第 3 章

信息处理与信息检索

【案例】

美国施乐公司作为世界复印机行业的巨人之一，于 20 世纪 60 年代在世界首次推出办公用复印机(型号为 Xerox914)，从而改变了人们的工作方式，施乐公司也因此垄断世界复印机市场长达 10 多年之久。后来，由于随着理光、佳能等日本企业先后进入复印机市场，该行业的竞争日益激烈，但是施乐公司忽视了全球性的竞争威胁情报研究，不能及时对经营战略进行调整，最后被迫进入防御状态。到 20 世纪 80 年代初，施乐公司的复印机全球市场份额由 82%下降到 35%。这时施乐公司才开始分析日本的产品和价格，结果令他们大吃一惊——日本佳能公司竟然采取了以施乐公司的成本价销售复印机。起初与其他美国企业一样，施乐公司怀疑日本产品质量差，但事实证明并非如此；施乐公司又认为日本产品采取低价倾销策略，价格如此之低肯定赚不到钱，结果又错了。经过对日本产品深入细致的竞争情报分析对比后，施乐公司才发现竞争对手企业在产品导入市场的时间和投入的人力都只有本公司的 1/2，而且设备安装时间仅是本公司的 1/3，这就是竞争对手可以大幅度降价的关键原因。

为了夺回已失去的市场份额，施乐公司加强了对竞争对手情报的搜集、处理和分析工作，决定以公司市场调研部为基础，成立了专门的竞争情报研究部门，协调和领导整个公司的竞争情报工作。为了时刻获得情报信息，施乐公司在 3 个层次上开展了竞争情报研究。

(1) 全球性的，由施乐公司的营业部负责搜集和分析影响公司长期计划或战略计划的信息。

(2) 全国性的，由美国顾客服务部收集美国国内的竞争情报。

(3) 地区性的，充分利用公司遍布在美国的 37 个销售服务网点，要求通过各自的市场经理收集和分析所在地区的信息，并在此基础上公司建立“竞争数据库”和“顾客数据库”。

而且为了实施竞争情报分析，施乐公司还成立了竞争评估实验室，组织实施反求工程(Reverse Engineering)，专门用以剖析竞争对手产品或有竞争威胁的产品。情报专家们通过合法渠道将这些产品买来并拆开，对其进行非常细致的分析，包括每一个细节、每一个特点、每一个优点和每一个缺点，尤其是公司可能面临的专利技术和秘密技术的应用及其特点，以了解竞争对手产品降低成本、提高质量的实用方法和制造原理，尔后将分析报告传送给设计师和工程师，使他们能够了解竞争对手的产品开发动态。这些竞争策略的实施使施乐公司最终从日本佳能公司那里夺回了其应有的市场份额。

案例分析：

实践证明，由于运用竞争情报，使施乐公司面对众多的对手，特别是不断地有强大的新对手加入，能够处之不惊，从大局着眼把握竞争形势，始终保持竞争的主动性。关于竞争情报的重要性，施乐公司的副总裁 M. Vezmar 认为："竞争情报应成为企业营销活动的一部分，每一项受竞争影响的活动都需要竞争情报，而且最重要的是要确保将正确的信息在正确的时间里传递给正确的人。"

因此，在当今信息时代，信息情报的获取和处理技术已成为企业和个人生存和发展的法宝与基本技能。

本章主要讲解有关信息和信息处理的知识和技能。

本章学习导航：

1. 信息处理基础知识
2. 信息检索
3. 互联网信息处理
4. 数字图书馆

知识点与能力目标：

1. 信息技术
2. 信息处理
3. 文字处理技术
4. 信息检索技术
5. 互联网信息处理技术
6. 数字图书馆

3.1　信息处理基础

信息的表现形式是多种多样的，计算机信息处理过程的范例不胜枚举，它不只限于算术运算处理，在语言、文字、声音、图像等信息的处理方面都得到了长足的发展。

3.1.1　信息处理基础知识

人类可以通过各种方式获取信息，最直接的就是用眼睛看、用鼻子闻、用耳朵听、用舌头尝；另外还可以借助各种工具获取更多的信息，例如用望远镜可以看得更远，用显微镜可以观察微观世界……

