

Web信息处理与应用

A thick, orange, slightly curved horizontal bar that spans most of the width of the slide, positioned below the title.

金培权

jpq@ustc.edu.cn

课程背景

全球市值最高的企业（2017）

	排名	公司	国家	行业	市值（亿美元）
Apple	1	苹果	美国	科技	7540
Google	2	Alphabet	美国	科技	5790
Microsoft	3	微软	美国	科技	5090
Amazon	4	亚马逊	美国	消费服务	4230
	5	伯克希尔哈撒韦	美国	金融	4110
Facebook	6	Facebook	美国	科技	4110
	7	埃克森美孚	美国	石油和天然气	3400
	8	强生	美国	卫生保健	3380
	9	摩根大通	美国	金融	3140
	10	富国银行	美国	金融	2790
Tencent	11	腾讯	中国	科技	2720
Alibaba	12	阿里巴巴	中国	消费服务	2690

以腾讯为例：

2009年市值\$130亿，2017年\$2720亿，8年间增长率1992%!

课程背景

我们已经处于Web时代，Web数据浪潮无法回避



中国网民数量：
6.49亿



手机网民：**5.57亿**
占网民总数：**85.8%**



中国注册网站数量：
364.7万个

“互联网化”浪潮来袭



月均网络交易：
16亿笔



发布的网页数量：
1899亿页
年增速：**~27%**



新浪微博



腾讯微博

微博月活跃用户数：
1.98亿



搜狐微博



网易微博

每日新发微
博数量：**1亿+条**

本课程讨论的问题

Search



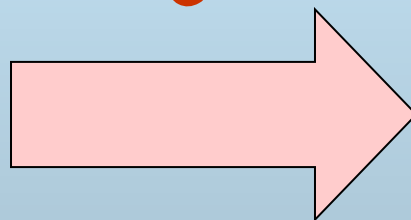
Extraction



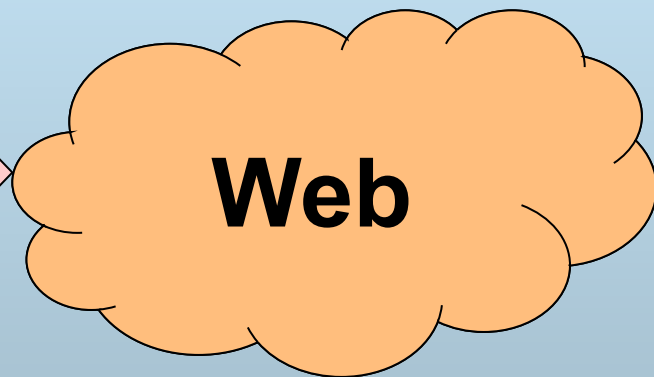
Mining



?




Web




- 非结构化信息
- 大数据
- 数据多样化

Example 1

DBLP FILTER Sign in

 Sort by ☐ relevance ☒ importance ☐ date

Scholar About 28 results (5.57sec)  (1998~2013)

Since Time

Since 2013

Since 2012

Since 2009

Custom range...

Sort By

Sort By Relevance

Sort By Importance

Sort By Date

A	EE	Scholar	A Data Model and Data Structures for Moving Objects Databases. (Luca Forlizzi and Ralf Hartmut G and ü) <i>ACM Conference on Management of Data (sigmod) [2000]</i> Cited by 353
A	EE	Scholar	Scientific Data Repositories: Designing for a Moving Target. (Etzard Stolte and Christoph von Praun and Gustavo Alonso) <i>ACM Conference on Management of Data (sigmod) [2003]</i> Cited by 43
A	EE	Scholar	A Data Model for Moving Objects Supporting Aggregation. (Bart Kuijpers and Alejandro A. Vaisman) <i>IEEE International Conference on Data Engineering (ICDE) [2007]</i> Cited by 20
B	EE	Scholar	Spatio-Temporal Data Types: An Approach to Modeling and Querying Moving Objects in Databases. (Martin Erwig and Ralf Hartmut G and ü) <i>GeoInformatica (GeoInformatica) [1999]</i> Cited by 367
B	EE	Scholar	A generic data model for moving objects. (Jianqiu Xu and Ralf Hartmut G and ü) <i>GeoInformatica (GeoInformatica) [2013]</i>
B	EE	Scholar	An Object-Field Perspective Data Model for Moving Geographic Phenomena. (Kyoung-Sook Kim and Yasushi Kiyoki) <i>Database Systems for Advanced Applications (DASFAA) [2010]</i>
C	EE	Scholar	Place: A Distributed Spatio-Temporal Data Stream Management System for Moving Objects. (Xiaopeng Xiong and Hicham G. Elmongui and Xiaoyong Chai) <i>International Conference on Mobile Data Management (MDM) [2007]</i> Cited by 18
C	EE	Scholar	An analytic solution to the alibi query in the space-time prisms model for moving object data. (Bart Kuijpers and Rafael Grimson and Walled Othman) <i>International Journal of Geographical Information Science (IJGIS) [2011]</i> Cited by 3
C	EE	Scholar	A Scaleless Data Model for Direct and Progressive Spatial Query Processing. (Sai Sun and Sham Prasher and Xiaofang Zhou) <i>International Conference on Conceptual Modeling (ER) [2004]</i> Cited by 2
C	EE	Scholar	Efficient Strip-Mode SAR Raw-Data Simulation of Fixed and Moving Targets. (Ozan Dogan and Mesut Kartal) <i>IEEE Geoscience and Remote Sensing Letters (LGRS) [2011]</i>
			Computational data modeling for network-constrained moving objects. (I aurvnas Snelcys and Christian S .Jensen and Augustas Kliavs) <i>GIS</i>

Jiang Du, Peiquan Jin, et al: **DBLP-filter: effectively search on the DBLP bibliography.** WWW 2014


Example 2

Teegoo Keywords In Location [Advanced Search](#) [Preferences](#)

☒ Content Time 2008-12-23 -- 2008-12-26
☐ Updated Time 0000-00-00 -- 9999-99-99

View 1955 - 2000

Result Lists Clustered Topics



Only show results circa:
E.g. "January 1975" or "1700-1750"

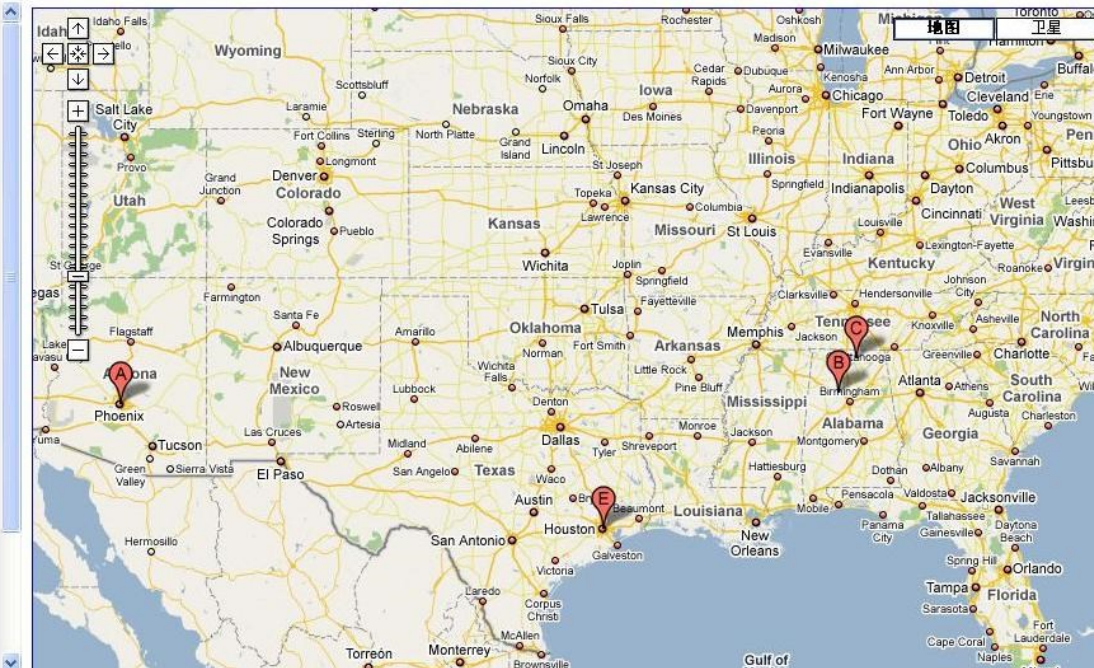
History of Space Exploration
A [Phoenix](#) - Mars Spirit Rover · Mars Opportunity Rover · New Horizons Pluto Kuiper Belt Flyby · MESSENGER · Phoenix Mars Lander. USSR/Russia Missions
www.solarviews.com

