信息检索研究：
过去三十年中我们走了多远

马少平，张敏
清华大学计算机系；
智能技术与系统国家重点实验室
2006年11月21日

缘起

- 过去三十年中，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 1971～2006年所有正式论文

主要内容

- 检索模型的发展
- 关键技术
- 检索任务的演变
- 人机交互与用户分析
- 信息检索的评价
- 信息检索中的自然语言处理
- 更多思考与讨论

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	片断理解	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	13	1	5	3	1	3			1								1	
89	2	2	1		1		3	5	8		4			1			3	1	1							
90	4	5	2	1			3	1	4		4			1		1	2					1				
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	1	2	1	2	5	7	4	2			2			3			2		2				3
94	1	2	2	1	2		1	3	4	3	2	2	1	2			5		3			3	2		1	4
95	2	4	2		3	1	1	2	3		2	1		2		2	3	3	3	1	2	2				4
96		3	3		2		2	3	5	1		1	4	1		1	5	1	3	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

检索模型的发展

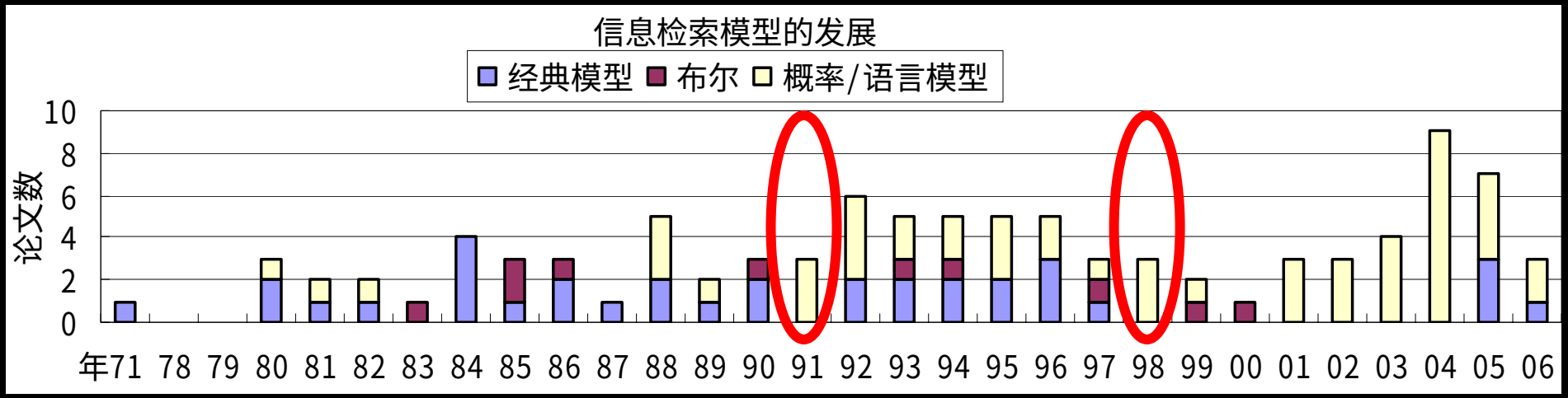
信息检索模型

- 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90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—— xml IR
  - 两条路逐渐呈现融合趋势

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

IR models

- 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 －－ 近两三年开始，标志IR进入新的阶段



关键技术

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
  - 最近
    - 垃圾 …
- 走出实验室
  - 面向海量数据、实时处理、真实网络环境…

年份	（半）结构化	一般方法/理论	经典模型	布尔模型	概率/语言模型	权值计算	相关反馈	索引实现	分布式系统
1971	8	5	1					1	
1978	4	2					1	1	
1979	1	9				1	1		
1980	7	2	2		1	3	1		
1981		9	1		1	2			
1982	6	5	1		1	1	2	1	1
1983	10	7		1		2			
1984	2	10	4			1		1	2
1985	3	10	1	2		1	1	2	1
1986	5	6	2	1		5	1	3	
1987	2	10	1			3		3	2
1988	5	7	2		3	3	1	5	
1989	2	2	1		1			4	1
1990	4	5	2	1				4	
1991	1	8			3		1	4	1
1992		7	2		4	1	2	2	
1993	1	2	2	1	2		4	2	
1994	1	3	2	1	2	2	3	2	
1995	2	4	2		3	1		2	3
1996		3	3		2	1	1		1
1997		1	1	1	1	1	2		1
1998					3	1	1	2	3
1999		6		1	1		1	1	4
2000		3		1			1		2
2001	2	5			3	1	1	2	1
2002		1			3	1		3	1
2003	4	1			4				3
2004	3	5			9	1	1	1	1
2005	3	3	3		4	4	4	2	4
2006	1	6	1		2	1	4	2	3