人们身边有大量的信息，不是所有的信息都是对人们有用的，因此要对获取的信息进行处理，首先是获取信息，然后再处理信息，最后输出信息。

依照信息处理的过程，人工处理信息是人们用眼睛、耳朵、鼻子、手等感觉器官直接获取外界的各种信息，经过大脑的分析、归纳、综合、比较、判断等处理后，产生更有价值的信息，并且采用说话、写字、动作、表情等方式输出信息。

其实，在很多时候人们不仅仅依靠自己的感觉器官来处理信息，而是利用各种设备帮助进行信息的处理。再从各种输出设备把处理的结果输出（显示器、打印机等）。例如气象工

作者借助于计算机处理卫星发回的大量数据,绘制出气象云图,可以及时地报出近期的天气趋势;从计算机诞生到现在,计算机已经成为信息处理的重要工具。

使用计算机进行信息处理有如下特点。

(1) 能高速度、高质量地完成各种数据加工任务。

(2) 提供友善的使用方法和多种多样的信息输出形式。

(3) 具有庞大的信息存储容量和极快的信息存取速度。

(4) 计算机网络使得世界变"小",距离不再是限制信息传播的屏障。

(5) 计算机在辅助开发新的信息处理应用方面能提供有力的支持。

总之,用计算机进行信息处理,具有极高的处理速度,多种多样的处理功能,友善的人机界面,几乎不受限制的存储容量,方便而迅速的计算机通信,高效率的计算机辅助开发手段,所有这些都决定了计算机在信息处理中具有最重要、最核心的突出地位。

3.1.2 文字信息处理技术

文字信息处理技术是计算机获得广泛应用的关键技术,它与数据库技术、操作系统技术同为最重要的三大基础核心软件技术。

处理文字的技术有手写、刻字、雕版印刷、活字印刷、机械式打字机和计算机。

无论怎么来看,文字信息加工总是可以分出两个层面来。

(1) 在文字本位层面,文字信息是用来记载和描述的,是为思想主动者服务的一种功能体现。

(2) 在文字修饰层面,文字信息是为了记忆和表现的,是为思想受众服务的一种功能体现。

有了这样的区分,相应的技术学习就不再迷惑。就面向大众的信息技术来说,主要进行内容方面的处理,通过方法与技术,达到使用文字更好地记载和描述事物的效果。在这一基础上,突出第二层面,即加强修饰,以便更好地呈现、记忆信息,从而提高受众的接受程度。在这个问题的理解上,还可以拿写书与出版这两个行为来对比:写书,是基本的文字信息处理,强调的是基本的编辑技术;而出版却不同,是要强调装帧、版式和美工。

实际上,在第二个层面上,文字信息已经开始向多媒体信息的性质靠近了。让文字美起来、格式化起来、图形化起来,甚至可以动起来,这些处理的确是在进一步把文字这种单媒体向多媒体效果去加工转化。因为人们已经不仅是在关注文字内在的信息,还要关注文字信息呈现的作用与效果。

3.1.3 与信息处理有关的法律法规

与信息处理有关的法律法规主要有以下几部,可以自行上网搜索详细内容。

- 《中华人民共和国计算机信息系统安全保护条例》
- 《互联网信息服务管理办法》
- 《计算机信息网络国际联网安全保护管理方法》
- 《互联网安全保护技术措施规定》

3.2　信息检索

3.2.1　信息检索的定义

信息检索(Information Retrieval)有广义和狭义之分。广义的信息检索全称为“信息存储与检索”,是指将信息按一定的方式组织和存储起来,并根据用户的需要找出有关信息的过程。狭义的信息检索为“信息存储与检索”的后半部分,通常称为“信息查找”或“信息搜索”,是指从信息集合中找出用户所需要的有关信息的过程。狭义的信息检索包括 3 个方面的含义:了解用户的信息需求、信息检索的技术或方法、满足信息用户的需求。

