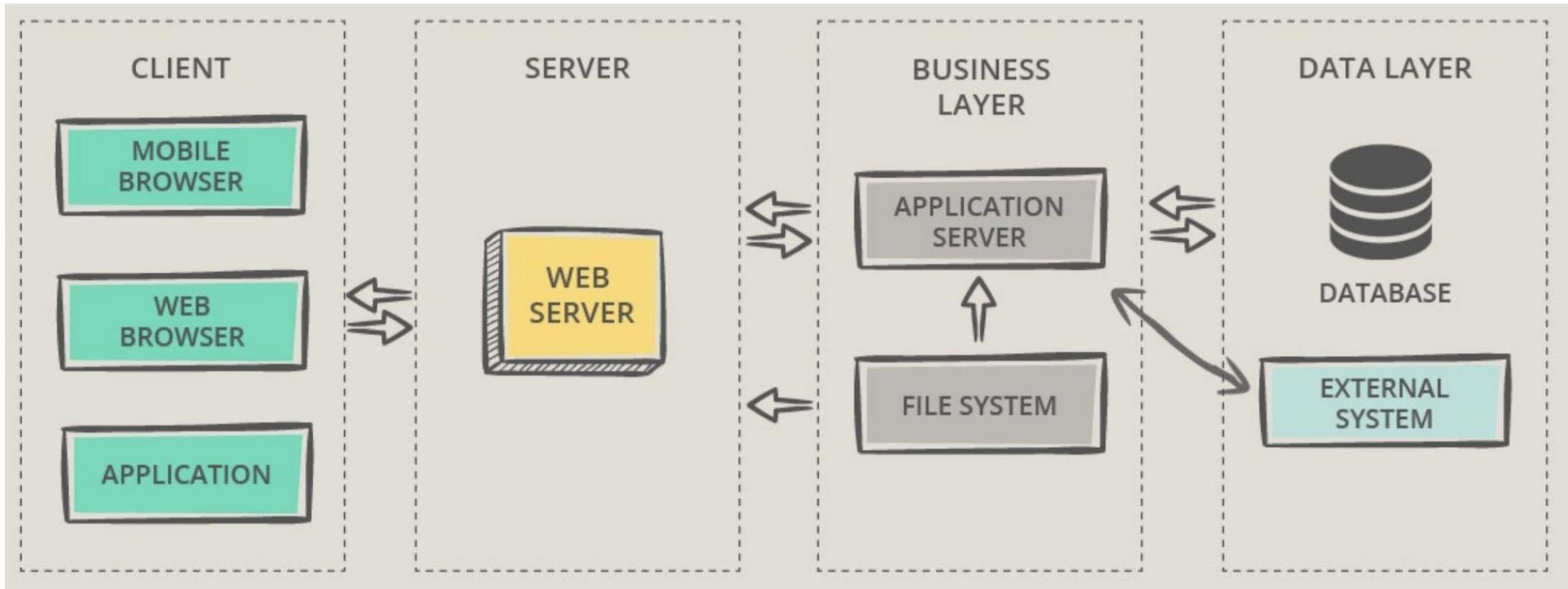
- Web站点上的信息是动态的、实时更新的

- **交互性**

- 用户提交需求，服务器根据需求反馈相应信息

Web特点

- Web系统的一般结构



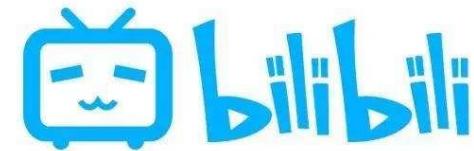
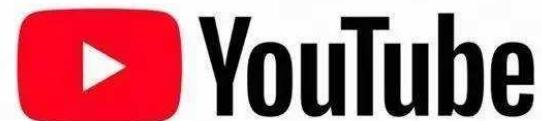
Web特点

- Web的信息流视角：Web 1.0时代



Web特点

- Web的信息流视角：Web 2.0时代



我们每个人，既是信息的生产者，也是信息的消费者

- Web的信息流视角：Web 3.0时代，会是怎样？



一种猜想：更加个性化、更加智能化、跨越平台与站点的信息大统一

- **Web 信息基础**

- Web信息起源

- **Web搜索发展史**

- Web搜索的挑战

- 信息检索概述

- 数据挖掘概述

- 搜索引擎的原型是怎样的?

电话簿 (黄页)



搜索起源

- 黄页的Web化：搜索引擎的雏形



The screenshot shows a search results page for 'Business & Economy'. The top navigation bar includes categories like 'Arts & Crafts', 'Body Building', 'Building Material', 'Chemicals', 'Communication', 'Consulting & Translation', 'Design', 'Electrical appliance', 'Employment', 'Fabrics & Textiles', 'Fair', 'Food & Medicine', 'Forest', 'Garment & Accessories', 'Hotels & Tourism', 'International Trade', 'Investment Projects', 'Law Firm', 'LCD Manufacture', 'LCD Material', 'Machinery & Tools', 'Magnet Material', 'Office Supplies', 'Packaging', 'Real Estate', 'Sports', 'Tourism', 'Travel', and 'Wholesale'. Below this is a 'Business & Economy' section with a grid of links. Further down are sections for 'Arts & Crafts' (listing companies like Shandong Linyi Ceramics Group Corp., Teloon Tennis Ball co., Ltd., Globe Glasses Factory Co., Ltd., and Xiaoqing Flowers Center) and 'Sports'.

由马云创立于1995年，用于传播中国新闻，并发布企业名录和信息

- 搜索引擎发展史：1990年，Archie
- Archie：一般公认最早的搜索引擎
 - 由麦吉尔大学的Alan Emtage等几位学生发明，用于搜索互联网上的匿名FTP
 - Archie依靠脚本文件搜索互联网上的匿名FTP（无需登录信息），然后根据用户需求反馈相应的文件，它的实质是一个可搜索的FTP文件名列表。
- 目前仍有少量提供Archie服务的网站

由波兰华沙理工大学提供的Archie

http://archie.icm.edu.pl/archie-adv_eng.html

Archie Query Form 

Search for:

Database: Worldwide Anonymous FTP Polish Web Index
Search Type: Sub String Exact Regular Expression
Case: Insensitive Sensitive

Do you want to look up strings only (no sites returned):
 NO YES

Output Format For Web Index Search: Keywords Only
 Excerpts Only Links Only

- 搜索引擎发展史：1993年，Wanderer

- Wanderer：最早的爬虫
 - 由MIT的学生Matthew Gray设计
 - 原意用于统计互联网上服务器的数量，而非为搜索引擎所设计
- Wandex：最早的网页索引计划
 - Wanderer后来发展为可以捕获网址，而为这些网址建立索引的计划就是Wandex