The Coalition for Space Exploration
B [America](#) - ... an inspiring agenda of human and robotic space exploration by NASA and America's space industry that spurs new technologies that improve our everyday lives.
C [Huntsville](#) - Taking it to the skies is exactly what's been done by the Space Hardware Club at the University of Alabama in Huntsville.
www.spacecoalition.com

NASA - Exploration Systems Mission Directorate
D [Nasa](#) - NASA - Exploration Systems Mission Directorate
spaceresearch.nasa.gov

NASA - 2nd Space Exploration Conference
E [Houston TX](#) - The 2nd Space Exploration Conference on December 4-6, 2006 in Houston Texas, will address how to make the Vision for Space Exploration a long-term reality.
www.nasa.gov

Space Exploration Alliance
F [Washington, DC](#) - 2009 SEA Legislative Blitz in Washington, DC. Join space advocates from around the country Feb 22-24



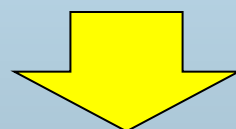
Example 3

中国科大主导研制的全球首颗量子科学实验卫星“墨子号”成功发射

2016-08-19

分享到:  QQ空间  新浪微博  腾讯微博  人人网  微信

2016年8月16日01时40分,由我校主导研制的世界首颗量子科学实验卫星“墨子号”在酒泉卫星发射中心用长征二号丁运载火箭成功发射升空。“墨子号”是中科院空间科学先导专项中首批确定立项研制的4颗科学实验卫星之一,它的成功发射和在轨运行,不仅将助力于我国广域量子通信网络的构建,服务于国家信息安全,还将开展对量子力学基本问题的空间尺度实验检验,加深人类对量子力学自身的理解。



When	Where	Who	Whom	What	How
2016/08/16 01:40	甘肃省酒泉 卫星发射中 心	中国科学 技术大学	墨子号	中国科学院先 导专项研制的 全球首颗量子 科学实验卫星	用长征二号丁 运载火箭成功 发射升空

Example 4

CoMiner Entity: Domain: Search

Competitors Results 1 - 5 of about 15 for Sony

① 竞争对手列表
按竞争度排序

② 竞争领域
以术语显示

③ 竞争依据
包含竞争信息的文档片段，正面或负面

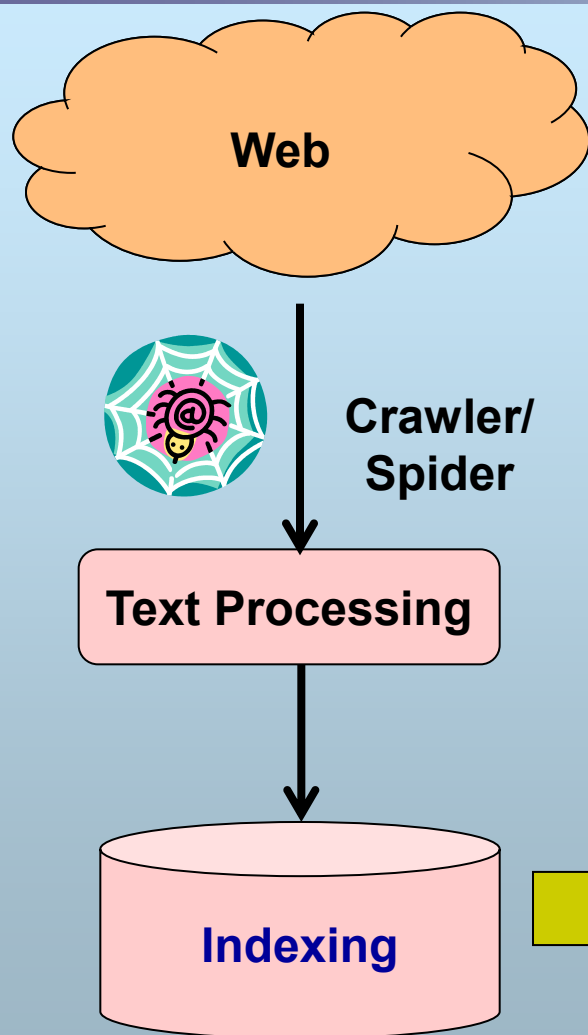
Microsoft **il**
Game, XBOX, Software, E3 Conference, PlayStation
PCWorld.com - Microsoft Eyes Deal With Sony in Digital Music Push
Bill Gates hints at possible partnership with giant electronics company.
URL: <http://www.pcworld.com/news/article/0,aid,119193,00.asp> <more evidences>

① Samsung **③ il** **②**
Mobile Phone, Cell Phone Accessory, Camcorders, LCD
Samsung, Sony join forces on LCDs | CNET News.com
Samsung, Sony join forces on LCDs | The consumer electronics giants form a 50-50 joint venture in Korea to produce liquid crystal displays for flat-panel ...
URL: http://news.com.com/Samsung,+Sony+join+forces+on+LCDs/2100-1041_3-5171753

Apple **il**
Music, Storage, Computer, IPOD, Technology
Apple, Sony sued over DRM in France | CNET News.com
Apple, Sony sued over DRM in France | Let the consumers choose, French consum association says. Two suits over companies' DRM are expected to be heard ...
URL: http://news.com.com/Apple,+Sony+sued+over+DRM+in+France/2100-1027_3-55754

S. Bao, et al. Competitor Mining with the Web: *IEEE Trans. On Knowledge & Data Engineering*, 20(10), 2008

本课程研究的问题（cont.）



- **Web网页如何获取？[Crawler]**
- **Web网页如何处理？[Text Processing]**
- **Web网页如何有效组织与存储？[Indexing]**
- **Web网页如何搜索？[Query/Ranking]**
- **如何从网页中获取信息？[Extraction]**
- **如何从网页中获取隐含的知识？[Mining]**

课程主要内容

■ PART 1: Web Search

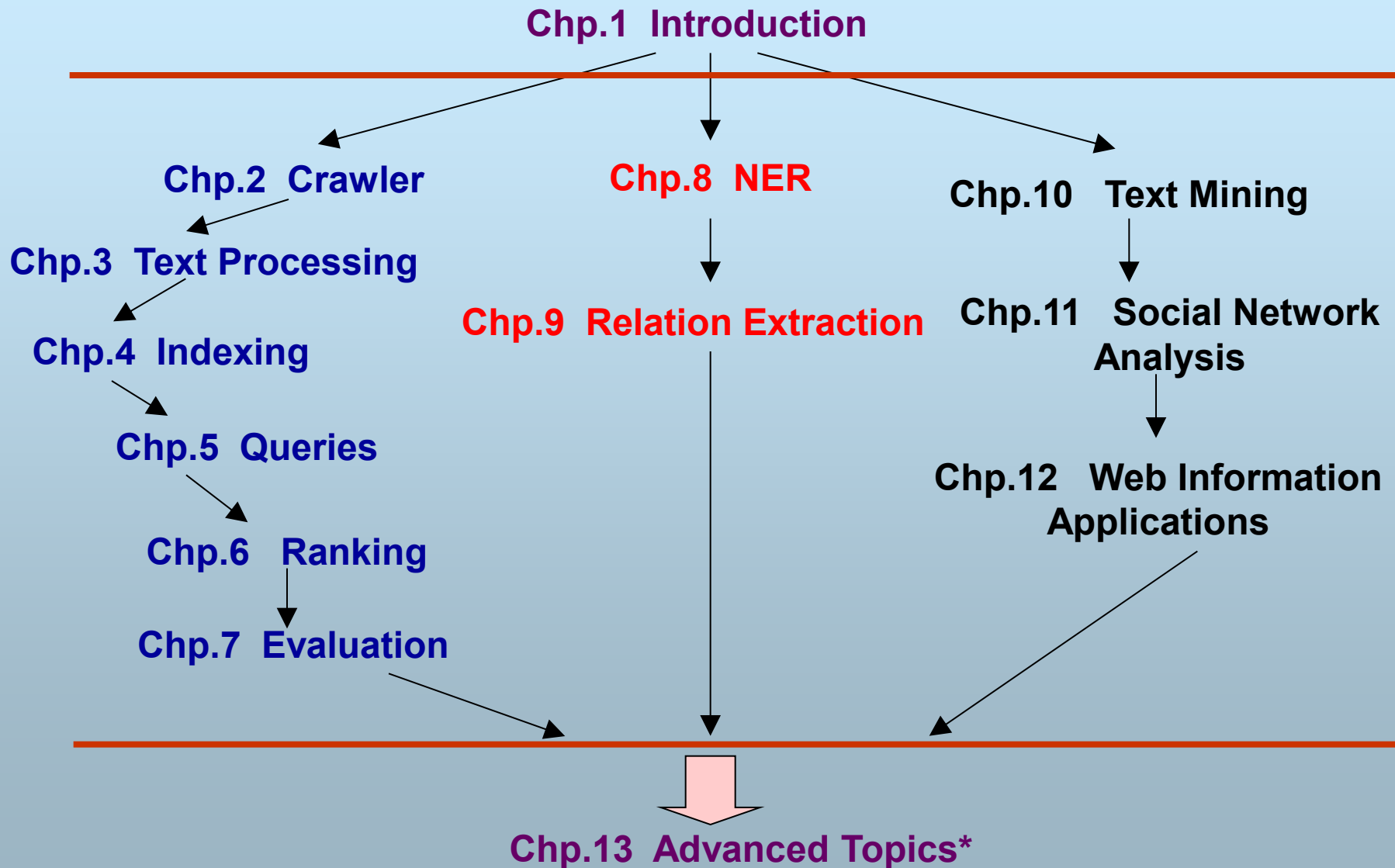
- Crawler
- Text Processing
- Indexing
- Query / Ranking
- Evaluation