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
    - CBIR
  - 最近
    - 区分不同主题
    - 区分不同词

年份	（半）结构化	一般方法/理论	经典模型	布尔模型	概率/语言模型	权值计算	相关反馈	索引实现	分布式系统
1971	8	5	1					1	
1978	4	2					1	1	
1979	1	9				1	1		
1980	7	2	2		1	3	1		
1981		9	1		1	2			
1982	6	5	1		1	1	2	1	1
1983	10	7		1		2			
1984	2	10	4			1		1	2
1985	3	10	1	2		1	1	2	1
1986	5	6	2	1		5	1	3	
1987	2	10	1			3		3	2
1988	5	7	2		3	3	1	5	
1989	2	2	1		1			4	1
1990	4	5	2	1				4	
1991	1	8			3		1	4	1
1992		7	2		4	1	2	2	
1993	1	2	2	1	2		4	2	
1994	1	3	2	1	2	2	3	2	
1995	2	4	2		3	1		2	3
1996		3	3		2	1	1		1
1997		1	1	1	1	1	2		1
1998					3	1	1	2	3
1999		6		1	1		1	1	4
2000		3		1			1		2
2001	2	5			3	1	1	2	1
2002		1			3	1		3	1
2003	4	1			4				3
2004	3	5			9	1	1	1	1
2005	3	3	3		4	4	4	2	4
2006	1	6	1		2	1	4	2	3

关键技术

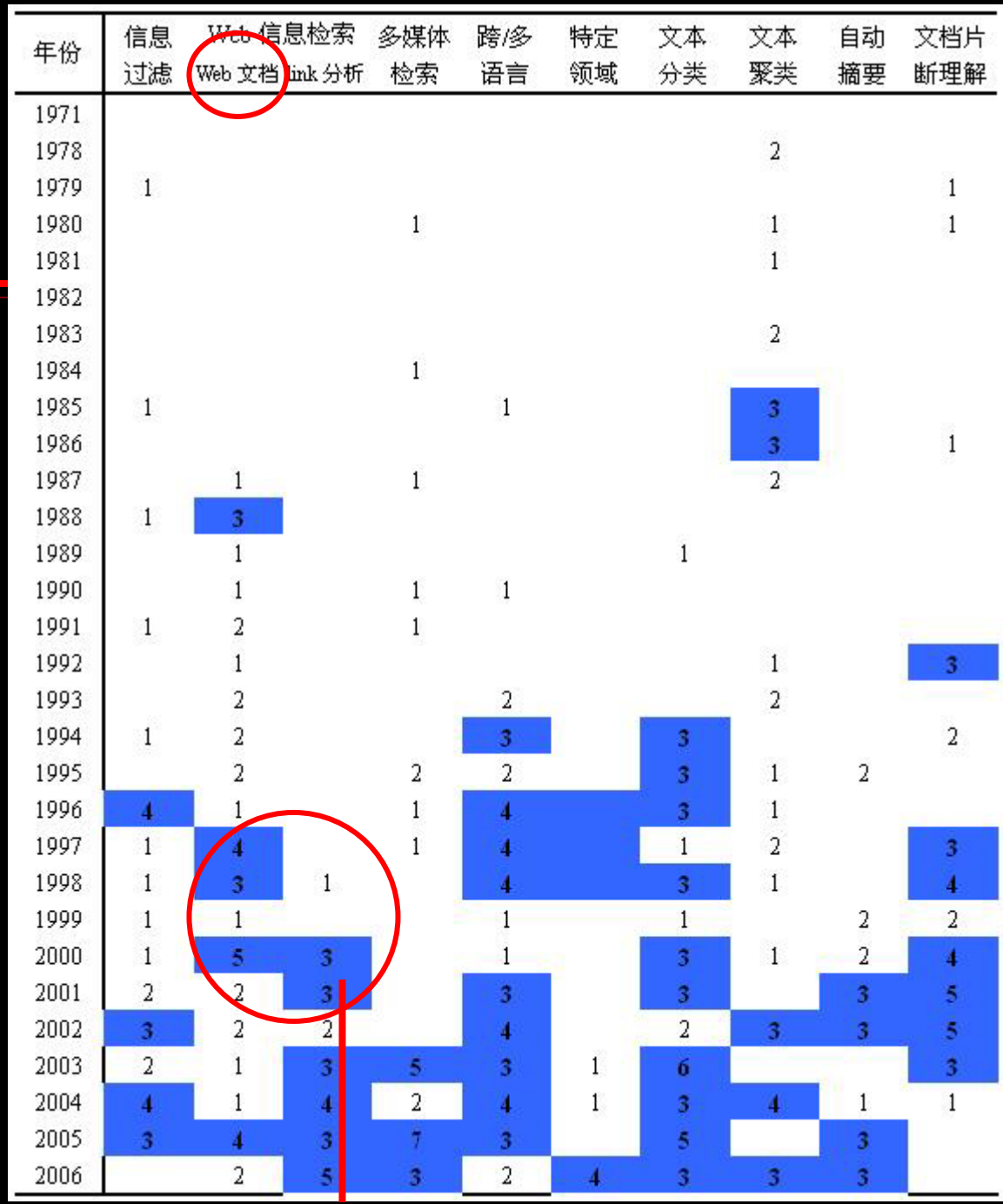
- 集中式不能满足要求
- 分布式系统架构
- 3个阶段
  - 早期：
    - 通用系统设计
  - 90年代中
    - 分布式
    - 大规模
    - 扩展性、效率
  - 最近
    - 自适应系统
    - 系统融合

