目录

第三章 网络文本的信息爆炸和搜索引擎的发明

3.1 网络文本的信息爆炸

网络文本到底有多少？我们可以通过下面的统计数据了解。截至2020年6月的统计显示，全球已建立索引的网页达到了56.2亿个[1]，如若按照全球人口来算，每人就有0.7个网页。相较而言，人类有史以来上万年间也只产生了大约1亿本书（李晓明等 2005）。互联网上到底有多少信息呢？我们难以计算，一个通行的答案是：海量信息。另据2019年互联网数据中心（IDC）的报告显示，预计到2020年全球数据总量将超过40ZB（相当于4万亿GB）[2]。而且，这一数据量还会继续增长。就像宇宙大爆炸一样，网络信息的爆炸式增长给人类带来了什么？机遇还是挑战？

初看上去，维基百科之类的知识平台已经帮助我们实现了亚历山大图书馆未竟的事业——将全世界的知识都储备起来。2001年，维基百科（Wikipedia）由吉米·威尔士（Jimmy Wales）和拉里·桑格（Larry Sanger）建立。Wiki是夏威夷语，意为“快点”，这符合互联网的节奏，它是一趟电

1 The size of the World Wide Web. https://www.worldwidewebsize.com/（2020年8月6日读取）。

2 ZB和GB都是计算机储存信息的单位。B指比特或字节（byte），它是基本单位。1ZB（Zettabyte）是十万亿亿字节，它等于2的70次方；1GB（Gigabyte）是通常所谓的“千兆”，它等于2的30次方。

子的高速列车，互联网也因此被叫作“信息高速公路”。维基百科不同于传统的《大不列颠百科全书》，这本最早最权威的百科全书是由专家撰写的，而维基百科是面向全体网民的，任何人都可以撰写和编辑词条，这就形成了一种交互验证的模式。从维基百科的发展规模和速度来看，这种新的模式有助于在短时间内以极低的成本创建综合性百科全书[1]。受其冲击，《大不列颠百科全书》在2012年3月宣布停止发行印刷版，全力发展在线版。

除了这种想把所有知识都汇总起来的百科全书式资源平台，互联网上还建立起了各种资源库，包括汇集了所有影视作品、音乐作品的多媒体库，以及专门搜集语言文本的语料库。例如，创建于1990年的互联网电影资料库（Internet Movie Database，简称IMDb），是一个关于电影演员、电影、电视节目、电视明星和电影制作的在线数据库。截至2020年6月，IMDb共收录了384,871,729项资料[2]。在这项统计信息的网页上，演员弗利西蒂·琼斯（Felicity Jones）说：“IMDb对所有想看电影和拍电影的人来说都是非常重要的资源。”[3]确实如此，我们现在不用去音像店一张一张地查找片子，只需在搜索栏中输入影片名称或明星的名字，就可以找到想要的影视作品。甚至，各种观影清单或指南也会给你推荐，如“周星驰作品大全”“李小龙经典作品”，等等。你不需要成为从事影视工作的专家，只要愿意，你可以搜寻到所有想要的片子（只要它已经被电子化并被收录到在线的数据库中）。这种富足感是十多年前的人们无法想象的，这些资源好像是无穷无尽的，唾手可得。但这也让许多人失去了淘片和收藏的乐趣。

在社交媒体崛起之后，人们随时随地都可以发表个人的看法。这些碎片化的信息无处不在。只要拿起手机，那些消息提醒就会去催促你跟进朋

1 参考自坎贝尔–凯利等（2014/2020）。

2 IMDb Statistics. https://www.imdb.com/pressroom/stats/?pf_rd_m=A2FGELUUNOQJNL&pf_rd_p=8baa8ab7-3e5f-42e8-93ad-002d341048c2&pf_rd_r=RRWV44M1S8G0AV2H79EG&pf_rd_s=right-2&pf_rd_t=60601&pf_rd_i=pressroom&ref_=fea_pr_pr_lk8（2020年8月6日读取）。

3 原文是“IMDb is such an invaluable resource for everyone who goes to see films and everyone who works in films.”。