■ PART 2: Web Information Extraction

- Named Entity Recognition (NER)
- Relation Extraction

■ PART 3: Web Data Mining

课程知识结构



课程安排

■ 讲课+实验

- 60学时讲授
- 30学时上机实验

■ 考核

- 期末60%，作业20%，实验20%

■ 教材

- **Search Engines: Information Retrieval in Practice**, W. Bruce Croft, et al. [搜索引擎：信息检索实践, 刘挺 等译, 机械工业出版社] ———for PART 1

■ 参考书

- **An Introduction to Information Retrieval**, Christopher D. Manning, et al., [信息检索导论, 王斌 译, 人民邮电]
- **大数据+互联网大规模数据挖掘与分布式处理**, Anand Rajaraman, Jeffrey David Ullman 著, 王斌 译, 人民邮电
- **Web数据挖掘：超文本数据的知识发现**, Soumen Chakrabarti著, 人民邮电
- **Some state-of-the-art papers from SIGIR, CIKM, WWW, etc.**

实验安排

- 实验内容见课程主页 [TBA]
 - PART 1: Web Search
 - PART 2: Web Information Extraction
 - PART 3: Web Mining
- 实验环境
 - Windows / Linux
- 开源工具
 - Lucene / ICTCLAS / Stanford NLP Tools
- 开发工具
 - Java为主

课程主页

<http://staff.ustc.edu.cn/~jpq/courses/webinfo.html>

Also linked in

<http://staff.ustc.edu.cn/~jpq>

第1章 Web信息处理概述



主要内容

- **Web搜索概述**
- **信息检索概述**

一、Web搜索

- Web的起源
- Web搜索的发展历史
- Web搜索的挑战

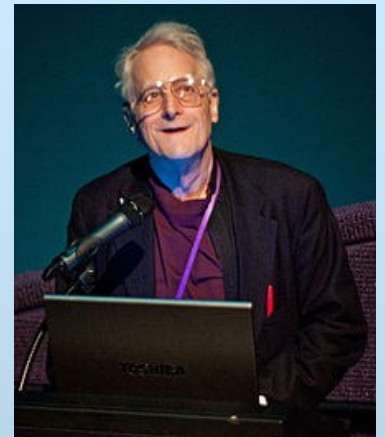
Web的起源

■ 1965年

- **Ted Nelson** 在1965年提出了超文本的概念.

- ◆ **Project Xanadu**

- ◆ 超文本传输协议(HTTP, HyperText Transfer Protocol)、超文本标注语言 (HTML) 的基础
- ◆ 思路来源于Vannevar Bush于1945年发表的“As We May Think”中提出的Memex (Memory Extender)



"wholly new forms of encyclopedias will appear, ready made with a mesh of associative trails running through them, ready to be dropped into the memex and there amplified".

-----from Vannevar Bush@Wikipedia

Web的起源

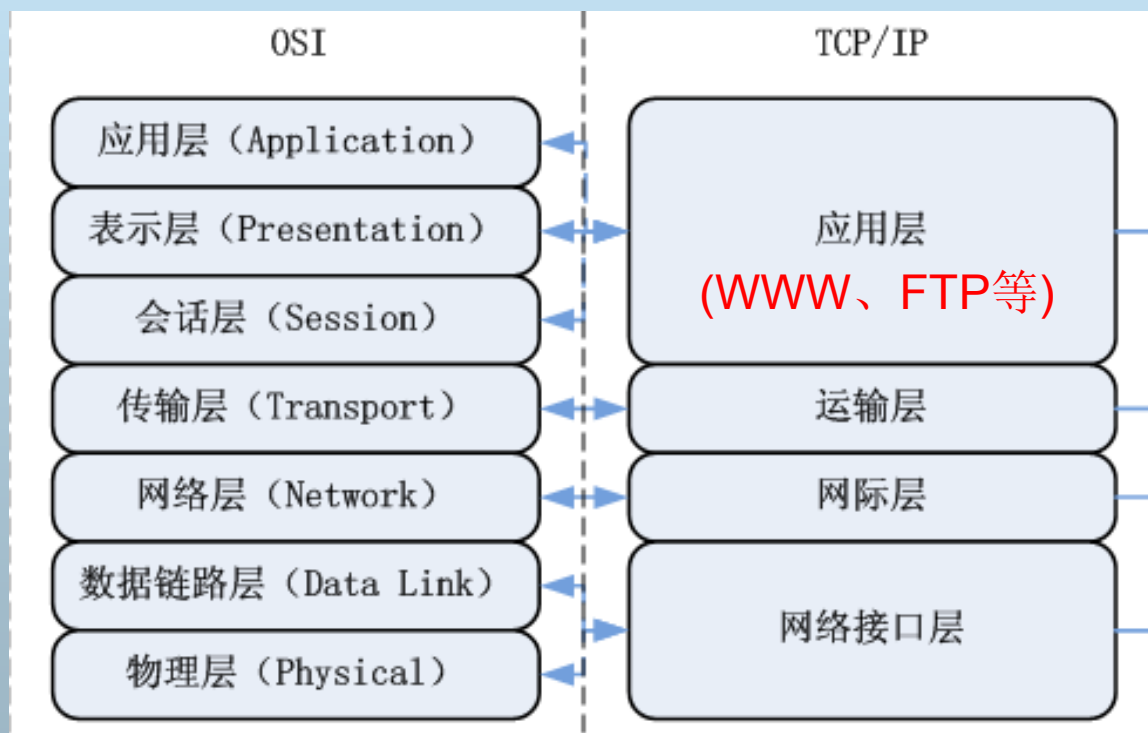
■ 1989年

- **Tim Berners Lee**（万维网之父）
在日内瓦欧洲离子物理研究所（**CERN**）开发计算机远程控制时首次提出了**Web**概念，并在**1990**年写出了第一个网页，并推出了第一个浏览器。
- 随后他设计出**HTTP**、**URL**和**HTML**的规范，使网络能够为普通大众所应用
- **W3C(World Wide Web Consortium)** 的创立者和现任主席