年份	（半）结构化	一般方法/理论	经典模型	布尔模型	概率/语言模型	权值计算	相关反馈	索引\实现	分布式系统
1971	8	5	1					1	
1978	4	2					1	1	
1979	1	9				1	1		
1980	7	2	2		1	3	1		
1981		9	1		1	2			
1982	6	5	1		1	1	2	1	1
1983	10	7		1		2			
1984	2	10	4			1		1	2
1985	3	10	1	2		1	1	2	1
1986	5	6	2	1		5	1	3	
1987	2	10	1			3		3	2
1988	5	7	2		3	3	1	5	
1989	2	2	1		1			4	1
1990	4	5	2	1				4	
1991	1	8			3		1	4	1
1992		7	2		4	1	2	2	
1993	1	2	2	1	2		4	2	
1994	1	3	2	1	2	2	3	2	
1995	2	4	2		3	1		2	3
1996		3	3		2	1	1		1
1997		1	1	1	1	1	2		1
1998					3	1	1	2	3
1999		6		1	1		1	1	4
2000		3		1			1		2
2001	2	5			3	1	1	2	1
2002		1			3	1		3	1
2003	4	1			4				3
2004	3	5			9	1	1	1	1
2005	3	3	3		4	4	4	2	4
2006	1	6	1		2	1	4	2	3

