信息检索与数据挖掘

第1章 绪论

提纲

1.1 信息检索的由来和这门课的意义

1.2 信息检索的历史和发展

1.3 信息检索与数据挖掘等其他学科的关系

1.4 信息检索的基本概念

1.5 课程要求和说明

提纲

信息检索的由来
为什么需要信息检索？
什么是信息检索？

什么是数据挖掘？
数据挖掘与信息检索有什么关系？

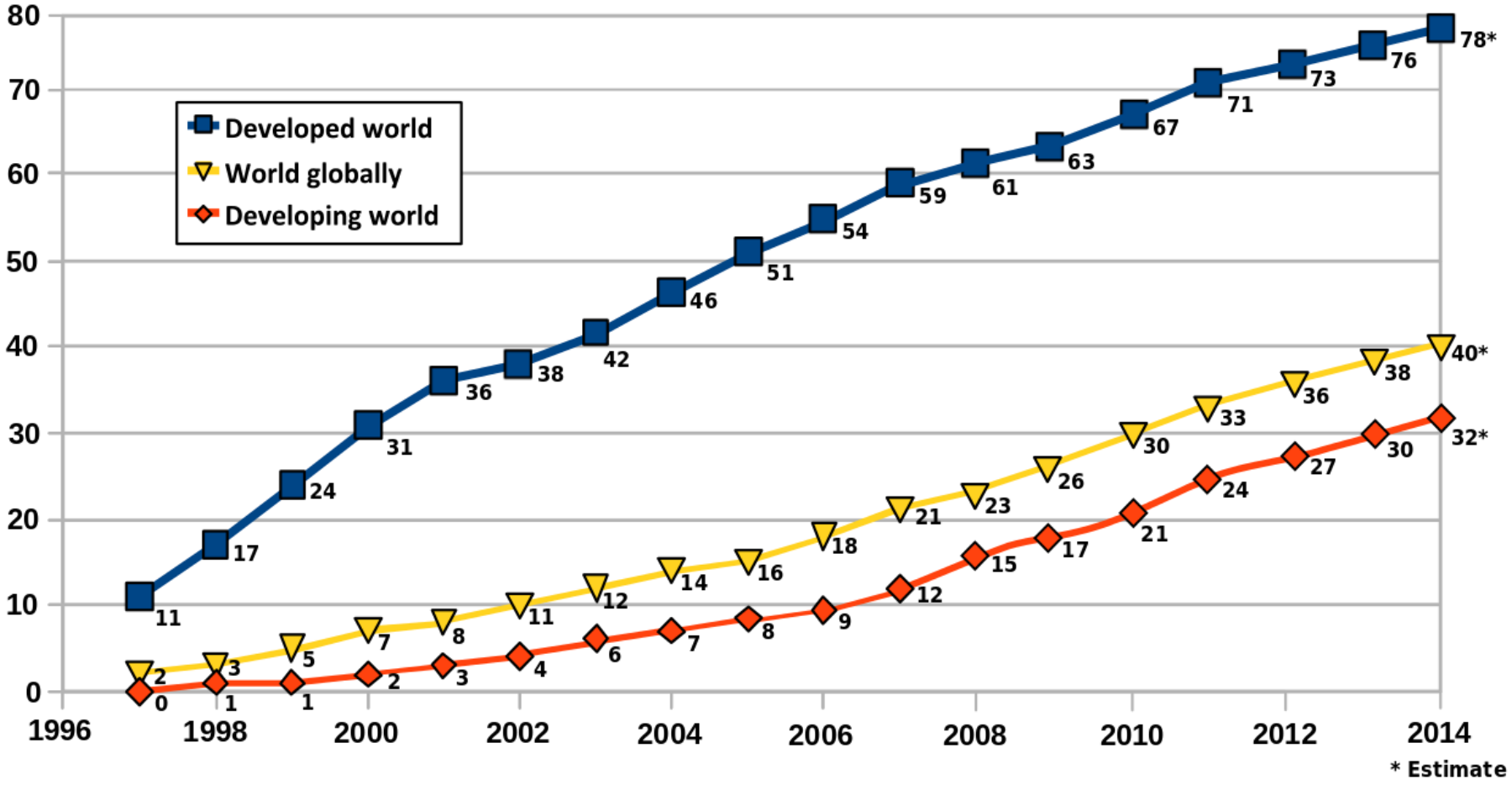
1.1.1 信息过载与大数据

2016互联网趋势报告:全球互联网用户数超30亿

1.1.1 信息过载与大数据

- **数据爆炸性的增长，而人的处理能力有限**

全球范围内平均每100人中的互联网用户数量



1.1.1 信息过载与大数据

- **数据爆炸性的增长，而人的处理能力有限**

2015年12月至2016年12月中国互联网基础资源对比

	2015 年 12 月	2016 年 12 月	年增长量	年增长率
IPv4（个）	336,519,680	338,102,784	1,583,104	0.5%
IPv6（块/32）	20,594	21,188	594	2.9%
域名（个）	31,020,514	42,275,702	11,255,188	36.3%
其中.CN 域名（个）	16,363,594	20,608,428	4,244,834	25.9%
网站（个）	4,229,293	4,823,918	594,625	14.1%
其中.CN 下网站（个）	2,130,791	2,587,365	456,574	21.4%
国际出口带宽（Mbps）	5,392,116	6,640,291	2,521,628	23.1%

1.1.1 信息过载与大数据

- 数据爆炸性的增长，而人的处理能力有限

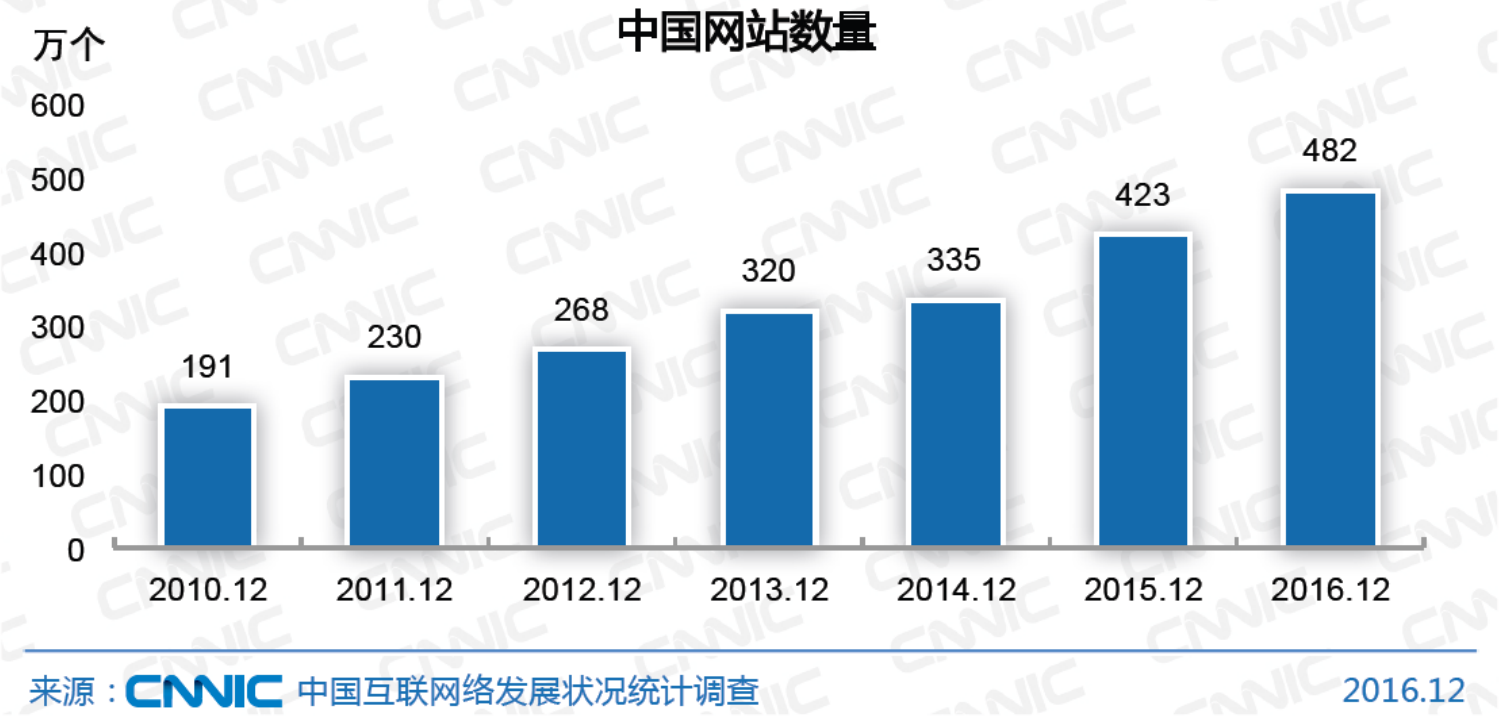


图 3　中国网站数量

注：数据中不包含.EDU.CN 下网站

1.1.1 信息过载与大数据

- 数据爆炸性的增长，而人的处理能力有限

1.1.1 信息过载与大数据

- 数据爆炸性的增长，而人的处理能力有限
  - 视频

1小时	YouTube上每秒上传视频的小时数
35万	YouTube平均每天上传视频的用户数量
超过10亿	YouTube用户数量

1.1.1 信息过载与大数据

- 数据爆炸性的增长，而人的处理能力有限
  - 图片

80亿	Flickr用户上传的图片总数
350万	Flickr用户每天上传的图片数量
2500亿	Facebook用户上传的图片总数
3.5亿	Facebook用户每天上传的图片数量

1.1.1 信息过载与大数据

截至2015年底国家/地区的互联网用户数量排行表（取前10位）

国家/地区	互联网用户数量	排序	互联网用户数占国家/地区总人口比例	排序
中国	692,152,618	1	50.30%	90
印度	340,873,137	2	26.00%	127
美国	239,882,242	3	74.55%	40
巴西	122,796,320	4	59.08%	71
日本	118,131,030	5	93.33%	9
俄罗斯	105,311,724	6	73.41%	43
尼日利亚	86,436,611	7	47.44%	95
墨西哥	72,945,992	8	57.43%	85
德国	70,675,097	9	87.59%	16
英国	59,538,545	10	92.00%	27

1.1.1 信息过载与大数据

Web2.0：信息爆炸

- 大众从内容消费者变成内容生产者

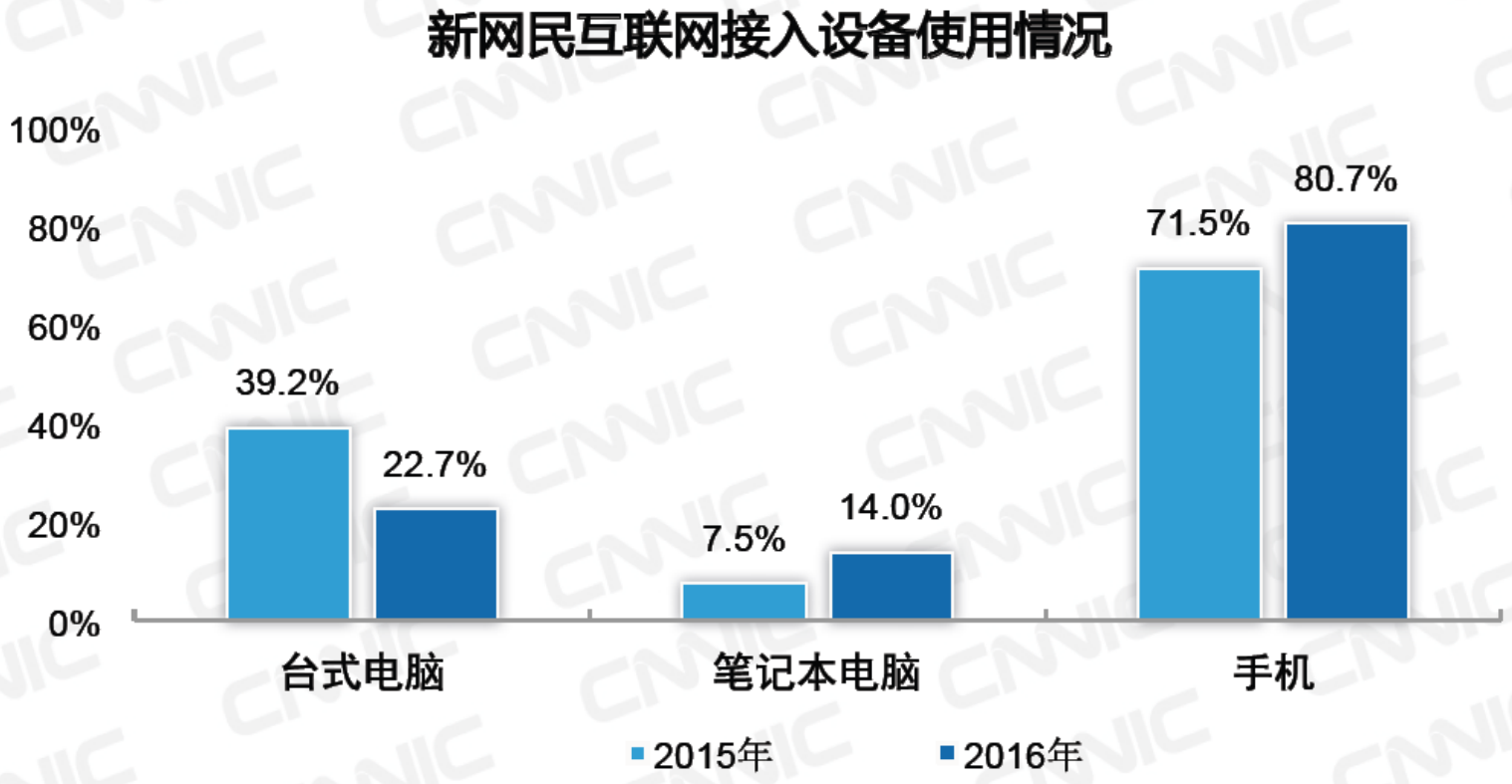
- 维基、论坛、博客
  SNS、Facebook、Twitter
  微博、微信等应用