由信息检索原理可知,信息的存储是实现信息检索的基础。这里要存储的信息不仅包括原始文档数据,还包括图片、视频和音频等,首先要将这些原始信息进行计算机语言的转换,并将其存储在数据库中,否则无法进行机器识别。待用户根据意图输入查询请求后,检索系统根据用户的查询请求在数据库中搜索与查询相关的信息,通过一定的匹配机制计算出信息的相似度大小,并按从大到小的顺序将信息转换输出。

3.2.2　信息检索的类型

1. 按存储与检索对象划分

(1) 文献检索

文献检索(Information Retrieval),是指将信息按一定的方式组织和存储起来,并根据信息用户的需要找出有关的信息过程,所以它的全称又叫信息的存储与检索(Information Storage and Retrieval),这是广义的信息检索。狭义的信息检索则仅指该过程的后半部分,即从信息集合中找出所需要的信息的过程,相当于人们通常所说的信息查询(Information Search)。

计算机信息检索是指以计算机技术为手段,通过联机等现代检索方式进行信息检索的方法。与手工检索一样,计算机信息检索应作为未来科技人员的一项基本功,这一能力的训练和培养对科技人员适应未来社会和跨世纪科研都极其重要,一个善于从电子信息系统中获取信息的科研人员,必定比不具备这一能力的人有更多的成功机会,美国报道生活新方式的期刊 POV 也将交互网络检索专家作为未来十大热门职业之一,这些情况都说明了计算机信息检索越来越重要,故值得大家对这一技术予以重视。

(2) 数据检索

数据检索(Data Retrieval)是将经过选择、整理和评价(鉴定)的数据存入某种载体中,并根据用户需要从某种数据集合中检索出能回答问题的准确数据过程或技术。按查询问题的要求,分为简单检索(即单一因素的检索)和综合检索(即综合条件检索)。数据文件组织方式不同,数据检索的技术方法亦不同。对顺序结构文件,常见方法有顺序检索、分块查找法、两分检索等。对随机结构文件,常采用直接地址法、杂凑(hash)法等。

(3) 事实检索

事实检索是情报检索的一种类型。广义的事实检索既包括数值数据的检索、算术运算、比较和数学推导,也包括非数值数据(如事实、概念、思想、知识等)的检索、比较、演绎和逻辑

推理。它要求检索系统不仅能够从数据(事实)集合中查出原来存入的数据或事实,还能够从已有的基本数据或事实中推导、演绎出新的数据或事实。例如,该系统中存储有如下事实:①李明是A校的学生;②A校的学生都学外语。如果该系统是一个事实检索系统,则它应当能回答某用户提出的"李明学外语吗?"这种问题。事实检索是情报检索中最复杂的一种。它要求系统中的数据和事实以自然语言或接近于自然语言的方式存储。不仅要存入各种数据或事实单元,还要存入各单元之间的语义关系、句法关系以及各种有关的背景知识。允许用户用自然语言提问,并能用自然语言作答。更重要的是,系统必须具有一定的逻辑推理能力和自然语言理解功能。

以上三种信息检索类型的主要区别在于数据检索和事实检索是要检索出包含在文献中的信息本身,而文献检索则检索出包含所需要信息的文献即可。

2. 按存储的载体和实现查找的技术手段为标准划分

(1) 手工检索

手工检索(Manual Retrieval)是一种传统的检索方法,即以手工翻检的方式,利用工具书(包括图书、期刊、目录卡片等)来检索信息的一种检索手段。

手工检索不需要特殊的设备,用户根据所检索的对象,利用相关的检索工具就可进行。手工检索的方法比较简单、灵活,容易掌握。但是,手工检索费时、费力,特别是进行专题检索和回溯性检索时,需要翻检大量的检索工具反复查询,花费大量的人力和时间,而且很容易造成误检和漏检。

(2) 计算机信息检索

计算机信息检索指利用计算机检索数据库的过程,优点是速度快,缺点是回溯性不好,且有时间限制。计算机检索、网络文献检索将成为信息检索的主流。网络信息检索,也即网络信息搜索,是指互联网用户在网络终端,通过特定的网络搜索工具或是通过浏览的方式,查找并获取信息的行为。信息检索的对象:①文献检索是以文献(包括题录、文摘和全文)为检索对象的检索。可分为全文检索和书目检索两种。②数据检索是以数值或数据(包括数据、图表、公式等)为对象的检索。③事实检索是以某一客观事实为检索对象,查找某一事物发生的时间、地点及过程的检索。

