文本检索会议简介

吴立德 黄萱菁
复旦大学计算机系，上海 200433
{ldwu,xjhuang}@fudan.edu.cn

摘要：由美国国家标准技术局和国防部高级研究计划局组织召开的文本检索会议（TREC）是文本检索领域最具权威的国际评测会议。本文从测试主题、语料库、评测方法和结果等方面介绍了2000年举行的第九次文本检索会议和其中的四个主要项目：问题回答、Web 检索、跨语言检索和文本过滤。

关键词：文本检索 问题回答 文本过滤 多媒体检索

Introduction to the Text Retrieval Conference

Wu Lide, Huang Xuanjing
Dept. of Computer Science, Fudan University, Shanghai 200433
{ldwu,xjhuang}@fudan.edu.cn

Abstract: Text Retrieval Conference (TREC), which is sponsored by the National Institute of Standards and Technology as well as the Defense Advanced Research Projects Agency, is the most authoritative international evaluation conference about text retrieval. This paper describes the Ninth Text Retrieval Conference held in 2000 and its four main tracks, which are Question Answering, Web Retrieval, Cross Language Information Retrieval and Text Filtering, from the aspects of test topics, corpus, evaluation metrics and results.

Keyword: Text Retrieval, Question Answering, Text Filtering, Multimedia Retrieval

1. 文本检索简介

随着互联网的发展和存贮技术的提高，计算机可读的文本信息也越来越多。据估计，截止到1999年，互联网上已约有15TB的信息容量，其中文字信息约为6TB[1]。然而，要有效地开发利用如此丰富的信息资源并不是轻而易举的事情，因为许多信息往往是规模巨大，实时性强，而且存贮分散；语言混杂，内容广泛；图文并茂，格式灵活，有时还含有一定的拼写错误或传输错误。而对于特定的用户而言，所需要的信息往往只占其中极小的一部分。要从如此规模的网络信息中抽取有用的信息资源，对信息处理的速度和精度将提出极为严格的要求，因而迫切需要对这种形式的混合语料进行更快速高效的处理。

在这种情况下，人们越来越多地依靠文本检索工具来寻找自己所需要的信息。文本检索指的是给定文本方式的检索需求，在电子文档库中查找出与指定表达式相匹配的文本，并将出现和包含这些文本的原文作为检索结果返回给用户。

文本检索的重要性早已得到了学术界、工业界和政府部门的重视。权威杂志《BYTE》早在1995年就已把大规模文本信息处理软件预测为未来最有价值的、能成为继办公自动化软件之后最流行的五种软件之一[2]；Google、AltaVista、Lycos和Yahoo等搜索引擎已经成为许多用户遨游互联网必不可少的重要工具。在国内，在2000年10月25日召开的全国中小学信息技术教育工作会议上[3]，教育部长陈至立指出："要把掌握和运用信息技术的能力作为与读，写，算一样重要的新的终身有用的基础能力。在知识经济时代，信息素养已成为科学素养的重要基础。"

在这种背景下，美国国家标准技术局（National Institute of Standards and Technology，简称 NIST）和国防部高级研究计划局（Defense Advanced Research Projects Agency，简称 DARPA）组织召开了一年一度的文本检索会议(Text REtrieval Conference，简称 TREC)。

TREC 会议的宗旨主要有三条[4]：通过提供规范的大规模语料（GB 级）和对文本检索系统性能的客观、公正的评测，来促进技术的交流、发展和产业化；促进政府部门、学术界、工业界间的交流和合作，加速技术的产业化；发展对文本检索系统的评测技术。

TREC 会议从 1992 年开始，迄今已举办了 9 次。参加单位包括许多著名的大学和公司，还有不少美国以外的文本检索领域的研究团体。TREC 不仅提供了一个标准文档库，而且还提出了一套较为科学的测试评价方法，为各种方法和系统提供了一个公平竞争的舞台，使 TREC 成为文本检索领域最权威的国际评测会议。

接下来，我们将介绍 2000 年举行的第九次文本检索会议[5]，并从测试主题、语料库、评测方法和系统排名等方面介绍本次会议的四个主要任务：问题回答、Web 检索、跨语言检索和文本过滤；最后再简要地介绍将于今年召开的最新一次的文本检索会议——TREC2001。