- 在人与人的交互中产生了大量的信息，甚至足以推翻一个总统（埃及总统辞职背后就有社交网络的推动），如何应对？

1.1.1 信息过载与大数据

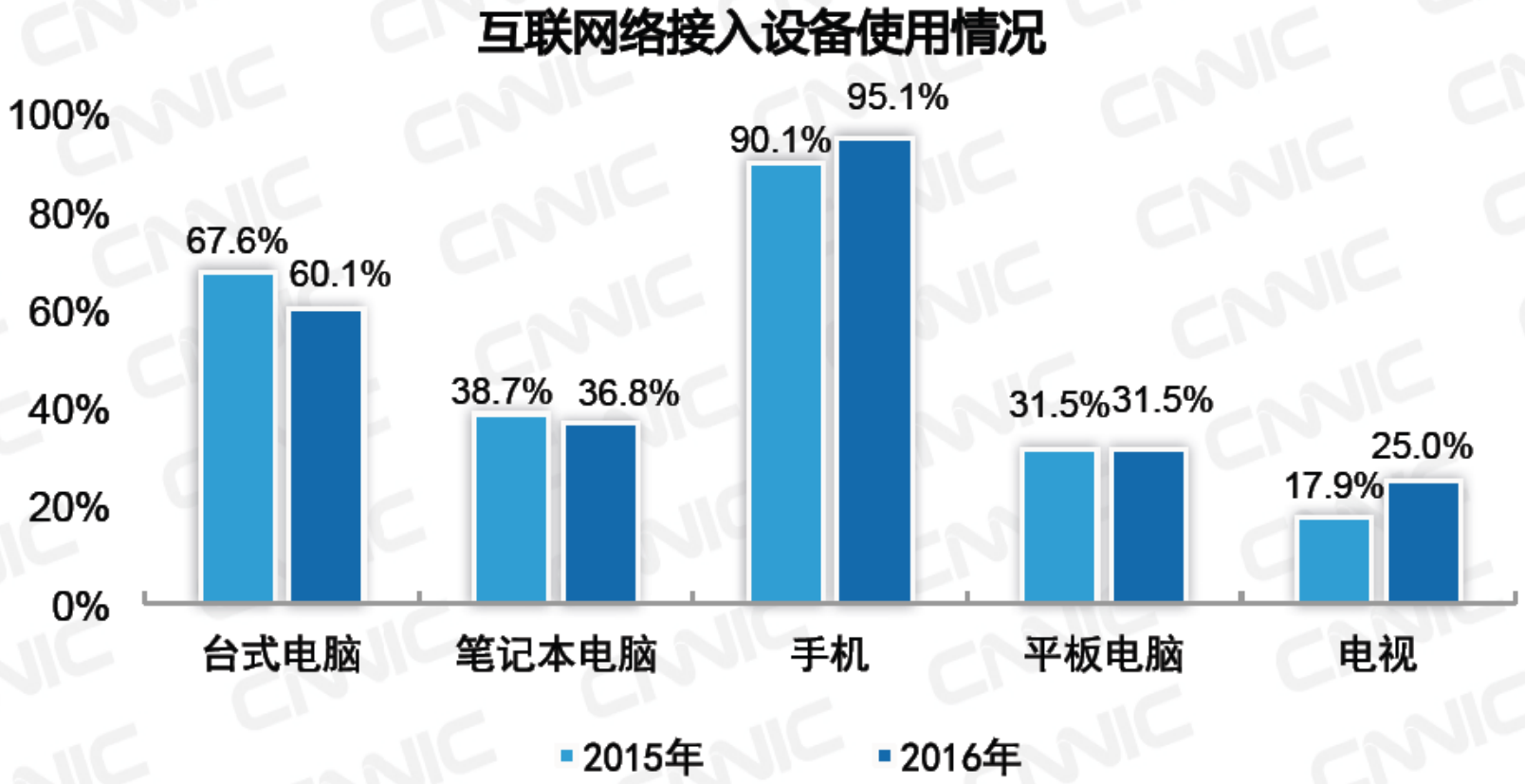
来源：CNNIC 中国互联网络发展状况统计调查 2016.12

1.1.1 信息过载与大数据



来源：CNNIC 中国互联网络发展状况统计调查 2016.12

1.1.1 信息过载与大数据

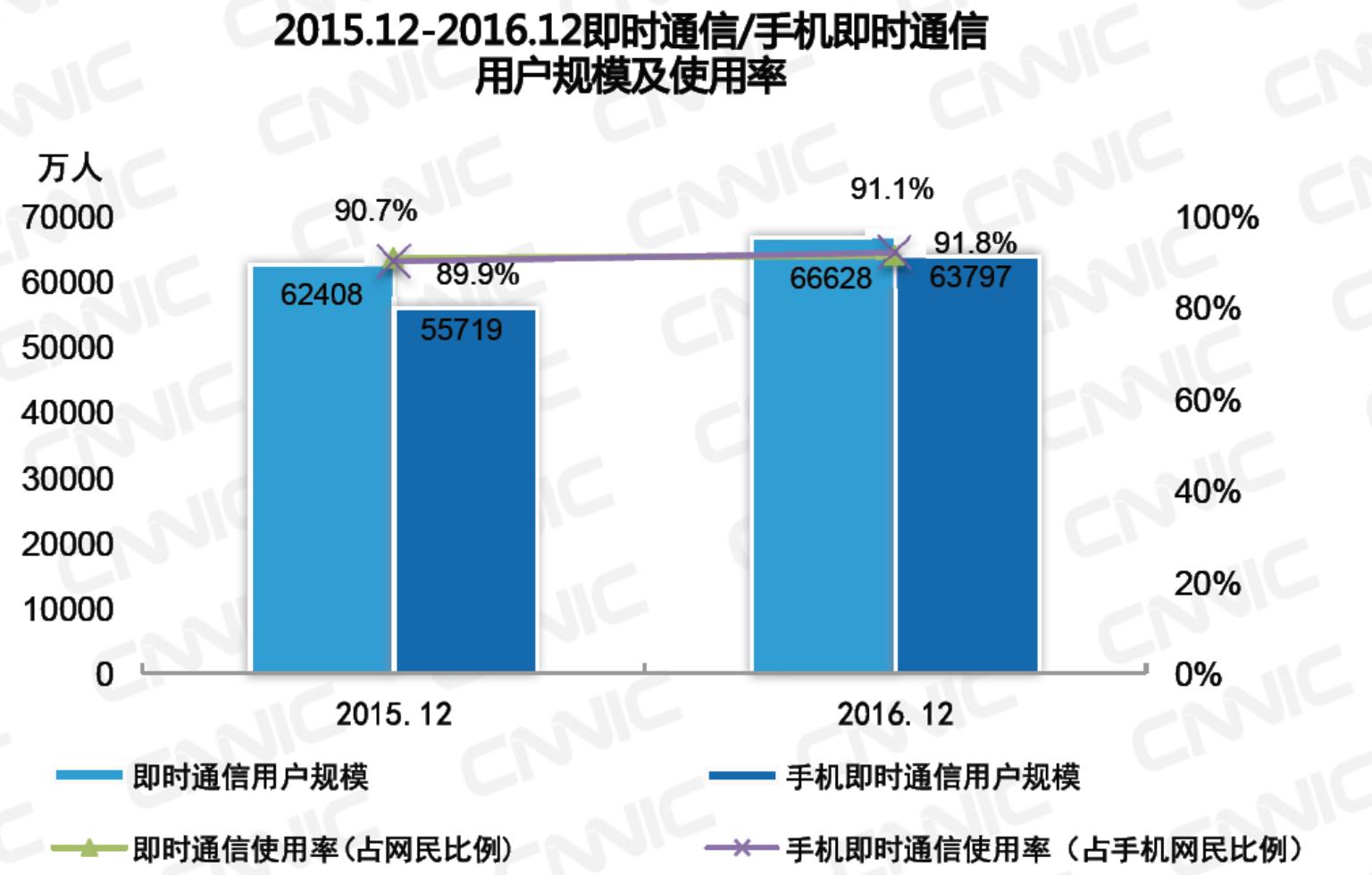


来源：CNNIC 中国互联网络发展状况统计调查 2016.12

1.1.1 信息过载与大数据

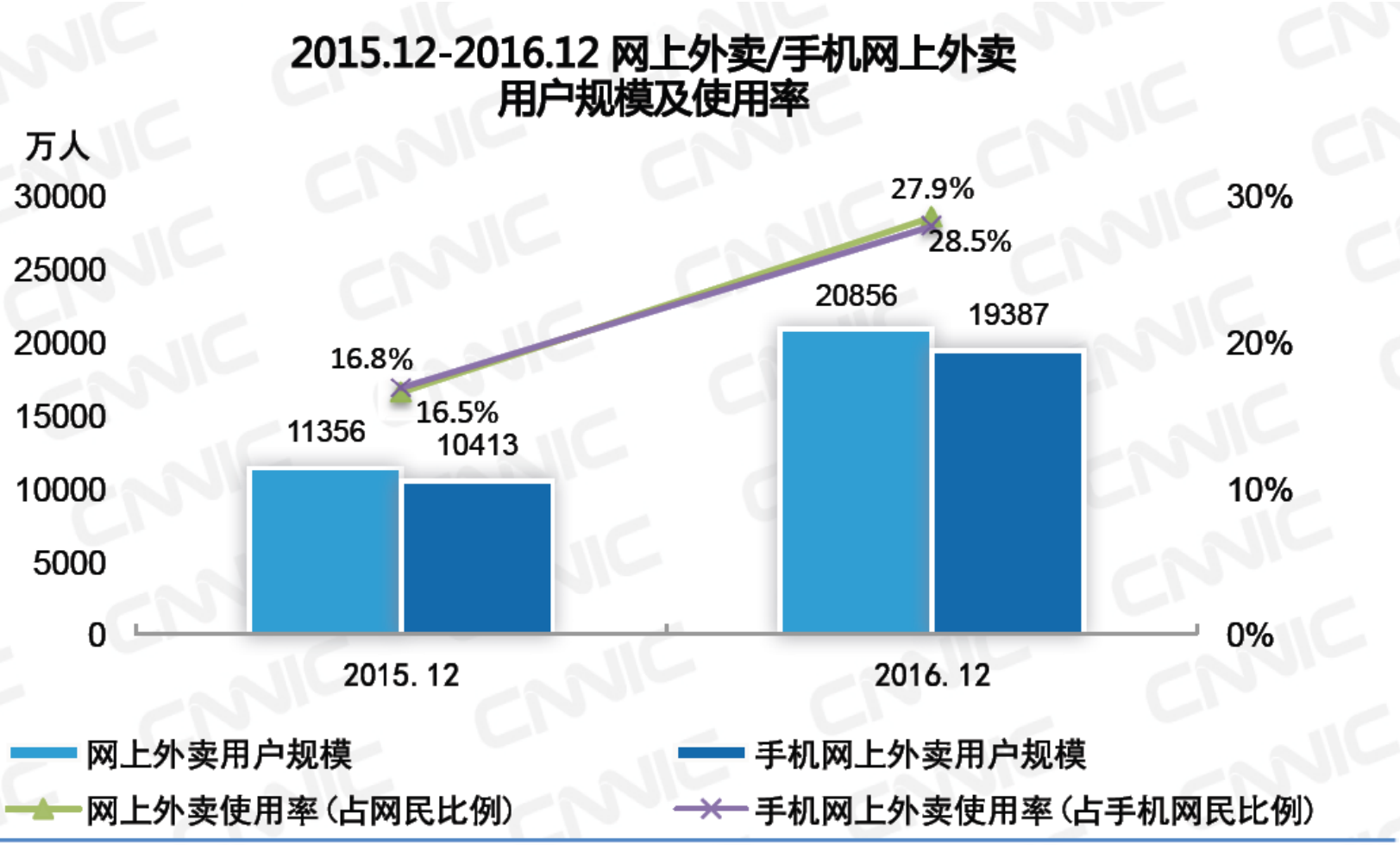
来源：CNNIC 中国互联网络发展状况统计调查 2016.12

1.1.1 信息过载与大数据



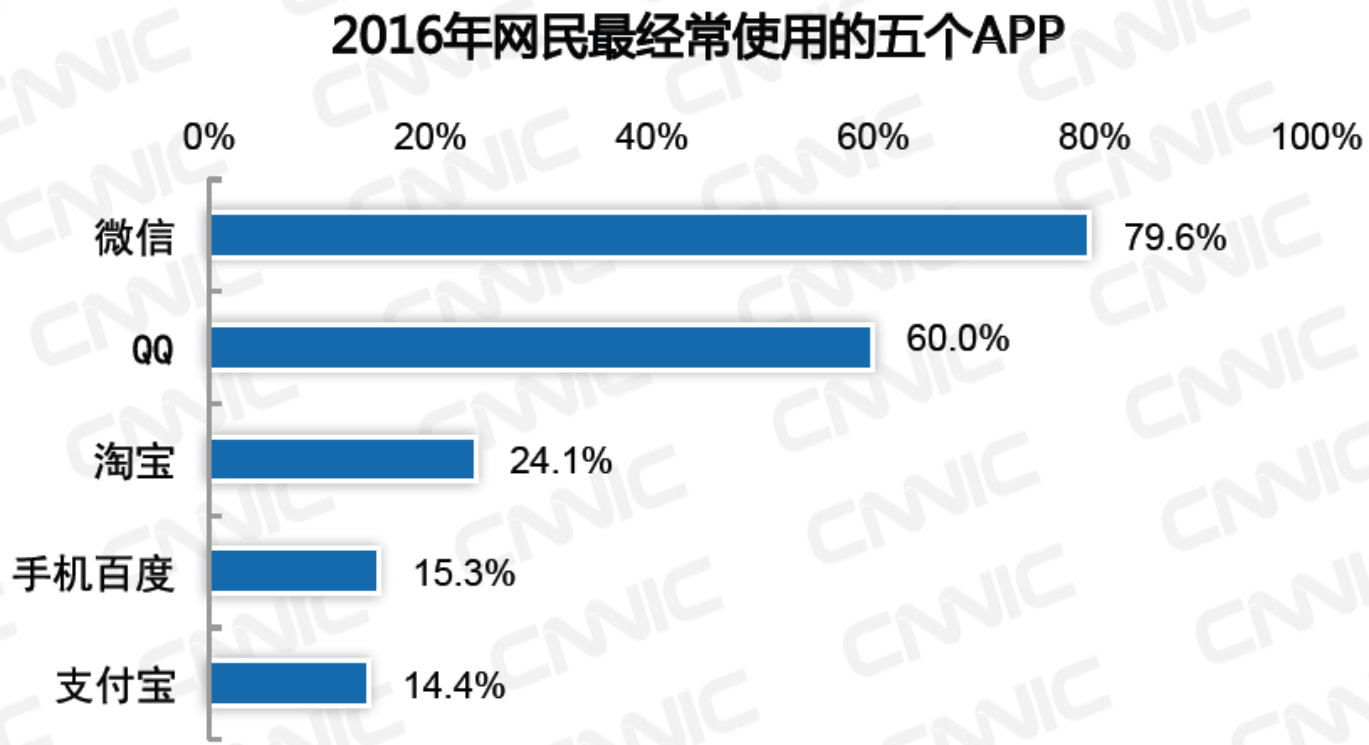
来源：CNNIC 中国互联网络发展状况统计调查 2016.12

1.1.1 信息过载与大数据



来源：CNNIC 中国互联网络发展状况统计调查 2016.12

1.1.1 信息过载与大数据



来源：CNNIC 中国互联网络发展状况统计调查 2016.12

1.1.1 信息过载与大数据

逐步走进大数据时代

- 1998年，美国前副总统戈尔提出数字化地球的概念。
- 1998年，江泽民总书记提出数字中国战略构想。
- 全世界各种数字化应用：数字图书馆、数字博物馆、数字电影、交互电视、会议电视、远程教育、遥感、GPS等产生大量文本和多媒体数据。
- 2005年11月17日，突尼斯， WSIS信息社会世界峰会，“无所不在的‘物联网’通信时代即将来临，世界上所有的物体从轮胎到牙刷、从房屋到纸巾都可以通过因特网主动进行交换。”