检索任务的演变

检索任务

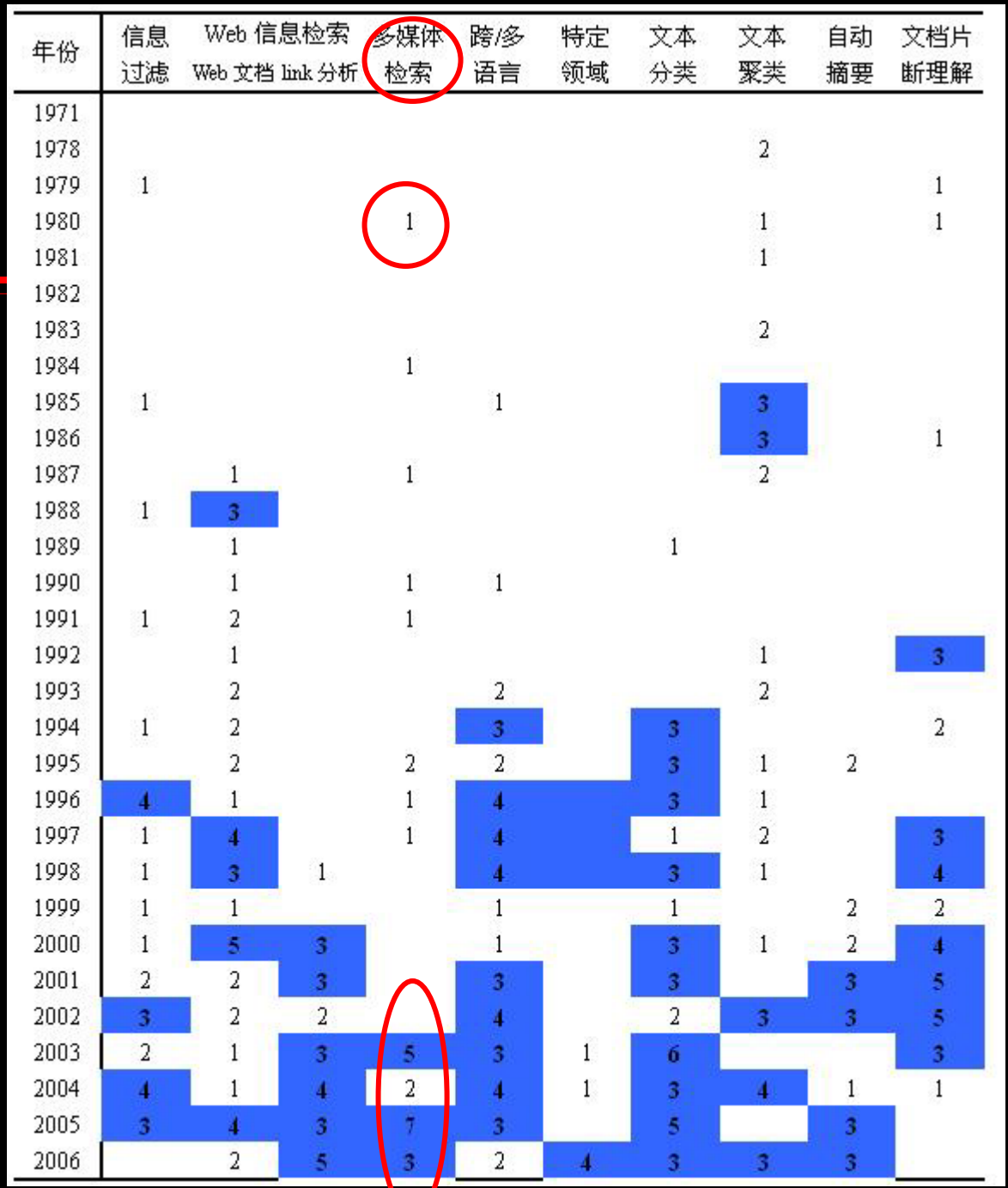
- Web IR
- 80年代末期
  - Webpage
  - Web与传统文本相区别的特性
- 1998年开始
  - Page, Kleinberg
  - 链接分析
  - 把Web作为完整的拓扑结构
- 2000年后
  - 更宏观——站点级
  - 更微观——Block级



年份	信息过滤	Web信息检索 Web文档	Web信息检索 link分析	多媒体检索	跨/多语言	特定领域	文本分类	文本聚类	自动摘要	文档片断理解
1971										
1978								2		
1979	1									1
1980				1				1		1
1981								1		
1982										
1983								2		
1984				1						
1985	1				1			3		
1986								3		1
1987		1		1				2		
1988	1	3								
1989		1					1			
1990		1		1	1					
1991	1	2		1						
1992		1						1		3
1993		2			2			2		
1994	1	2			3		3			2
1995		2		2	2		3	1	2	
1996	4	1		1	4		3	1		
1997	1	4		1	4		1	2		3
1998	1	3	1		4		3	1		4
1999	1	1			1		1		2	2
2000	1	5	3		1		3	1	2	4
2001	2	2	3		3		3		3	5
2002	3	2	2		4		2	3	3	5
2003	2	1	3	5	3	1	6			3
2004	4	1	4	2	4	1	3	4	1	1
2005	3	4	3	7	3		5		3	
2006		2	5	3	2	4	3	3	3	

