网络信息检索的发展趋势（戴莉）

[作者] 戴莉
[单位] 中共海南省委党校图书馆
[摘要] 介绍了网络信息检索的内涵，分析了当前网络信息检索的概况，并对网络信息检索的未来发展趋势进行了初步的探讨。
[关键词] 信息检索，网络，发展趋势

随着网络技术的飞速发展，信息检索工作已经由传统的手工文献检索发展到智能检索。认清网络信息检索的发展趋势，掌握先进的网络信息检索技术，从浩如烟海的信息中找到所需的信息，已成为当前重要而迫切的研究课题。

1、什么是网络信息检索

网络信息检索是由网络站点、网页浏览器和搜索引擎以及网络支撑组成的检索系统，其中的核心部分，不是众多站点，而是网络浏览器和具有收集、检索功能的搜索引擎。众多站点、网页上的信息是网络信息的基本组成部分。在网络发展初期，浏览器和简单的搜索引擎即可帮助人们检索所需的文献信息。浏览器相当于提供了一个信息总目，提供用户对各个网站进行直接点击、浏览，通过超文本链接，选择自己所需的信息。浏览虽然方法简易、直接，但随机性强，耗时费用较高，因此，更科学的方法是借助搜索引擎。搜索引擎是网络信息的检索工具，它可以帮助用户快速搜索所需信息及其相关资料。搜索引擎是因特网上的一种特殊类型的站点，通过用户输入所需信息的关键词，经由检索服务器处理内部数据库，匹配相关资料并整理后输出，通过网络传给用户使用。

2、网络信息检索技术的发展现状

网络信息检索开始于 20 世纪 90 年代初。1991 年思维机等公司、明尼苏达大学、欧洲高能粒子协会分别推出了因特网上的检索工具 WAIS、GOTHER 和 WWW。目前，WWW 因其集文本、图像、声音等多媒体信息于一体的巨大优点，已占信息服务的主导地位，基于 Web 的搜索引擎已成为最重要的信息检索工具。著名的有 Yahoo、Lycos、Infoseek、Excite 等。信息检索发展到今天，已呈现联机检索、光盘检索和网络检索三者并存的局面。

2.1 文本信息检索技术

文本信息检索技术包括传统文本检索、全文检索等二个方面。

(1)传统文本检索

传统文本检索已发展了几十年。它是以文本，特别是二次文献为检索信息源。信息的检索模型有布尔检索模型、向量空间模型、概率模型、模糊集合模型、扩展布尔检索模型等几种。它的检索方式有逐一比较、二分法、随机检索等。具体检索技术有布尔检索、截词检索、限制检索、加权检索、聚类检索等。信息的存储方式有顺排资料档和倒排资料档两种。顺排

资料档检索采用表变换法。倒排资料档检索通常采用逆波兰法处理提问式，并使用倒排算法等查找文本数据库中的叙词等倒排文档来获取信息。目前常用的是倒排文档的检索技术。流行较广泛的文本检索软件有联合国教科文组织开发的 CDS/ISIS 等。

(2)全文检索

全文检索是以全文本信息为主要检索对象，允许用户以布尔逻辑等和自然语言，根据资料内容而不是外在特征来实现检索的先进的检索技术。1959 年美国匹兹堡大学建立了世界上第一个全文检索系统——法律信息检索系统。全文检索系统标引方式有词典法标引、单汉字标引、特殊标引等。检索技术有后控检索、原文检索、期望值与加权检索等，检索功能强大。以全文检索为核心技术的搜索引擎已成为因特网时代的主流技术之一。著名的全文检索系统有 WAIS 等。如何最大限度的提高查全率和查准率是全文检索系统一直在努力研究的内容，这需要从全文检索系统的标引和检索技术两方面配合进行研究。在全文检索领域中，还包括超文本检索和基于概念信息检索两方面的研究内容。

①超文本检索。超文本检索技术是以超文本网络为基础的信息检索技术。在超文本检索系统中正文信息是以节点而不是以字符串为信息的基本单元，节点间以链连接。在检索时，节点间的各种链接关系可以动态的选择激发，通过链从一个节点跳到另一个节点，实现联想式检索。1945 年美国计算机科学家范尼瓦*布什首先提出了超文本思想。1965 年泰得*纳尔逊(TedNelson)提出了超文本(Hypertext)概念。1967 年布朗大学研制成功世界上第一个超文本系统——超文本编辑系统(Hypertext Editing System)。因特网上的搜索引擎代表了超文本检索技术的发展水平，有的还有自动分类、自动文摘、自动索引等功能。著名的超文本检索系统有 Yahoo、Lycos、Infoseek 等。

②概念信息检索。又称基于知识信息检索，是通过对文献中的原文信息进行语义上的自然语言处理，析取各种概念信息，由此形成一个知识库。然后根据对用户提问的理解，检索知识库中相关信息，以提供直接的回答。概念信息检索的理论框架最早由美国著名的人工智能专家 Schank， Kolodner 和 Dejong 在 1981 年发表的《概念信息检索》一文中建立。自 1981 年以来一些概念信息检索系统相继推出，它们具备了一些智能检索的特性，有较强的分析和理解能力。Excite 搜索引擎即是采用概念检索技术的数据库。

