大海捞针亦有道

——中文信息检索技术的现状与挑战

在知识经济与信息化蓬勃发展的当今时代，信息与知识的占有和利用程度，越来越成为衡量一个国家和民族进步和发展水平的重要标准。

为了解决信息资源利用的问题，将传统的情报检索技术与计算机应用实际相结合的现代信息检索系统应运而生。信息检索系统的目的，是快速、准确和最大程度地从海量信息资源中定位并获取高质量的相关信息。

中文是联合国的工作语言之一，全球以中文为母语的人数有 10 多亿，遍布中国、新加坡、马来西亚、印尼以及世界各地的华人华侨地区。随着中国经济和技术的发展，中文具有越来越重要的地位，中文文档特别是中文网页的数目也在迅速增长。随之而来的，多种多样的中文信息检索工具也应运而生，中文信息检索技术及其实现产品——中文网络搜索引擎的发展，已经为华人访问信息资源提供了巨大的便利。最近的中国互联网络发展状况统计报告[1][2]指出，中国搜索引擎用户已占互联网用户的 95.2%，绝对用户数超过 1 亿人，包括搜索引擎在内的信息检索工具已经成为当今获取信息的主要手段之一。

从数据处理对象的角度分析，信息检索系统面对的处理对象则包括信息资源与检索用户两方面的内容。对中文信息检索技术而言，中文信息资源与中文检索用户的特点决定了其发展的方向，针对这两方面特点所进行的努力也构成了最近三十余年来的中文信息检索研究的主线。

本文中，我们将首先针对传统信息检索与中文信息检索技术的发展历史进行简单回顾；随后针对中文信息检索技术的若干关键问题与发展现状进行分析；最后根据我们自己的理解，对中文信息检索技术的发展方向给出展望。

一、信息检索技术的发展

1945 年，在二次世界大战即将胜利之际，曼哈顿工程和美国自然科学基金的创立人 Vannevar Bush 提出了这样一个想法：在 2010 年左右，世界上应该有一种工具，它能够使人们最方便快捷的获取所有图书馆中藏有的知识，这个构想对应的就是信息检索工具的雏形。在 Bush 的支持下，自 1950 年起美国政府就开始了对信息检索相关研究的支持。1951 年，Calvin Mooers 首次提出了“信息检索（Information Retrieval, IR）”这一概念，给出了信息检索的主要任务：协助信息的潜在用户将信息需求转换为一张文献来源信息列表，而这些文献包含有对其有用的信息。随着计算机技术的发展和迅速普及，信息检索作为应对信息爆炸问题的主要手段而迅速发展起来，其研究领域也由最初的科学技术领域扩展到人类活动的各个方面。

互联网的出现和计算机硬件水平的提高使得人们存储和处理信息的能力得到巨大的提高，从而加速了信息检索研究的进步，并使其研究对象从图书资料和商用数据扩展到人们生活的方方面面。伴随互联网爆炸性的发展，普通网络用户想找到所需的资料简直如同大海捞针，这时为满足大众信息检索需求的专业搜索网站便应运而生了。现代意义上的搜索引擎的祖先，是 1990 年由蒙特利尔大学学生 Alan Emtage 发明的 Archie。虽然当时 World Wide Web 还未出现，但网络中的文件传输已经相当频繁。大量的文件散布在各个分散的 FTP 主机中，查询起来非常不便，因此 Archie 应运而生。随着 WWW 的出现，搜索网站的查找对象从单纯的文件扩展到网页。最早现代意义上的搜索引擎出现于 1994 年 7 月，Michael Mauldin 首次将网络爬虫程序与文本索引程序相结合，创建了现在仍在提供服务的 Lycos 搜索引擎（http://www.lycos.com/）。1995 年，斯坦福大学的两名博士生，David Filo 和杨致远共同创办了基于目录索引结构的 Yahoo!搜索引擎，并成功地使网络搜索概念深入人心，从此搜索引擎进入了高速发展时期。

