信息检索研究：过去三十年中我们走了多远*

马少平，张敏

{msp,z-m}@tsinghua.edu.cn

（清华大学计算机科学与技术系 北京 100084）

摘要： 本文以对信息检索领域最顶级的国际会议 SIGIR 三十年来论文主题的分析为切入点，结合近来年对相关领域的研究和实践，对信息检索研究的发展变化历程和趋势进行总结和分析。

关键词：信息检索，SIGIR

回顾现代信息检索方法与技术发展，从 1971 年第一次信息检索相关的国际会议 SIGIR 的召开到现在，已经经历了超过三十个年头的历程。在这过去的三十年里，人们从最初开始探讨什么是信息检索，尝试设计信息检索系统的基本体系架构，研究应该如何高效地存储文档，如何判断文档与用户查询的相似性等最基本的问题开始，到在互联网上帮助人们进行信息查找的搜索引擎的出现并得到广泛应用，再到后来更多更高级更深层次的技术的提出，直到今天，人们开始感慨“搜索无处不在”，开始讨论什么是“下一代搜索引擎”。三十年的时间里，我们都做了些什么？有哪些问题是研究者们所一直关注着的问题？又有哪些问题经过历史的发展已经渐渐被人们遗忘和冷落？有哪些是随着社会和技术的进步所提出的新的课题？

在这里，我们在近年来对信息检索的研究和实践的基础上，以分析信息检索领域最顶级的国际会议 SIGIR 在这三十年中所收录的论文及其主题的发展变化为切入点，尝试由此映射出整个信息检索领域相关研究发展变化的历程和趋势，一起回顾和总结一下在信息检索研究的道路上，我们已经走了多远？我们还有多远的路要走？

一、 信息检索模型的发展与变革

进行信息检索方法研究需要解决的第一个问题就是检索模型和结构。

从最初起，信息检索的系统设计就是沿着两条路来走的：一是借鉴结构化数据处理的基础和知识，借助数据库等已较为成熟的技术来实现文档的全文检索；二是不拘泥于已有的技术思路，而从文档本身的特点出发，设计并实现专门用于信息检索的体系结构。在开始的十年里，在这两条道路上不同的研究者们都分别进行了充分地尝试，研究并设计出了不同的信息检索体系结构。可以说，这时面向结构化数据的信息检索还占据了相当重要的地位。但是当我们把目光转入 90 年代之后，就会发现与结构化数据存储和检索相关的研究在人们的视线中只是偶尔出现了。直到进入 2000 年，这条分支似乎又有所复苏 —— 但是事实上，与二十年前相比，人们已经转变了思路，或者说开辟了一条新的道路——基于以 xml 为代表的半结构化数据的检索。于是，最初形成的两条分支开始呈现出一定程度的融合趋势。

在信息检索的一般方法和理论层面，研究者们的努力从来都没有停懈过，只是研究的关注点会随着时间的推移而有所改变。在文档信息量还不是如此巨大的时代，人们还是比较关心检索文档的召回率，希望能够找到的信息越多越好。但是随着信息的爆炸式增长，信息量已经不是问题，而相反地，如何能够找到更准确的信息，如何提高系统的鲁棒性，如何进行高效的文档压缩，如何提取出文档中最丰富最有效的信息，如何进行特征降维等问题开始受

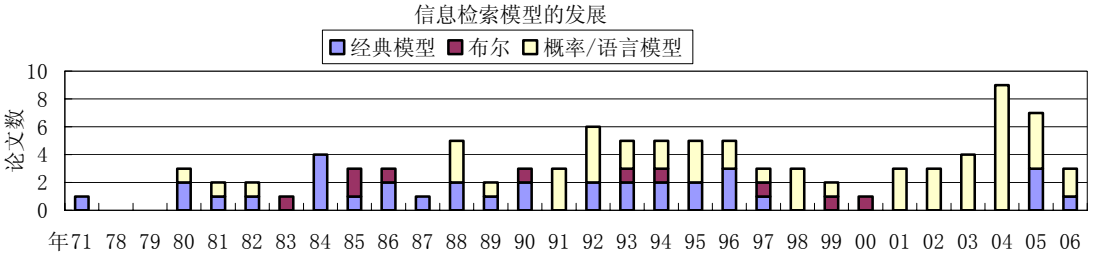
* 基金项目：得到国家重点基础研究（973）(2004CB318108)、自然科学基金（60223004, 60321002, 60303005, 60503064）和教育部科学技术研究重点项目（批准号：104236）资助。

到关注，一些相关理论也得到了稳定的发展，例如潜在语义索引（Latent Semantic Index）从八十年代被提出，到九十年代开始有一些进展，而进入最近十年之后，则几乎每年都会报告出新的进展。再例如近几年来，高精度检索，高鲁棒性检索，对用户查询的分析与预分类等问题也都开始成为研究的热点。