友们的最新情况。还有各种新闻平台的推送，世界各地的大事小情都为你呈上。你绝不会孤陋寡闻，而是知晓天下事、身边事。有人将这些信息称为“洪流”（flood），这些过量的、碎片化的信息会淹没有趣的灵魂，人们没有时间去放空，而是被不断高涨的洪流裹挟着。“信息过载”“信息过量”“信息焦虑”“信息疲劳”这些词语也随之而来，体现出人们真实的感受。戴维·福斯特·华莱士（David Foster Wallace）将现代人的这种困境称为“全噪声”（total noise）。想象一下香农的噪声信道中，只有噪声，没有信息——有效的信息。诗人托马斯·斯特恩斯·艾略特（Thomas Stearns Eliot）在《磐石》（*The Rock*）中提出的一连串疑问也适用于这里，“我们在生活中丢失的生命何在？ /我们在知识中丢失的智慧何在？ /我们在信息中丢失的知识何在？”[1]。

跟信息的爆炸式增长相对应的是，人类的大脑容量在上万年间并没有大的变化。受限于心智的有限性，人们对信息的获取，尤其是有效信息的获取更为困难了。对于海量数据而言，获取有效信息好比“大海捞针”，人们需要借助过滤器和搜索引擎。过滤器指的是编辑和评论式的推荐。在生活中，我们会参考米其林指南去选择餐厅；在网络上，我们也需要过滤器帮助我们过滤掉无用的垃圾邮件，或者去阅读推荐列表。例如，雅虎最初便是靠提供网页的索引列表起家的。同样地，搜索引擎可以帮助人们更为精准地获取信息，它就像我们熟悉的图书馆一样，对书目进行登记和归档，然后通过搜索引擎给网页建立索引。只有被编入索引的网页才能被人们找到，否则，即使它是存在的，也没有人知道。但是对于网络上的信息而言，大部分是非结构化的数据，包括文本、图像、影音等，其形式标准也多种多样。跟印刷文本的规范性相比，网络文本的形式和内容的随意性较强，这都为数据的获取和索引的建立提出了挑战。

1 摘自《磐石》中的合唱词，原文是 “Where is the life we have lost in living?/Where is the wisdom we have lost in knowledge?/Where is the knowledge we have lost in information?”，译文参考了艾略特（1969/2012：197）。

3.2 信息检索与搜索引擎

信息检索并不是互联网催生的新事物，从智人想要找到最新鲜的果子开始，人类的大脑就一直在进行着信息的搜集、筛查和选取的工作。从狭义上来看，传统的信息检索是指图书文献管理员所做的工作，针对专业领域的文献进行检索，不限于大学和科研机构的图书馆，还包括其他专业领域如新闻、法律和医疗单位图书馆或资料室。这类信息检索跟专业领域的关系更为密切。但近几十年来，个人计算机逐步普及，尤其是互联网迅速发展，产生了诸多影响。一方面，网络上涌现出大量的非结构化数据，这些个人刊发的观点形式多样、数量巨大；另一方面，人们对检索质量和响应速度的要求越来越高。这两个方面像一柄双刃剑推动着信息检索技术向前发展，例如，布尔模型、向量空间模型、网页排名等技术不断在进行更新。下面，我们首先介绍信息检索的概念与类型，在此基础上重点讲解网络搜索引擎的工作原理，继而在3.3节详细说明几个主流的搜索引擎技术。

3.2.1 信息检索的概念与类型

信息检索（information retrieve，简称IR）指在大量（通常是储存在计算机中）的数据中，查找满足信息查询需求的具有非结构性（通常是文本）的资料（通常是文件）（Manning *et al.* 2008）。信息检索的对象是非结构化数据，即那些不具有清晰的、语义显性的，以及不易于计算机处理的结构的数据，如电子邮件、聊天记录、XML或HTML网页、各类文档、数据报表、图片、图像及音视频。这些数据的格式和标准都是多种多样的，不具有完整和规则的数据结构，而且难以用二维逻辑来表示。相较而言，结构化数据，例如关系型数据库利用二维表结构来组织数据。不管在企业中还是互联网上，非结构化数据不仅占比高，且获取难度大。由此，我们需要信息检索系统来帮助用户找到所需的资料。