2. 第九次文本检索会议

2000年举行的第九次文本检索会议（TREC-9）是最近一次的文本检索会议，共有来自世界各地的100多个单位报名，实际提交结果的有70个单位，包括许多著名的大学，如：CMU; Johns Hopkins U.; Syracuse U.; U. Of California Berkeley; U. Of Cambridge; U. Of Maryland; U. Of Massachusetts等；也包括许多著名的公司，如：AT&T; BBN; Fujitsu; IBM; Microsoft; NTT; RICOH; SUN; XEROX等。特别地，有四家参加单位来自中国：复旦大学、香港中文大学、台湾大学、微软中国研究院。值得注意的是，复旦大学和微软中国研究院是第一批来自中国大陆的参加单位。

每届文本检索会议都针对当前文本检索会议的最新热点，设置若干个评测主题。早期的主题是标准的文本检索，由两个主要的研究任务组成。一个是被称为"常规

检索”（Routing）的任务。它是这样被定义的：已知用户的检索需求和对应于该检索需求的训练文档集中的相关文档，用自动或人工方式构造被测系统的查询实例来查询测试文档集。这个任务主要是测试系统使用训练文档集构造此文档集“模板”（Profile）的能力，和在相关文档已知的条件下尝试新的检索算法。另一个被称为Ad hoc的任务。它被定义为：已知一文档集和新的检索需求，自动或手动构造查询实例来搜寻该文档集的相关文档。这种方法就象读者在图书馆中进行书目检索。

近年来，随着文本检索技术的不断发展和成熟，文本检索会议逐渐把评测主题转移到更新的研究方向上（称为项目）。下表按参加单位数量从多到少的顺序，列出了TREC-9设置的7个评测项目，并特别地列出了中国参加单位的情况。

表 1：TREC-9 的 7 个项目

项目名	中文含义	参加单位数	中国参加单位
QA	问题回答	28	复旦,台湾
Web	网页检索	23	/
CLIR	跨语言检索	16	复旦,香港中文,台湾,微软
Filtering	文本过滤	15	复旦
Interactive	交互检索	6	/
Query	查询处理	6	/
SDR	语音文本检索	3	/

表中的前四个项目参加单位较多，而后三个项目只有少数的单位参加，其中的Query 和 SDR 两个项目在 TREC-2001 已经取消了。在下面的几个小节中，我们将从任务、语料、评价方法和评价结果几个方面对前四个项目作简单介绍。

2.1 QA（问题回答）

通常意义下的文本检索输入的查询是关键词，返回的是相关文本。而问题回答（Question Answering，简称QA）输入的查询是问题，希望返回的不是整篇文本，而是文本中的相关片断。这个项目是第二次进入文本检索会议，却吸引了最多的参加单位。

具体地，问题回答要求对输入的问题，从给定语料库中返回排了序的 5 段答案。答案的长度有两种，分别为 50 和 250 字节。评价方法如下：如在返回的第一段中即包含问题的答案，得 1 分；如第一段中无答案，而第二段中有，得 1/2 分，…。依此类推，分别为 1/3、1/4 和 1/5 分。如全部 5 段都无答案，得 0 分。对全部问题取平均，其值即为评价结果，称为 Mean Reciprocal Rank。

测试问题共有 682 个，都是客观的或称为“基于事实的”，每个问题都确保在文档集中有答案。问题的几个例子如下：

Who invented the paper clip?

How much folic acid should an expectant mother get daily?

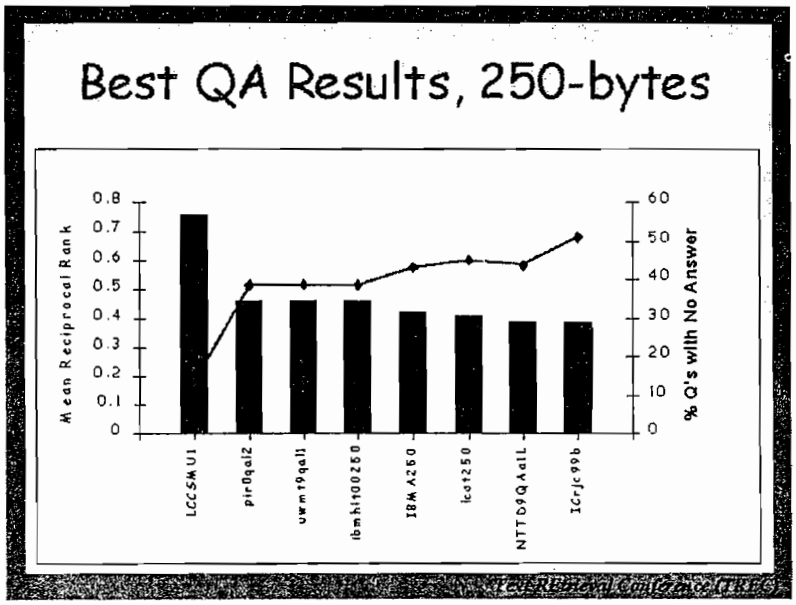
Name a file in which Jude Law acted?