而信息检索模型的发展则经历了三个不同阶段：

首先是以文档的向量表示、以及 TFIDF 等为代表的经典信息检索模型的提出，并在整个八十年代里始终是人们研究的重点；从八十年代末开始，概率模型（特别是以 Okapi 系统为代表的 BM25 系列算法）出现并逐渐分享了经典模型在信息检索模型领域的地位，成为新兴的且功能强大表现越来越出色的模型；到九十年代末期，在自然语言处理、语音识别领域已经受到广泛应用的语言模型开始被应用于信息检索（确切地说，应该是以 1998 年有两篇关于语言模型的论文同时在 SIGIR 上发表为里程碑），打破了数十年来在检索模型上没有大的变革的局面，从此，概率模型和语言模型成为信息检索研究领域最常用的两种方法，受到其影响，传统的检索模型逐渐受到冷落。在整个期间内，布尔模型以及扩展布尔模型也时有研究出现，但是始终没有形成与其他三种模型一样的广泛而持续的影响。

从近两三年开始，新的局面又悄悄出现了：在模型和理论研究上，人们的注意力开始向这已有的三种主流模型以外开始扩展，逐渐提出了更多的新的框架、理论和方法，也开始探索信息检索模型的本质。我们认为，从这时开始，信息检索已经进入了一个新的阶段，并可以期待能够有更具突破性的进展。



图一 信息检索模型相关研究在过去三十年中被 SIGIR 收录的情况分析

二、 信息检索关键技术的发展

下面的表一中列出了在历届 SIGIR 会议中，信息检索的模型方法和关键技术等研究主题被收录论文的情况。窥一斑而见全身，结合我们近年来的研究与实践，由此不难分析出其阶段性的发展历程和变化。

倒排索引(inverted index)的提出解决了信息检索体系结构上的一个核心问题，针对这一方面的研究主要集中在 1985 到 1995 年期间。从 2000 年以后，人们开始更多关注大规模数据处理以及检索效率等问题，并针对这些问题对索引和文档压缩进行更进一步的处理。在 2006 年有一篇关于 spam 的论文，则更是体现了真实环境下的信息检索研究（虽然早在 2003 年，Google 的研究者就已经呼吁研究届对大规模数据中的垃圾和有害信息问题进行关注，并将其称为搜索引擎所面临的挑战之一）。这些发展和变化，表明了信息检索的研究，已经开始走出实验室，而面向海量数据、实时处理、真实网络环境下存在的挑战问题。

相关反馈在信息检索研究领域是一个经久不衰的话题。从信息检索刚提出到现在，始终有人在研究如何通过相关反馈来改进检索的性能。其发展过程大致可以分为三步：一是在研究初期，关注如何建立信息检索的反馈机制；二是 90 年代中期开始，由于多媒体信息检索越来越受到关注，特别是基于内容的图像信息检索（Content-Based Image Retrieval, CBIR）的提出，使得相关反馈技术有了进一步的发展，并成为图像信息检索中一个相当重要的环节

和热点话题；第三是近几年来人们对信息检索的质量的追求，要求系统能够返回更精准的检索结果，这也促进了研究者们开始重新思考如何进行有效的相关反馈，于是提出了更有针对性的 topic by topic 甚至是 term by term 的反馈方法（即针对不同的用户查询，选取不同数量的相关文档中的特征词进行反馈，甚至对特征词的选择也根据查询的特性而采用更精细的算法），并获得了一定的成功。

在检索系统设计实现方面，随着互联网上数据的迅猛增长，集中式的检索系统已经不可能满足实际的需求，分布式系统成为海量数据处理发展的必然趋势。因而从九十年代中期以来，分布式信息检索的相关研究有了长足的进展。特别是针对系统的可扩展性和系统运行效率，人们对系统进行了深入的优化和改进，涉及到索引体系结构的变化，根据用户查询进行自适应检索，以及各种先进的结果融合技术等。

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

表一 信息检索模型及关键技术在历届 SIGIR 会议所收录文章中的发展变化情况

三、 信息检索任务的变化

下面的表二中了列出了过去三十年来曾经或者正在成为热点的检索任务及其发展变化。

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**	

表二 热点信息检索任务在历届 SIGIR 会议所收录文章中的发展变化情况

从八十年代末期开始，Web 信息检索产生并发展起来 —— 这也是互联网逐渐繁荣的时期。当时研究者们主要关注超链接文本的特性，特别是与传统的普通文本（plain text）相比 Web 文档所具有的特征，例如文本内的 html 标记信息及其相关含义、链接等。特别值得一提的是，从 1998 年开始，Web 信息检索，事实上是整个信息检索领域进入了一个全新的阶段，其里程碑就是链接分析技术的提出。1998 年，在 WWW 会议上，Kleinberg 和 Page 同