中文文本信息检索最早见于“748 工程”中的汉字情报检索。1974 年 8 月，我国启动了包括汉字通信、汉字情报检索和汉字精密照排研究在内的“748 工程”科研项目。80 年代中期后，由于计算机处理能力的大大提高和应用的广泛普及，中文文本信息检索的研究开始进入黄金期，各种汉字文本索引方法、检索算法以及实用化系统开始出现，各种全文检索商用系统的出现就是这个阶段的成果，如清华大学的《中国学术期刊(光盘版)》、北大方正的 MIRS、易宝北信的自由词全文检索系统等。中文网络信息检索工具的发展历史则可以追溯到 1997 年，当时北大网络实验室推出了在教育网内的 Web 信息导航服务，即后来的天网搜索引擎。2001 年，百度搜索面世并开始逐渐成为中文搜索引擎市场的领头羊。从 2003 年开始，中文网络信息服务的四大门户网站（新浪、搜狐、网易、腾讯）陆续推出了自己的搜索引擎服务，搜索引擎市场的激烈竞争方兴未艾。中文情报检索、图书期刊的全文检索、中文搜索引擎等的广泛使用大大促进了中文信息检索技术的发展。

二、中文信息检索技术的现状与挑战

1．中文语言特性与信息检索

在自然语言处理中，词是最小的能够独立使用的有意义的语言成分，但中文以字作为其基本书写单位，词语之间没有明显的区分标记，从形式上看，汉语中没有“词”这个单位。因此，进行中文自然语言处理通常都是先将汉语文本中的字符串切分成合理的词语序列，然后再在此基础上进行其它分析处理。将中文字符串切分成合理的词语序列的过程，称为中文分词。