1.1.1 信息过载与大数据

逐步走进大数据时代

- 2009年8月，温家宝总理在无锡视察时，提出了“感知中国”，自此，物联网被正式列为国家五大新兴战略性产业之一，写入“政府工作报告”。
- 2011，巴塞罗那，移动通信世界大会，移动互联网这一正在迅速扩大的新兴市场，已经成为通信产业链各个企业的着力方向
- 国际数据公司（IDC）的研究结果表明，2008年全球产生的数据量为0.49ZB，2009年为0.8ZB，2010年为1.2ZB，2011年为1.82ZB，2012年更是增长为2.8ZB.（1ZB = 1024EB，1EB = 1024PB，1PB = 1024TB，1TB = 1024GB).

1.1.1 信息过载与大数据

逐步走进大数据时代

- 2012年，随着《大数据时代：生活、工作与思维的大变革》一书的出版，大数据时代的概念进入公众的视野并得到极大的关注
- 2012年3月22日，奥巴马政府宣布投资2亿美元拉动大数据相关产业发展，将“大数据战略”上升为国家战略。奥巴马政府甚至将大数据定义为“未来的新石油”。
- 2016年12月18日，国家工业和信息化部，编制了《大数据产业发展规划（2016－2020年）》。

1.1.1 信息过载与大数据

大数据时代的“4 V”特征

- Volume
  - 数据量大，全球每年产生的数据总量已经达到ZB级别
- Variety
  - 数据种类繁多，如文本、图片、视频、地理信息、各种传感器信息等
- Velocity
  - 数据流动速度快，对数据处理的时效性要求高
- Value
  - 大数据蕴含着巨大的价值，可以帮助人们解决数据量不足时所不能解决的问题

1.1.1 信息过载与大数据

信息过载的负面影响

信息过载是信息时代信息极大丰富的负面影响之一。

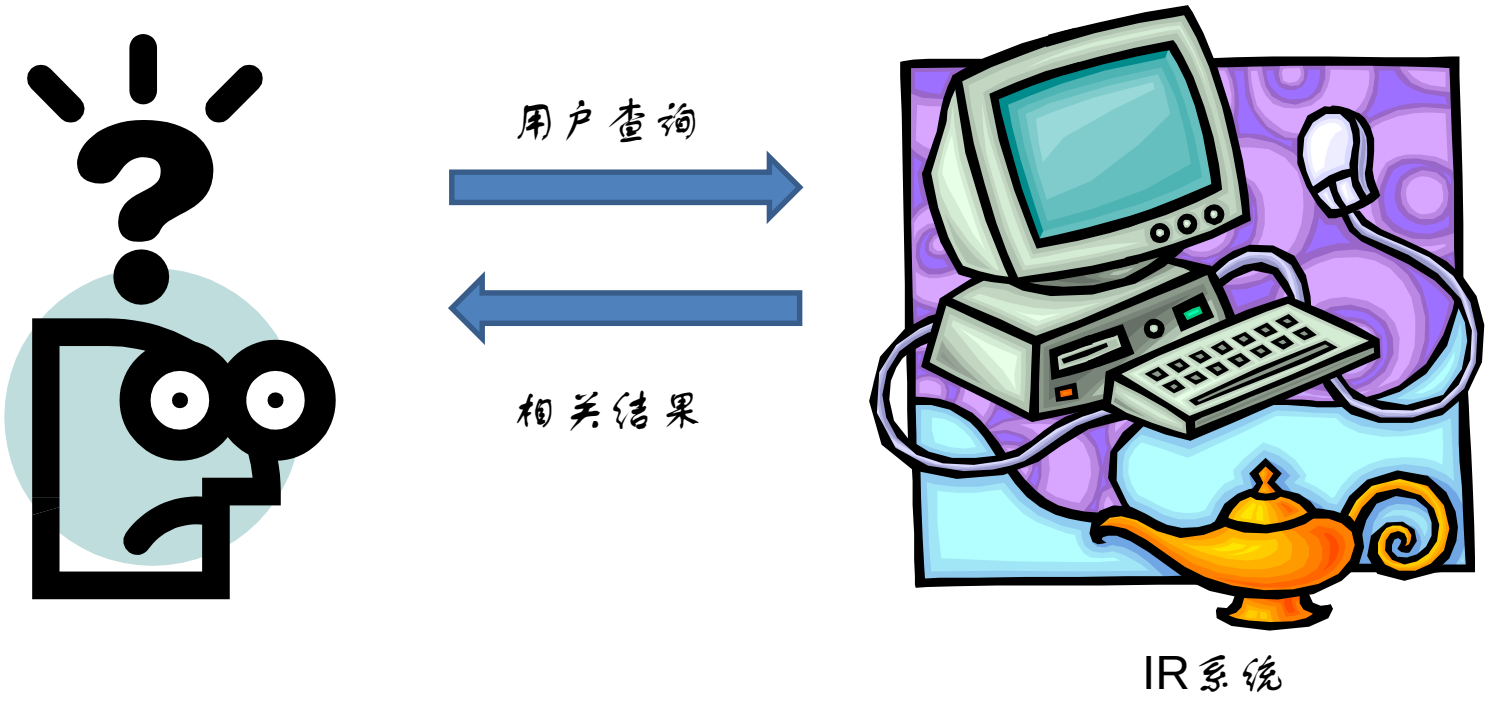
信息过载指的是社会信息超过了个人或系统所能接受、处理或有效利用的范围，并导致故障的状况。主要表现为：

1. 受传者对信息反映的速度远远低于信息传播的速度；
2. 大众媒介中的信息量大大高于受众所能消费、承受或需要的信息量；
3. 大量无关的、没用的、冗余的信息严重干扰了受众对相关有用信息的准确分析和正确选择。

1.1.1 信息过载与大数据

信息过载，
如何解决？

- **信息检索**可以帮助人们从海量的数据中快速的找到有用的信息
- **数据挖掘**可以从大数据中提取出隐含的、先前未知的并有潜在价值的信息



1.1.2 信息检索

- Information Retrieval这个术语产生于 Calvin Mooers 1948年在MIT的硕士论文。
- Information Retrieval(IR)：从大规模非结构化数据（通常是文本）的集合（通常保存在计算机上）中找出满足用户信息需求的资料（通常是文档）的过程。
- 作为一门学科，是研究信息的获取(acquisition)、表示(representation)、存储(storage)、组织(organization)和访问(access)的一门学问。

1.1.2 信息检索

- 信息检索可以看成计算机科学(Computer Science)和图书情报学(Library & Info. Science)的交叉学科。
- 以计算机为手段，处理信息对象和其他学科也融合：语言学、认知科学
- 检索来自英文单词Retrieval，有些人把它翻译成获取。其本义是“获得与输入要求相匹配的输出”。
- 注意：和我们平时所理解的搜索意义上的检索不一样。（这里提醒一下：与Search的区别）

1.1.2 信息检索

- IR不仅仅是搜索，IR系统也不仅仅是搜索引擎。
  - 例1：返回与信息检索相关的网页 ->搜索引擎(Search Engine,SE)
  - 例2：曾哥是狮子座的吗？->问答系统(Question Answering,QA)
  - 例3：返回Ipad的各种型号、配置、价格等 ->信息抽取 Information Extraction, IE)
  - 例4：使用Google Reader订阅新闻，并获取推荐->信息过滤(Information Filtering)、信息推荐(Information Recommending)

1.1.2 信息检索

IR系统甚至可以是试衣间！（视频）

1.1.3 数据挖掘

- 数据挖掘（Data Mining）从*大量的*、*不完全的*、*有噪声的*、*模糊的*、*随机的*实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程
- 数据挖掘的基本内容
  - 特征提取、分类、聚类
  - 话题检测、自动摘要
  - 智能问答等

1.1.3 数据挖掘

- 数据挖掘与信息检索的关系

数据挖掘可以认为是信息检索的一个模块

挖掘出知识可以更好的检索

特别是针对有多媒体信息必须要数据挖掘来**跨越“语义鸿沟”**

为什么要开这门课？

1.1.4 本课程的意义

市场发展的需求

- **用户需要信息检索技术：**互联网的信息量巨大，寻找有用信息耗时耗力
- **公司需要信息检索技术：**搜索引擎改变了人们获取信息的方式，并成为了各大科技公司的一项基本服务，Google、Microsoft、Baidu，还有如Tencent、Sina、Sohu、360等都加入到搜索技术的竞争当中
- **人才的竞争：**搜索相关人才紧缺，成为各大科技巨头争相抢夺的重要资源

1.1.4 本课程的意义

对信息类研究生的基本要求

- 信息检索将会成为一门计算机专业和信息处理学科的基础方向
  - 搜索的三个层次：
    - 应用层次
    - 中间工具层次
    - 核心层次
- 目前国外已经开课多年，国内已经有些大学在本科阶段就开始上信息检索课，中国科大要加大步伐跟进。

1.1.4 本课程的意义

课程特点

- 不是教如何使用信息检索工具（学校有专门的课程），而是了解信息检索工具背后的基本原理和技术，并且能够进行深层的研究或开发相关的应用。

- 基本原理+广泛实践

提纲

1.2.1 信息检索的历史

历史分段

- 计算机出现以前
- 计算机出现以后
- Internet出现以后

1.2.1 信息检索的历史

- 计算机出现以前：
  - 约4000年前，人类就开始有目的地组织信息，一个典型的例子就是图书中的目录。
  - 随后，逐渐出现索引的概念，即从一些词和概念指向相关信息或者文档的指针。
  - 计算机问世以前，人们主要通过手工方式来建立索引。
  - 例子：词典（拼音检字、部首笔画检字等）

1.2.1 信息检索的历史

- 1948年：
  - C. N. Mooers在其MIT的硕士论文中第一次创造了“Information Retrieval”这个术语。
- 1960－70年代：
  - 人们开始使用计算机为一些小规模科技和商业文献的摘要建立文本检索系统。
  - 产生了布尔模型(Boolean Model)、向量空间模型(Vector Space Model)和概率检索模型(Probabilistic Model)。康奈尔大学的Salton领导的研究小组是该领域研究的佼佼者。
  - 伦敦城市大学的Robertson及剑桥大学的Sparck Jones是概率模型的倡导者。

1.2.1 信息检索的历史

- 1980年代：
  - 出现了一些商用的较大规模数据库检索系统
  - Lexis-Nexis
  - Dialog
  - MEDLINE

1.2.1 信息检索的历史

- 1986：Internet正式形成
- 1990’s:
  - 第一个网络搜索工具：1990年加拿大蒙特利尔麦吉尔(McGill)大学开发的FTP搜索工具Archie
  - 第一个WEB搜索引擎：1994年美国CMU开发的Lycos
  - 1995：斯坦福大学博士生开发的Yahoo
  - 1998：斯坦福大学博士生开发的Google，提出PageRank计算公式。
  - 1998：基于语言模型的IR模型提出。

1.2.1 信息检索的历史

- 1990年代的其他重要事件：
  - 评测会议
    - NIST TREC
  - 推荐系统的出现
    - Ringo
    - Amazon
    - NetPerceptions
- 文本分类和聚类的使用

1.2.1 信息检索的历史

- 2000’s
  - 信息抽取
    - Whizbang
    - Fetch
    - Burning Glass
  - 问答系统
    - TREC Q/A track
  - 2000年，百度成立

1.2.1 信息检索的历史

- 2000以来的其他重要事件：
  - 多媒体 IR
    - Image
    - Video
    - Audio and music
  - 跨语言 IR
    - DARPA Tides
  - 文本摘要
    - DUC评测

1.2.2 信息检索在工业界的发展

1993	W3Catalog	Launch
	Aliweb	Launch
	JumpStation	Launch
1994	WebCrawler	Launch
	Go.com	Launch
	Lycos	Launch
1995	AltaVista	Launch
	Daum	Founded
	Open Text Web Index	Launch [2]
	Magellan	Launch
	Excite	Launch
	SAPO	Launch
	Yahoo!	Launched as a directory

1995年yahoo!**作为一个目录导航系统发布**，网站收录/更新都要靠人工维护，所以在信息量剧增的条件下，就不是非常实用

1.2.2信息检索在工业界的发展（续）

1996	Dogpile	Launch
	Inktomi	Launch
	HotBot	Founded
	Ask Jeeves	Founded
1997	Northern Light	Launch
	Yandex	Launch
1998	Google	Launch
	MSN Search	Launch
1999	AlltheWeb	Launch
	GenieKnows	Founded
	Naver	Launch
	Teoma	Founded
	Vivisimo	Founded
2000	Baidu	Founded
	Exalead	Founded
2002	Inktomi	Acquired by Yahoo
2003	Info.com	Launch