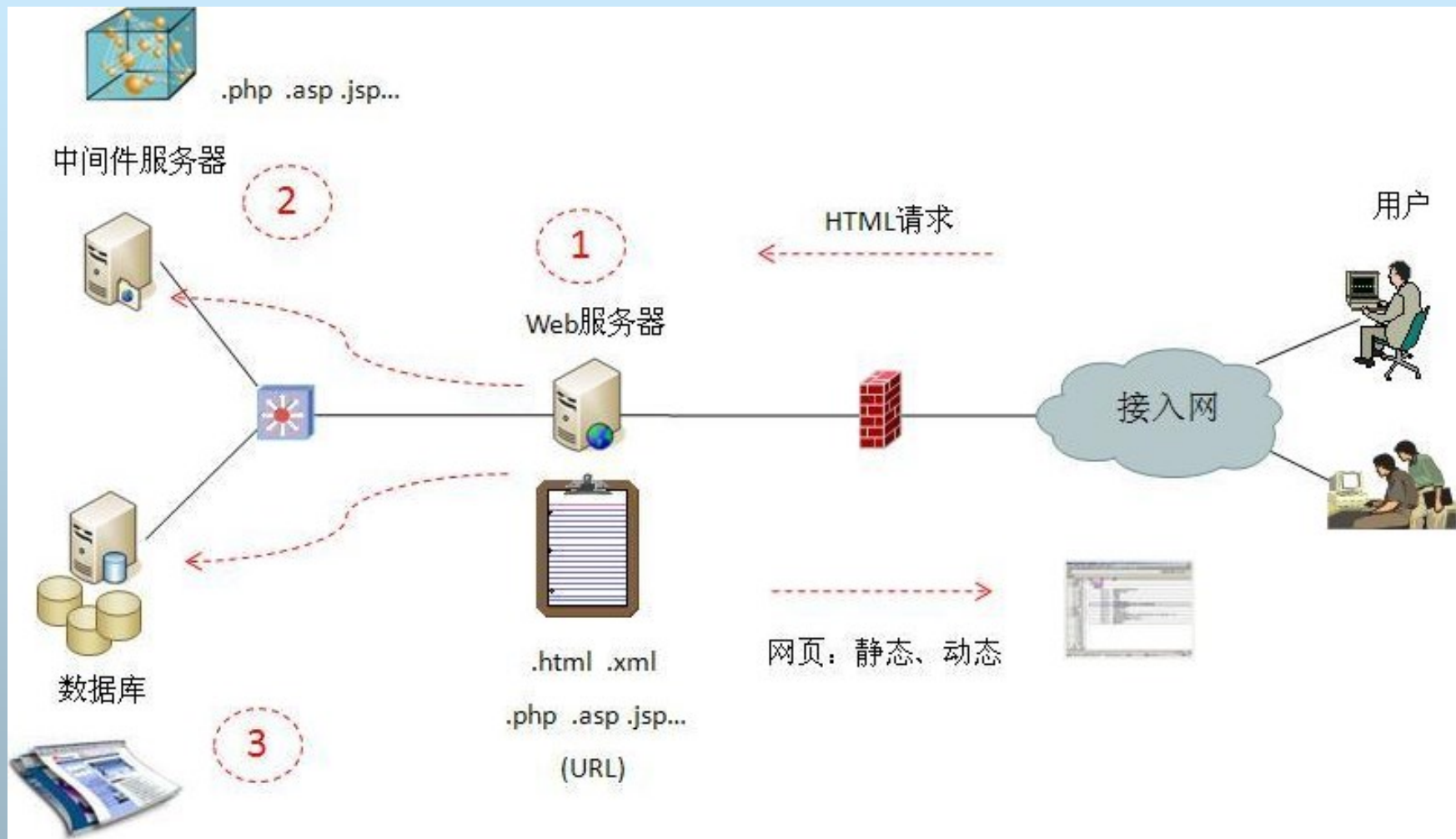


Web的特点

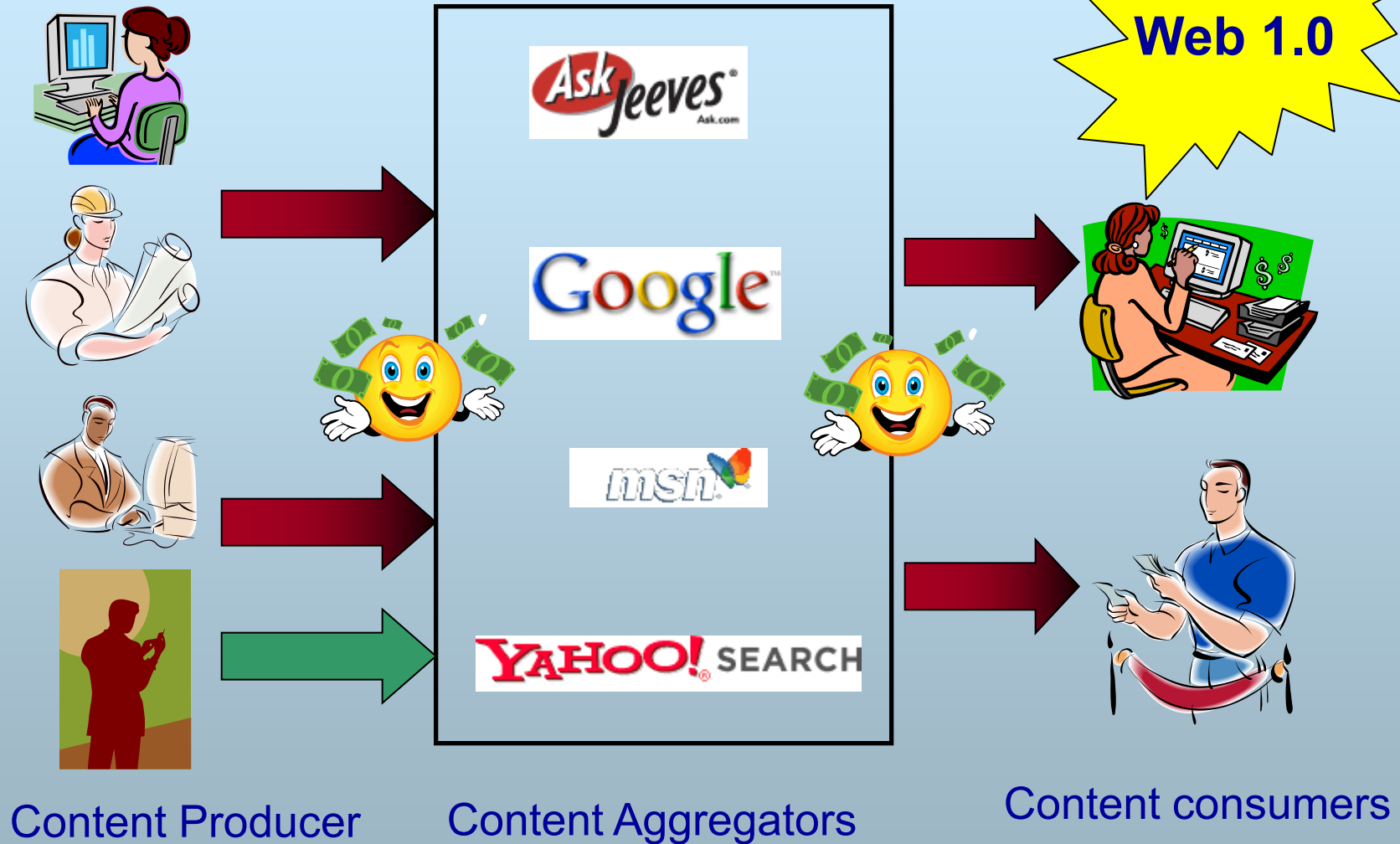
- 图形化
- 平台无关
- 分布式
- 动态性
- 交互性



Web系统的一般结构



Web结构的信息流视角

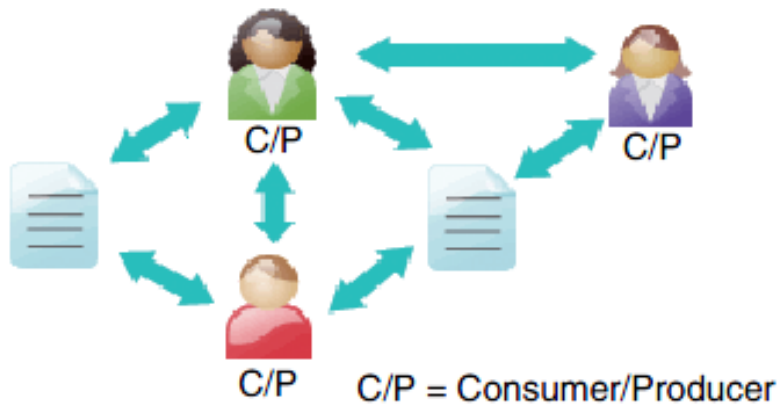


Web结构的信息流视角

Web 1.0



Web 2.0




Web 2.0



facebook



一、Web搜索

- Web的起源
- Web搜索的发展历史 
- Web搜索的挑战

Web搜索的发展历史

■ 1990

- **Archie: 最早的搜索引擎**
- 由Montreal的McGill University（麦吉尔大学）学生**Alan Emtage**、**Peter Deutsch**、**Bill Wheelan**发明的**Archie(Archie FAQ)**
- 实际上是一个可搜索的**FTP**文件名列表