根据信息检索系统的应用范围，我们将其分为三类，即网络检索、个

人信息检索和企业、机构及特定领域检索。基于网络的检索要处理储存在数百万计算机上的数十亿的文本。庞大的数据量对建立文本的索引提出了更高的要求，要保证效率，还要能够处理一些特殊问题，如超链接文本的挖掘和欺诈性网页排名的辨别等。相较于网络搜索这种数据量巨大的检索系统，个人操作系统中的信息检索模块的数据量要小得多。不过，虽然个人操作系统中检索模块的数据规模较小，但是数据类型较多，而且还要求它对系统造成的负担要小。个人电脑上常用的电子邮件系统不仅提供搜索功能，还可以进行文本分类，如垃圾邮箱就是一个文本分类器。第三类就是数据量介于前两者之间的面向企业、机构和特定领域的检索，常见于企业内部文本库、医院的患者数据库，以及旅游问答系统。这些系统一般将文本集中储存在中央文件系统中，再由一些专门的系统来进行检索（Manning *et al.* 2008）。总之，这三类检索系统在数据量、处理方法、实时性、安全性、个性化、智能化要求和实现难易程度等方面具有很大的差异。尤其是，网络检索主要应用于基于互联网的开放数据库，并利用搜索引擎来生成索引和提供检索服务。

3.2.2 网络搜索引擎的工作原理[1]

信息检索系统一般要求用户输入一个查询表达式，系统回复一个包含所需信息的文档列表。Manning & Schütze（1999）将这类检索问题称为点对点的检索。为了解决这类检索问题，主要采用两种检索方法，即精确匹配模型和相关匹配模型。对于精确匹配模型来说，要求检索系统返回的结果与用户的要求精确匹配。例如，基于二元逻辑的布尔查询系统返回的结果要么相关，要么不相关。这类模型主要用于面向企业内部的文本系统，处理的是封闭的数据库。相较而言，相关匹配模型要求系统按照用户查询要求与查询文档之间的相关度返回查询结果，这更适于面向互联网的开放式数据库。这是因为，网络具有不可预测性，而且网络数据的数量级巨

1 本节内容参考了李晓明等（2005）。

大，数据类型缺乏一致性，加上参与者的背景和动机复杂（Manning *et al.* 2008），精确匹配结果难以满足不同用户的个性化需求。

早期的网络搜索主要有两种类型：一种是全文本索引的检索，另一种是将网页分类，如雅虎。前者是一种基于关键词的检索策略，运用倒排索引和排序机制来获得检索结果。后者则是利用网页的不同类型构建出范畴的层级树，这种类似于传统的检索编辑系统的形式难以满足网络文本不断增长的数据量，并且对于用户而言，其检索效率不高。因此，虽然雅虎在刚开始时很受欢迎，且建立了具有1000多个不同节点的分类体系（Manning *et al.* 2008），但是后来还是被基于关键词的搜索引擎所取代。下面，我们来详细讲解这类搜索引擎的工作原理。

搜索引擎是一个网络应用软件系统，它能够按照用户通过浏览器提交的查询词或短语，在一个可以接受的时间内返回一个和该用户查询匹配的网页信息列表。并且，该列表的每一个条目至少包含标题、网址链接和摘要这三个元素。例如，我们在百度的搜索栏中输入“诚信的故事”[1]，得到的检索结果如图3.1所示。

其中，每个条目的第一行是标题，每个标题关联一个链接的网页，标题下面是摘要，有些条目的摘要前面还有发布时间，这一般出现在新闻类网页中，最后，还有网页的出处。如图中题为“诚信小故事大全”的出处是网页地址，在出处后面还有一个链接是“百度快照”，这个链接对应的是建立索引时的网页，而这个网页有可能在用户检索时发生变化，甚至是不存在了。如果用户在点击标题的链接时找不到对应的网页，则可以在“百度快照”的链接中找到跟摘要一致的内容。

面对海量的动态信息，网络搜索引擎一般采用“网页搜集—预处理—查询服务”的三段式工作流程。首先，搜索引擎要处理的第一个问题是如何有效搜集网页，并对网页数据库进行维护。需要明确的一点是，