- 其他诞生于1993年的Robots
 - ALIWEB（Archie-like Index of WEB，发表于首届WWW会议）
 - WWW Worm，收集了海量多媒体文件，并可通过关键词检索



Wandex

- 搜索引擎发展史：1994年，Yahoo！
- 1994年，最老的“分类目录”搜索引擎之一Yahoo诞生
 - 由美籍华人Jerry Yang（杨致远）与David Filo所共同创造
 - 最早的Yahoo的数据是手工输入的，实际上只是一个可搜索的目录
 - 1995年，Yahoo网站正式上线

1995年的Yahoo！



- 搜索引擎发展史：1994年， Lycos
- 1994年诞生，搜索引擎中的元老，是最早提供信息搜索服务的网站之一
- 通过前缀匹配与字符近似匹配，提供网页自动摘要和相关性排序，数据量较大，整合了搜索数据库、在线服务和其他互联网工具，
- 2000年被西班牙网络集团收购后，目前是全世界最大的西班牙语搜索引擎

早期的Lycos



现在的Lycos



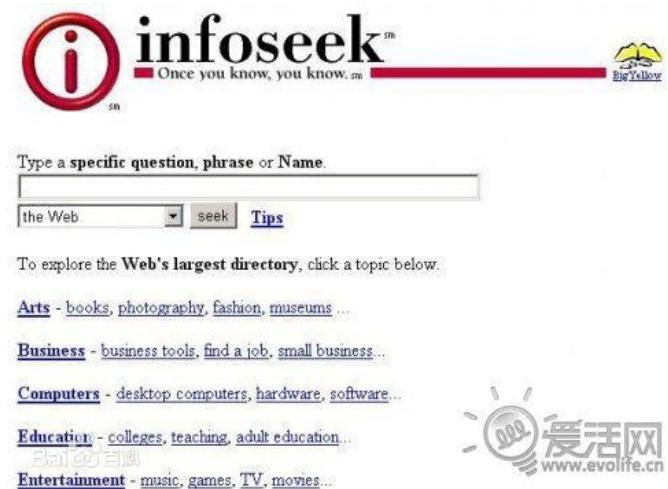
- 搜索引擎发展史：1994年，Infoseek

- 1994年诞生，沿袭了Yahoo! 与Lycos 的概念。
- 1995年，与网景公司（Netscape）的战略性协议实现强强联合
- 2001年2月，Infoseek改用Overture的搜索结果
- 李彦宏曾担任Infoseek核心工程师，主导了Infoseek的革新换代

折戟沉沙
的网景



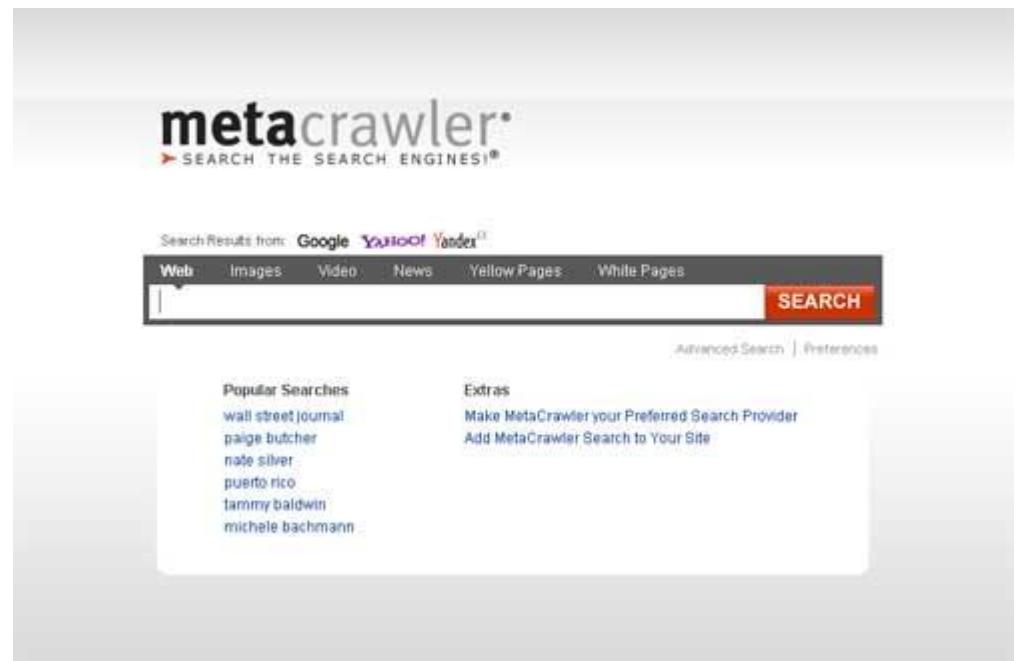
Infoseek



- 搜索引擎发展史：1995年，Metacrawler

- 1995年，第一个元搜索引擎诞生，由华盛顿大学的两位硕士生共同开发
- 元搜索引擎的概念（Meta Search Engine Roundup）
 - 用户提交搜索后，由元搜索引擎负责转换处理，然后提交给多个预先选定的独立搜索引擎
 - 各独立搜索引擎返回查询结果后，再集中处理并返回给用户

Metacrawler



- 搜索引擎发展史：1995年，Metacrawler

中国的元搜索：

曾经的百Google度



在他们之间平均**85%**链接均不相同

百Google度搜索

内容来自**百度** 和 Google中文 | 把 BaiGoogledu 设为首页

Baigoogledu.com ©2005-2013 建议和联系

京ICP备11039618号-2

现在的曾经的三百搜 (360+百度+搜狗)



两个搜索引擎之间平均 **85%** 链接均不相同

 NEW 网盘 网页 微信 微博 新闻 图片 视频 购物 音乐
 网盘搜索 手写

[网站导航](#) [我的网址](#) [常用查询](#) [热门搜索](#)

搜索起源

- 搜索引擎发展史：1995年，Altavista

- 第一个支持自然语言搜索的引擎
- 第一个实现高级搜索语法的引擎
 - AND、OR、NOT等
- 2003年，Altavista被Overture收购，后者是Yahoo的子公司