测试语料库是 TREC 主语料库的一个子集。主语料库包括 5 张光盘，总计 4.5G

文本。所用文档有的来自财经报道，如华尔街日报，也有的来自新闻报道，如美联社新闻，还有来自计算机领域的文章，如 Computer Selects articles，还有公文，如 Federal Register 等，内容十分广泛。所有的文档都用 SGML 语言进行了极简单的标记。问题回答项目选用了主语料库中新闻财经类的全部文本作为语料库，包括美联社新闻、华尔街日报、圣何塞水星报、金融时报、洛杉矶时报和国外广播信息（FBIS）六个部分；而答案的评测全部由人工进行。

图 1 说明了最好的一批 QA 系统，要求的答案长度是 250 字节。图中的曲线表示系统回答不出的问题的比例。图中取得第一名的系统来自 SMU（Southern Methodist University）。这个系统在 50 字节的子任务中也同样获得了第一名。获得第二和第三名的分别是 Queens College(CUNY)和 Waterloo University。

图 1：最好的一批 QA 系统



2.2 WEB（网页检索）

Web检索项目试图模拟在WWW上的信息检索，此次也是第二次进入文本检索会议，共有23个系统参加。这也从一个侧面反映了文本检索会议捕捉检索领域研究热点的敏锐程度。

Web 检索有两个子任务。主要的子任务（称为 Main Web Task）的目标一方面是建立一个标准的测试库，另一方面是考察网页中的链接信息是否可提高检索的性能，以及现有的系统是否能很好地处理异质的、动态变化的文档集。所用的语料为从 WWW 中选出的一个子集，容量为 10GB。所用的查询一共有 50 个，由 title（标题）、description（描述）和 narrative（叙述）三个部分组成，其中的标题部分是真实的查询语句，由 Alta Vista 等检索引擎提供，例如：

<title> do beavers live in salt water

<desc> Describe the normal habitat for beavers; note exceptions, if any.

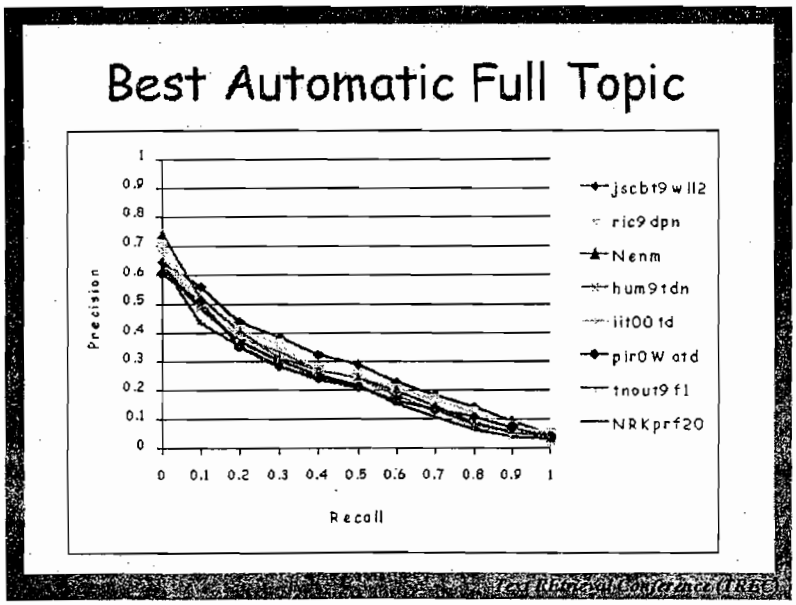
<narr> Relevant documents describe the habitat range as well as references to specific areas and bodies of water.