其中发展比较迅速的计算机检索是"网络信息检索",网络信息检索(Network Information Retrieval,NIR)一般指因特网检索,是通过网络接口软件,用户可以在一终端查询各地上网的信息资源。这一类检索系统都是基于互联网的分布式特点开发和应用的,即数据分布式存储,大量的数据可以分散存储在不同的服务器上;用户分布式检索,任何地方的终端用户都可以访问存储数据;数据分布式处理,任何数据都可以在网上的任何地方进行处理。

网络信息检索与联机信息检索最根本的不同在于网络信息检索是基于客户机/服务器的网络支撑环境的,客户机和服务器是同等关系,而联机检索系统的主机和用户终端是主从关系。在客户机/服务器模式下,一个服务器可以被多个客户访问,一个客户也可以访问多个服务器。因特网就是该系统的典型,网上的主机既可以作为用户的主机里的信息,又可以作为信息源被其他终端访问。

3.2.3　信息检索原因

1. 信息检索是获取知识的捷径

美国普林斯顿物理系一个年轻大学生名叫约翰·菲利普，在图书馆里借阅有关公开资料，仅用 4 个月时间，就画出一张制造原子弹的设计图。他设计的原子弹，体积小（棒球大小）、重量轻（7.5kg）、威力大（相当广岛原子弹 3/4 的威力），造价低（当时仅需 2000 美元），致使一些国家（法国、巴基斯坦等）纷纷致函美国大使馆，争相购买他的设计图。

20 世纪 70 年代，美国核专家泰勒收到一份题为“制造核弹的方法”的报告，他被报告精湛的技术设计所吸引，惊叹地说：“至今我看到的报告中，它是最详细、最全面的一份。”但使他更为惊异的是，这份报告竟出于哈佛大学经济专业的青年学生之手。

2. 信息检索是科学研究的向导

美国在实施“阿波罗”登月计划中，对“阿波罗”飞船的燃料箱进行压力实验时，发现甲醇会引起钛应力腐蚀，为此付出了数百万美元来研究解决这一问题。事后查明，早在十多年前，就有人研究出来了，方法非常简单，只需在甲醇中加入 2%的水即可，检索这篇文献的时间是 10 多分钟。在科研开发领域里，重复劳动在世界各国都不同程度地存在。据统计，美国每年由于重复研究所造成的损失，约占全年研究经费的 38%，达 20 亿美元之多。日本有关化学化工方面的研究课题与国外重复的，大学占 40%、民间占 47%、国家研究机构占 40%，平均重复率在 40%以上；中国的重复率则更高。

3. 信息检索是终身教育的基础

学校培养学生的目标是学生的智能：包括自学能力、研究能力、思维能力、表达能力和组织管理能力。UNESCO（联合国教育、科学及文化组织）提出，教育已扩大到一个人的整个一生，认为唯有全面的终身教育才能够培养完善的人，可以防止知识老化，不断更新知识，适应当代信息社会发展的需求。

3.2.4　信息检索方法

信息检索方法包括普通法、追溯法和分段法。

（1）普通法是利用书目、文摘、索引等检索工具进行文献资料查找的方法。运用这种方法的关键在于熟悉各种检索工具的性质、特点和查找过程，从不同角度查找。普通法又可分为顺检法和倒检法。顺检法是从过去到现在按时间顺序检索，费用多、效率低；倒检法是逆时间顺序从近期向远期检索，它强调近期资料，重视当前的信息，主动性强，效果较好。

（2）追溯法是利用已有文献所附的参考文献不断追踪查找的方法，在没有检索工具或检索工具不全时，此法可获得针对性很强的资料，查准率较高，查全率较差。

（3）分段法是追溯法和普通法的综合，它将两种方法分期、分段交替使用，直至查到所需资料为止。

3.2.5　信息检索系统

信息检索系统（Information Retrieval System）是指根据特定的信息需求而建立起来的一种有关信息搜集、加工、存储和检索的程序化系统，其主要目的是为人们提供信息服务。