Altavista



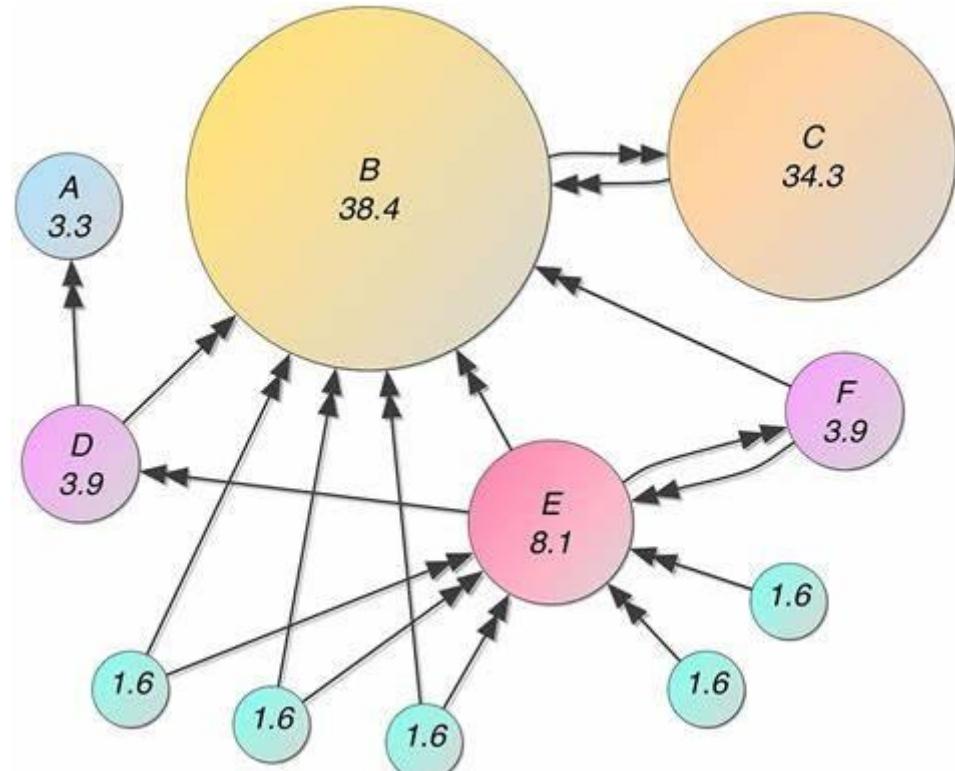
Business Services Submit a Site About AltaVista Help

© 2003 Overture Services, Inc.

- 搜索引擎发展史：1997年，Google

- 1997年，全球最大的搜索引擎
Google诞生。
 - 1995年，Larry Page (PageRank因此得名) 来到斯坦福攻读博士，并开始研究网络链接项目
 - 他与Sergey Brin提出了PageRank技术，并用于搜索引擎，从而改写了搜索引擎的定义
 - 1997年，Google.com域名被注册，1998年，Google公司正式成立

Google赖以起家的PageRank技术



- 搜索引擎发展史：1997年，天网
- 国内第一个基于网页索引搜索的搜索引擎，见证了中国互联网发展史
 - 由北京大学网络实验室研究开发，是国家重点科技攻关项目"中文编码和分布式中英文信息发现"的研究成果。
 - 于1997年10月29日正式在CERNET上向广大互联网用户提供Web信息搜索及导航服务
 - 教育网优势，FTP搜索功能强大

北大天网搜索



- 搜索引擎发展史：2000年，百度
- 2000年，由前Infoseek资深工程师李彦宏创立
 - 专注于中文搜索领域，目前是最大的中文搜索引擎
 - 2003年，根据某在线调查，百度已超越Google成为中国网民首选的中文搜索引擎

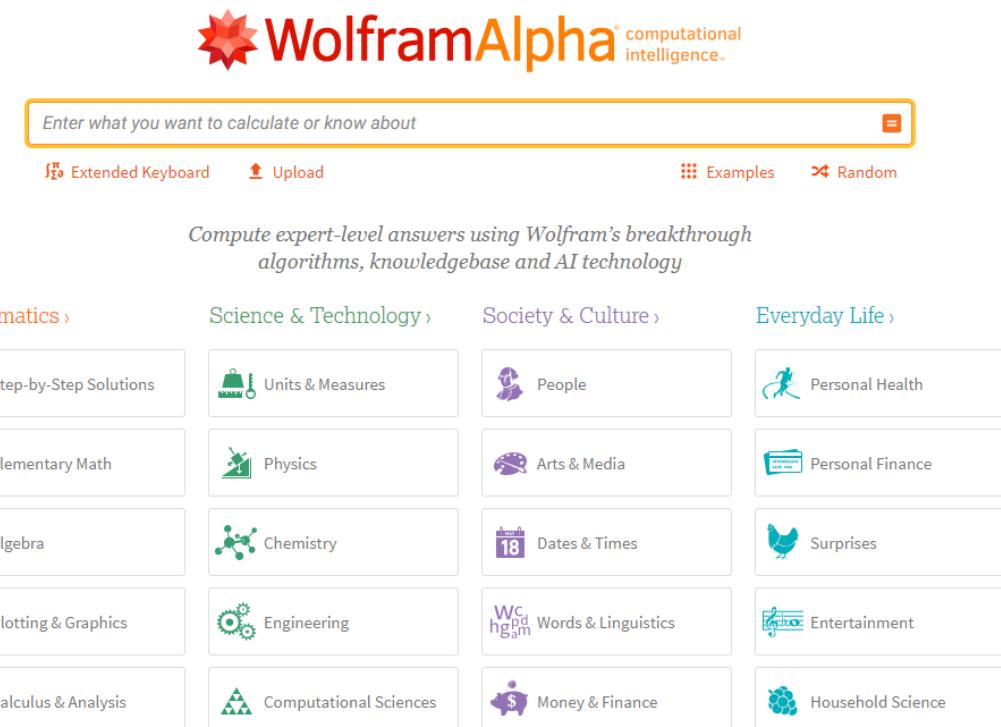


【萌】你不会百度吗？

P.S. 骂人是不对的.....