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5年
  - 视频
  - 音乐
  - …



年份	信息过滤	Web 信息检索 Web 文档	link 分析	多媒体检索	跨/多语言	特定领域	文本分类	文本聚类	自动摘要	文档片断理解
1971										
1978								2		
1979	1									1
1980				1				1		1
1981								1		
1982										
1983								2		
1984				1						
1985	1				1			3		
1986								3		1
1987		1		1				2		
1988	1	3								
1989		1					1			
1990		1		1	1					
1991	1	2		1						
1992		1						1		3
1993		2			2			2		
1994	1	2			3		3			2
1995		2		2	2		3	1	2	
1996	4	1		1	4		3	1		
1997	1	4		1	4		1	2		3
1998	1	3	1		4		3	1		4
1999	1	1			1		1		2	2
2000	1	5	3		1		3	1	2	4
2001	2	2	3		3		3		3	5
2002	3	2	2		4		2	3	3	5
2003	2	1	3	5	3	1	6			3
2004	4	1	4	2	4	1	3	4	1	1
2005	3	4	3	7	3		5		3	
2006		2	5	3	2	4	3	3	3	

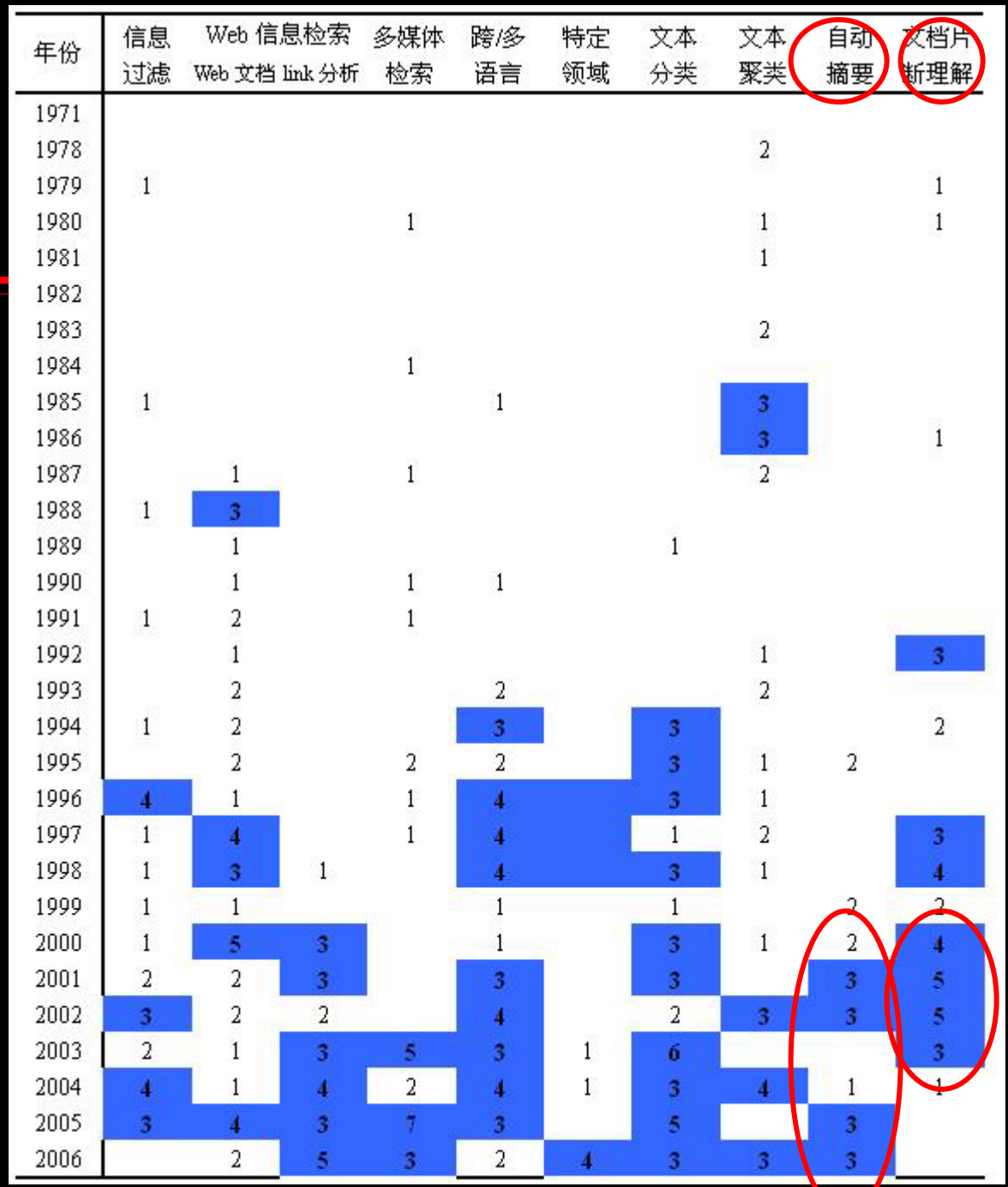
检索任务

- 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

年份	信息过滤	Web 信息检索 Web 文档	Web 信息检索 link 分析	多媒体检索	跨/多语言	特定领域	文本分类	文本聚类	自动摘要	文档片断理解
1971										
1978								2		
1979	1									1
1980				1				1		1
1981								1		
1982										
1983								2		
1984				1						
1985	1				1			3		
1986								3		1
1987		1		1				2		
1988	1	3								
1989		1					1			
1990		1		1	1					
1991	1	2		1						
1992		1						1		3
1993		2			2			2		
1994	1	2			3		3			2
1995		2		2	2		3	1	2	
1996	4	1		1	4		3	1		
1997	1	4		1	4		1	2		3
1998	1	3	1		4		3	1		4
1999	1	1			1		1		2	2
2000	1	5	3		1		3	1	2	4
2001	2	2	3		3		3		3	5
2002	3	2	2		4		2	3	3	5
2003	2	1	3	5	3	1	6			3
2004	4	1	4	2	4	1	3	4	1	1
2005	3	4	3	7	3		5		3	
2006		2	5	3	2	4	3	3	3	

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
  - ...