1 检索网址：http://www.baidu.com（2020年8月16日读取）。

图 3.1 "诚信的故事"检索结果（检索日期：2022 年 4 月 15 日）

网页并不是即时搜集的，而是要预先搜集好的。这是因为如果在用户查询的时候去抓取网页，再来分析处理，就无法满足搜索引擎对响应时间的要求。尤其是面对大量的用户查询需求，不可能来一个需求就搜索一次。所以，大规模的搜索引擎的基础是由一批预先搜集好的网页构成的网页数据库。由于网络具有动态性和复杂性，因此如何保证系统所含内容的时新性，是在搜集网页和维护数据库的过程中要重点考虑的。比如说，Google在早期有段时间是每隔28天批量搜索一次。这样的搜集策略的时新性不高，还难以避免重复性搜索。另一种增量搜索策略则是在开始时搜集一批，往后只搜集新出现的网页，或者搜集那些上次搜集后有变化的网页，再者是删除那些上次搜集后不存在了的网页。这种策略的时新性较高，但是对系统的要求比较高。具体的网页搜集过程可以比作一个爬虫程序，即将网页的抓取看作是一只蜘蛛在蜘蛛网(即网络)上爬行。由此，也将这一过程叫作"爬取"。我们将互联网上的网页集合看成是一个有向图，从一个起

始的URL集合S开始，沿着网页中的链接，按照先深、先宽或其他遍历策略，不停地从S中移除URL，下载相应的网页，继而解析出网页中的超链接URL，看是否已经被访问过，将未被访问过的那些URL加入集合S。

在搜集到的海量原始网页集合的基础上，我们需要进行第二步，对网页进行预处理，并为建立索引做好准备。预处理主要包括关键词的提取，镜像网页或转载网页的消除、链接分析和网页的重要度计算。具体来看，我们需要提取出网页源文件的内容部分所含的关键词。以中文网页为例，我们需要根据一个词典从网页文本中找出该词典所含的词语。这样，一篇网页就可以由一组词来近似地表示，如$p=\{t1, t2, \cdots\cdots, tn\}$。由于互联网网页具有动态变化的特点，因此，还需要删除那些内容重复或主题重复的网页。此外，网页上大量的HTML标记虽然也给网页的预处理造成了困难，但也相应地提供了一些线索。比如说，<title>和</title>之间的信息一般是文章的标题，这类标签性信息就比较重要。再者，在预处理阶段还会涉及对网页重要程度的计算。由于搜索引擎返回的是查询结果的列表，那么人们自然希望能够满足查询需求的条目列在前面，这些条目就是较为重要的。那么，如何来评价其重要性呢？有一个普遍的认识是“被引用多的就是重要的”，被引用则可以通过超链接来进行追踪。这样，我们在一个原始的网页集合S的基础上，经过预处理后会得到S中每一个元素的内部表示，包括原始网页文档、URL和标题、编号、关键词集合和重要程度等其他指标。

接下来要做的工作是，如何从这个集合中生成一个列表，因为在用户的搜索端看到的是列表，而不是集合。为此，查询服务主要包括三个方面，即查询方式和匹配、结果排序和文档摘要。由于网络用户复杂的背景和需求，现有的查询方式不可能满足所有人的要求。目前，主流的查询模式是用一个词或短语来表达查询需求，并希望返回的网页中含有该词或该短语，这可以通过查询它跟文档的词表是否匹配来实现。在得到了跟用户查询相关的文档集合的基础上，需要按照列表的形式来呈现。列表，就是

要按照某种评价标准，来确定集合中元素的排列顺序。早期按照词的出现频率的方法并不可靠，目前搜索引擎主要采用网页排名（PageRank，简称PR）的方法，即为每篇网页形成一个独立于查询词的重要性指标，将它跟其他相关性指标结合以形成最终的排序。最后，每个条目还包括从网页正文中生成的摘要。不同于文献生成系统，搜索引擎中摘要的生成比较简单，要求不高，只需要将查询词周围的文字摘出即可。这需要在预处理阶段分词的时候记住每个关键词在文档中出现的位置。