- 搜索引擎发展史：下一代搜索引擎？

- 2009年，Wolfram Alpha上线
 - 搜索引擎？计算知识引擎！
 - 直接向用户返应回答，而不是返回网页链接
 - 倘若输入“抛10次，4次正面向上”，它可以回答抛硬币的概率问题。甚至连某地下一次日食的时间，或者国际空间站现在的位置，它都能给你答案



- 搜索引擎发展史：下一代搜索引擎？
- 2016年，微软小冰读心术
 - 如何明确用户的检索需求？
 - 通过若干连续问题确认用户的真实意图，避免歧义干扰
 - 本质是[决策树](#)的应用（详见第十二周课程）
 - 背后有庞大的数据库支撑
 - 如何设计提问策略是核心问题



我想



车智澈
青瓦台天文学家

- **Web 信息基础**

- Web信息起源
 - Web搜索发展史
 - **Web搜索的挑战**
-
- 信息检索概述
 - 数据挖掘概述

挑战来源

- 来自三方面的挑战：数据、用户、利益



飞速增长的海量数据

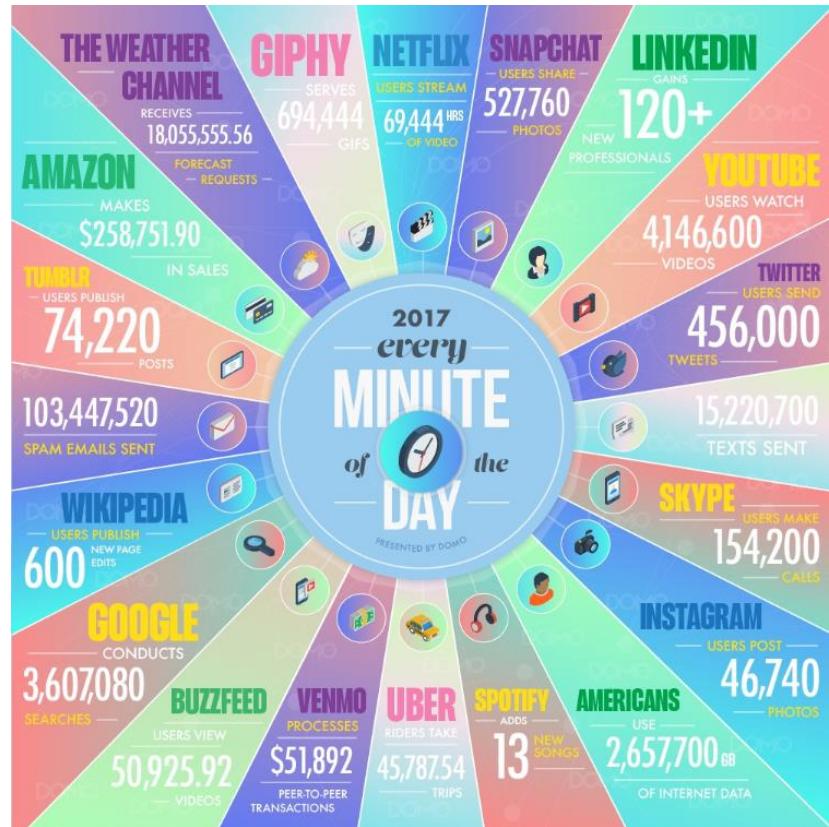
复杂多变的用户需求



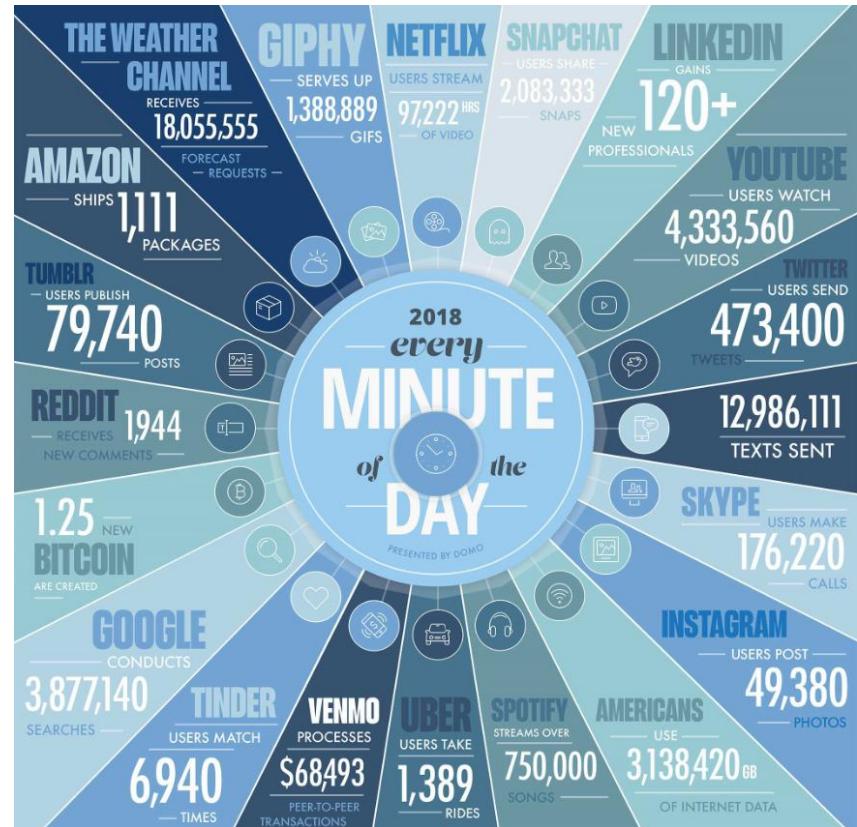
技术让位于利益

- 来自数据的挑战：海量数据

2017年版



2018年版



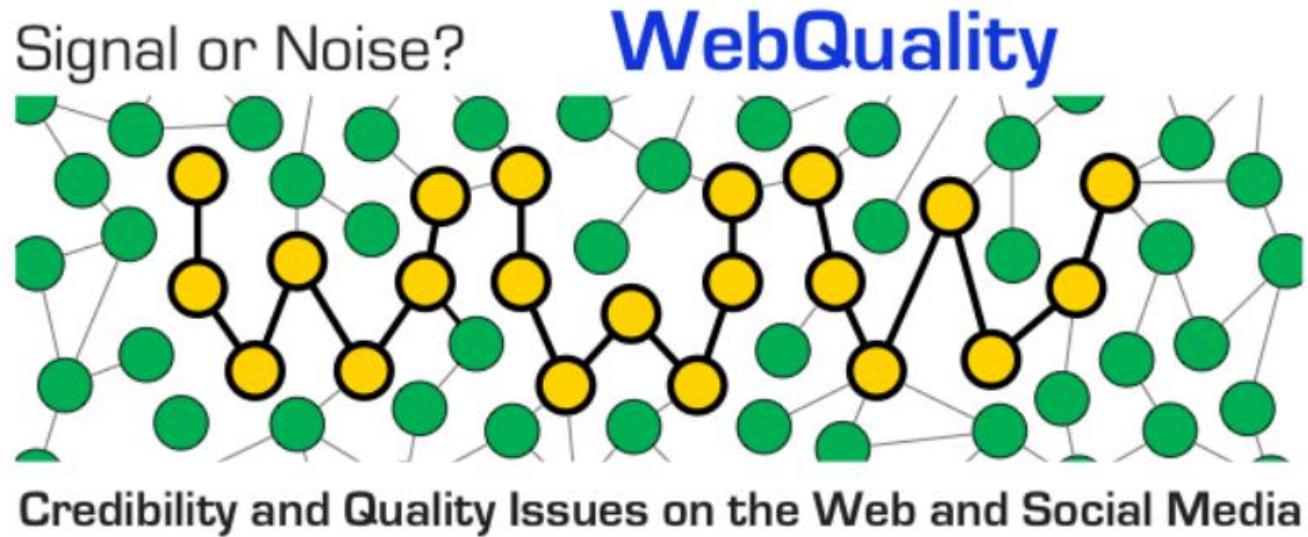
数据的积累，无论绝对增长还是相对增速都是惊人的数字

数据挑战

- 来自数据的挑战：异构数据

无论是网页结构的不同，还是数据模态的不同，都对Web信息的有效处理带来了挑战

- 来自数据的挑战：数据质量



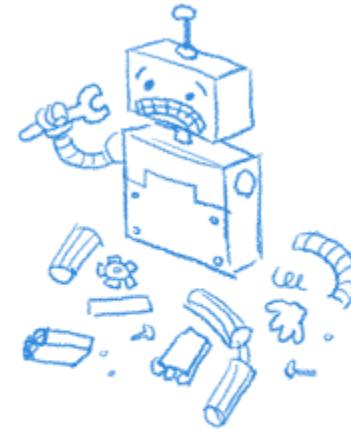
Web中包含大量未经编辑处理或权威确认的信息，
可能导致错误、无效或误导

- 来自数据的挑战：数据不稳定性



404. That's an error.

The requested URL /404notfound was not found on this server. That's all we know.



许多网站和文档快速的添加和消亡，导致大量死链的存在

- 来自数据的挑战：数据不稳定性（续）

The image displays two side-by-side screenshots of a Baidu search results page for the query "吴亦凡".
Left Screenshot (General Result):
- Title: 吴亦凡(中国内地男演员、...)
- Summary:
 职业: 歌手
 生日: 1990年11月06日
 个人信息: 187 cm/73 kg/天蝎座/O型
 简介: 吴亦凡 (Kris), 1990年11月6日出生于广东省广州市, 加...
- Photo: A portrait of Wu Yifan in a black suit.
- Source: baike.baidu.com/
Right Screenshot (Accused Result):
- Title: 吴亦凡(涉强奸罪的加拿大籍男艺人)-百度百科
- Summary:
 职业: 歌手、演员、音乐制作人
 生日: 1990年11月06日
 个人信息: 187 cm/73 kg/天蝎座/O型
 简介: 吴亦凡 (Kris), 1990年11月6日出生于广东省广州市, 加...
- Photo: A portrait of Wu Yifan in a black suit.
- Source: baike.baidu.com/