Web搜索的发展历史

■ 1993年: Wanderer

- MIT 的学生**Matthew Gray**开发了**World Wide Web Wanderer**，它是世界上第一个利用网页之间的链接关系来监测**Web**发展规模的机器人（**Robot**）程序。
- 最开始只是用来统计互联网上的服务器数量，之后发展为也能捕获网址。
- **1993年还发明了另一些Robots**
 - ◆ **ALIWEB** (Archie-Like Index of the WEB, presented@1st WWW conference)
 - ◆ **WWW Worm** (Oliver McBryan@Univ. of Colorado, indexed 300000 multimedia objects which can be searched via keywords)

Web搜索的发展历史

■ 1994年: Yahoo!

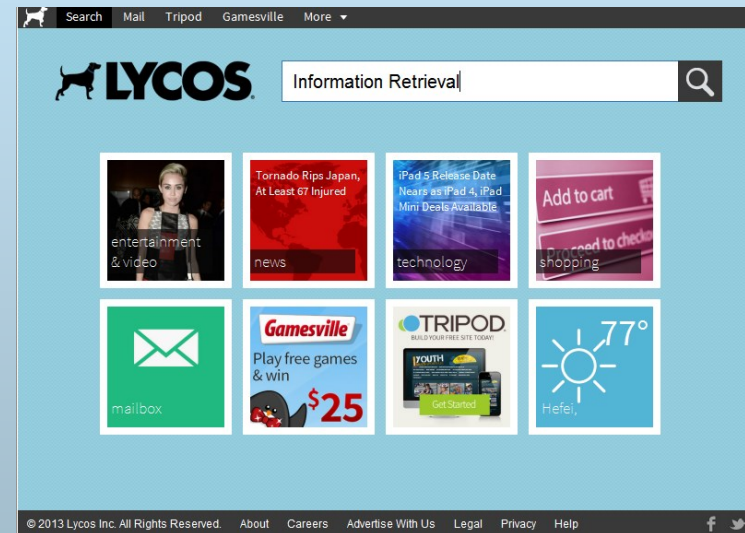
- 1994.4美籍华人Jerry Yang(杨致远)和David Filo完成了一套搜索软件
- 最初Yahoo的数据是手工输入的，实际上只是一个可搜索的目录
- 1995年1月，正式成立Yahoo网站



Web搜索的发展历史

■ 1994年: Lycos

- Michael Mauldin@CMU 将John Leavitt的蜘蛛程序接入到其索引程序中，创建了Lycos.
- 前缀匹配和字符近似匹配、网页自动摘要、相关性排序、数据量相对较大
- 2000年以125亿美元被西班牙的公司收购



Web搜索的发展历史

■ 1994年: Infoseek

- Infoseek推出，沿袭Yahoo!和Lycos的概念
- 1995.12与Netscape的战略性协议使它变得很强势
- 2001年2月，Infoseek改用Overture的搜索结果

Web搜索的发展历史

■ 1995年：Metacrawler

- 第一个元搜索引擎，是Washington大学硕士生 Eric Selberg 和 Oren Etzioni开发的 Metacrawler（1995）
- 元搜索引擎(A Meta Search Engine Roundup)。
 - ◆ 用户提交搜索后，由元搜索引擎负责转换处理后提交给多个预先选定的独立搜索引擎，并将从各独立搜索引擎返回的所有查询结果，集中起来处理后再返回给用户。

Web搜索的发展历史

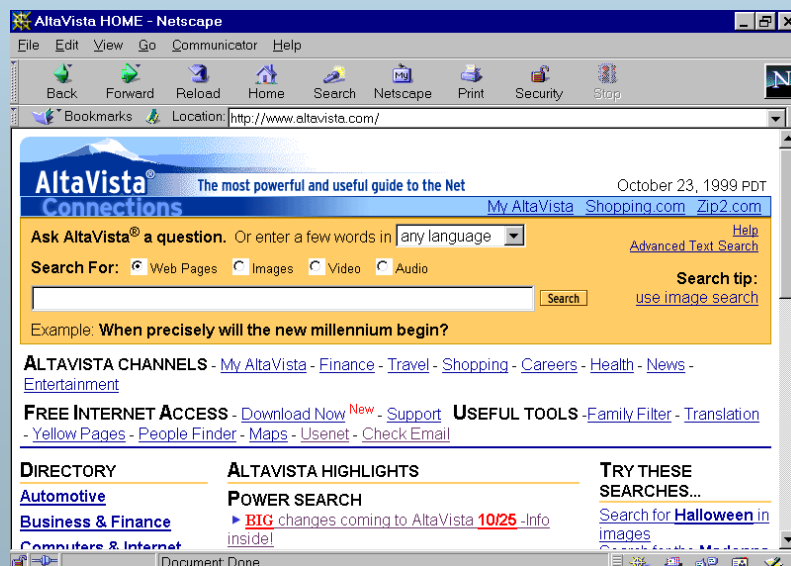
■ 元数据搜索引擎



Web搜索的发展历史

■ 1995年：Altavista

- 第一个支持自然语言搜索的搜索引擎
- 第一个实现高级搜索语法的搜索引擎（如**AND**, **OR**, **NOT**等）
- 2003年AltaVista被Overture收购，后者是Yahoo的子公司



Web搜索的发展历史

■ 1997年：Google

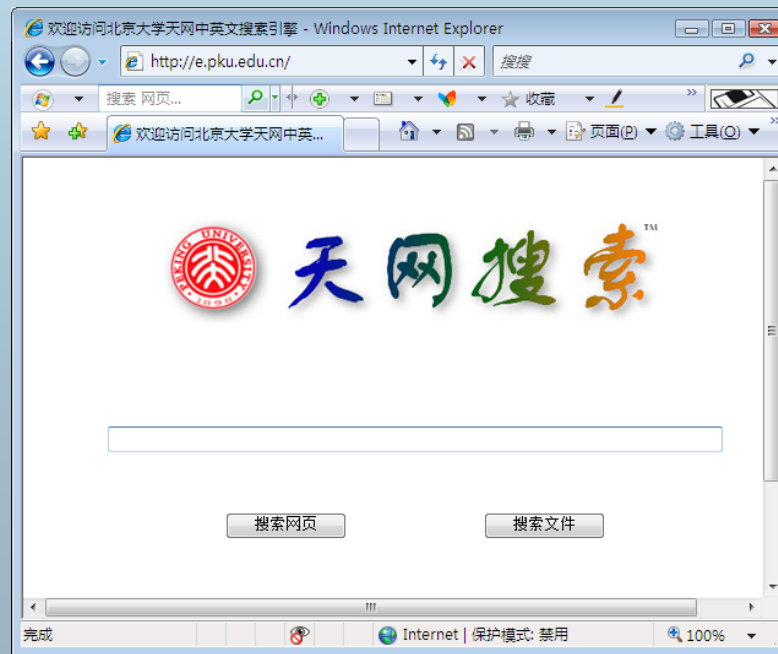
- 1995年，**Larry Page**来到斯坦福读博士，开始网络链接结构方面的研究项目**BackRub**
- 之后，他和**Sergey Brin**提出了**PageRank**技术，用于对网页评级
- 之后用于搜索引擎，改写了搜索引擎的定义，建立了**Google**
- 1997年9月注册了**Google.com**域名，1998年9月创立**Google**公司，2000年开始与**Yahoo**合作，2004年7月上市时市值**250**亿美元。增长速度超过**Microsoft**