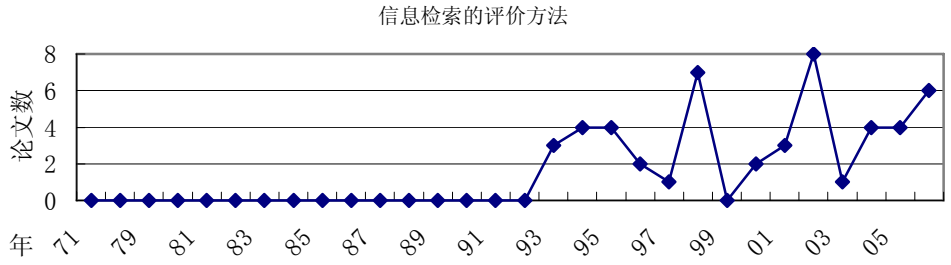
时发表了两篇论文，与以往不同，他们所关心的不再只是简单的网页特性（html 标记、入链接、出链接等），而是把整个 web 看作一个完整的拓扑结构，在一个统一的整体上区分每个网页在整个 Web 环境中的重要性——该重要性在 Kleinberg 提出的 HITS 算法中用权威度（authority）和集成度（hub）来描述，而在 Page 提出的算法中则用 PageRank 值来描述 —— 从此时开始，人们公认 Web 信息检索开始进入了新一代搜索引擎阶段。也是从此时开始，对 Web 上的链接分析如火如荼地展开。研究者们从开始的关注 PageRank、HIT 算法的各种改进，到提出基于站点（site）而非网页（page）的链接分析方法，到研究网页内部的块（block）结构及其上的改进 PageRank 算法，再到后来关心 Web 环境中一个网页的代表性和网页质量描述等等，Web 上的各种信息得到了充分了利用和研究。直到今天，这一领域的研究仍然在持续进行着，并不断有新的想法出现。

另一个热点任务是多媒体信息检索。虽然多媒体信息检索的概念几乎是与文本信息检索同时被提出的，但是在整个八十年代，其大多数的研究都集中在图像检索领域。同时由于多媒体信息检索问题的复杂性，如特征空间的高维度问题，多媒体信息的底层特征与用户查询语义描述之间所存在的鸿沟问题（这比文本检索中的同样问题表现得更加突出），多媒体数据的处理能力等问题，都造成了完全依靠多媒体技术和底层特征进行检索的研究往往局限在较小的数据集合上，因而大多只是实验室研究成果。而在实际应用中，则还是要依赖文本信息检索的技术作为其基础和核心。在最近五年中，更多的研究者们将注意力扩展到视频检索、音频/音乐检索等领域，因此多媒体信息检索也开始进入了一个逐渐繁荣的阶段。

跨语言/多语言信息检索的兴起，与文本信息检索国际标准评测 TREC（Text REtrieval Conference）是分不开的。TREC 在九十年代中后期，分别举办了日语、汉语、阿拉伯语等语言的非英语信息检索或跨语言信息检索任务，与此同时，相关研究领域也开始成为热点。直到今天 NTCIR 的亚洲多语言信息检索评测的举行，也仍然为这一研究领域的发展继续做出贡献。查询翻译技术和自然语言处理技术成为这一研究领域的关键点。

事实上，国际标准评测的召开，对整个信息检索研究的发展中都起着至关重要的影响，有相当多的检索任务都是由这些评测所提出和带动的，例如 TREC 的 Genomics 任务推动了特定领域信息检索方法研究，而文档片断理解的相关研究发展（包括相关段落检索、句子级别的新信息查询、话题检测与跟踪、文档片断的情感分析等）则离不开 TDT（Topic Detection and Tracking）评测、TREC 的 novelty、HARD 等评测任务的作用。