综上，我们可以归纳出搜索引擎的体系结构，如下图所示。其中，网页搜集、预处理和查询服务分别对应于搜集器、索引器和检索器三个部分。控制器是为了控制搜集的效率而引进的。这样，我们从网络的云端通过搜集器爬取网页，并通过控制器的协调来决定爬取的策略，形成一个网页的原始数据库。在此基础上，我们通过索引器对网页建立索引，如倒排索引，构建索引网页库。继而，设立面向用户的检索系统。这一系列操作就构成了从搜集到预处理，再到检索的全过程。

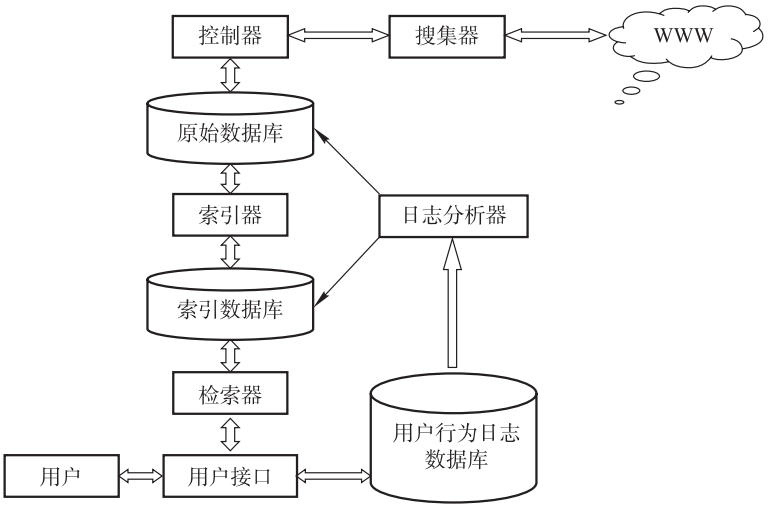

图 3.2　搜索引擎的体系结构（李晓明等 2005）

3.3 搜索引擎技术

在搜索引擎的体系结构中，索引和检索都是非常重要的环节。实际上，有两个重要的技术参数可以反映搜索引擎系统的优劣，分别是索引的建立和相关度计算。而且，不管是在网页文档端还是查询语句端，一般以词作为索引的基本单位。本节主要讲解这类基于关键词匹配的搜索引擎技术，主要包括布尔检索和倒排索引、超链接分析与网页排序。

3.3.1 布尔检索和倒排索引

布尔检索是信息检索模型中采用的最为经典的模型。其中，查询词Q由布尔表达式表示，候选查询文档由关键词的逻辑组合表达式表示，如AND(与)、OR(或)及NOT(否)这类逻辑操作符。比如说，我们想在《莎士比亚全集》中找到包含Brutus和Caesar，但不包含Calpurina的作品，我们可以逐字逐句的来检索每一部作品，但是这种方式的效率极低，尤其是数据量很大的情况下就难以适用了。这时候，我们可以采用另一种非线性的扫描方式，就是为文档建立索引(index)。就是说，对于莎士比亚全集中的每一部作品(即每一篇文档)，我们都提前记录好其中是否包含词表中的某一个词。我们将这种词项与文档的对应关系用如下的词项—文档关联矩阵表示。

表 3.1 词项—文档关联矩阵（Manning *et al.* 2008）

文档 / 词项	*Antony and Cleopatra*	*Julius Caesar*	*The Tempest*	*Hamlet*	*Othello*	*Macbeth*
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0