国内Web搜索发展

■ 1997年：天网

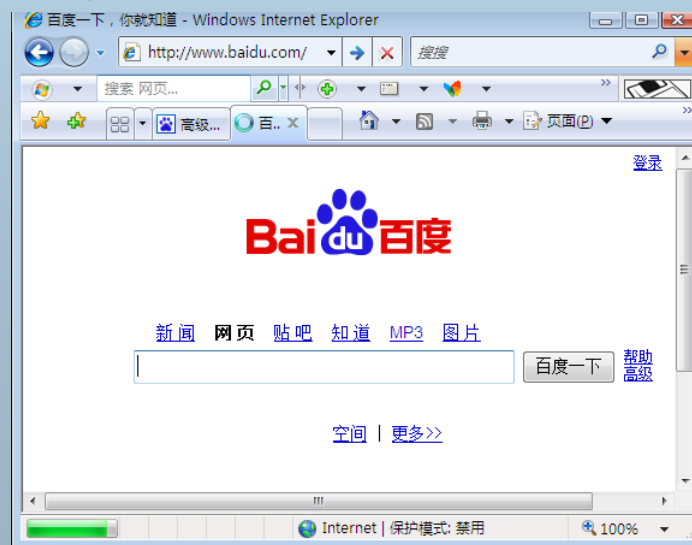
- 由北大开发，于1997年10月29日正式在CERNET上提供服务。
- 利用教育网优势，有强大的FTP搜索功能。



国内Web搜索发展

■ 2000年：百度

- 2000.1李彦宏创立了百度。
 - ◆ 前Infoseek资深工程师
- 2001.8发布百度测试版。
- 目前是最大的中文搜索引擎
- 专注于中文



国内Web搜索发展

百度发展
很快!

Baidu 百度 新闻 网页 贴吧 知道 音乐 图片 视频 地图 文库

百度一下 帮助

百度首页 > 产品大全

新上线^{新!}

百度安全管家 杀病毒，防骚扰，省流量	百度任务平台 任务发布、创意征集	百度众测 用户体验中心	百度手机助手 最新最全安卓手机
百度壁纸 天天换壁纸，时时好心情	百度杀毒 轻巧、免费、干净	百度云 文件备份、分享、同步工具	百度翻译App 您的掌上翻译专家
百度云ROM 稳定流畅的手机系统	百度魔拍 手机拍照，一键美颜		

搜索服务

网页 搜索海量网络资料、资源	百度翻译 轻松解决语言差异困扰
地图 搜索功能完备的网络地图	新闻 搜索浏览最热新闻资讯

Baidu 百度 新闻 网页 贴吧 知道 音乐 图片 视频 地图 文库 更多»

百度一下


第三个字是风的成语：

放诞风流、骨化风成、扯顺风旗、见事风生、颐指风使、跌宕风流、林下风气、烟花风月、雪虐风饕、止谈风月、八方风雨、日暖风恬、月露风云、风言风语、言论风生、想望风惠、议论风发、雾鬓风鬟、电照风行、一代风流、沦落风尘、雨散风流、铁窗风味、饶有风趣、流言风语、大煞风趣、雨霾风障、林下风范、线断风筝、风张风势、雨偃风愁、论议风生、雨沐风餐、雷厉风行、迅雷风烈