Main Web子任务的评价指标是历次文本检索会议通用的指标。对每一查询，要求系统从指定的语料库中返回按相关程度排了序的1000篇文本，然后分别在计算在0.0、0.1、……和1.0召回率下的准确率，并进行平均，称为平均准确率（MAP）。这里的准确率较容易计算，而召回率的评价就比较困难，因为需要一个答案的全集，这在语料库很大的情况下几乎是无法获得的。TREC对此采用Pooling法加以解决：每个测试系统对文档集进行检索后，将它们返回的最相关的前面若干篇文档合并，对这个文档集进行人工评价，并把选出的相关文档作为答案的全集。

另一个子任务（称为 Large Web Task）则试图模拟真正的 Web 检索。考察的指标一方面是检索性能评价，即考察检索出的前 20 篇文本的准确率，此外还包括硬件环境、索引速度、索引容量和查询处理速度等。所用的语料有 100G，包括 1.85 亿个网页，查询语句共有 10000 个，例如“*the chicken heart*”和“*dundee scottland*”。

图2说明了最好的一批Web系统（Main Web）。第一名和第二名是日本的两个公司JustSystem和Ricoh，第三名的则是瑞士的University De Neuchatel。

图 2：最好的一批 Web 系统



2.3 CLIR（跨语言检索）

CLIR (Cross-Language Information Retrieval)，即跨语言信息检索，共有16个单

位参加。CLIR要求对一种语言的查询语句，从其它语种的语料库中找出相关的文件；实际上要考察的是系统进行面向文本检索的机器翻译，以及合并来自多语种的相关文本集的能力。前几次文本检索会议CLIR所用的语言都是西方语言，如英语、法语、德语、意大利语等。本次会议CLIR的查询语言是英文，检索的目标语料是中文，因此吸引了全部中国的研究小组参加这个项目。

CLIR的评价指标和Web检索是相同的。语料是繁体中文。包括98、99年的香港商报、大公报、香港日报，共约260M，语料编码为Big5。而测试主题一共有25个，由title（标题）、description（描述）和narrative（叙述）三个部分组成，例如：

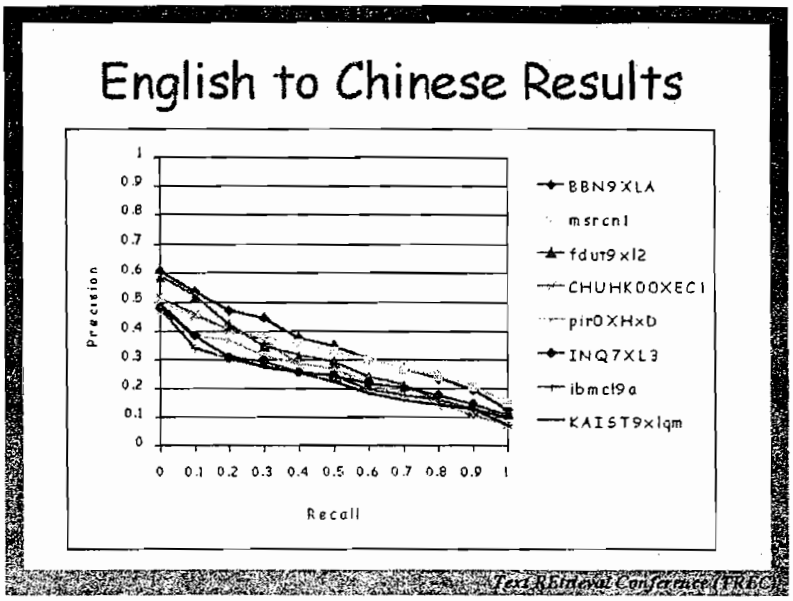
<title> World Trade Organization membership

<desc> Description: What speculations on the effects of the entry of China or Taiwan into the World Trade Organization (WTO) are being reported in the Asian press?

<narr> Narrative: Documents reporting support by other nations for China's or Taiwan's entry into the World Trade Organization (WTO) are not relevant.

图 3 说明了最好的一批 CLIR 系统。获得第一名的是美国的 BBN 公司，第二名的是微软中国研究院，获得第三名的是复旦大学，获得第四名的是香港中文大学。

图 3：最好的一批 CLIR 系统



2.4 Filtering（文本过滤）

Filtering 即文本过滤，共有 15 个单位参加。它试图模拟这样一个过程：在用户信息需求相对固定的情况下，给定一个动态变化的文本流，要求系统自适应地、实时地向用户推荐相关文档，并根据用户对相关性的判断结果自适应地改进系统。这里的用户信息需求则由预定义的主题描述和训练文档来表示。根据训练文档数量的

多少，分别称为自适应过滤（只有 2-4 篇训练文档）和批过滤（第一年作为训练文档）。此外还有一个较为次要的 Routing 子任务，引入它的目的是为了和过去的文本检索会议兼容。