甚至，已有网页内容也在不断地发生更新

用户挑战

- 来自用户的挑战：查询需求的表达

钢铁锅含眼泪喊修瓢锅这是什么歌词

 我来答  分享  举报

6个回答

#热议# 王嘉尔夹走王一博香菜，王嘉尔生活中什么性格？

 苏冰堰2012 Lv13
推荐于2018-03-23

关注

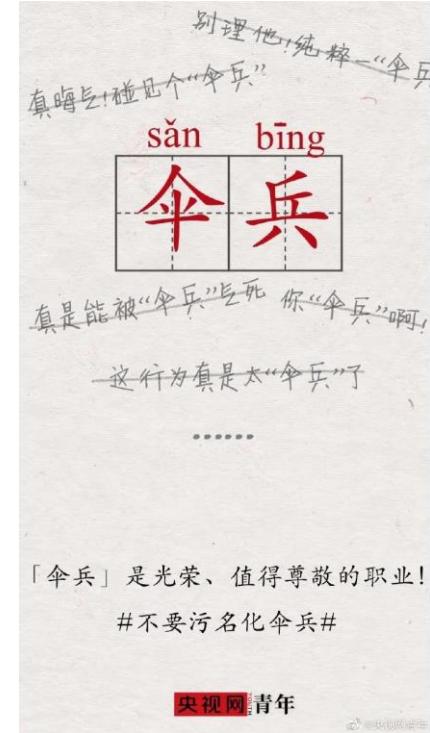
你好！是《海阔天空》粤语的音译。

1. “今天我，寒夜里看雪飘过” 粤语发音音译为：钢铁锅，含眼泪喊修瓢锅。
2. 也是网上有人的恶搞翻唱。
3. 你可以去听听，希望对你有帮助。

 5 |  评论  分享  举报

用户可能无法采用规范、清晰的方式表达其查询需求

- 来自用户的挑战：查询需求的表达（续）



用户表达的非规范性 × 语义演化的日新月异

用户挑战

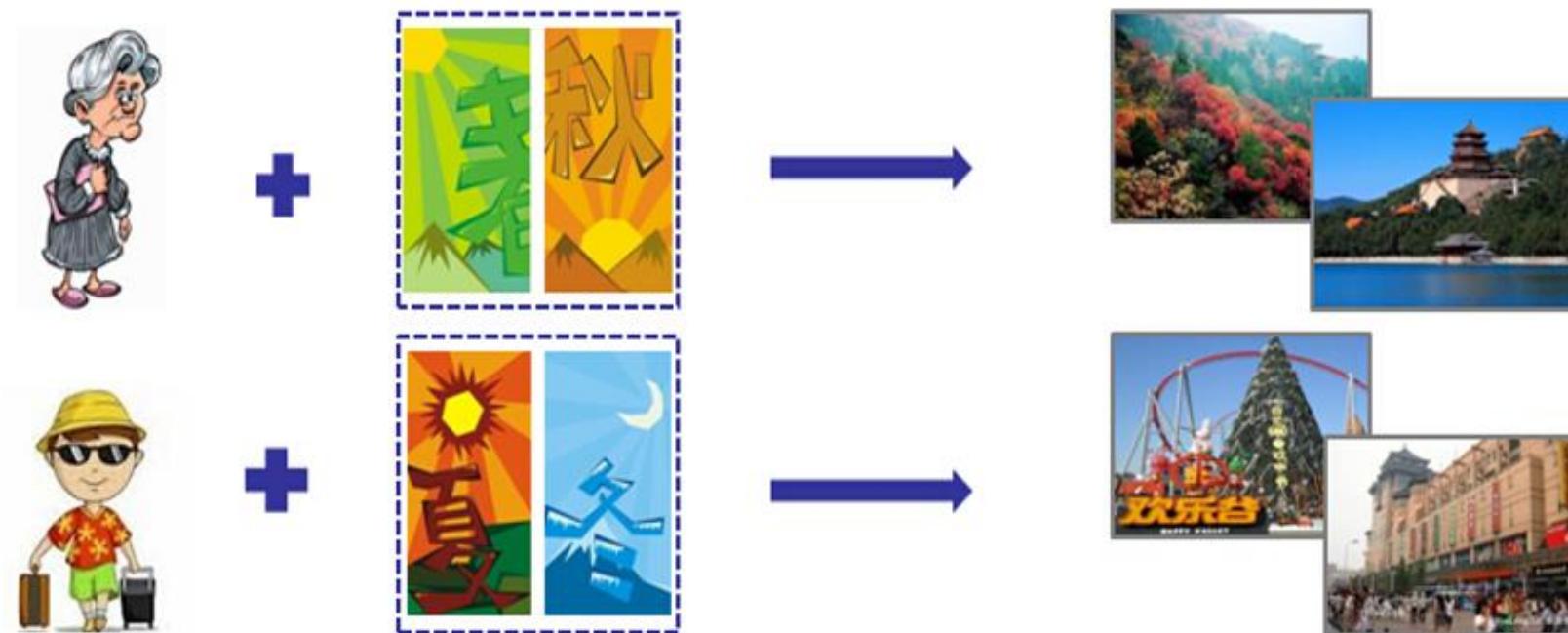
- 来自用户的挑战：知识需求与直观表达



技术使人们缺乏耐心，希望直接从搜索引擎获得答案，而不是通过阅读文档自行得到答案

- 来自用户的挑战：个性化需求

输入查询关键词：北京旅游



大众化的信息需求被个性化、差异化的信息需求所取代

- 来自利益的挑战：SEO对于搜索的干扰



搜索引擎优化（Search Engine Optimization），可能提升网站效率，也可能因滥用搜索算法而影响正常使用

- 来自利益的挑战：竞价排名对于搜索的干扰

Baidu search results for "土耳其签证" (Turkey visa):

- 申请土耳其签证 流程简单**
申请土耳其签证首选e-Visa Turkey签证中心,简化流程,24H贴心服务,轻松搞定电子签证!我们经验丰富的专家全天随时都可为您服务,如有任何疑虑,请通过电子邮件联系我们!
www.turkey-onlinevisa.org 2019-09 - 评价 广告
- 在线申请土耳其签证 电子资料**
在线申请土耳其签证首选e-Visa Turkey签证中心,简化流程,24H贴心服务,轻松搞定电子签证!我们经验丰富的专家全天随时都可为您服务,如有任何疑虑,请通过电子邮件联系我们!
www.turkey-onlinevisa.org 2019-09 - 评价 广告
- 土耳其共和国电子签证申请系统 官网**
提交相关信息后,您将可以进行电子签证申请。您可使用Mastercard、Visa 或 UnionPay 信用卡/银行卡来支付。支付完成后,您的电子签证下载链接将发送...
www.evisa.gov.tr

基于广告改变排序，对使用者产生误导

- 来自利益的挑战：低质内容的滥觞——洗稿

2013年度新浪政务微博报告：12月26日，人民网舆情监测室联合新浪共同发布《2013年新浪政务微博报告》（以下简称“报告”）。
目前新浪认证的政务微博总数超过10万个，较去年同期增加4万余个，增长率约为67%。
报告显示，今年新浪政务微博发展亮点不断。



2013年新浪政務微專申報宣佈：12月26日，國民網輿情監測室結合新浪配合宣佈《2013年新浪政務微專申報》（以下簡稱申報）。今朝新浪認證的政務微專總數跨越10萬個，較客歲同期增長4萬餘個，增加率約為67%。申報表現，本年新浪政務微專成長明面賡續。