1 2 3 下一页>

报错

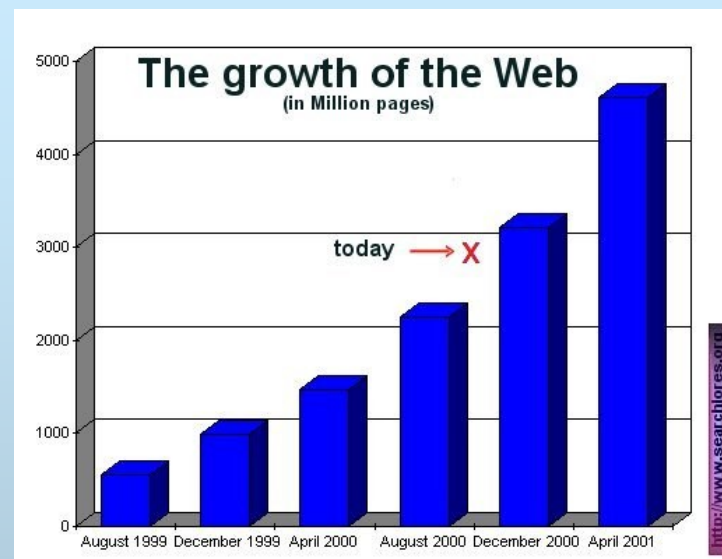
一、Web搜索

- Web的起源
- Web搜索的发展历史
- Web搜索的挑战 

Web搜索的挑战

■ 大规模

- 网络数据量的指数增长，由此引发了一系列难以处理的规模问题。



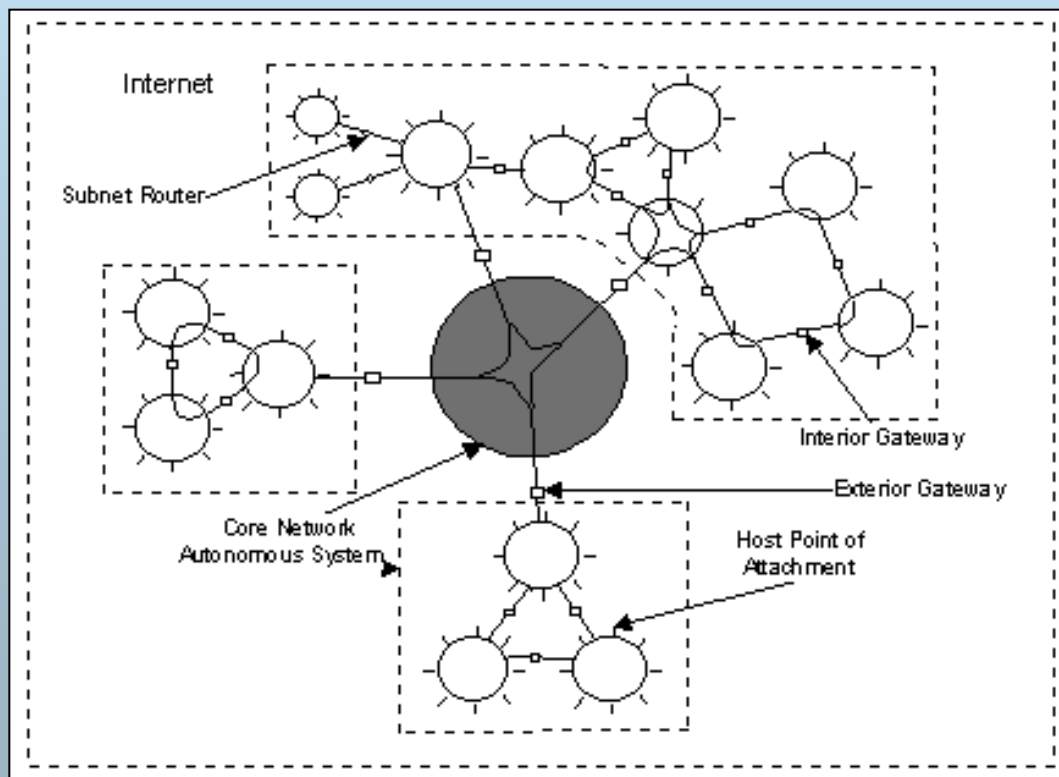
未来每18个月产生的数据量等于有史以来的数据量之和

-- Jim Gray 1998图灵奖获奖演说

Web搜索的挑战

■ 数据的分布性

- 文档散落在数以百万计的不同服务器上，没有预先定义的拓扑结构相连



Web搜索的挑战

■ 数据的不稳定性

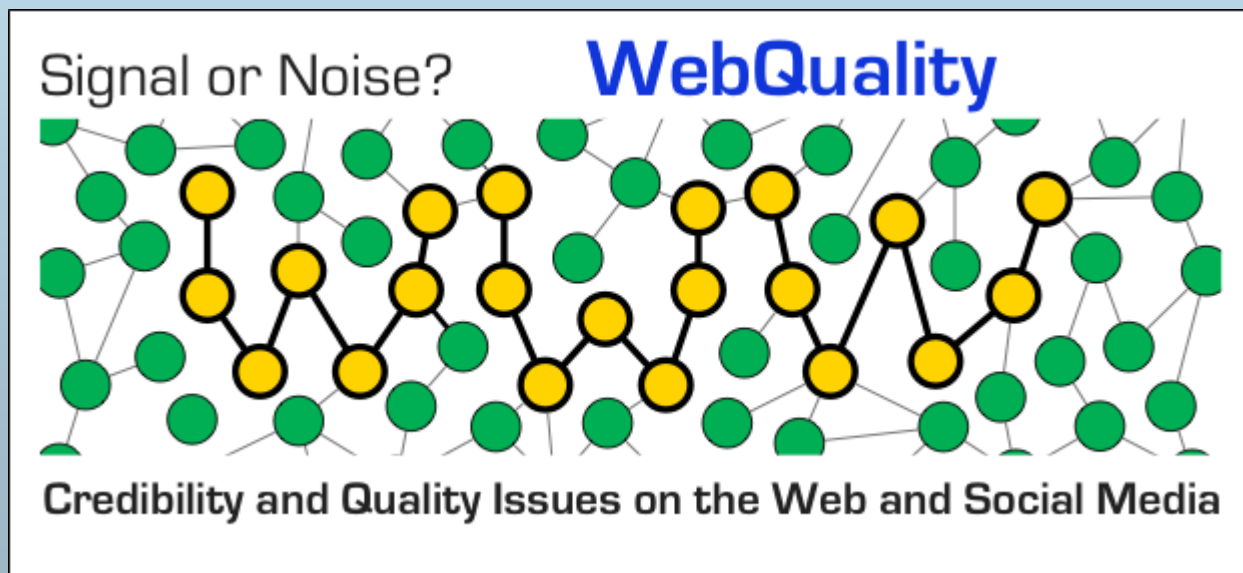
- 许多网站和文档迅速地添加或删除 (e.g. dead links).



Web搜索的挑战

■ 数据的质量

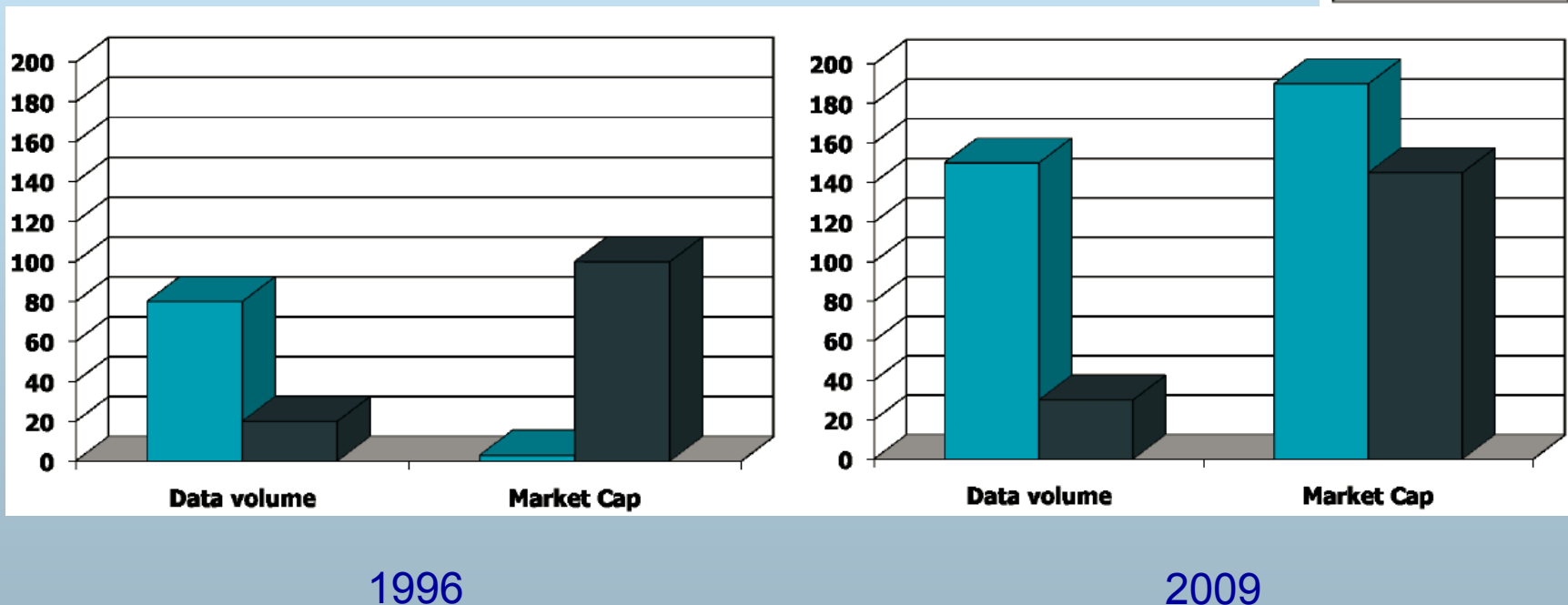
- 许多内容没有经过编辑处理，数据可能是错误的，无效的。错误来源有录入错误，语法错误等。



Web搜索的挑战

■ 无结构信息

● 每个HTML页面没有统一的结构



Web搜索的挑战

■ 异构数据

- 多媒体数据(images, video, VRML), 语言, 字符集等.



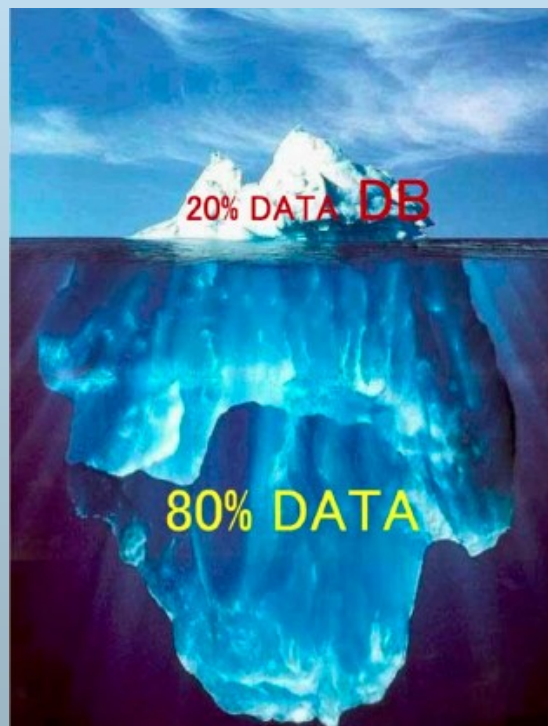
Web搜索的挑战

■ 高价值信息的发现

- **Web**中蕴含着丰富的信息，但现有搜索只能发现冰山一角

Surface Web →

Deep Web/
Dark Web →



主要内容

- Web搜索概述

- 信息检索概述 

二、信息检索概述

■ Information Retrieval

Given a **query** and a **corpus**,
find **relevant documents**.