中文的这一特性对于中文信息检索系统的设计，包括文本预处理、索引建立、查询等多个模块都产生直接的影响，这也构成了中文信息检索系统与传统信息检索系统的最大差异所在。

信息检索系统（如果没有特别说明，本文中的“信息检索系统”指狭义的信息检索系统，即文本信息检索系统）的基本架构和运行示意如下图所示：

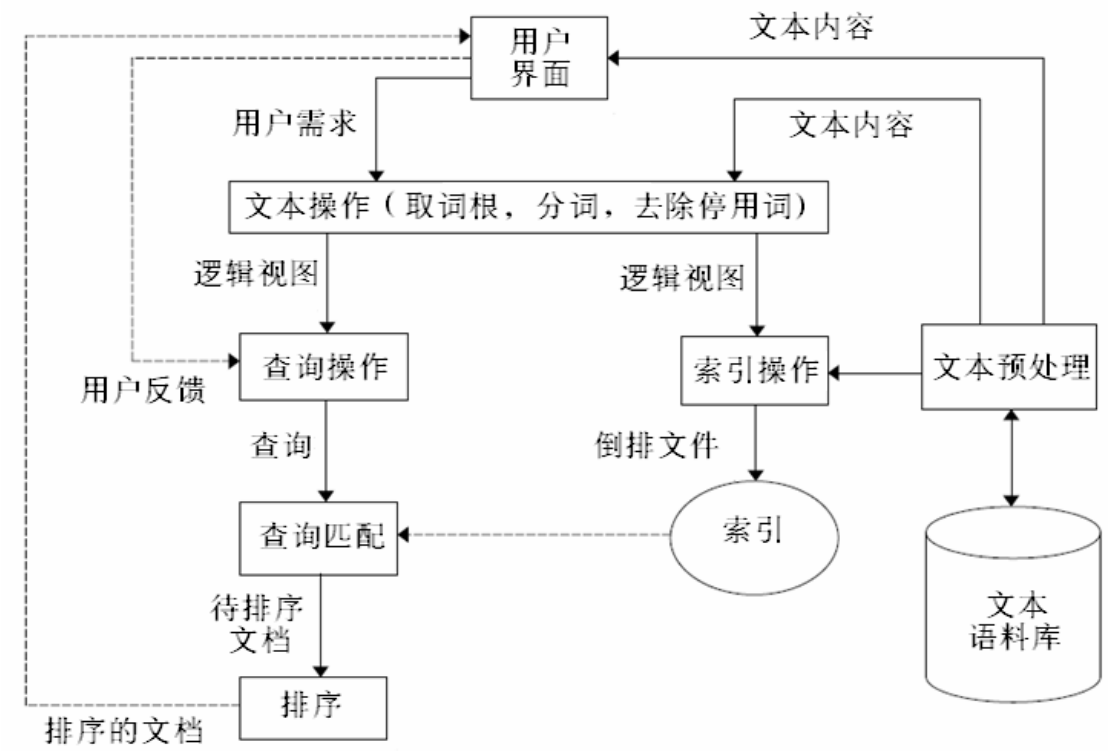

图 1 信息检索系统架构及运行原理示意图，修改自[3]

如图 1 所示，传统意义的文本信息检索系统是由文本处理、内容索引、查询处理、用户界面等模块组成的。具体的运行流程上，系统将信息资源中的文本经过取词根（英文）、分词（中文）、去除停用词等预处理操作（Text Preprocess）后输入进索引模块，索引模块以词项（term）为中心组织倒排索引（inverted index），从而完成系统准备工作。用户进行检索时，首先通过用户界面输入其查询需求（query），查询需求经过取词根（英文）、分词（中文）、去除停用词等操作后通过查询模块在倒排索引中定位相关文档集合，再依据相关文档与查询需求的相似度对相关文档进行排序（ranking），并通过用户界面反馈给用户检索结果。

对于中文信息检索系统而言，由于其索引结构需要以词项（term）为中心组织，因此无论是信息资源文本、还是用户输入的查询需求文本都需要进行分词操作，转化成为“词项序列”方能被检索系统处理。

尽管各种中文自然语言处理应用都需要进行分词，但不同应用对应的分词需求是不同的。针对中文信息检索系统需要的分词方式，国内外都有相当多的研究，

作为文本信息检索领域标准评测的TREC（Text Retrieval Conference，文本信息检索会议）还在 1996 和 1997 两年分别组织了针对中文的检索评测。但是不同的研究得到的结论却大相径庭，如参与TREC 5 与TREC 6 评测的小组中，有 7 个小组对应该基于字还是词为单位组织中文检索系统的索引进行了研究，其中 3 个小组认为基于字为单位更好，其中另外 3 个小组得到相反的结论，还有 1 个小组认为两者的效果类似[4]。

经过近十年的研究和讨论，目前比较公认的针对面向检索系统的中文信息处理算法的结论包括：

[1] 为满足大规模文本检索的需要，中文信息检索系统对应的分词算法的时间性能要好，因此基于复杂规则分析的方法是不可行的。

[2] 以词或以字（包括多元字符串）作为词项单位建立中文索引都是可行的，它们有不同的适用范围。其中以词为单位建立索引可以获得比较高的准确率，而以字为单位建立索引可以获得比较高的召回率。结合实际需求，将两种不同的词项切分方法相结合，往往能取得最好的检索效果。

[3] 索引系统与查询处理系统采用的词项切分方法应当一致，以取得更好的检索效果，即文本信息资源和用户查询需求要进行相同的中文信息处理操作，如分词、专有名词识别等。

尽管中文信息检索系统与传统文本信息检索系统相比，需要多经过词项切分等操作，但不少相关研究[4, 5]都证明，中文检索的效果并不会因为这些操作比传统检索系统有明显的降低。但是如何针对具体需求设计合理的中文文本索引方式，仍旧是中文信息检索系统设计的重点和难点。

2．中文信息资源与信息检索

中文信息资源是中文信息检索系统的处理对象，除中文语言文字的特性之外，信息资源本身的特点也对检索系统的设计和运行产生直接的影响。上个世纪最后二十年以来，网络信息迅速成为社会成员获取知识和信息的主要渠道之一，而中文互联网资源也成为目前相当一部分中文信息检索系统面对的处理对象，本节中，我们将重点讨论中文互联网信息资源对中文检索系统带来的影响。