同样，信息检索的评价方法及测试集构建的相关研究发展也应归功于评测的举办。

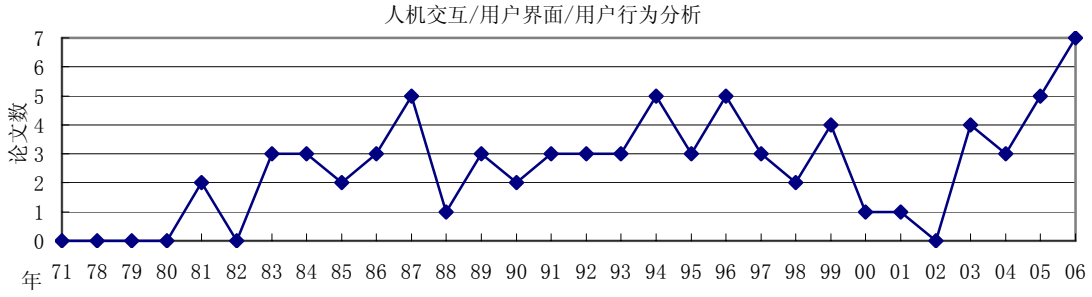


图二 信息检索的评价方法被 SIGIR 收录的情况分析

四、 人机交互与用户行为分析

在信息检索研究领域中，人们对人机交互和用户的研究始终青睐有加，这一主题也几乎在每一年的 SIGIR 大会上都占有一席之地，如下图三所示。不过这个研究主题下，人们的关注点也发生着变化，例如最初的检索本身的研究，后来研究文档和用户信息的可视化表示，

再到自然语言的交互界面，特别是从 2002 年以后，搜索日志和用户行为研究被给予了高度的重视，人们开始从更深的层次，甚至社会网络等方面来研究更友好的人机交互，并力图使能够具备从交互中进行快速学习的能力。

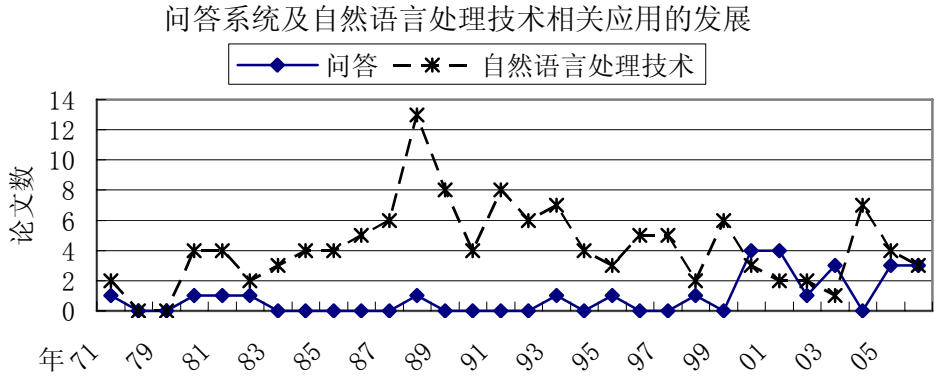


图三 人机交互及用户行为分析课题在 SIGIR 中所反映的情况

五、 自然语言处理在信息检索中的应用和发展

有趣的是，我们原以为，问答系统应该是近十年才出现的新的检索方式，但是回溯到信息检索最初的发展源头，我们发现这个问题从最初就被提出了，并且其研究始终贯穿着整个信息检索发展的历史。

同样地，从有信息检索技术的那一天起，人们就同时在思考着，如何能够把自然语言处理的相关技术引入信息检索中——如何使机器自动检索的效果能够达到人类自己进行检索一样的准确。然而不得不承认的是，在这条路上，我们虽然走了很长时间，但是走得还不够远。目前为止，自然语言处理在信息检索中的应用还仅仅集中在分词、词典的使用、词义消歧、命名实体识别等方面，信息检索与自然语言处理技术的结合还只是通过较松散的预处理、后处理等方式进行的。不过，值得高兴的是，研究者们已经开始考虑如何将更深层次的自然语言处理技术应用到信息检索中来，例如句子的完整性重建(sentence completion)等。特别是 2005 年，在 SIGIR 上有两篇文章同时提出了不同的将自然语言处理的信息融合到检索模型的语言模型中去的方法，也使相关研究整体上向前迈进了一步。



图四 问答系统和自然语言处理技术相关应用研究在过去三十年中被 SIGIR 收录的情况分析

六、 结论

事实上，纵览信息检索研究的发展历史，其推动力通常有两个方面：一是实际应用需求，由在真实环境中运行的系统所面临的挑战和困难所提出的研究课题，例如分布式系统的发展，系统设计上的可扩展性、鲁棒性发展，Web 信息检索与链接分析，搜索日志分析等；二

是由各种国际标准评测所提出和带动的研究关注点，典型的例子有跨语言检索的发展、信息检索的评价与测试集的构建，话题检测与跟踪，新信息发现等。当然在信息检索领域中，还有不少研究方向的发展则是两种推动力共同作用的产物，也受到了二者共同的启发，例如QA，检索模型的发展等，这也是研究界所最期望的。

当我们回顾历史时，很高兴地看到，信息检索的研究如同一棵已经开始成长和壮大的树，经过过去三十年的培育和发展，枝叶已经越来越茂密和伸展，根也开始越扎越深。在信息检索研究的道路上，我们已经走了很远；我们相信未来还有很长的路要走，而未来的路也会越走越宽阔。

参考文献：

SIGIR 论文集 1971~2006，Annual ACM Conference on Research and Development in Information Retrieval.