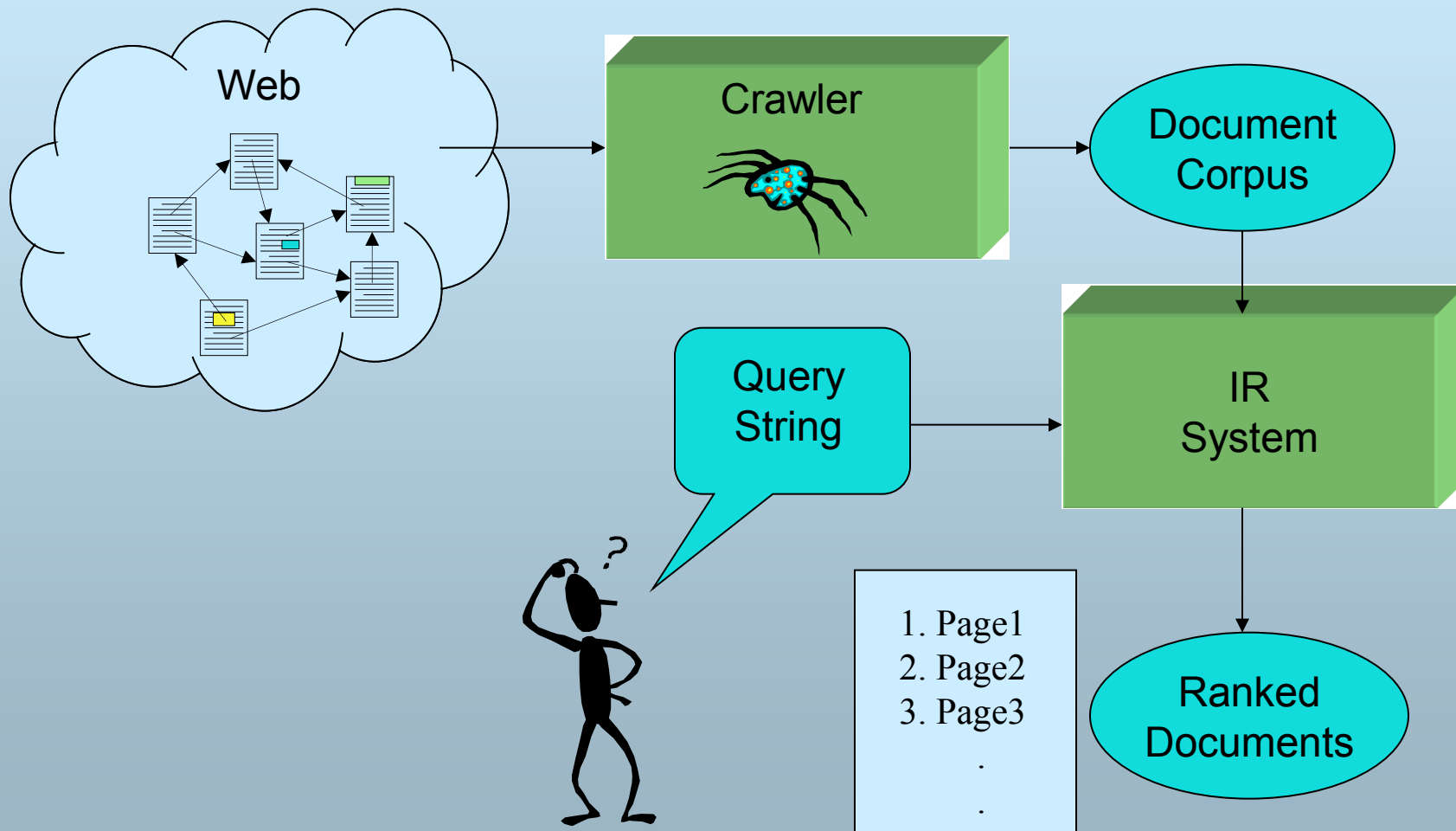
- ◆ **query**: user's expression of the information need
- ◆ **corpus**: the repository of retrievable items
- ◆ **relevance**: satisfaction of the information need

Gerard Salton's definition (1968):

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。

Web搜索与IR

Web Search: given a **keyword** and a **web crawler**, find **relevant URLs**.



IR的其它例子

■ Image Search

- given a **keyword** and an **image database**, find **relevant images**.

■ Web问答系统(Question Answering, QA)

- given a **question** and some **available texts**, **rules**, and **logics**, find an **answer**.

■ Job Search

- given a **resume** and some **job advertisements**, find **relevant jobs**.

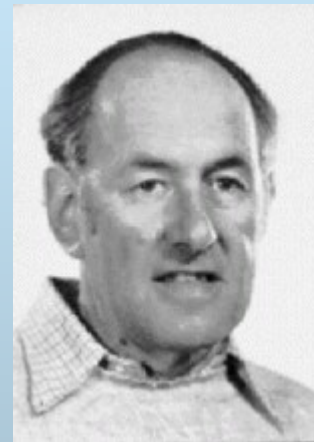
IR的发展历史

■ 1950

- Calvin Mooers @ Univ. Minnesota 提出了 “information retrieval” 一词

■ 1960s

- Gerard Salton@Cornell研发了SMART系统，IR鼻祖
- SMART: Salton's Magic Automatic Retriever of Text



■ 1970s

- SIGIR成立; IR领域的旗舰会议.

■ 1980s

- 商用IR系统出现; 1983设立SIGIR Gerard Salton Award

■ 1990s

- TREC会议(1992), 开始标准评测、Web搜索等研究

IR vs. DB

	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>

- 文本数据往往被认为是典型的非结构化数据，但是如果考虑文本中隐含的语言结构信息，那么它们也不能算是“非结构化数据”。
- 现实中的大部分文本仍然都有其他结构，如文本的标题、段落、脚注等，这些结构往往通过显式的标记来体现（如网页中的格式标签）。
- 我们也把网页这种具有格式标记的数据称为“半结构化数据”（**semi-structured data**）。例如对于新闻报道。报道有一些属性，比如标题和新闻来源，但重要的内容是报道本身。

IR的应用

■ 通用搜索

- **Web**搜索是信息检索最常见的应用

■ 垂直搜索（**Vertical search**）

- 是网络搜索的特殊形式，搜索被限制在特殊的主题上

■ 企业搜索(**Enterprise search**)

- 是在散布在企业内部网中的大量计算机文件中寻找所需的信息

■ 桌面搜索(**Desktop search**)

- 是企业搜索的个人版，信息源是存储在一台个人电脑中的文件集合，包括那些被浏览过的邮件和网页

■ P2P搜索(**Peer-to-peer search**)

- 是在节点机或计算机构成的网络中搜寻信息，但没有任何集中式的控制

IR的任务

■ 随机搜索(Ad hoc Search)

- 因为查询的范围巨大而且事先没有约定) 是搜索引擎研究的主要任务

■ 信息过滤 (Information Filtering)

- 从动态的信息流中选取满足用户兴趣的信息——IR+用户模型

■ 分类 (Classification)

■ 问答 (Question Answering)

- “珠穆朗玛峰的高度是多少？”、“亚马逊河流有多长？”

表1-1 信息检索的维度

内容实例	应用实例	任务实例
文本	网络搜索	特殊搜索
图像	垂直搜索	过滤
视频	企业搜索	分类
扫描文档	桌面搜索	问答
音频	P2P搜索	
音乐		

IR的基础性问题

■ 相关性计算 (Relevance)

- 相关性是信息检索中的基本概念。相关文档包含用户把查询发给搜索引擎后他想要找的信息。
- 对查询和文档进行简单的比较，寻找精确的匹配，那结果的相关性一定很差。
 - ◆ “Apple Price”
 - ◆ “今天 北京 航班”

IR的基础性问题

■ 检索模型(Retrieval Model)

- 是对查询与文档匹配过程的形式化表示，它是排序算法(**ranking algorithm**)的基础，搜索引擎利用排序算法生成文档的有序列表

- ◆ abstractly represent the documents

- ◆ abstractly represent the queries

- ◆ model the relationship between query and document representations

● 比如布尔检索模型

- ◆ 文档：词汇集合；查询：词汇集合

- ◆ 检索：返回包含查询词集的无序文档集合

IR的基础性问题

■ 评价(Evaluation)

- 文本排序的质量依赖于该文本与用户期望的匹配程度
- 评价指标：准确率(Precision)、召回率(Recall)、F值(F-measure)等
- 基准测试集：TREC(<http://trec.nist.gov/>)测试集是会议提供的测试集
- 检索模型和搜索引擎的评测是一个非常活跃的领域

IR的基础性问题

■ 信息需求(Information Need)

- 信息需求是人们向搜索引擎发送查询的背后动因。
- 用户是搜索质量的终极判定者。人们怎样与搜索引擎之间进行交互，帮助用户表达他们的信息需求
 - ◆ “最近东北地区有没有Nike运动鞋打折的？”
- 查询建议(query suggestion)、查询扩展(query expansion)和相关性反馈 (relevance feedback) 等

IR的基础性问题

■ 检索性能(Efficiency)

- 如何快速响应用户的信息检索需求?
- 查询分词技术
- Indexing
- 快速计算相关性得分

IR与Web搜索领域的出版物

■ 国际会议

- **A类：SIGIR、WWW、SIGKDD、SIGMOD/VLDB/ICDE等**
- **B类：CIKM、WSDM、ICDM、SDM、ECIR、DASFAA等**
- **C类：DEXA、APWEB/WAIM、WISE、WebDB等**
- **中国信息检索会议CCIR、中国数据库学术会议NDBC**

■ 国际期刊

- **A类：TODS、TOIS、VLDB Journal、TKDE等**
- **B类：DKE、Information Systems、Information Retrieval、TWEB、KIS等**

可参考CCF计算机国际会议与期刊排名

本章小结

- Web搜索概述
- 信息检索概述