1998年Google诞生，凭借独特的PageRank技术，使它很快后来居上，成为当前全球最受欢迎的搜索引擎；2000年Baidu诞生，国内垄断的搜索引擎

1.2.2 信息检索在工业界的发展（续）

2004	Yahoo! Search	Launched own web search
	A9.com	Closed
	Sogou	Launch
2005	Ask.com	Launch
	GoodSearch	Launch
	SearchMe	Founded
2006	wikiseek	Founded
	Quaero	Founded
	Ask.com	Launch
	Live Search	Launched as rebranded MSN Search
	ChaCha	Launch
	Guruji.com	Launch
2007	wikiseek	Closed
	Sproose	Closed
	Wikia Search	Launched
	Blackle.com	Launched

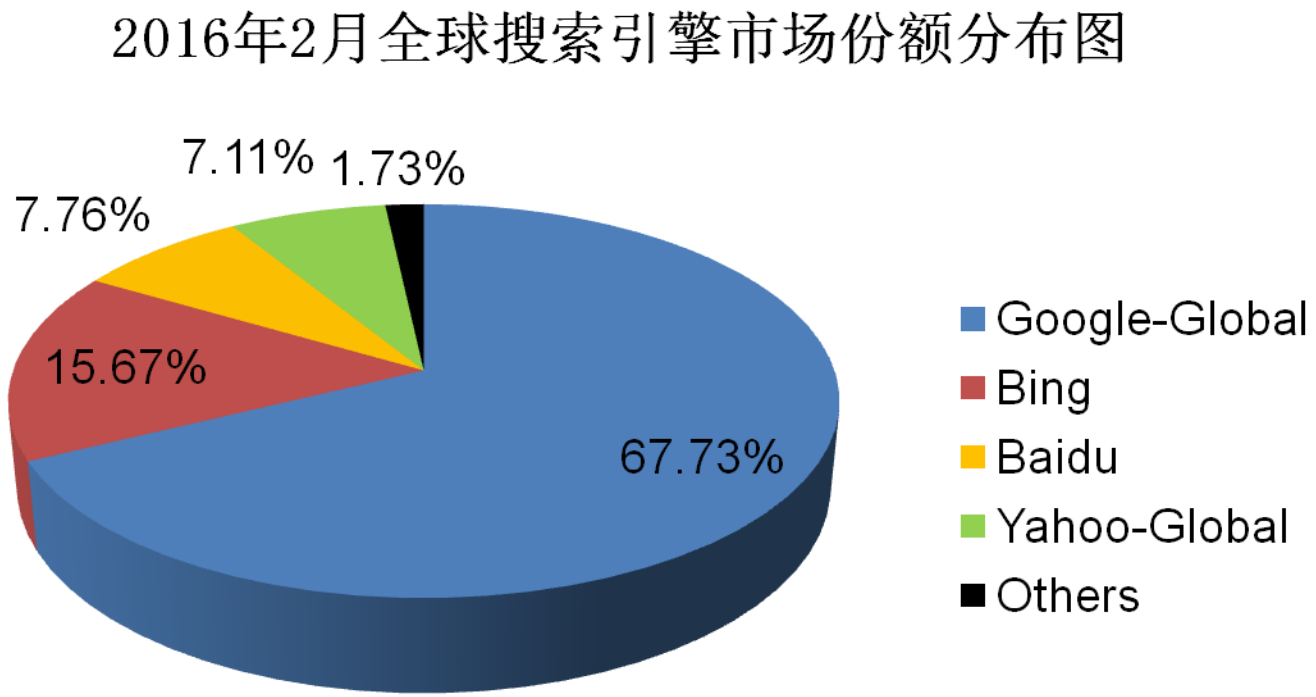
2002年开始很多公司受搜索市场前景和Google神话的吸引，积极进入搜索引擎市场，谋求一席之地。但是大都难以撼动Google的地位。

1.2.2 信息检索在工业界的发展（续）

2008	Powerset	Acquired by Microsoft
	Picollator	Closed
	Viewzi	Closed
	Cuil	Launched
	Boogami	Launched
	LeapFish	Beta Launch
	Forestle	Launched
	VADLO	Launched
	Duck Duck Go	Launched
2009	Bing	Launched as rebranded Live Search
	Yebol	Beta Launch
	Mugurdy	Closed due to a lack of funding
	Goby	Launched
2010	Yandex	Launched global (English) search
	Cuil	Closed
	Blekko	Beta Launch
	Viewzi	Closed due to a lack of funding
	Yummly	Launched
2011	Interred	Active

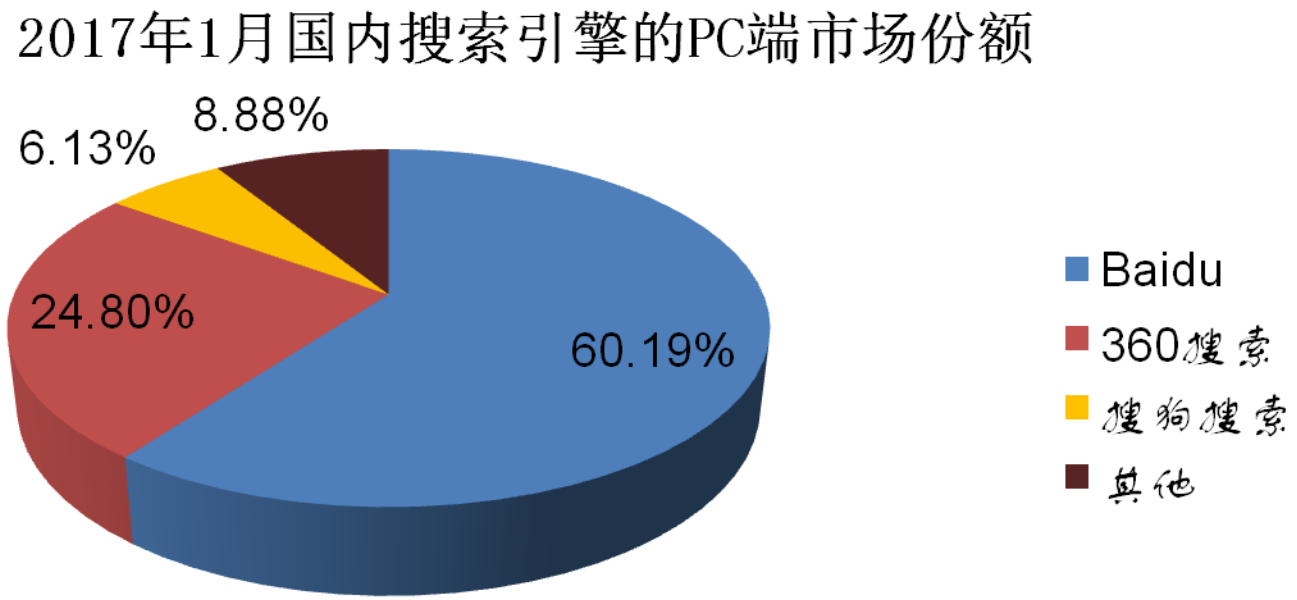
其中一些搜索引擎建立的时候号称“google杀手”，没有杀掉Google，自己反而倒闭，如Cuil。

1.2.2 信息检索在工业界的发展（续）



数据来源：IDC评述网

1.2.2 信息检索在工业界的发展（续）



数据来源：StatCounter Global Stats

1.2.3 信息检索在学术界的发展

- 从SIGIR（IR顶级会议）看信息检索发展
  - 过去四十多年，我们在信息检索的路上走了多远？
  - 在IR舞台上，什么是长盛不衰的？

    哪些已经渐渐谢幕？

    哪些即将登场？
  - SIGIR 1971～2011年所有正式论文

1.2.3 信息检索在学术界的发展（续）

年	结构化	通用	模型	布尔	概率/LM	问答	NLP	概念	概念/NLP	反馈	索引/实现	权值计算	过滤	Web	link分析	多媒体	交互/用户/界面	分布式	分类	聚类	摘要	跨/多语言	片段理解	特定领域	LSI	评价
71	8	5	1			1	1	1	2		1															
78	4	2								1	1									2						
79	1	9								1		1	1										1			
80	7	2	2		1	1		4	4	1		3				1				1			1			
81		9	1		1	1		4	4			2					2			1						
82	6	5	1		1	1	1	1	2	2	1	1						1								
83	10	7		1				3	3			2					3			2						
84	2	10	4				1	3	4		1	1				1	3	2								
85	3	10	1	2				4	4	1	2	1	1				2	1		3		1				
86	5	6	2	1			2	3	5	1	3	5					3			3			1			
87	2	10	1				1	5	6		3	3		1		1	5	2		2						
88	5	6	2		3	1	6	7						3											1	
89	2	2	1		1		3	8			4			1			3									
90	4	5	2	1			3	1			4			1			2									
91	1	8			3		2	6	8	1	4		1	2		1	3	1								
92		6	2		4		3	3	6	2	2	1		1			3			1			3		1	
93	1	2	2		2	1	2	5	7	4	2			2			3					2				3
94	1	2	2		2		1	3	3	3	2			1			5					3	2		1	4
95	2	4	2		3	1	1							1		2	3		3	1	2	2				4
96		3	3		2		2	3		1				1					1			4				2
97		1	1	1	1		1	4	5	2		1	1	4		1	3	1	1	2		4	3			1
98					3	1	1	1	2	1	2	1	1	3	1		2	3	3	1		4	4			7
99		4		1	1		3	3	6	1	1		1	1			4	4	1		2	1	2		2	
00		2		1		4	1	2	3	1			1	5	3		1	2	3	1	2	1	4		1	2
01	2	5			3	4	1	1	2	1	2	1	2	2	3		1	1	3		3	3	5			3
02		1			3	1	1	1	2		3	1	3	2	2			1	2	3	3	4	5			8
03	4				4	3	1		1				2	1	3	5	4	3	6			3	3	1	1	1
04	3	2			9		7		7	1	1	1	4	1	4	2	3	1	3	4	1	4	1	1	3	4
05	3	1	3		4	3	3	1	4	4	2	4	3	4	3	7	5	4	5		3	3			2	4
06	1	5	1		2	3	1	2	3	4	2	1		2	5	3	7	3	3	3	3	2		4	1	6
07		2	4		4	4	3		3		6	6	3	3	3	8	6	3	4	2	3	2		3	3	9
08		5	4		9	3				6	4	7	5			2	6		7	6	4	3	3	3		7
09		6	9		3	3		3	3	3	3	6	6	6		6	3	4	2	1	3		3	3		6
10	4	4	4		3					4		7	4	10	4	7	6		3	6	6	6	3			7
11		5	6		5	4	4		4	4	4	4	4	8	1	6	16	1	3	3	3	3		3	3	7

1.2.3 信息检索在学术界的发展

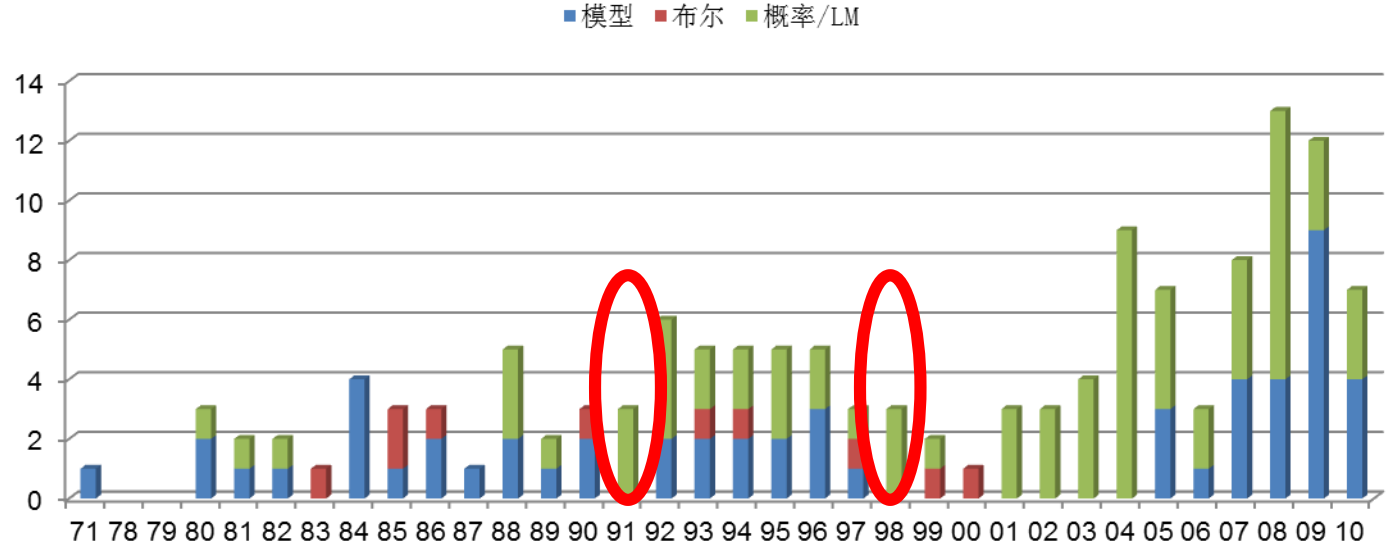
结构化数据 vs 非结构化数据

- 从一开始就沿两条路发展
  - 来源于结构化数据处理的灵感
    - e.g. 数据库
  - 直接从自由文本处理的角度
- 前10年，并驾齐驱，结构化方法占有一定的主导地位
- 进入1990年代之后，结构化数据存储相对沉寂
- 进入2000年，开始复苏
  - 思路转变——基于以xml为代表的半结构化数据的检索
  - 两条路逐渐呈现融合趋势

年	结构化	通用
71	8	5
78	4	2
79	1	9
80	7	2
81		9
82	6	5
83	10	7
84	2	10
85	3	10
86	5	6
87	2	10
88	5	6
89	2	2
90	4	5
91	1	8
92		6
93	1	2
94	1	2
95	2	4
96		3
97		1
98		
99		4
00		2
01	2	5
02		1
03	4	
04	3	2
05	3	1
06	1	5
07		2
08		5
09		6
10	4	4
11		5

1.2.3 信息检索在学术界的发展
检索模型

- 自由文本模型——三个阶段
  - 向量空间模型 ——80年代初的重点
  - 概率模型 —— 80年代末兴起，90年代逐渐成为主流
  - 基于语言模型的检索 —— 1998年，里程碑
  - 更多模型 —— 大约从1999年开始，标志IR进入新的阶段