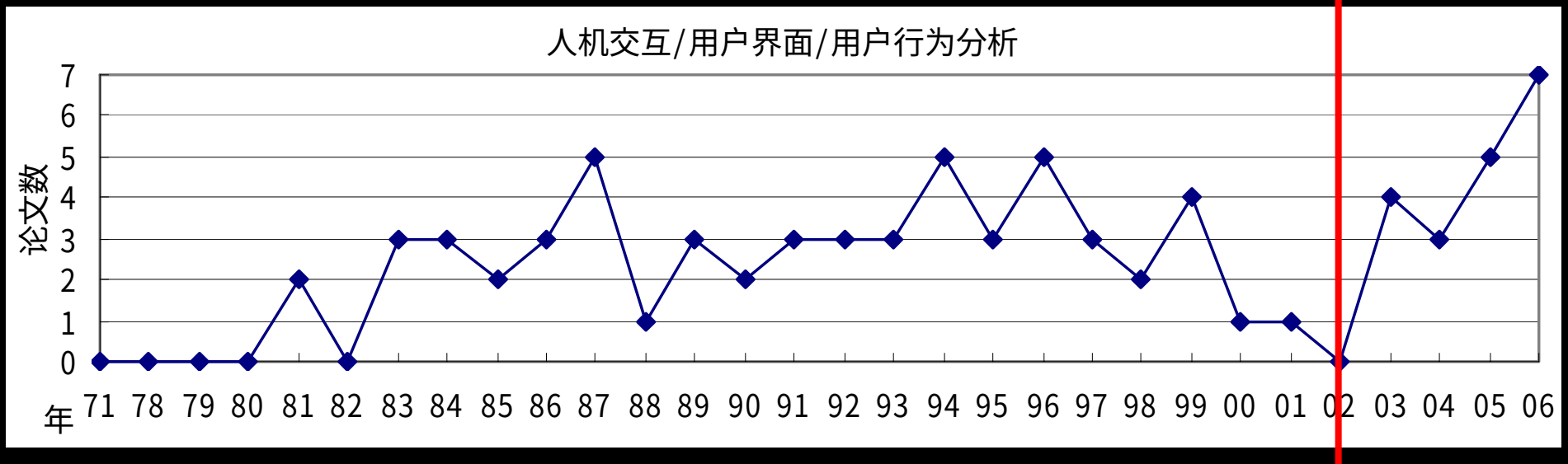


年份	信息过滤	Web 信息检索 Web 文档 link 分析	多媒体检索	跨/多语言	特定领域	文本分类	文本聚类	自动摘要	文档片断理解
1971									
1978							2		
1979	1								1
1980			1				1		1
1981							1		
1982									
1983							2		
1984			1						
1985	1			1			3		
1986							3		1
1987		1	1				2		
1988	1	3							
1989		1				1			
1990		1	1	1					
1991	1	2	1						
1992		1					1		3
1993		2		2			2		
1994	1	2		3		3			2
1995		2	2	2		3	1	2	
1996	4	1	1	4		3	1		
1997	1	4	1	4		1	2		3
1998	1	3	1	4		3	1		4
1999	1	1		1		1		2	2
2000	1	5	3	1		3	1	2	4
2001	2	2	3	3		3		3	5
2002	3	2	2	4		2	3	3	5
2003	2	1	3	5	3	1	6		3
2004	4	1	4	2	4	1	3	4	1
2005	3	4	3	7	3		5		3
2006		2	5	3	2	4	3	3	3

人机交互与用户分析

人机交互与用户分析

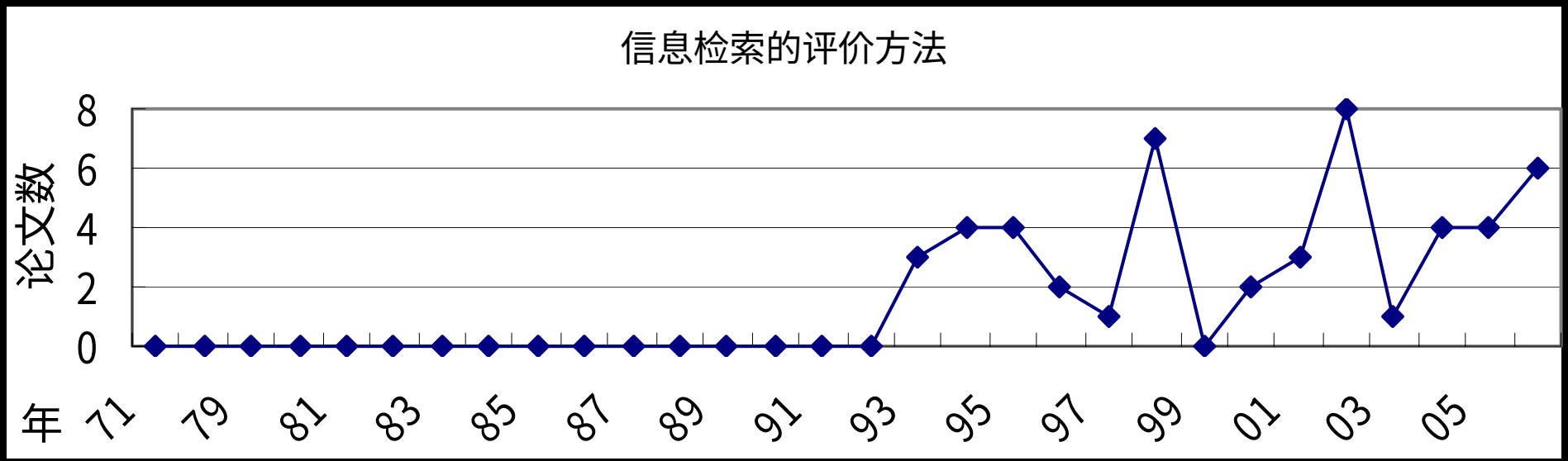
- 人们始终青睐有加的研究领域
- 早期：可视化表示 （查询、文档的可视化）
- 自然语言交互界面
- 2002年以后：
  - 用户日志分析，Social Network，快速学习能力