信息检索系统的3个基本要素:人、检索工具(包括设备)和信息资料。信息检索系统的作用就是对数据系统进行有效管理和利用。

信息检索的3个主要环节如下。

(1) 信息内容分析与编码,产生信息记录及检索标识。

(2) 组织存储,将全部记录按文件、数据库等形式组成有序的信息集合。

(3) 用户提问处理和检索输出。

中文文献检索技术就是对中文文献进行存储、检索和各种管理的方法和技术。中文文献检索技术出现在1974年,20世纪80年代得到了快速增长,20世纪90年代主要研究支持复合文档的文档管理系统。中文信息检索在20世纪90年代之前都被称为情报检索,其主要研究内容有:包括布尔检索模型、向量空间模型和概率检索模型在内的信息检索数学模型;如何进行自动录入和其他操作的文献处理;进行词法分析的提问和词法处理;实现技术;对查全率和查准率研究的检索效用;标准化;扩展传统信息检索的范围等。中文信息检索主要是书目的检索,用于政府部门、信息中心等部门。

总体上,信息检索系统可分为4个部分:数据预处理、索引生成、查询处理、检索。下面分别对各个部分加以介绍。

(1) 数据预处理。目前检索系统的主要数据来源是Web,格式包括网页、Word文档、PDF文档等,这些格式的数据除了正文内容之外,还有大量的标记信息,因此从多种格式的数据中提取正文和其他所需的信息就成为数据预处理的主要任务。

(2) 索引生成。对原始数据建立索引是为了快速定位查询词所在的位置,为了达到这个目的,索引的结构非常关键。目前主流的方法是以词为单位构造倒排文档表,每个文档都由一串词组成,而用户输入的查询条件通常是若干关键词,因此如果预先记录这些词出现的位置,那么只要在索引文件中找到这些词,也就找到了包含它们的文档。

(3) 查询处理。用户输入的查询条件可以有多种形式,包括关键词、布尔表达式、自然语言形式的描述语句甚至是文本,但如果把这些输入仅当作关键词去检索,显然不能准确把握用户的真实信息需求。很多系统采用查询扩展来克服这一问题。各种语言中都会存在很多同义词,比如查"计算机"的时候,包含"电脑"的结果也应一并返回,这种情况通常会采用查词典的方法解决。但完全基于词典所能提供的信息有限,而且很多时候并不适宜简单地以同义词替换方法进行扩展,因此很多研究者还采用相关反馈、关联矩阵等方法对查询条件进行深入挖掘。

(4) 检索。最简单的检索系统只需要按照查询词之间的逻辑关系返回相应的文档就可以了,但这种做法显然不能表达结果与查询之间的深层关系。为了把最符合用户需求的结果显示在前面,还需要利用各种信息对结果进行重排序。目前有两大主流技术用于分析结果和查询的相关性:链接分析和基于内容的计算。

3.2.6 信息检索中的需求表达

在信息检索中,要获得目标值,提高检索的准确度,检索策略占据很重要的位置,而检索策略的制定取决于对信息需求的正确分析。对信息检索人员来说,信息需求往往不是其自身研究的课题,而是外在需求,因此,正确分析这种信息需求就变得十分重要。

由此可知,获取需求是一个确定和理解不同需要与限制的过程。需求获取是在问题及

其最终解决方案之间架设桥梁的第一步。需求获取可能是最困难、最关键、最易出错及最需要交流的方面。因此，当我们对信息需求进行分析时，必然要分析清楚需求间的逻辑关系，对所获取的需求进行优先级的排列，就能探索出描述这些需求的多种解决方案，否则，将费时费力。

情报检索不是一蹴而就的，应该是采用“总体规划，分步实施”的原则推进。在需求获取的过程中，可能会发现各层次信息化需求之间的逻辑关系，包括因果关系、依赖关系、主次关系等。只有当人们确立了信息化需求的逻辑结构，需求分析结果才能真正地为情报检索提供依据。