1.2.3 信息检索在学术界的发展

关键技术

- 实现
  - 早期
    - 倒排索引的提出与研究
  - 2000后
    - 大规模检索和检索效率
  - 最近
    - 垃圾信息应对 …
- 走出实验室
  - 面向海量数据、实时处理、真实网络环境…

年	结构化	通用	概率/LM	相关反馈	索引/实现	分布式
1971	8	5			1	
1978	4	2		1	1	
1979	1	9		1		
1980	7	2	1	1		
1981		9	1			
1982	6	5	1	2	1	1
1983	10	7				
1984	2	10			1	2
1985	3	10		1	2	1
1986	5	6		1	3	
1987	2	10			3	2
1988	5	6	3	1	5	
1989	2	2	1		4	1
1990	4	5			4	
1991	1	8	3	1	4	1
1992		6	4	2	2	
1993	1	2	2	4	2	
1994	1	2	2	3	2	
1995	2	4	3		2	3
1996		3	2	1		1
1997		1	1	2		1
1998			3	1	2	3
1999		4	1	1	1	4
2000		2		1		2
2001	2	5	3	1	2	1
2002		1	3		3	1
2003	4		4			3
2004	3	2	9	1	1	1
2005	3	1	4	4	2	4
2006	1	5	2	4	2	3
2007		2	4		6	3
2008		5	9	6	4	
2009		6	3	3	3	4
2010	4	4	3	4		
2011		5	5	4	4	1

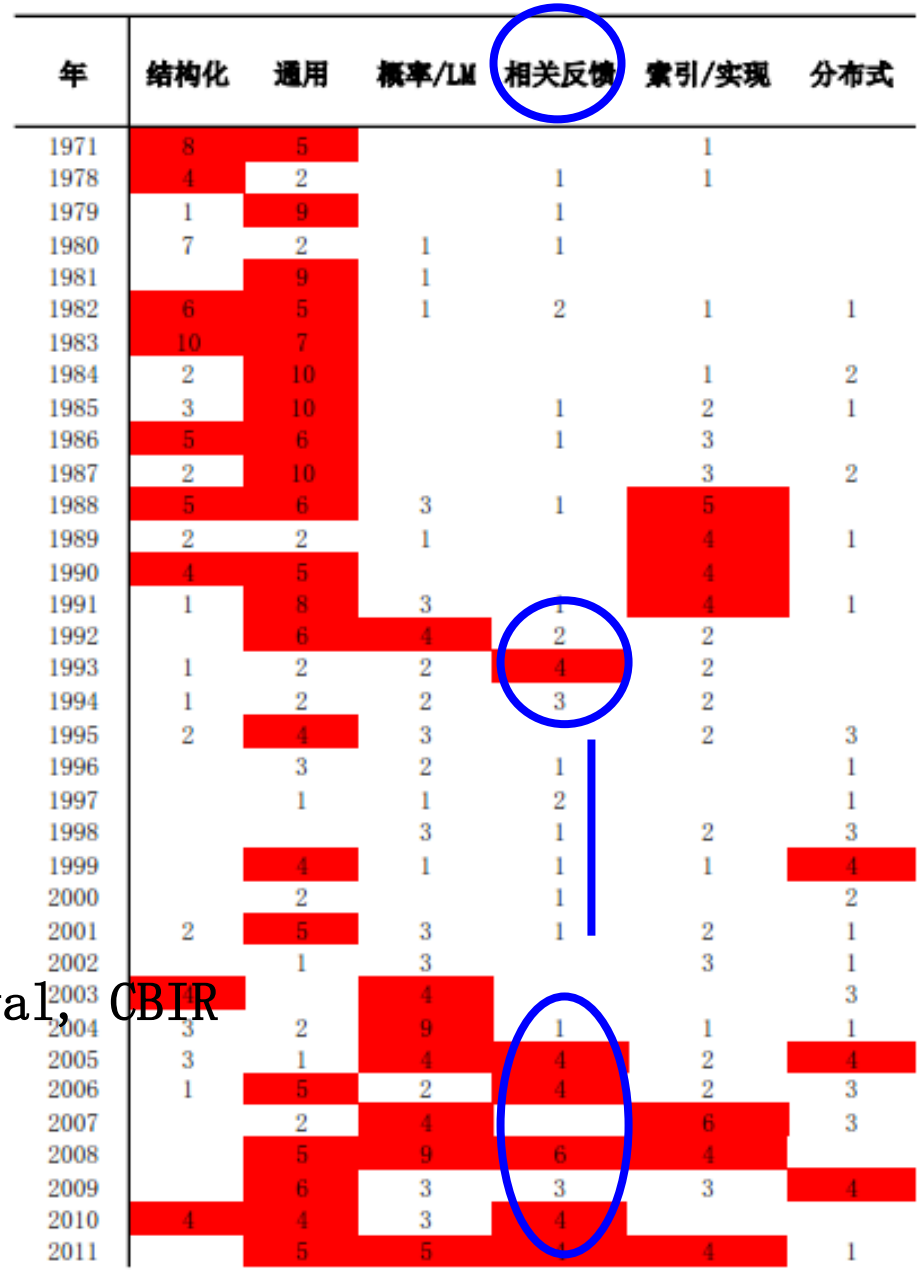
1.2.3 信息检索在学术界的发展

关键技术

- 相关反馈
- 经久不衰的话题
- 3个阶段
  - 早期
    - 建立反馈机制
  - 90年代中
    - 基于内容的图像信息检索

    Content-Based Image Retrieval, CBIR

    近来（对信息检索质量）
    - 区分不同主题
    - 区分不同词



年	结构化	通用	概率/LM	相关反馈	索引/实现	分布式
1971	8	5			1	
1978	4	2		1	1	
1979	1	9		1		
1980	7	2	1	1		
1981		9	1			
1982	6	5	1	2	1	1
1983	10	7				
1984	2	10			1	2
1985	3	10		1	2	1
1986	5	6		1	3	
1987	2	10			3	2
1988	5	6	3	1	5	
1989	2	2	1		4	1
1990	4	5			4	
1991	1	8	3	1	4	1
1992		6	4	2	2	
1993	1	2	2	4	2	
1994	1	2	2	3	2	
1995	2	4	3		2	3
1996		3	2	1		1
1997		1	1	2		1
1998			3	1	2	3
1999		4	1	1	1	4
2000		2		1		2
2001	2	5	3	1	2	1
2002		1	3		3	1
2003	4		4			3
2004	3	2	9	1	1	1
2005	3	1	4	4	2	4
2006	1	5	2	4	2	3
2007		2	4		6	3
2008		5	9	6	4	
2009		6	3	3	3	4
2010	4	4	3	4		
2011		5	5	4	4	1

1.2.3 信息检索在学术界的发展

关键技术

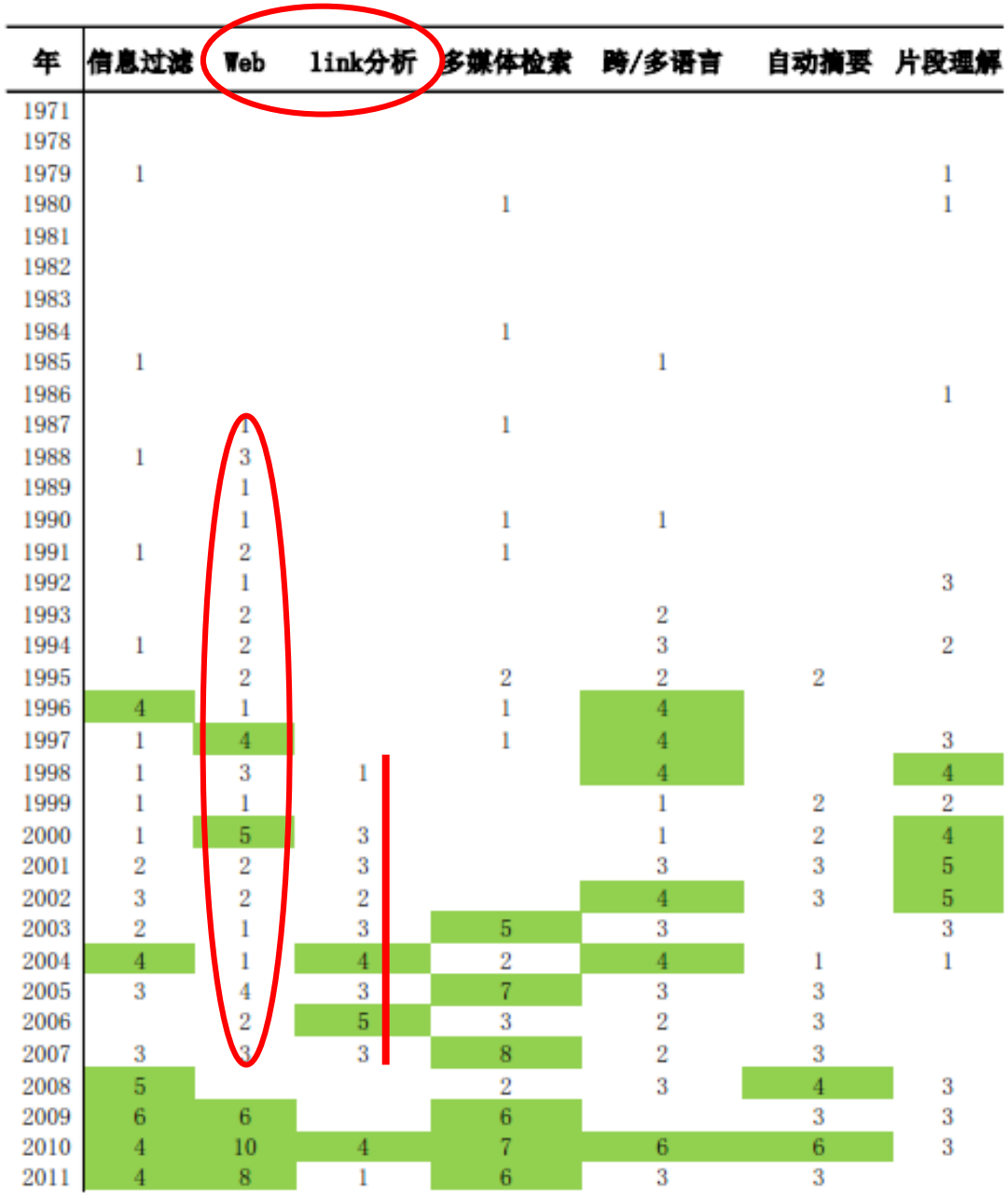
- 集中式不能满足要求
- 分布式系统架构
- 3个阶段
  - 早期：
    - 通用系统设计
  - 90年代中
    - 分布式
    - 大规模
    - 扩展性、效率
  - 最近
    - 自适应系统
    - 系统融合

年	结构化	通用	概率/LM	相关反馈	索引/实现	分布式
1971	8	5			1	
1978	4	2		1	1	
1979	1	9		1		
1980	7	2	1	1		
1981		9	1			
1982	6	5	1	2	1	1
1983	10	7				
1984	2	10			1	2
1985	3	10		1	2	1
1986	5	6		1	3	
1987	2	10			3	2
1988	5	6	3	1	5	
1989	2	2	1		4	1
1990	4	5			4	
1991	1	8	3	1	4	1
1992		6	4	2	2	
1993	1	2	2	4	2	
1994	1	2	2	3	2	
1995	2	4	3		2	3
1996		3	2	1		1
1997		1	1	2		1
1998			3	1	2	3
1999		4	1	1	1	4
2000		2		1		2
2001	2	5	3	1	2	1
2002		1	3		3	1
2003	4		4			3
2004	3	2	9	1	1	1
2005	3	1	4	4	2	4
2006	1	5	2	4	2	3
2007		2	4		6	3
2008		5	9	6	4	
2009		6	3	3	3	4
2010	4	4	3	4		
2011		5	5	4	4	1

1.2.3 信息检索在学术界的发展

检索任务

- Web IR
- 80年代末期
  - Webpage
  - Web与传统文本相区别的特性
- 1998年开始
  - Page, Kleinberg
  - 链接分析
  - 把Web作为完整的拓扑结构
- 2000年后
  - 更宏观——站点级
  - 更微观——Block级



年	信息过滤	Web	link分析	多媒体检索	跨/多语言	自动摘要	片段理解
1971							
1978							
1979	1						1
1980				1			1
1981							
1982							
1983							
1984				1			
1985	1				1		
1986							1
1987		1		1			
1988	1	3					
1989		1					
1990		1		1	1		
1991	1	2		1			
1992		1					3
1993		2			2		
1994	1	2			3		2
1995		2		2	2	2	
1996	4	1		1	4		
1997	1	4		1	4		3
1998	1	3	1		4		4
1999	1	1			1	2	2
2000	1	5	3		1	2	4
2001	2	2	3		3	3	5
2002	3	2	2		4	3	5
2003	2	1	3	5	3		3
2004	4	1	4	2	4	1	1
2005	3	4	3	7	3	3	
2006		2	5	3	2	3	
2007	3	3	3	8	2	3	
2008	5			2	3	4	3
2009	6	6		6		3	3
2010	4	10	4	7	6	6	3
2011	4	8	1	6	3	3	

1.2.3 信息检索在学术界的发展

检索任务

- 多媒体检索很早被提出
- 语义鸿沟问题
  - 图像检索
  - 实验室结果
  - 利用文本信息
- 最近年
  - 视频
  - 音乐
  - …

年	信息过滤	Web	link分析	多媒体检索	跨/多语言	自动摘要	片段理解
1971							
1978							
1979	1						1
1980				1			1
1981							
1982							
1983							
1984				1			
1985	1				1		
1986							1
1987		1		1			
1988	1	3					
1989		1					
1990		1		1	1		
1991	1	2		1			
1992		1					3
1993		2			2		
1994	1	2			3		2
1995		2		2	2	2	
1996	4	1		1	4		
1997	1	4		1	4		3
1998	1	3	1		4		4
1999	1	1			1	2	2
2000	1	5	3		1	2	4
2001	2	2	3		3	3	5
2002	3	2	2		4	3	5
2003	2	1	3	5	3		3
2004	4	1	4	2	4	1	1
2005	3	4	3	7	3	3	
2006		2	5	3	2	3	
2007	3	3	3	8	2	3	
2008	5			2	3	4	3
2009	6	6		6		3	3
2010	4	10	4	7	6	6	3
2011	4	8	1	6	3	3	