技术的滥用导致洗稿工具等手段盛行
拉低内容质量，恶化用户体验

- 来自利益的挑战：低质内容的滥觞——洗稿

?

经由多年积淀，中国科学手艺大学汇聚了一批国际一流的基础学科领武士才，形成了一个具有国际学术视野、创新能力强、能率领本学科攀缘学科岑岭的带头人群体。

彩蛋：猜猜这是啥？

- Web 信息基础
- 信息检索概述
- 数据挖掘概述

- 信息检索 (Information Retrieval)
- 基本含义：给定用户需求，从数据库中寻找并反馈相关的文档
 - Query：用户的查询需求
 - Corpus：待检索的数据库
 - Relevance：文档满足查询需求的程度
- 信息检索是关于信息的结构、分析、组织、存储、搜索 (Search) 和获取 (Retrieval) 的领域

—— Gerard Salton, 1968



- 信息检索的发展历史
- 1950年，明尼苏达大学的Calvin Mooers提出了“信息检索”这一概念
- 1960年代，康奈尔大学的Gerard Salton研发了SMART系统，被视作信息检索的鼻祖
- 1970年代，SIGIR成立，信息检索领域的旗舰学术会议由此开始
- 1980年代，商用IR系统开始出现
- 1990年代，TREC会议于1992年起始，开始标准测评、Web搜索等研究



Gerard Salton

- 信息检索 vs. 数据库

- 数据库属于标准的结构化数据，而信息检索往往面临文本、图像、视频等非结构化或半结构化数据。
- 数据库依赖精确的查询条件，而信息检索的查询词更加自由，匹配也相对粗疏
- 数据库对排序并不强调，而信息检索的效果关键在于相关性排序

	DB	IR
Data	<i>Structured</i>	<i>Semi-structured</i>
Fields	<i>Clear Semantics</i>	<i>Free text</i>
Queries	<i>Structured</i>	<i>Free Text</i>
Matching	<i>Exact</i>	<i>Imprecise</i>
Ranking	<i>None</i>	<i>Important</i>

信息检索

- 信息检索的应用场景
- 通用搜索
 - 一般的Web搜索，IR最常见的应用
- 垂直搜索（Vertical Search）
 - 搜索被限定在特定的主题和领域上
- 内部搜索
 - 内部网络甚至个人电脑中的搜索引擎
- P2P搜索（Peer-to-peer Search）
 - 由节点构成的网络中寻找信息，但没有集中式的控制

不同领域的垂直搜索

58同城·房产

请输入房源相关信息

搜房源

租房 二手房 商铺 生意转让 写字楼 厂房 仓库 土地 车位

合肥58同城 > 合肥房产信息 > 合肥二手房

二手房 小区 地图找房 新房 经纪人

区域 ▲

地铁 ▼

总价：不限 30万以下 30-40万 40-50万 50-60万 60-80万 80-100万 100-120万 120-160万 160-200万 200万以上