3.2.7　信息检索技术

计算机检索的基本检索技术有如下几种。

1. 布尔检索

利用布尔逻辑运算符进行检索词或代码的逻辑组配，是现代信息检索系统中最常用的一种方法。常用的布尔逻辑运算符有 3 种，分别是逻辑或 OR、逻辑与 AND、逻辑非 NOT。用这些逻辑运算符将检索词组配构成检索提问式，计算机将根据提问式与系统中的记录进行匹配，当两者相符时则命中，并自动输出该文献记录。

下面以“计算机”和“文献检索”两个词来解释三种逻辑运算符的含义。

"计算机"AND"文献检索"，表示查找文献内容中既含有“计算机”又含有“文献检索”词的文献。

"计算机"OR"文献检索"，表示查找文献内容中含有“计算机”或含有“文献检索”以及两词都包含的文献。

"计算机"NOT"文献检索"，表示查找文献内容中含有“计算机”而不含有“文献检索”的那部分文献。

检索中逻辑运算符使用是最频繁的，对逻辑运算符使用的技巧决定检索结果的满意程度。用布尔逻辑表达检索要求，除要掌握检索课题的相关因素外，还应在布尔运算符对检索结果的影响方面引起注意。另外，对同一个布尔逻辑提问式来说，不同的运算次序会有不同的检索结果。布尔运算符使用正确但不能达到应有检索效果的事情是很多的。

2. 截词检索

截词检索就是用截断的词的一个局部进行的检索，并认为凡满足这个词局部中的所有字符(串)的文献，都为命中的文献。按截断的位置来分，截词可有后截断、前截断、中截断 3 种类型。

3. 原文检索

“原文”是指数据库中的原始记录，原文检索即以原始记录中的检索词与检索词间特定位置关系为对象的运算。原文检索可以说是一种不依赖叙词表而直接使用自由词的检索方法。

原文检索可以弥补布尔逻辑检索、截词方法检索的一些不足。运用原文检索方法，可以增强选词的灵活性，部分地解决布尔检索不能解决的问题，从而提高文献检索的水平和筛选能力。但是，原文检索的能力是有限的。从逻辑形式上看，它仅是更高级的布尔系统，因此

存在着布尔逻辑本身的缺陷。

3.3 网络信息检索

3.3.1 网络信息检索的定义

网络信息检索,简单地说,就是网络环境下的信息检索。网络信息检索在信息资源组织和管理、系统工具的开发和设计思想、检索方法与策略等方面同传统的信息检索理论与方法体系有着密切的联系。但是,网络信息检索与传统信息环境下的检索又有很大的不同。网络信息检索是传统信息检索理论与方法体系的扩展与革新。这种检索方式可同时使用网上多个主机,甚至所有主机的某种资源而并不需要用户预先知道它们的具体地址,极大地拓宽了信息检索的空间和信息量。

3.3.2 网络信息检索工具的类型

网络信息检索工具是指在因特网上提供信息检索服务的计算机系统,其检索对象是存在于因特网信息空间中各种类型的网络信息资源。按检索资源的类型,可分为两大类:Web资源检索工具和非Web资源检索工具。

1. Web资源检索工具

(1) 搜索引擎。搜索引擎使用自动索引软件来发现、收集并标引网页,建立数据库;以Web形式提供给用户一个检索界面,供用户输入检索关键词、词组或短语等检索项;代替用户在数据库中找出与提问相匹配的记录,并将返回结果按相关度排序输出。使用此类工具的检索方法被称为"关键词搜索",可以在主机查询,也可以在类目下查询。此类检索工具的优点是信息量大且新,检索快;缺点是准确性较差。百度、谷歌等都是著名的搜索引擎。

(2) 目录型检索工具。目录型检索工具是按照某种分类体系编制的一种可供检索的等级结构式目录。分类方法以学科分类为主,也有采用图书分类方法的。使用此类工具的检索方法被称为"分类搜索",这是一种"自顶向下,逐步细化"的搜索方法。此类检索工具的优点是检索质量较高,缺点是检索到的信息数量有限、新颖性不够。