根据 2007 年 6 月CNNIC发布的《第 20 次中国互联网络发展状况统计报告》[2]，我国所开设的网站数量已经达到 131 万个，并以超过 60%的年增长率飞速增长。而在中文网页数目方面，CNNIC2005 年底估计当时的中文网页数目已经超过 26 亿个，搜狗搜索则在 2007 年初声称其索引量达到 100 亿页面，并表示这只占中文网页数目的一半左右[1]。

可以说，中文网络信息资源的数量已经极大丰富，并且还在以非常快的速度

[1] http://www.donews.com/Content/200701/93090e3265084a78a32fea44ba250530.shtm

增长。然而，网络信息资源的质量却并不能让人满意。根据[6]的统计，2005 年底统计到的 26 亿中文网页中，完全镜像冗余的页面就占 8%左右。而所有统计涉及到的网站中，每天的全站访问量不足 50 次的冷门网站就占近 40%；站点之间的相互引用比较国际互联网的情况也少得多，完全不含指向其他站点的超链接的网站竟占了 37%以上。这说明，尽管从绝对数量上已经达到了一个比较高的水平，但中文网络信息资源的质量还有待进一步提高。

中文网络信息资源质量堪忧的现状从垃圾页面泛滥的现象中也可以看出来。所谓垃圾页面，是指那些利用不正当手段获取在搜索引擎查询结果列表中的较高排名的网页。由于搜索引擎在用户获取信息的过程中发挥的重要作用，一个网站想要被用户点击浏览，最好的办法就是出现在搜索引擎的比较有利排序位置的检索结果中。我们从搜索引擎用户查询日志[7]中选择了 750 个查询词，并进行这些查询所对应查询结果页面的标注，结果发现搜索引擎返回的前二十位结果中有 3.4%的结果是垃圾页面，进一步的实验则发现查询频度较高的查询词对应的垃圾比率更高，甚至达到了 6%以上。其中部分垃圾网页的样例如下图所示：

图 2. 垃圾网页样例（左图为一个堆砌热门关键词的垃圾网页，右图为一个利用 JavaScript 脚本把作弊内容显现给用户的垃圾网页）

可以说，庞大的数据规模与良莠不齐的数据质量是中文信息资源带给信息检索工具的最大难题，而针对这方面问题的研究也正是最近一个阶段的信息检索研究热点方向之一。

3. 中文用户行为与信息检索

中文检索用户是中文信息检索工具的服务对象，信息检索工具的核心目的，就是根据用户提交的信息需求查找信息资源中的相关内容，再反馈给用户。可以说，信息检索的整个处理流程就是围绕用户来进行的。对于中文信息检索系统而言，最大的用户群体来自于中文网民。截至 2007 年 6 月，中国网民总人数达到 1.62 亿，仅次于美国 2.11 亿的网民规模，位居世界第二。与 2006 年末相比，新

增网民 2500 万，增长异常迅速。考虑到我国网络普及率仍然偏低的事实（只有 12.3%，低于全球 17.6%的平均水平），网民数量多、增长快的状况将长期存在。

中文网络用户中，网络信息检索工具（即中文搜索引擎）使用的普及率相当高，据 CNNIC 统计[2]，超过 9 成（90.4%）的网民表示，需要信息时，首先想到的就是去互联网上寻找，尤其是高收入、高文化程度网民群体中接近 9 成（89.2%）都将互联网作为主要信息渠道，报纸和电视作为信息渠道的提及比例分别比互联网的覆盖比例低了近 15 个和 22 个百分点。这说明了中文用户对于网络信息检索工具的青睐程度。