上表中，行对应于词项，如Antony、Brutus、Caesar等，列对应于文档，如*Antony and Cleopatra*、*Julius Caesar*、*Hamlet*等。当一个词项t在作品d中存在时，矩阵元素（t，d）的值就为1，否则为0。每一行的词项都可以看作是由0和1组成的二进制数，由此，我们就可以为其进行布尔逻辑运算。此例中，要查找的表达式是“Brutus AND Caesar NOT Calpurina”，其中“Brutus=110100”，“Caesar=110111”，而“Calpurnia=010000”。当我们将这几个词项代入前面的表达式时，可以将NOT变为AND，其后的Calpurnia的向量求反，变为“101111”。接着，建立起相应的二进制表达式“110100 AND 110111 AND 101111”，最终结果是“100100”。这表示向量中的第1个和第4个元素为1，也就是说文档1和文档4符合我们的查询条件。这两部作品就是*Antony and Cleopatra*和*Hamlet*。可见，使用词项—文档矩阵来进行布尔检索简单有效。但是，这种矩阵的缺点是数据极为稀疏，其中有很多元素为0。如果将数据量变大，可以发现用这种方法来储存大数据会浪费大量的存储空间。由此引出一种改进的方法，就是倒排索引（inverted index）。

倒排索引是对每一个词项t，记录所有包含t的文档列表，每篇文档都用一个唯一的编号（docID）来表示，如1、2、3……这样，我们将上面的词项—文档矩阵变换为下面的形式：

表 3.2　倒排索引示例（Manning *et al.* 2008）

Brutus	1	2	4	11	31	45
Caesar	1	2	4	5	6	16
Calpurnia	2	31	54	101		

上表中最左侧一列叫作词项词典，右侧的表格就是倒排记录表，表中的每个数字记录的是词项在文档中存在的信息。这个倒排记录表按照文档

的编号进行排序，例如，Brutus这个词项在文档1、文档2、文档4、文档11、文档31和文档45中存在；而Calpurnia则存在于文档2、文档31、文档54和文档101中。

如果我们要查询“Brutus AND Calpurnia”，就可以使用上表中的倒排索引进行如下的操作：1）在词典中定位Brutus；2）返回其倒排记录：“1、2、4、11、31、45”；3）在词典中定位Calpurnia；4）返回其倒排记录：“2、31、54、101”；5）对两个倒排记录求交集，结果为“2、31”。

这里，逻辑操作符AND是取交集，如果是OR就是取并集，NOT就是从第一个倒排记录中排除第二个倒排记录的内容。对于AND的交集操作，可以参考下图中的算法：

```
INTERSECT (p1, p2)
 1  answer ← < >
 2  while p1 ≠ NIL and p2 ≠ NIL
 3  do if doc ID (p1) = doc ID (p2)
 4        then ADD (answer, doc ID (p1))
 5                p1 ← next (p1)
 6                p2 ← next (p2)
 7        else if doc ID (p1) < doc ID (p2)
 8                then p1 ← next (p1)
 9                else p2 ← next (p2)
10  return answer
```

图 3.3　两个倒排记录的合并算法（Manning *et al.* 2008）

根据上述算法，我们比较docID（p1）和docID（p2）这两个倒排记录中文档的docID的值，如果两个ID的值相同，则输出这个ID，然后比较下一个文档的ID；如果二者不同，那么更大的ID不变，再继续跟更小的ID的下一个ID去作比较。显而易见，这种检索方法比遍历式的检索方法具有更高的效率。而且，我们还可以根据词项的文档频率按照从小到大的顺序来排序，以对查询进行优化。

但是，布尔模型只关注查询词是否存在于文档中，并没有告诉用户哪些文档与之相关，哪些不相关。而对于搜索引擎来说，最重要的两个评价指标就是正确率和召回率[1]。对于布尔模型来说，如果想检索一篇题为《语言是一个复杂适应系统》的文章，那么所有的空格都识别为AND，这样只能返回跟标题完全匹配的那篇文档，但是没有其他任何的信息。也就是说，正确率很高，但是召回率很低。为此，信息检索领域还提出了其他模型，如向量空间模型、概率模型和语言模型等。

3.3.2 超链接分析与网页排序

在面向网络的搜索引擎中，除了建立索引的模型起到了重要的作用，对网页本身的可靠性的分析也非常重要。互联网上充斥了大量重复的、低质量的网页。很多网页的内容无法反映出它的性质，比如IBM公司的主页中没有包含词项computer（Manning *et al.* 2008）。这就要求我们从网页内容之外寻找办法来获取高质量的网页。受到文献计量学（bibliometrics）的启发[2]，也可以将超链接看作是一个网页对另一个网页的权威度的认可。当然，超链接的数量并不是越多越好，因为会有一些人为的作弊链接。