1.2.3 信息检索在学术界的发展

检索任务

- 多语言检索
- TREC
  - 日语
  - 汉语
  - 阿拉伯语
- NTCIR
  - 亚洲多语言
  - 英文
- 主要技术
  - 自然语言处理技术
  - 词语翻译技术

年	信息过滤	Web	link分析	多媒体检索	跨/多语言	自动摘要	片段理解
1971							
1978							
1979	1						1
1980				1			1
1981							
1982							
1983							
1984				1			
1985	1				1		
1986							1
1987		1		1			
1988	1	3					
1989		1					
1990		1		1	1		
1991	1	2		1			
1992		1					3
1993		2			2		
1994	1	2			3		2
1995		2		2	2	2	
1996	4	1		1	4		
1997	1	4		1	4		3
1998	1	3	1		4		4
1999	1	1			1	2	2
2000	1	5	3		1	2	4
2001	2	2	3		3	3	5
2002	3	2	2		4	3	5
2003	2	1	3	5	3		3
2004	4	1	4	2	4	1	1
2005	3	4	3	7	3	3	
2006		2	5	3	2	3	
2007	3	3	3	8	2	3	
2008	5			2	3	4	3
2009	6	6		6		3	3
2010	4	10	4	7	6	6	3
2011	4	8	1	6	3	3	

1.2.3 信息检索在学术界的发展

检索任务

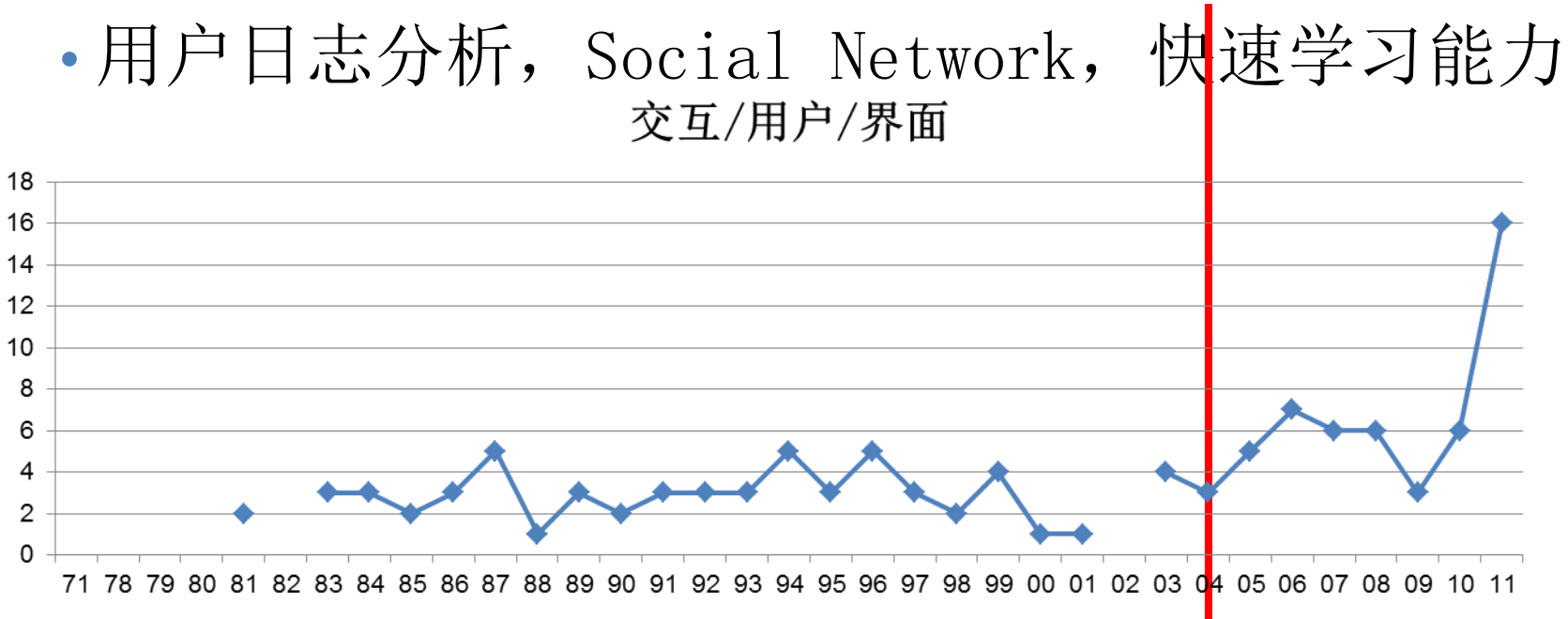
- 由国际标准评测提出，有效推动了信息检索研究的发展
- TDT
- TREC
  - Novelty
  - HARD
  - Genomics
  - Blog
  - Legal
  - …

年	信息过滤	Web	link分析	多媒体检索	跨/多语言	自动摘要	片段理解
1971							
1978							
1979	1						1
1980				1			1
1981							
1982							
1983							
1984				1			
1985	1				1		
1986							1
1987		1		1			
1988	1	3					
1989		1					
1990		1		1	1		
1991	1	2		1			
1992		1					3
1993		2			2		
1994	1	2			3		2
1995		2		2	2	2	
1996	4	1		1	4		
1997	1	4		1	4		3
1998	1	3	1		4		4
1999	1	1			1	2	2
2000	1	5	3		1	2	4
2001	2	2	3		3	3	5
2002	3	2	2		4	3	5
2003	2	1	3	5	3		3
2004	4	1	4	2	4	1	1
2005	3	4	3	7	3	3	
2006		2	5	3	2	3	
2007	3	3	3	8	2	3	
2008	5			2	3	4	3
2009	6	6		6		3	3
2010	4	10	4	7	6	6	3
2011	4	8	1	6	3	3	

1.2.3 信息检索在学术界的发展

人机交互与用户分析

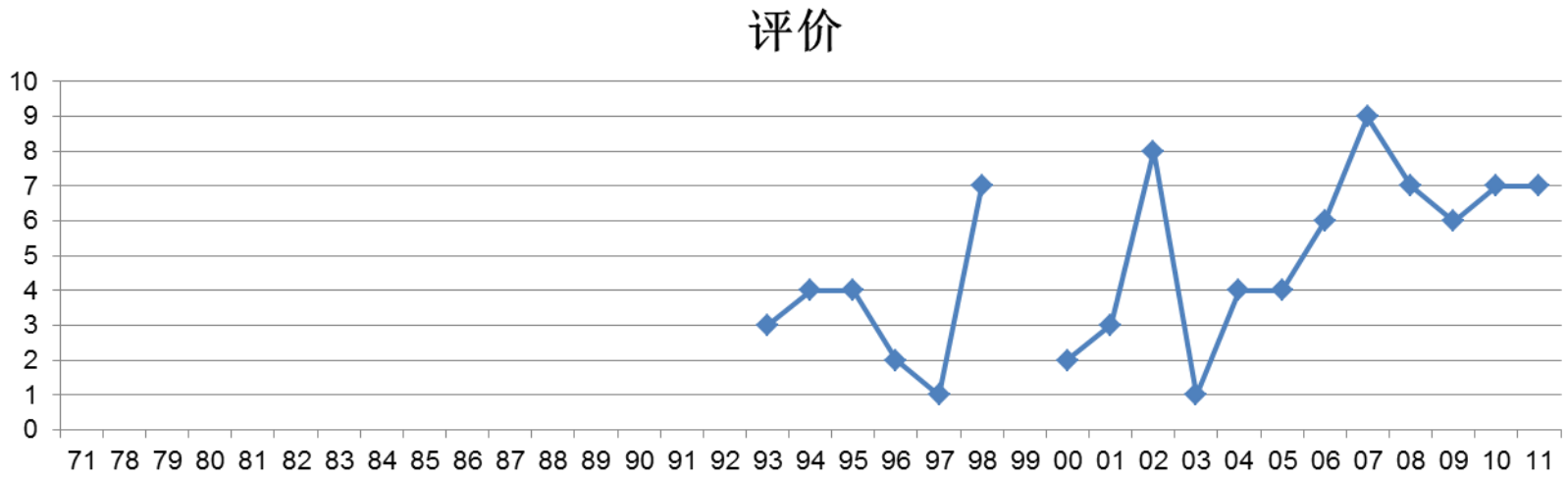
- 人们始终青睐有加的研究领域
- 早期：可视化表示 （查询、文档的可视化）
- 自然语言交互界面
- 2002年以后：
  - 用户日志分析，Social Network，快速学习能力



1.2.3 信息检索在学术界的发展

检索的评价

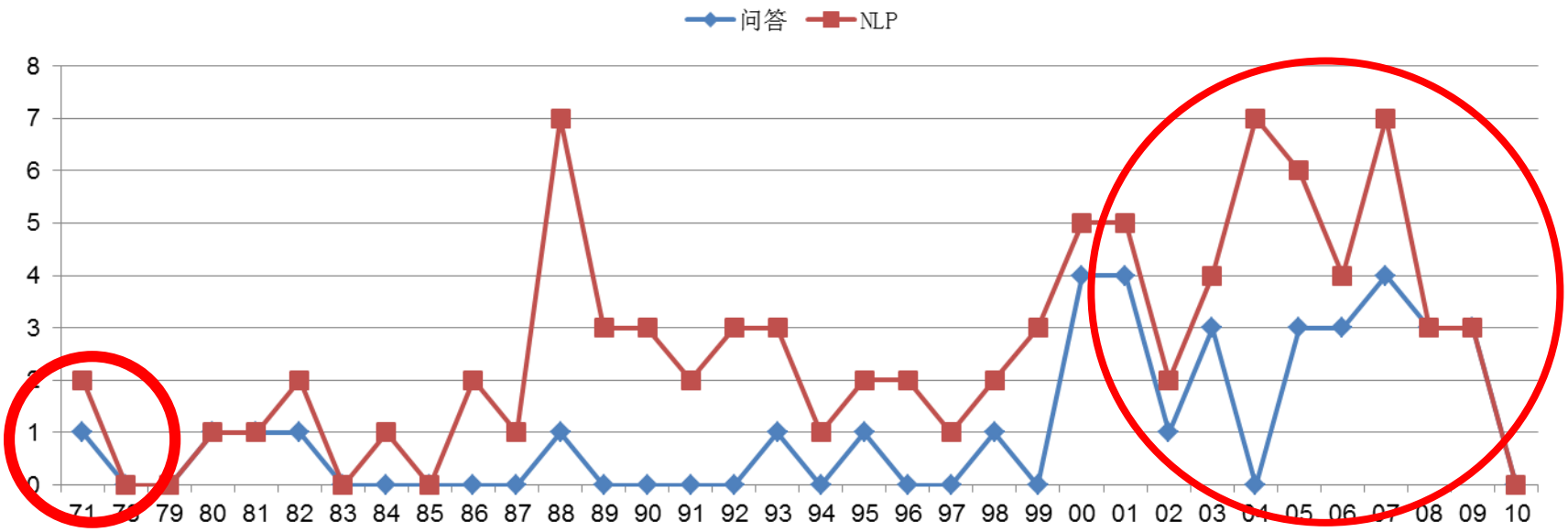
- TREC
- Pooling技术
- 更紧接本质的评价技术
- 评价与技术的共同发展



1.2.3 信息检索在学术界的发展

NLP and IR

- 最早被提出的问题之一
- Stemming，分词，词典使用，词义消歧，命名实体…
- 近年来：更深层次的使用
  - 句子完整性重构（更自然的语言表达）
  - 2005年，将NLP信息融合到检索的语言模型中



1.2.4 国际著名研究机构和代表人物

- 美国康奈尔大学 Salton (1927-1995)
  - 现代信息检索的奠基人，倡导向量空间模型
  - SMART的完成人
  - 第一任Salton奖(1983年)得主，ACM Fellow

- 英国剑桥大学 Sparck Jones (1935-2007)
  - 概率检索模型的提出者之一
  - NLP和IR中的先辈
  - 曾获ACL终身成就奖和1988年Salton奖

1.2.4 国际著名研究机构和代表人物

- 微软英国剑桥研究院、伦敦城市大学 Robertson
  - 概率检索模型的先驱和倡导者
  - 开发了OKAPI检索系统
  - 2000年Salton奖得主

- 美国 UMass CIIR W. B. Croft，ACM Fellow
  - 基于统计语言建模IR模型的提出者和倡导者
  - 和CMU共同开发了Lemur工具
  - 2003年Salton奖得主

1.2.4 国际著名研究机构和代表人物

- 英国Glasgow大学 Rijsbergen， ACM Fellow
  - 信息检索逻辑推理学派的提出者和倡导者
  - 现在试图用量子的方法解决IR问题
  - 2006年Salton奖得主

- 微软美国研究院 Susan Dumais
  - 隐性语义索引LSI的提出者
  - 2009年Salton奖得主

1.2.4 国际著名研究机构和代表人物

—些活跃的华裔学者

- 加拿大蒙特利尔大学聂建云教授
  - 跨语言检索
  - IR模型
- 美国UIUC 翟成祥(Chengxiang Zhai博士)
  - IR模型、主题模型(Topic Model)
- 美国CMU 杨颐明(Yiming Yang)教授
  - 文本分类领域最著名的学者之一

提纲

1.1 信息检索的由来和这门课的意义

1.2 信息检索的历史和发展

1.3 信息检索与数据挖掘等其他学科的关系

1.4 信息检索的基本概念

1.5 课程要求和说明

1.3 信息检索与其他学科的关系

相关研究领域

- 数据挖掘（Data Mining）
- 图书情报学(Library & Info. Science)
- 数据库管理(Database Management)
- 人工智能(Artificial Intelligence)
- 自然语言处理(Natural Language Processing)
- 机器学习(Machine Learning)

1.3 信息检索与其他学科的关系

信息检索与数据挖掘的关系

- 数据挖掘（Data Mining）从*大量的*、*不完全的*、*有噪声的*、*模糊的*、*随机的*实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。
- 信息检索的一些内容如检索模型中的概率模型、语言模型，还有文本相关的一些处理如文本聚类、文本分类都涉及到数据挖掘相关知识。
- 多媒体检索中的图像、语音、视频检索更是需要数据挖掘来得到图像、语音、视频中的知识才能进行检索。

1.3 信息检索与其他学科的关系

图书情报学(Library and Information Science, LIS)

- IR最初起源于LIS
- LIS主要关注IR中的用户方(人机交互、用户界面、可视化)
- LIS关注人类知识的高效分类
- LIS关注文献的引用分析(citation analysis)和文献计量(bibliometrics)
- 近年来数字图书馆方面的工作使得LIS和IR日益融合（liser.ustc.edu.cn）

1.3 信息检索与其他学科的关系

数据库管理系统（Database Management, DM）

- DM主要面向关系表中的结构化数据而非自由文本。
- DM主要集中于高效解决形式化语言(如SQL)定义的查询。
- DM中不论是查询还是数据都具有明确的语义。
- 近年来半结构化的XML数据的出现使DM和IR逐渐融合。

1.3 信息检索与其他学科的关系

人工智能(Artificial Intelligence, AI)

- AI关注知识的表示、推理和智能行为。
- AI中知识的形式化表示
  - 一阶谓词逻辑(First Order Predicate Logic)
  - 贝叶斯网络(Bayesian Networks)
- 近年来Web本体及智能信息Agent方面研究使得IR和AI相互融合。

1.3 信息检索与其他学科的关系

自然语言处理(Natural Language Processing, NLP)

- NLP关注自然语言文本的语法(syntactic) 、语义(semantic)及语用(pragmatic)分析。
- NLP可以分析短语结构和语义，使得IR可以在短语上、或者从语义上进行处理，而不是仅仅基于单个关键词。NLP和IR天生就是融合的。

1.3 信息检索与其他学科的关系

NLP和IR融合的其他方面

- 通过上下文词义消歧(word sense disambiguation)来确定一个词在某个特定上下文的语义。
- 通过一些NLP方法来获得文档中的一个语言片断(information extraction).
- 通过NLP方法可以从文档集合中返回一些问题的答案(question answering)

1.3 信息检索与其他学科的关系

机器学习(Machine Learning)

- ML关注通过对经验的学习来提高计算机系统的性能。
- 从标注好的例子中学习相关概念，然后进行自动分类(有监督的学习，supervised learning)
- 将未标注的例子自动聚集到有意义的不同集合中(无监督的学习，unsupervised learning).