面积：不限 50m²以下 50-70m² 70-90m² 90-110m² 110-130m² 130-150m² 150-200m² 200-300m² 300-500m² 500m²以上

厅室：不限 一室 二室 三室 四室以上

其他：朝向不限 楼层不限 产权不限 类型不限 装修不限 房龄不限

拉勾

首页

公司

校园招聘

new

言职

课程

登录

职位(500+)

公司(0)

Java

搜索

相关搜索：java后端 java web java大数据 java分布式 java服务端 java后端实习 java实习 java架构师 后端

工作地点：全国 北京 上海 深圳 广州 杭州 成都 南京 武汉 西安 厦门 长沙 苏州 天津

更多 ▾

工作经验：不限 应届毕业生 3年及以下 3-5年 5-10年 10年以上 不要求

学历要求：不限 大专 本科 硕士 博士 不要求

融资阶段：不限 未融资 天使轮 A轮 B轮 C轮 D轮及以上 上市公司 不需要融资

公司规模：不限 少于15人 15-50人 50-150人 150-500人 500-2000人 2000人以上

更多 ▾

行业领域：不限 移动互联网 电商 金融 企业服务 教育 文娱 | 内容 游戏 消费生活 硬件

- 信息检索的基础问题（1）：查询理解

- 信息需求是人们发送查询的背后的动因
- 准确理解查询需求是信息检索的前提
- 用户是搜索质量的终极判定者，需要通过与用户的交互，帮助用户表达他们的信息需求
 - 通过上下文信息去除歧义影响
- 查询建议、查询扩展等应用



• 信息检索的基础问题（2）：相关性计算

- 相关性是判断是否满足需求的基础，基于相关性的排序决定了文档呈现顺序
- 单纯依赖查询和文档的简单匹配，未必能够得到所需的结果

文献全部分类 主题 Web信息 检

主题:Web信息 查看 Web信息 的指数分析结果

分组浏览: 主题 发表年度 研究层次 作者 机构 基金

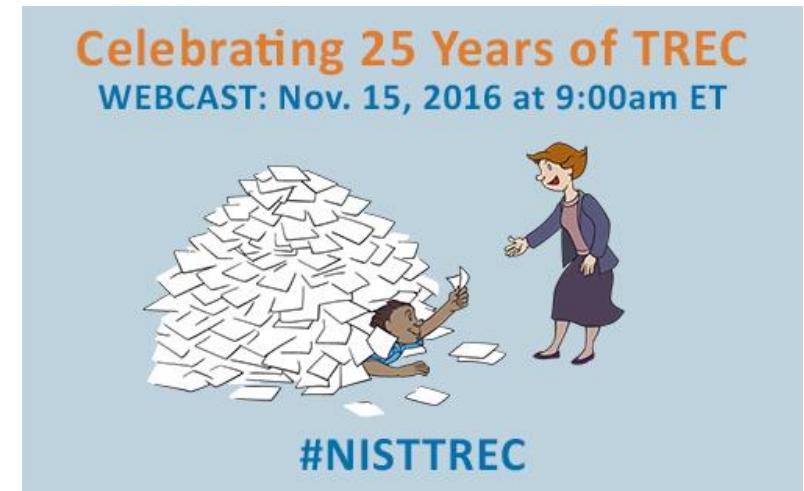
Web (5246) WEB (837) 计算机应用 (766) XML (509) 信息抽取 (507) Web服务 (460) Web2.0 (415) 数据库系统 (415)
管理信息系统 (379) 管理信息系统 (375) 情报工作 (367) 信息检索 (340) 数据库服务器 (333) 数据挖掘 (317) >>

排序: 相关度 发表时间 被引 下载 中文文献 外文文献 列表

序号	题名	作者	来源	发表时间	数据库
1	基于web的培训信息管理系统设计与实现	陈艳君;安然	中国石油学会 2019年物探技术 研讨会论文集	2019-09-09	中国会议
2	基于深度学习的Web信息抽取模型研究与应用	俞鑫;吴明晖;	计算机时代	2019-09-03	期刊
3	Intellectual Information System Of Subjects Methodological Support Based On The Web	D. Grinchenkov;D. Kushchii;A.		2019-08-31	Atlantis Press

- 不同类型的检索模型（Retrieval Model）导致了不同假设的相关性计算

- 信息检索的基础问题（3）：效果评估
- 信息检索的质量取决于反馈文档与用户期望的匹配程度
- 常见的评价指标：准确率（Precision）、召回率（Recall）、F值（F-measure）等
- 活跃的基准测试项目：TREC
 - <https://trec.nist.gov/>
 - 围绕问答、特定领域检索、主体识别等项目展开测评



- 信息检索的基础问题（4）：检索性能

- 如何快速响应用户的检索需求？
- 如何利用索引减少检索所需时间？
- 如何对检索条件和规则进行模块化，以实现效果与效率的均衡？



- Web 信息基础
- 信息检索概述
- **数据挖掘概述**

- 数据挖掘 (Data Mining)
- 基本含义：从海量数据中提取或挖掘潜在的知识和规律，用于支持当前的判断或未来的决策
 - 数据准备：筛选、清理并整合有待挖掘的数据
 - 数据建模：使用智能方法建模数据并提取规律
 - 知识表示：以用户能理解的方式展示所得知识
- 随着数据捕获、传输和存储技术的快速发展，用户将更多地需要采用新技术来挖掘数据的价值

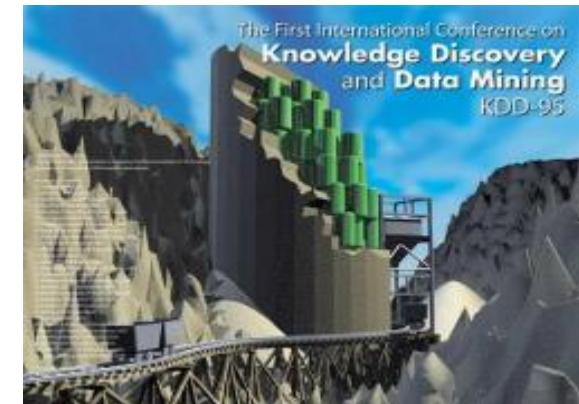
—— Gartner Group, 2011



- 数据挖掘的发展历史

- 1989年8月，第11届国际人工智能联合会议（IJCAI）组织了知识发现专题讨论会（IJCAI-1989 Workshop on KDD）
 - 首次出现了KDD这一术语，第二个D仍指Database
 - 本次讨论会参与人数30人
 - 1990年第二届讨论，参与人数46人
- 1995年，首届知识发现与数据挖掘国际会议召开于加拿大蒙特利尔召开
 - 第二个D已由Database改为Data Mining
 - 1993年，国家自然科学基金首次资助KDD相关研究项目

KDD 1995

会议
论文集

- 数据挖掘 vs. 数据库 vs. 信息检索
- 以一家大型超市为例



数据库：
进货单、价目表.....