用户行为分析是信息检索技术得以前进的重要基石，也是能够在商用搜索引擎中发挥重要作用的各种算法的基本出发点之一。早在网络信息检索研究开始之前，用户相关反馈在文本检索中的研究就被认为是提高检索精度的主要途径之一。面对复杂的用户需求与更加复杂浩繁的文档集合，通常只有几个字词组成的用户查询成为了影响用户与检索系统之间信息传递的瓶颈，而相关反馈则成为了更明确表达用户意图的有效方案。尽管由于增加了使用成本，用户相关反馈并没有能够在文本检索中得到大规模的应用，但这方面的工作后来却成为了伪相关反馈（pseudo relevance feedback）、查询扩展（Query expansion, QE）等信息检索核心技术的基础。

与此同时，对网络用户的行为分析自从WWW得到大规模应用以来，也得到了足够的关注。Silverstein等在[8]中总结出了一系列用户行为分析结果，如：用户需求中有大量的重复项、绝大部分用户在进行检索的过程中不对其查询进行修改等。这些分析结果对检索系统设计有着深远的影响。我们在 2006 年基于对搜狗搜索引擎的分析，也总结出了中文检索用户（尤其是网络信息检索用户）的一些行为特点[9]。这些结果中既有与英文检索用户类似的特征，如大部分用户只翻看搜索引擎返回结果的第一个页面等。也包括一些异于英文用户的特征，如：中文检索用户提交的查询中只有 0.73%是使用了高级检索功能，远低于英文用户行为分析中 20%的比例；中文检索用户提交的查询中重复率比英文检索用户高得多，少数查询出现总数占了总查询数的绝大部分等。

检索用户的行为分析一直是推动信息检索研究向前发展的重要动力，对于中文信息检索而言更是如此，目前不少中文搜索引擎根据用户行为习惯设计的查询辅助手段如查询提示、相关查询扩展等功能都大大便利了中文用户获取信息的过程，也拓展了检索系统的交互渠道。在可以想见的未来，用户行为分析仍将视中文信息检索研究中的重要方法和研究方向。

三、结语

以上，我们针对中文信息检索与传统信息检索的特点进行了分析，并进一步

从中文信息资源和中文检索用户行为的角度，对中文信息检索当前研究的主要热点进行了介绍。

值得欣慰的是，在产业界和研究人员的共同努力下，中文信息检索的研究和应用都已达到了相当的水平，为中文用户在信息爆炸的时代中获取信息资源提供了有力的保障。但同时我们也必须看到，中文语言固有的一些特点、中文信息资源特殊的繁杂性质和中文检索用户的行为习惯等共同给检索系统的设计与实现带来了巨大的挑战。如何能够更好的应对这些挑战，把一个在信息、知识的汪洋大海中“捞针”的有效工具提供给用户，我们仍旧任重道远。

参考文献

1. 中国互联网络信息中心(CNNIC). 2005. 第 16 次中国互联网络发展状况统计报告, 2005 年 7 月.
2. 中国互联网络信息中心(CNNIC). 2007. 第 20 次中国互联网络发展状况统计报告, 2007 年 7 月.
3. Ricardo Baeza-Yates, Berthier Ribeiro-Neto. Modern Information Retrieval. Addison Wesley Longman Publishing Co. Inc. 1999.
4. 王思力，面向大规模信息检索的中文分词技术研究，中国科学院硕士学位论文，2007 年 1 月。
5. 昝红英，基于实体属性的中文网页检索研究，北京大学博士论文，2004 年 5 月。
6. 中国互联网络信息中心(CNNIC). 2006. 2005 年中国互联网络信息资源数量调查报告, 2006 年 3 月.
7. 搜狗实验室查询日志：http://www.sogou.com/labs/q.html.
8. Silverstein, C., Marais, H., Henzinger, M., and Moricz, M. 1999. Analysis of a very large web search engine query log. SIGIR Forum 33, 1 (Sep. 1999), 6-12.
9. 余慧佳，刘奕群，张敏，茹立云，马少平. 2007. 基于大规模日志分析的网络搜索引擎用户行为研究. 中文信息学报 Vol. 21(1): pp. 109-114, 2007.