1.3 信息检索与其他学科的关系

ML和IR融合的方面

- 文本分类(Text Categorization)
  - 自动层次分类(如Yahoo目录)
  - 自适应过滤或推荐(Adaptive filtering/recommending)
  - 垃圾过滤(Spam filtering)
- 文本聚类(Text Clustering)
  - IR结果的自动聚类
  - 层次型类别体系的自动构建(如Yahoo!目录)

提纲

1.4.1 信息检索的基本概念

- 用户需求(User Need，UN)：用户需要获得的信息
  - 严格地说，UN只存在于用户的内心，但是通常用文本来描述，如 查找与2016美国总统大选相关的新闻，有时也称为主题(Topic)
  - UN提交给检索系统时称为查询(Query)，如 2017特朗普总统，对同一个UN，不同人不同时候可以构造出不同的Query，比如上述需求也可表示2017年美国特朗普总统新闻，Query在IR系统中往往还有内部表示

1.4.1 信息检索的基本概念

- 文档(Document)：检索的对象
  - 可以是文本，也可以是图像、视频、语音等多媒体文档，text retrieval/image retrieval/video retrieval /speech retrieval/multimedia retrieval
  - 可以是无格式、半格式、有格式的
- 文档集合(Collection)：所有待检索的文档构成的集合
  - 也称为Repository，Corpus

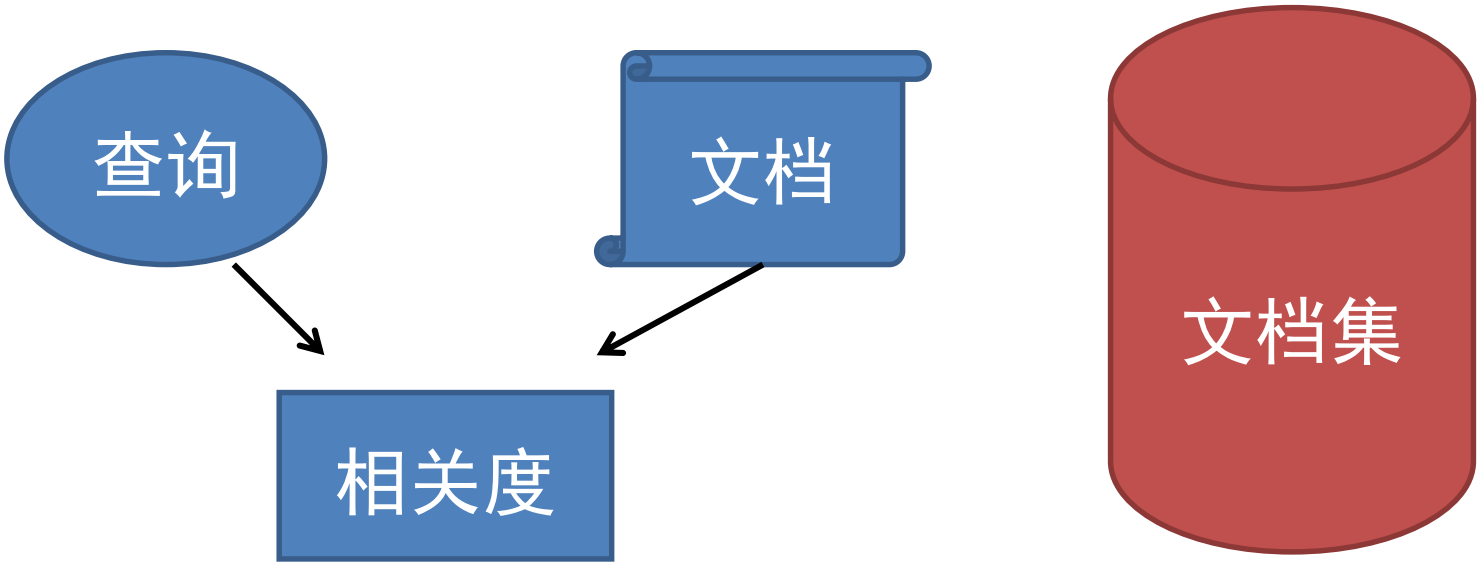
1.4.1 信息检索的基本概念

- 相关(relevant）、相关度（relevance)
  - 相关取决于用户的判断，是一个主观概念
  - 不同用户做出的判断很难保证一致
  - 即使是同一用户在不同时期、不同环境下做出的判断也不尽相同

1.4.1 信息检索的基本概念

- 定义“相关性”的两个角度：
  - 系统角度：系统输出结果，用户是信息的接受者。这种理解置用户于被动的地位，基于这种理解，研究的重心落在系统本身。主题相关性：检索系统检出的文档的主题即核心内容与用户的信息需求相匹配。系统角度相关并不和用户脱节。系统角度定义的相关简单可以计算。
  - 用户角度：观察用户对检索结果的反应，是系统输出向用户需求的投射。相关性被认为是用户方面的属性。用户角度定义的相关目前仍然难以计算。
- 现代信息检索研究中仍然主要采用系统角度定义的主题相关性概念，当然也强调考虑用户的认知因素。

1.4.1 信息检索的基本概念



1.4.1 信息检索的基本概念

- 形式上说，信息检索中的相关度是一个函数R，输入是查询Q、文档D和文档集合C，返回的是一个实数值
  - R=f(Q,D,C)
- 信息检索就是给定一个查询Q，从文档集合C中计算每篇文档D与Q的相关度并排序(Ranking)。
- 相关度通常只有相对意义，对一个Q，不同文档的相关度可以比较，而对于不同的Q的相关度不便比较
- 相关度的输入信息可以更多，比如用户的背景信息、用户的查询历史等等
- 现代信息检索中相关度不是唯一度量，如还有：重要度、权威度、新颖度等度量。或者说这些因子都影响“相关度”。
  - Google中据说用了上百种排名因子

1.4.1 信息检索的基本概念

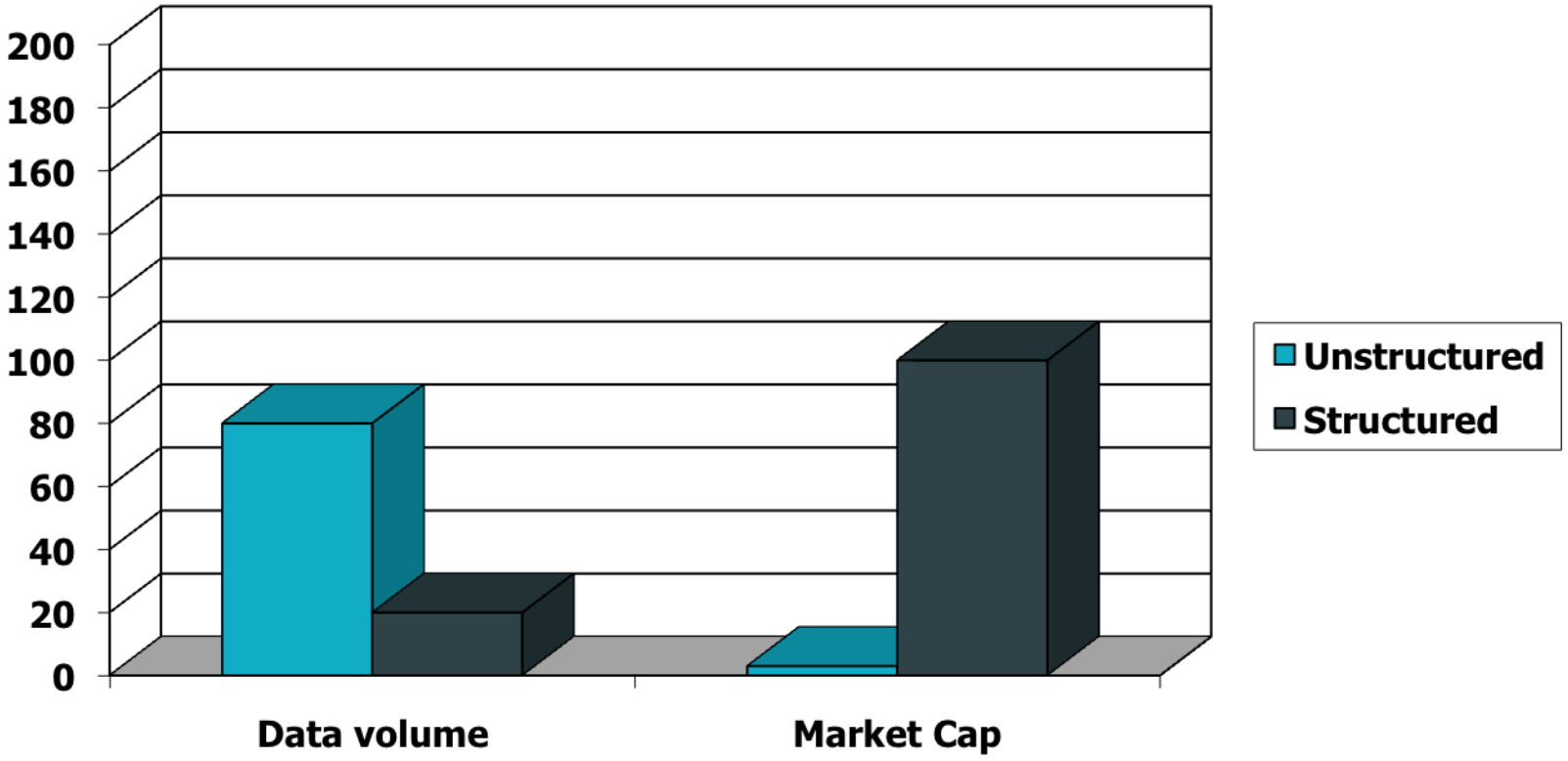
信息检索与数据库检索

	信息检索	数据库检索
检索对象	无结构、半结构数据 如网页、图片……	结构化数据 如：员工数据库
检索方式	通常是近似检索 如：每个结果有相关度得分	通常是精确检索 如：姓名==“李明”
检索语言	主要是自然语言 如：查与超女相关的新闻	SQL结构化语言