2.2 基于内容检索技术

(1)基于内容检索。即多媒体信息检索，20 世纪 90 年代初国际上就开始了这方面的研究。它是直接对图像、视频、音频等多媒体信息进行分析，抽取特征和语义，利用这些内容特征建立索引，然后进行检索。目前，大量的原型系统已推出，典型的系统有 IBM 公司的 QBIC 系统、美国哥伦比亚大学的 Visual SEEK 系统等。

(2)超媒体检索。这是超文本检索的自然扩展，检索对象由文本扩展为多媒体信息。它的检索方法与超文本检索是一样的。目前，超媒体检索正向智能超媒体检索和协作超媒体检索方向发展。WWW 是第一个全球性分布式超媒体系统。其它超媒体检索系统有 Harmony，英国南安普敦大学的 Microcosm 等。

2.3 万维网信息检索技术

(1)检索方式。万维网是利用搜索引擎为检索手段，它的检索方式有分类目录式(网站级)检索、全文(网页级)检索等几种方式。分类目录式检索即超文本检索。在全文检索方式中，搜索引擎使用网络信息资源自动采集机器人程序(也称网络蜘蛛、爬虫软件)，动态访问各站

点，收集信息，建立索引，并自动生成有关资源的简单描述，存入数据库中供检索。但这种机器人程序的查准率有待提高。目前全球最大的搜索引擎是 Google，访问量排在雅虎之后居世界第二。台湾网擎公司启动了 www.openfind.com 全球搜索引擎网站，资料量是 Google 的 1.7 倍，达 35 亿网页，向 Google 发起挑战。

(2)元搜索引擎。又称多元搜索引擎或集成搜索引擎，是网络检索的后起之秀，是多个单一搜索引擎的集合。它没有独立的数据库，主要依靠系统提供的统一界面，构成一个一对多的分布式且具有独立功能的虚拟逻辑机制。主要的元搜索引擎有 W3 Search Engines、Savvy Search、Metacrawler 等。

(3)网络智能检索。包括智能搜索引擎、智能浏览器、智能体(Agent)等。智能搜索引擎可以预期用户的需求，并可有效地抑制关键词的多义性。比较成功的智能搜索引擎有 FSA、Eloise 和 FAQFinder。智能浏览器是基于机器学习理论设计的智能系统，经过训练后，可成为某个领域中熟练的搜索专家。两个比较成功的实验原型是 Web-Watcher 和 Letizia。智能体是一个具有控制问题求解机理的计算单元，网络中的智能体通常是一个专家系统、一个模块等。它在经用户指导后，可在不用用户干预的情况下，找到所需信息。有些智能体使用神经网络与模糊逻辑而不是关键词来识别信息的模式。例如 Brower Buddy 是一个基于规则的智能体。

3、网络信息检索的发展趋势

3.1 可视化趋势将会更加明显

可视化是将数据库中不可见的语义关系用图像方式显示， 并表达用户检索过程。可视化检索有许多优点，主要表现在:对文献或检索式内部语义关系的理解有助于用户判断一个检索中的相关文献；一个透明的检索过程使检索更容易更有效；可视化的环境可以为用户提供更丰富和更直观的信息；相关性在传统的信息检索中只指检索结果、检索式相关，而在可视化检索中则指检索结果之间的相关度；使得用户可以进行交互式输入，允许在信息空间进行动态移动，允许用户修改数据的显示方式，使他们理解数据的个人偏好可视化；减少了理解检索结果的时间，可以对相关信息进行聚类分析(Clusters Analysis)，而聚类分析可帮助人们发现新的学科点，也可作为反馈的工具；操纵检索的内部过程；提高检索系统与人之间的交互性；检索结果可以模仿网络环境形成拓扑结构图，在拓扑结构图中所有相关文献或其他类型资源将被归为同类。

可视化技术如今在地理信息系统(Geographic Information System)、产品设计(Product Design)、城镇建设与规划（Urban Construction and Plan）等领域得到了广泛的应用。可视化信息检索系统也已经出现，如中国气象局设置了网上极轨气象卫星资料可视化检索页面。

3.2 智能化进程将会跨上新台阶

智能化是网络信息检索未来的主要发展方向。智能检索是基于自然语言的检索形式，机器根据用户所提供的以自然语言表述的检索要求进行分析而后形成检索策略进行搜索。近年来，因特网上不断涌现的人工智能产品，如智能搜索引擎、智能浏览器、智能代理等，它们将提高网络信息检索的智能化程度，促进智能信息检索的发展。

智能搜索引擎有 3 个主要的特征：网络蜘蛛的智能化、为特定用户提供相关信息、搜索引擎人机接口的智能化。它可以在因特网中导引用户，不仅在用户进行搜索、浏览时给予直接的支持，而且能够提供具有独立搜索功能的智能体的幕后支持。它还可以预期用户的需求，并有效地抑制关键词的多义性。目前，比较有名的智能搜索引擎有 FSA、Eloise 和 FAQFinder 等。