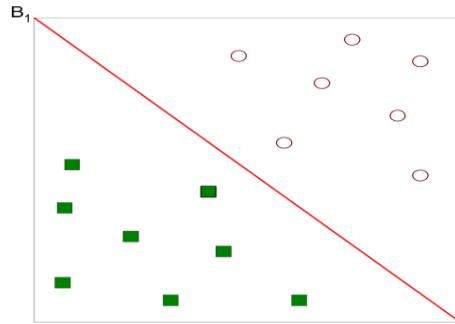


信息检索：
最符合顾客需求的是？

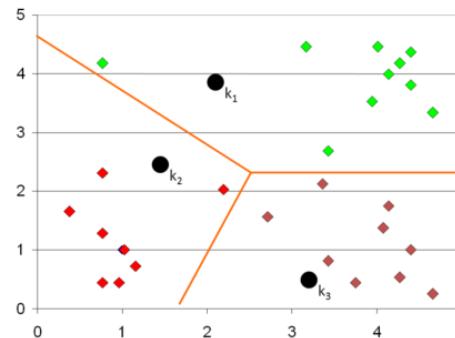


数据挖掘：
啊！ 啤酒和尿布！

- 数据挖掘的基本方法



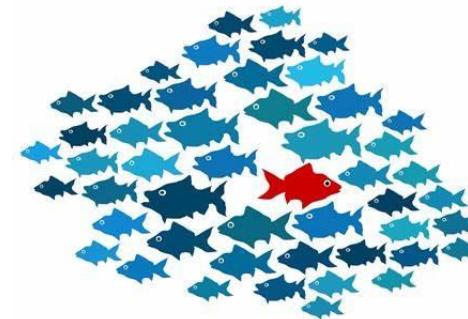
分类



聚类

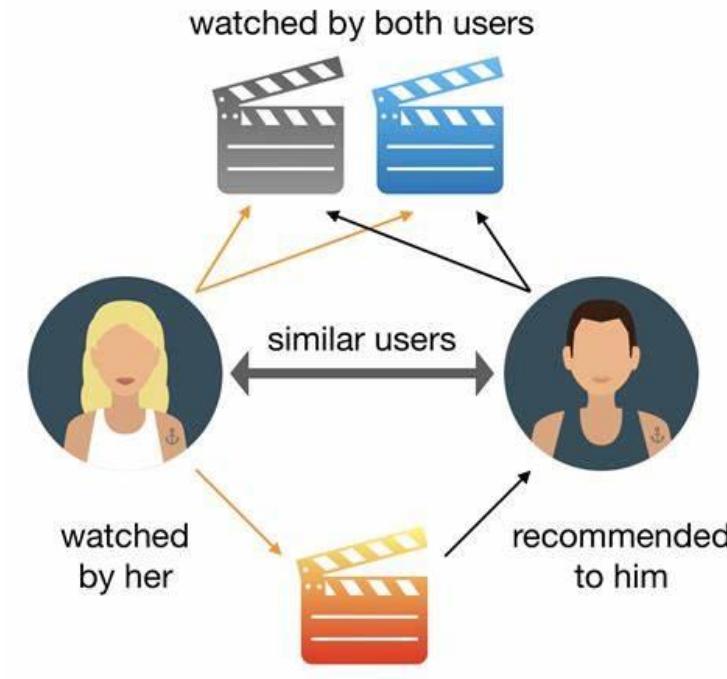
TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

关联规则

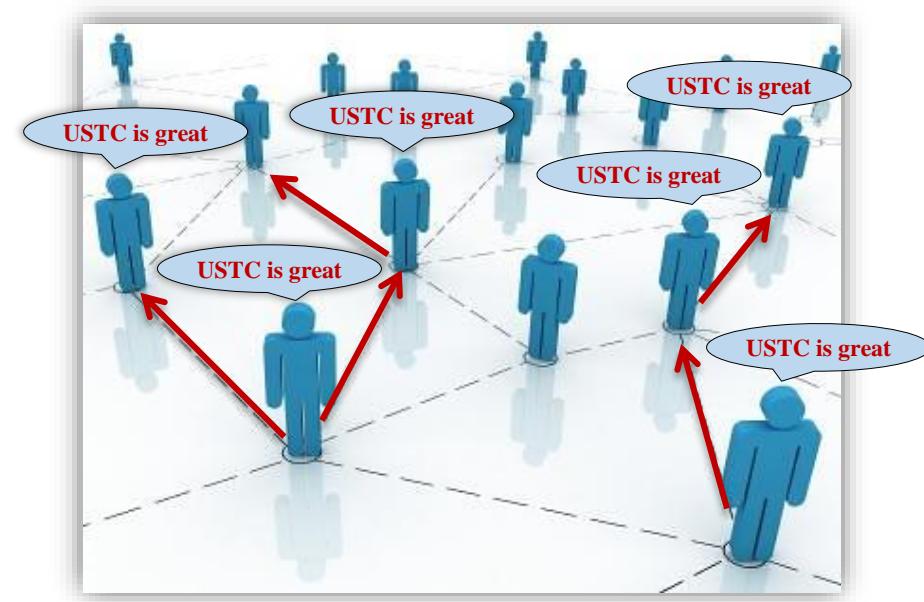


离群检测

- 数据挖掘的典型应用场景



推荐系统



社会网络分析

- 信息检索与数据挖掘领域的重要国际学术期刊与会议
- 国际学术会议
 - A类: SIGIR、WWW、SIGKDD、ICDE等
 - B类: CIKM、WSDM、ICDM、SDM、DASFAA等
- 国际学术期刊
 - A类: TOIS、TKDE、TKDD、TODS、VLDB Journal等
 - B类: TWEB、DMKD、Information Systems、KIS等

可参考CCF、CSRank等计算机国际会议与期刊排名

本章小结

Web信息概论

- 课程背景、问题与挑战
- Web信息基础
 - Web与Web搜索起源
 - Web搜索面临的挑战
- 信息检索概述
- 数据挖掘概述

tongxu@ustc.edu.cn