近年来，随着XML的出现，两种检索已经逐渐融合，边界越来越不明显。

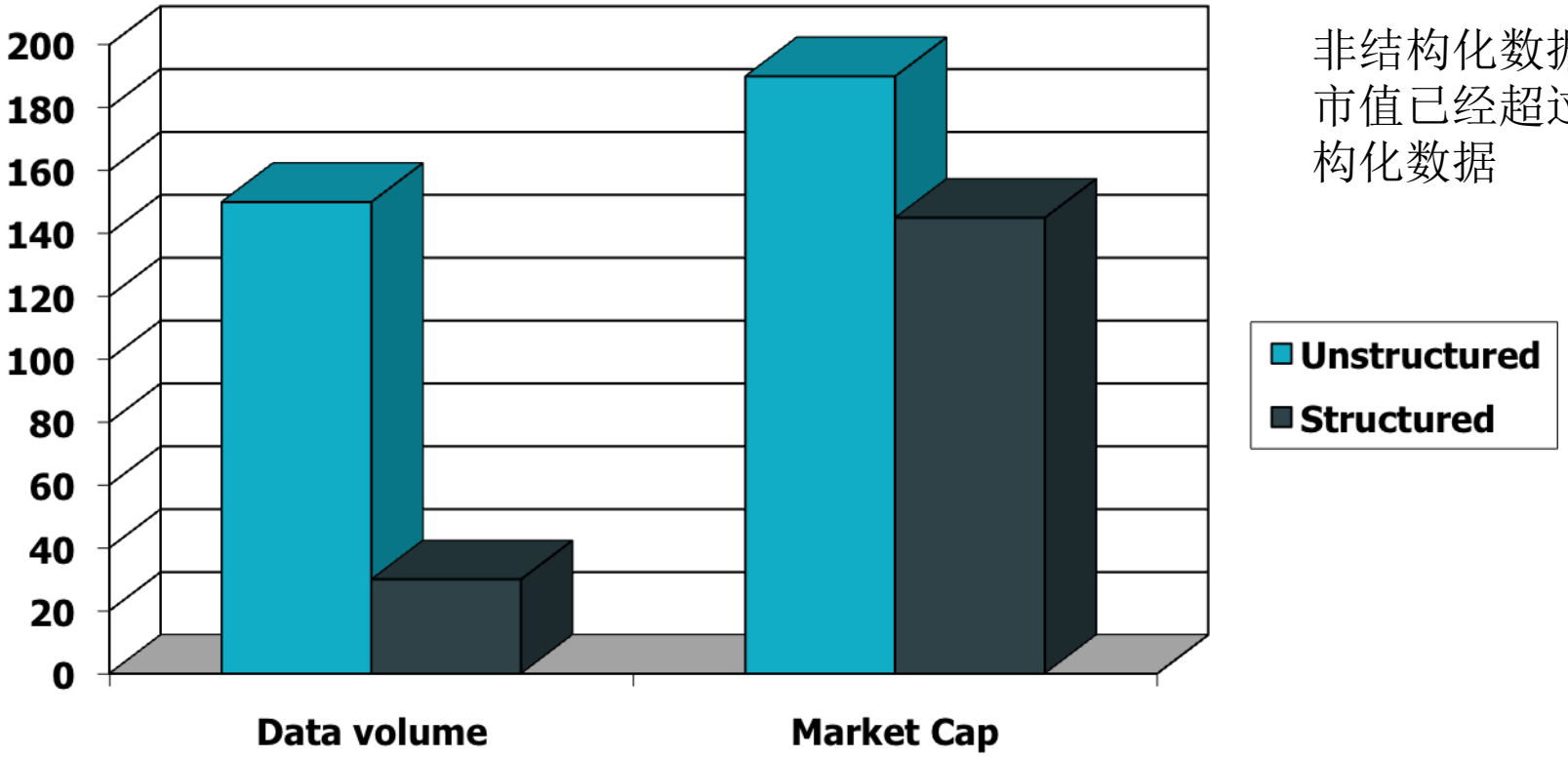
1.4.1 信息检索的基本概念

Unstructured (text) vs. structured (database) data in 1996



1.4.1 信息检索的基本概念

Unstructured (text) vs. structured (database) data in 2009



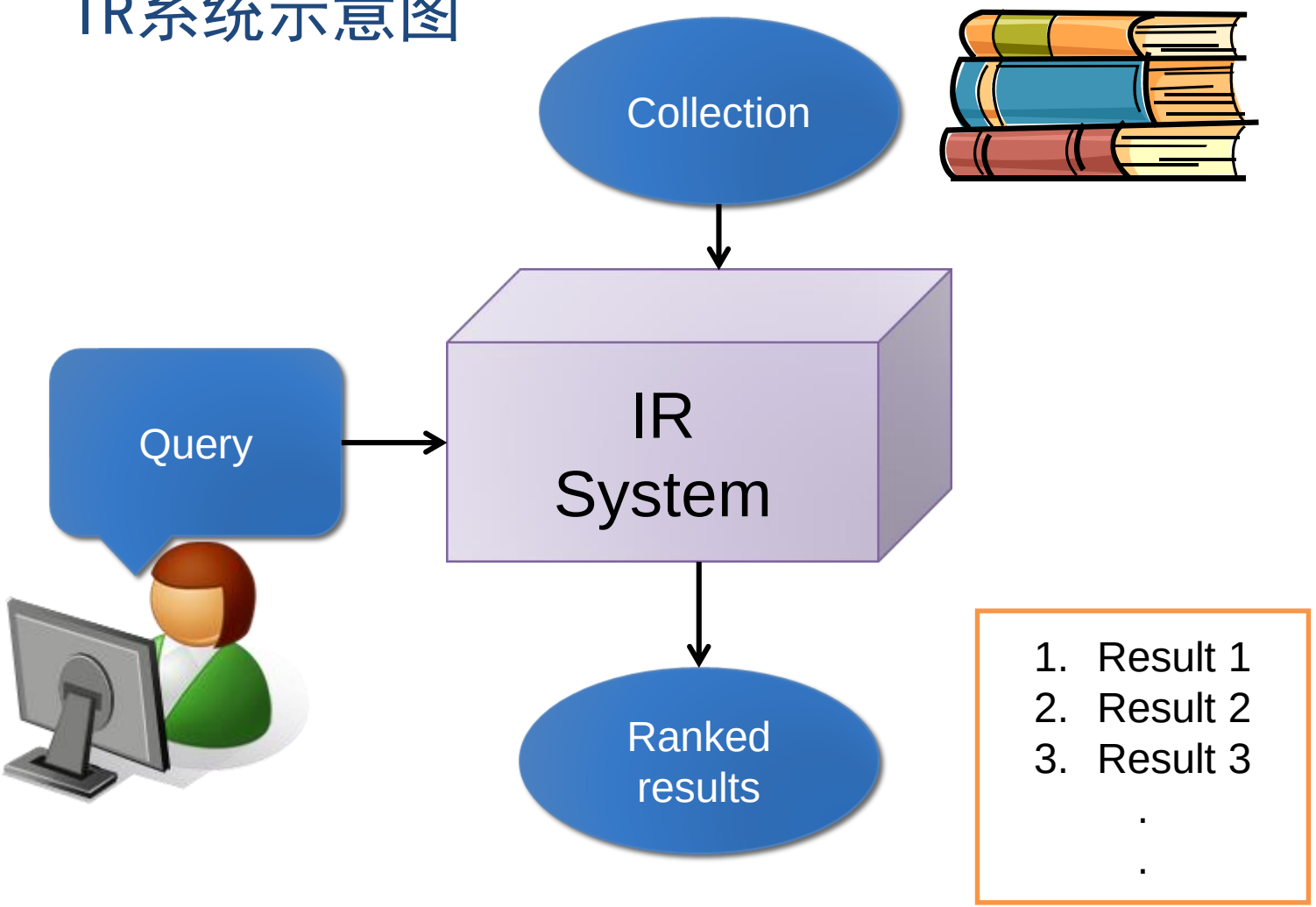
非结构化数据的市值已经超过结构化数据

1.4.1 信息检索的基本概念

- 本课程将主要介绍面向文本对象的检索，即文本检索（text retrieval）。
  - 文本是人们表达知识(论文)、交流(口语)的最常用的形式。
  - 文本可以用于描述其他媒体。
  - 其他媒体形式的检索的方法往往借鉴自文本检索。
- 信息检索的两种研究方式：
  - 以计算机为中心：IR的工作主要是建立索引、对用户查询进行处理、排序算法等等
  - 以用户为中心：IR的主要工作是考察用户的行为、理解用户的需求、这些行为和需求如何影响检索系统的组织
  - 本课程主要研究以计算机为中心的IR问题，目前是主流

1.4.2 信息检索系统的基本组成

IR系统示意图



1.4.2 信息检索系统的基本组成

- 用户接口(User Interface)：用户和IR系统的人机接口
  - 输入查询(Query)
  - 返回排序后的结果文档(Ranked Docs)并对其进行可视化（Visualization)
  - 支持用户进行相关反馈(Feedback)
- 用户的两种任务：retrieval 或者 browsing
- IR的两种模式： pull (ad hoc) 或者 push (filtering)
  - Pull: 用户是主动的发起请求，在一个相对稳定的数据集合上进行查询
  - Push:用户事先定义自己的兴趣，系统在不断到来的流动数据上进行操作，将满足用户兴趣的数据推送给用户

1.4.2 信息检索系统的基本组成

- 文本处理(Text Operations)：对查询和文本进行的预处理操作
  - 中文分词(Chinese Word Segmentation)
  - 词干还原(Stemming)
  - 停用词消除(Stop-word removal)
- 查询处理(Query operations)：对经过文本处理后的查询进行进一步处理，得到查询的内部表示(Query Representation)
  - 查询扩展(Query Expansion)：利用同义词或者近义词对查询进行扩展
  - 查询重构(Query Reconstruction)：利用用户的相关反馈信息对查询进行修改
- 文本索引(Indexing)：对经过文本处理后的文本进行进一步处理，得到文本的内部表示(Text Representation)，通常基于索引项(Term)来表示
  - 向量化、概率计算
  - 组成成倒排表进行存储

1.4.2 信息检索系统的基本组成

- 搜索(Searching)：从文本中查找包含查询中索引项的文本
- 排序(Ranking)：对搜索出的文本按照某种方式来计算其相关度
- Logical View：指的是查询或者文本的表示，通常采用一些关键词或者索引项(index term)来表示一段查询或者文本。

提纲

1.1 信息检索的由来和这门课的意义

1.2 信息检索的历史和发展

1.3 信息检索与数据挖掘等其他学科的关系

1.4 信息检索的基本概念

1.5 课程要求和说明

1.5.1 授课老师介绍

- 授课教师
  - 俞能海、陈晓辉

1.5.2 授课方式

- 每周5课时，共60学时
- 研讨会

1.5.3 授课内容

- 第一章　绪论

信息检索的典型应用。信息检索的基本概念和发展历史。信息检索和其他相关学科(自然语言处理、机器学习、概率统计、模式识别、数据库、数据挖掘等等)的关系。信息检索系统的基本构架和一般流程。

- 第二章 布尔检索及倒排索引

字符串匹配及倒排索引。布尔查询处理及其优化。扩展的布尔操作。短语查询的处理。布尔检索模型及其扩展。

- 第三章 词典查找及扩展的倒排索引

支持词典快速查找的数据结构(哈希表、二叉树等)。支持通配查询处理的索引结构。支持拼写或发音纠错处理的索引结构。

1.5.3 授课内容

- 第四章 索引构建和索引压缩

文本预处理。一般构建过程。基于块排序的构建过程。单遍内存式扫描构建方法。分布式及动态索引方法。词项的统计特性。词典的压缩。倒排记录表的压缩。

- 第五章 向量模型及检索系统

向量空间模型及词项权重计算机制。检索中的快速实现方法。检索系统的一般构成。隐性语义索引方法。基于开源工具搭建简单搜索引擎。

- 第六章 检索的评价

效率和效果的评价。查全率和查准率。其他效果评价方法。用户体验及结果摘要。相关评测语料和评测会议。

1.5.3 授课内容

- 第七章 相关反馈和查询扩展

相关反馈和伪相关反馈。查询扩展及重构。全局方法及局部方法。

- 第八章 概率模型

概率排序原理。回归模型。二值独立概率模型。OKAPI BM25公式。

- 第九章 基于语言建模的检索模型

查询似然模型。其他语言模型。语言模型的相关反馈。

- 第十章 文本分类

文本分类的概念及评价方法。文本分类中的特征选择方法。

1.5.3 授课内容

- 第十一章 文本聚类

文本聚类的概念及评价方法。文本聚类算法。检索结果聚类的标签生成。

- 第十二章 Web搜索

Web结构。信息采集。网页查重方法。链接分析算法(PageRank和HITS)。

- 第十三章 多媒体信息检索

自动图像标注，语义距离的度量，图像搜索，视频概念检测

- 第十四章 其他应用简介

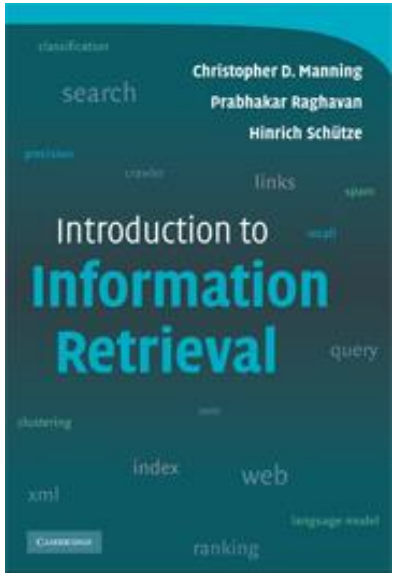
数字图书馆，过滤及推送系统、XML检索、跨语言检索、信息抽取、问答系统、互联网广告系统等等。

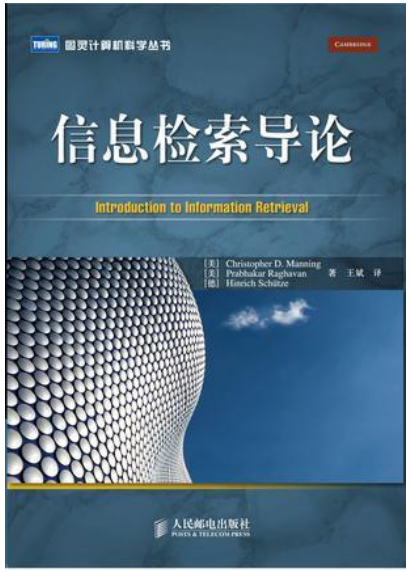
1.5.4 考核方式

- 不定期考勤
- 课堂提问
- 研讨会
- 平时作业（5-10%）
- 实验（15-20%）
- 期中考试（30-40%）
- 期末考试（30-40%）

1.5.5 其他信息

- 教材：
  - 原版：Introduction to Information Retrieval，Cambridge University Press，C.D. Manning.
  - https://nlp.stanford.edu/IR-book/
  - 中文翻译版：信息检索导论，人民邮电出版社，王斌译





1.5.5 其他信息

- 参考资料
  - Wiki，google
  - 其他大学或科研机构的相关课程主页
    - Stanford：http://www.stanford.edu/class/cs276/index.html
  - 其他书籍
    - ***Managing Gigabytes***，by I. Witten, A. Moffat, and T. Bell.
    - ***Modern Information Retrieval***，by R. Baeza-Yates and B. Ribeiro-Neto.
- 国际会议：
  - SIGIR、ACL、WWW、SIGKDD、WSDM、ICML
  - CIKM、EMNLP、COLING
  - TREC、NTCIR评测会议
  - ECIR、AIRS
- 国内会议：
  - 全国信息检索学术会议(1年一届)
  - 全国计算语言学联合会议(2年一届)
  - 搜索引擎和WEB挖掘学术会议(1年一届，上半年)

1.5.5 其他信息

- 重要工具
  - emur、Indri：包含各种IR模型的实验平台，C++
  - SMART：向量空间模型工具，C编写
  - Weka：数据挖掘工具，Java编写
  - Lucene：开源检索工具，Java版本受维护，存在各种语言编写的其他版本
  - Nutch：开源爬虫，Java版本
  - Sphinx：开源检索工具，C++
  - Larbin：采集工具，C++
  - Firtex：检索平台，C++，计算所开发
  - 更多http://www.searchtools.com/tools/tools-opensource.html

思考题

- 信息检索的定义？
- 信息检索中的用户需求、查询、相关度都是什么含义？
- 信息检索和其他相关学科是什么关系？
- 信息检索系统由哪些部分组成？各部分的功能是什么？

本章小结

- 信息检索是一门交叉学科，不仅仅是搜索
- 信息检索中的用户需求、查询、文档、文档集、相关度概念
- 信息检索和其他学科领域的关系
- 信息检索的组成和流程


谢谢大家！