文本过滤的语料库采用了医学文献语料库 OHSUMED。这是著名的美国国家医学图书馆（National Library of Medicine）的 MEDLINE 医学文献库的一个子集，由 1987－1991 年的医学文摘组成，共含文本近 35 万篇，来自 270 种医学期刊，总容量为 400M。其中 87 年的文摘将作为训练语料，而 88－91 年的文摘将作为测试语料。相应地，用户的信息需求则表现为疾病或症状描述，共有 63 个，由标题和描述两个部分组成，例如：

<title> 60 year old menopausal woman without hormone replacement therapy
<desc> Are there adverse effects on lipids when progesterone is given with estrogen replacement therapy

文本过滤的主要评价指标是一个基于准确率的指标，称为 T9P：$T9P = R^+ / \max(MinD, (R^+ + N^+))$。这里的 R^+和 N^+分别表示检出的相关和不相关文本。这个指标强调过滤的准确率，但要求检出的文本数不少于一个下限 MinD（4 年 50 篇）。对每个主题的 T9P 进行平均，就得到全局的平均准确率。

文本过滤项目的评测结果如下：对最主要的自适应过滤任务，英国伦敦的城市大学与微软英国研究院的联合小组获得了文本过滤的第一名，美国的 CMU 和复旦大学分获二、三名。在批过滤任务上，前三名分别是复旦大学、微软英国研究院和 CMU 的另一个小组[6]。具体数值如下表所示：

表 2：最好的一批文本过滤系统

Routing: P@50		Batch filtering: T9P		Adaptive filtering: T9P	
ICDC	37.0	Fudan	31.7	Microsoft	29.4
Microsoft	33.6	Microsoft	30.5	CMU-LTI	27.9
Nijmegen	28.2	CMU-Y	26.1	Fudan	26.5
OHSU topics Filtering runs optimised for T9P Best runs from best 3 group					

3. 总结与展望

表 3 对上述 4 个项目的排名作了一个简单的小结。表中的第二到第五列是 4 个项目（括号里是参加单位数量），第一列是各项目主要任务排名位于前列的单位。

2001 的文本检索会议已经于今年 2 月份启动。今年的文本检索会议一共设 6 个项目，其中 CLIR、Filtering、Interactive、QA 和 Web 这 5 个项目是原先就有的，但在具体任务上则有所改变。例如今年 CLIR 的语料将为阿拉伯语，而 QA 则增加了一个新的子任务，问题的答案不能简单地从一篇文档中获得，而要求从若干个文档

中总结出。此外，还增加了一个新的任务：Video（视频检索），要求从数字视频中检索出和给定的用户需求相关的视频片断。

表 3：各个项目的领先单位

单位	QA(28)	Web(23)	CLIR(16)	Filtering（13）
BBN Technologies			1	
CMU (卡内基一梅隆大学)				2
Fudan University（复旦大学）			3	3
Queens College, CUNY	2			
Justsystem Corp.		1		
MSRCN （微软中国研究院）			2	
MSRBR （微软英国研究院）				1
University De Neuchatel		3		
Ricoh（理光公司）		2		
Southern Methodist U.	1			
Waterloo University（滑铁卢大学）	3			

总之，经过多年的实践，TREC 已经建立了在文本检索会议的国际权威地位，吸引了世界各地越来越多的高水平的参加单位，也发展了一套较为成熟的评测方法。同时，TREC 还逐渐地把评测领域从单纯的文本检索扩大到多媒体检索，如已经结束的语音文本检索和刚刚起步的视频检索。目前国内对 TREC 感兴趣的研究单位也越来越多，希望本文能对推动国内更多的单位参加 TREC 以及在国内举办中文文本检索会议起到一定的作用。

参考文献

[1] S.Lawrence, C.Lee Giles, Accessibility of Information on the Web, Nature, 400(1999.7.8), 107-109

[2] Tomorrow's Top Five Software Categories，BYTE，September 1995，68—69

[3] 陈至立，抓住机遇，加快发展，在中小学大力普及信息技术教育，在全国中小学信息技术教育工作会议上的报告（2000.10.25）

[4] E. M. Vooheers, Overview of the Seventh Text Retrieval Conference (TREC-8), http://trec.nist.gov，2000 年

[5] E. Voorhees，Presentation to the Text REtrieval Conference (TREC-9) *November 9-12, 2000 Gaithersburg, MD, USA*

[6] S. Robertson, D. Hull, the TREC-9 Filtering Track Final Report，Proceeding of the Ninth Text Retrieval Conference（TREC-9），2001 年 2 月