信息检索的评价

检索的评价

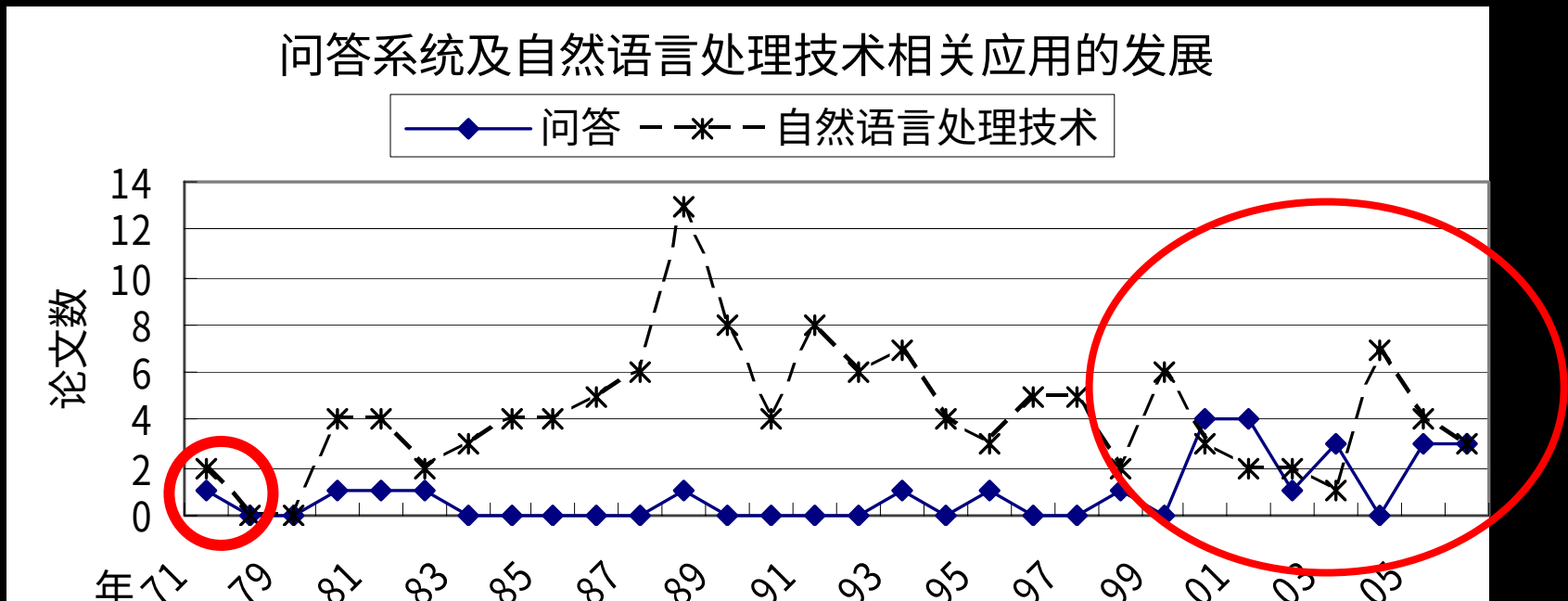
- TREC
- Pooling技术
- 更紧接本质的评价技术
- 评价与技术的共同发展



信息检索中的自然语言处理

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



更多思考与讨论

IR 的发展

- 来源之一：实际应用
  - 分布式系统
  - 系统设计与实现的可扩展性、鲁棒性
  - **Web IR, 链接分析**
  - 用户分析：搜索日志分析

IR 的发展

- 来源之二：国际标准评测
  - 跨语言检索
  - 信息检索的评价与测试集的构建
  - 话题检测与跟踪
  - 新信息发现

IR 的发展

- 来源之三：二者共同推动
  - QA
  - 检索模型发展
  - Spam
  - Intranet信息检索
  - Blog检索与情感分析
  - …

总结

- 缘起
- 信息检索模型
- 关键技术的发展
- 检索任务的演化
- 人机交互/用户分析
- 检索的评价
- 信息检索与自然语言处理
- 其他思考—— 关于IR的发展

谢谢！