智能浏览器则是基于机器学习理论而设计的智能系统，经过一定的训练后，它可以成为某个领域中熟练的搜索专家，帮助用户在网络中查找信息。如 Web-watcher 能不断地给用户推荐一系列站点并建立超链接。它可以记录数以万计的用户数据来训练自己，从而不断更新知识；它会对成功检索的每一个超链接用代表用户兴趣的关键词加以注释，并存入知识库。网络中的智能代理通常是一个专家系统、一个过程、一个模块或一个求解单元。

智能代理可以获得用户的信息需求，自动检索信息和推送检索结果。多智能代理系统还具有信息发现、信息筛选、信息推送和信息导航功能，可满足专业研究人员的特定需求，实现网络信息检索与服务的智能化。

3.3 个性化服务将进一步提高

个性化是指各网站针对不同的用户需求提供有特色的服务内容。个性化服务的实质在于提供真正适应用户需要的产品。事实上，网上已经开始出现专门收录某一领域信息的网站，尤其是在一些热门领域，如 Stock Site 提供股市分析文章、股票分析工具、公司研究文章及与商业和金融相关的新闻。一些大型的搜索引擎已注意到个性化信息服务的提供，如 Yahoo 的 My Yahoo、Lycos 的 My Start Page； Google 的 My Preference 中可对检索用语种、网站语种进行设置，还可将检索范围限制在商业网站、教育网站、政府网站等域名中；Altavista 的个性化定制选项覆盖 9 个方面:描述语、URL、最近更新日期、网页大小、网站语言、翻译、该站点的更多页面、相关页面、公司情况；Northern light 的特色之一便是除了可对流行信息进行检索外，还有一个经过人工筛选、分析、标引的专门资源，并提供专门资源中的文献传递服务。一些中文搜索引擎也开始推出“ 跟踪式 ”信息检索服务或提供用户定制功能。目前支持个性化信息服务所需的支撑技术已经基本成熟，如 Web 数据库技术、数据推送技术、网页动态生成技术和智能代理技术。可以预见，将来网络的“ 个性化 ”功能将得到进一步加强。用户可以预先选择自己的信息源，向自己感兴趣的、值得自己信赖的信息源提问，索取特定类型的信息，用户还可以在一定程度上改变检索结果显示的格式。

3.4 多样化趋势将进一步增强

多样化趋势主要表现在：(1) 可以检索的信息形态多种多样，有文本、声音、图像、动画等。目前网络信息检索的主体是文本信息，基于内容的检索技术和语音识别技术的发展，将使多媒体信息的检索变得逐渐普遍。(2) 检索工具向全球化、多语种化方向发展。网络的迅速发展，使得整个世界变成了地球村，世界各地上网人数的不断增多，语言障碍越来越明显。许多搜索引擎已认识到此问题，正在研发多语种引擎以减轻语言不同所带来的障碍。Altavista 不仅提供了包括中文在内的 25 种语言检索，还提供了 5 种拉丁语系的语言与英语互译的功能。Google、Yahoo、Lycos、Excite 都在世界各地设立了分支机构，使检索服务本地化，增加服务器，分流用户，提高上网查询速度。它们的本地化服务站点，也以本地语言提供当地的主要信息。(3) 网上检索工具的服务多样化。网上检索工具已不仅仅是单纯的检索工具，正在向其他服务范畴扩展，提供天气预报、新闻报道、股票行情、机构名录、

交通旅游等服务内容，并以各种形式满足大众的信息需要。

另外，网络信息检索可以间接地服务于其他行业。例如数据挖掘技术可用于分析历史数据的变化趋势，预测未来发展方向，发现大量数据中潜在的模式规律，为投资、科研、项目评估等提供有力的依据，还可以系统地、定量地分析目前较为热门的研究发展领域及查询频繁的文献资料种类，可使图书情报部门不断调整信息资料的收集工作，以市场为导向建立一套更为科学的管理方式。

参考文献

1、向桂林.复合型 Web 信息检索系统.情报学报，2003(5)
2、王启云.如何利用搜索引擎检索网络信息.现代图书情报技术，2001(4)
3、郭少友.Web 环境下分布式信息检索模式.情报科学，2003(6)
4、丛石.三种信息检索语言的功能及其应用.图书情报知识，2003(3)
5、张喜年.网络信息检索工具的检索功能述略.图书馆理论与实践，2003(2)
6、陆承兆.试论计算机情报检索途径和技术发展趋势.图书馆论坛，2002，6(3)
7、徐莉，胡维青.WWW 信息检索系统评介.晋图学刊，2003(3)
8、张燕飞，彭燕云.基于 WWW 中文网络信息检索工具的比较研究.江西图书馆学刊，2003(1)
9、郑德权等.提高 W e b 信息检索精度的多步策略.哈尔滨商业大学学报，2003(3)