2. 非Web资源检索工具

非Web资源检索工具是以FTP、Telnet、Gopher为检索对象。

(1) FTP类的检索工具。这是一种实时联机检索工具,用户首先要登录到对方的计算机,登录后可以进行文献搜索及文献传输等有关操作。使用FTP(文件传输协议)几乎可以传输任何类型的文本文件、二进制文件、图像文件、声音文件等。在这类检索工具中,Archie是最常用的。Archie是自动标题检索软件,它借助FTP来访问。用户只需告诉其要检索的文件名的有关信息便可获得文件所在的主机名和路径。与一般的检索软件不同的是,它不用主题来实现相应的检索,而只能根据文件名和目录名进行检索。它是获取免费软件和共享软件资源不可缺少的工具。

(2) Telnet类检索工具。它指的是借助远程登录(Remote Login),在网络通信协议Telnet的支持下,在远程计算机上登录,使自己的计算机暂时成为远程计算机的终端,进而可以实时访问,使用远程计算机中对外开放的资源。Hytelnet是用于Telnet信息资源检索

的工具。它以超文本形式分门别类地汇集并罗列了数量相当多的 Telnet 信息资源，在远程登录后，对方系统往往设有专门的检索工具，以方便用户查找和利用。

3.3.3　搜索引擎

1. 搜索引擎的定义

搜索引擎(Search Engines)是一个对互联网上的信息资源进行搜集整理，然后供查询的系统，它包括信息搜集、信息整理和用户查询 3 部分。

搜索引擎是一个为使用者提供信息“检索”服务的网站，它使用某些程序把互联网上的所有信息归类，以帮助人们在茫茫网海中搜寻到所需要的信息。

早期的搜索引擎是把互联网中的资源服务器的地址收集起来，由其提供的资源的类型不同而分成不同的目录，再一层层地进行分类。人们要找自己想要的信息可按它们的分类一层层进入，就能最后到达目的地，找到自己想要的信息。这其实是最原始的方式，只适用于互联网信息并不多时。随着互联网信息按几何级数增长，出现了真正意义上的搜索引擎，这些搜索引擎知道网站上每一页的开始，随后搜索互联网上的所有超链接，把代表超链接的所有词汇放入一个数据库。这就是现在搜索引擎的原型。

随着 Yahoo 的出现，搜索引擎的发展也进入了黄金时代，相比以前其性能更加优越。现在的搜索引擎已经不只是单纯地搜索网页的信息，它们已经变得更加综合化、完美化，以搜索引擎权威 Yahoo 为例，从 1995 年 3 月由美籍华裔杨致远等人创办 Yahoo 开始，到现在，他们从一个单一的搜索引擎发展到现在有电子商务、新闻信息服务、个人免费电子信箱服务等多种网络服务，充分说明了搜索引擎的发展从单一到综合的过程。

然而由于搜索引擎的工作方式和互联网的快速发展，使其搜索的结果让人越来越不满意。例如，搜索“电脑”这个词汇，就可能有数百万页的结果。这是由于搜索引擎通过对网站的相关性来优化搜索结果，这种相关性又是由关键字在网站的位置、网站的名称、标签等公式来决定的。这就是使搜索引擎搜索结果多而杂的原因。而搜索引擎中的数据库因为互联网的发展变化也必然包含了死链接。

2. 搜索引擎的基本检索功能

搜索引擎之所以短短几年时间发展如此迅速，最重要的原因是搜索引擎为人们提供了一个前所未有的查找信息资料的便利方法，从一定意义上说改变了人们查找信息的习惯。

21 世纪是信息时代，信息的选择和分析是企业和个人决胜市场与社会的最重要手段。因此搜索引擎技术的出现为人们提供了新的机会和手段。

搜索信息技术最重要功能也是最基本功能就是搜索信息的及时性、有效性和针对性。

(1) 及时性。能否迅速及时查找到尽可能多的信息是搜索技术第一位的功能。

(2) 有效性。有了信息的及时性还不够，还要有信息的有效性，无效的信息不但没用，还可能造成错误。

(3) 针对性。信息搜索除了要满足及时性和有效性外，还需要有针对性。

人们通过搜索引擎查找信息除了查找需要访问的网站外，还有一个重要的信息查找需求，就是主题内容的查找，通常是技术、经济、文化和市场方面主题内容搜寻和查找，这就是信息查找的针对性。

搜索引擎这 3 项基本功能是搜索引擎存在和发展的价值依靠与保证。