我们在下面重点介绍一种基于链接结构的评分和排序方法，即谷歌的创办者拉里·佩奇（Larry Page）和谢尔盖·布林（Sergey Brin）提出的网页排名算法。他们利用网络爬虫采集反向链接的数据，并根据按照重要性排名的反向链接开发出网页排名算法。该算法的核心是，提供链接的网站越重要，对链接所指向网站的排名影响就越大（坎贝尔–凯利等2014/2020）。这类似于社会网络中的名人效应。互联网可以看作是由超链接构成的有向图，每个网页是图的顶点，每个超链接表示一个有向边。

1 根据Manning *et al.*（2008），正确率（也译作“准确率”）指返回的结果中真正跟信息需求相关的文档所占的百分比。召回率指所有和信息需求真正相关的文档中被检索系统返回的百分比。

2 文献计量学通过分析文献之间的引用模式来量化分析学术论文的影响力。

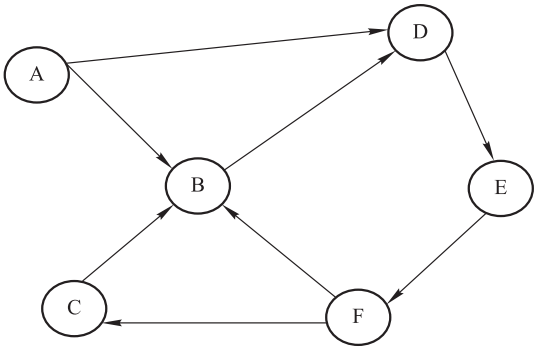

图 3.4　一个小型网络图的例子（Manning *et al.* 2008）

图3.4是一个有向图，它包括6个节点和8条边。每个节点表示一个网页，用字母A到F来标记。箭头表示链接的方向，箭头指向的终点是起点所引用的网页。网页A引用了网页B和D，而网页B引用了网页D。指向某个网页的链接称为入链，而从某个网页指出去的链接称为出链。简单来说，网页排名算法根据网页节点的入链集合的加权影响力之和来计算一个网页的影响力。图3.4中，网页B的入链数量最多，它分别被网页A、网页C和网页F引用，其PageRank的值（简称PR值）为：PR（B）=PR（A）+PR（C）+PR（F）。不过，网页A还指向了网页D，网页F还指向了网页C。如果说每个网页只有一次投票机会，那么指向两个节点的网页只能算作半票，其PR值应为PR（B）=PR（A）/2+PR（C）+PR（F）/2。也就是说，还要根据出链总数计算出一个页面的PR值，用L（X）表示X页面的出链总数，那么可以将PR值的计算统一为PR（B）=PR（A）/L（A）+PR（C）/L（C）+PR（F）/L（F）。由此，可以将公式归纳为[1]：

$$PR_i=\sum_{j\in Bi}\frac{PR_j}{L_j}$$

1　详见https://blog. csdn.net/weixin_43378396/article/details/90322422（2020年8月11日读取）。

其中，PR_i表示网页i的PageRank值，PR_j表示网页j的PageRank值，L_j表示网页j的出链数，B_i表示链接到网页i的网页集合。

上面的例子是理想的情况，实际的网络访问是随机的，也就是说，在随机过程中访问越频繁的网页也越重要。但是，网页排名算法仅利用网页的超链接，无法判断网页内容的相关性，这就可能会导致主题漂移的问题。再有，由于新网页的入链相对较小，即使内容上质量较高，也需要很长的时间来获得较高的PR值。为此，针对PageRank算法也有许多改进的算法。总之，该算法为网络搜索引擎奠定了基础，也成为谷歌——目前最成功的互联网公司之一——的基石。

需要说明的是，谷歌提供的搜索结果不仅包括由搜索引擎技术产生的相关链接，还包括商业赞助的链接。企业可以购买与特定关键字相关的赞助链接，当用户点击这些链接时，谷歌就会获得一小笔收入(坎贝尔–凯利等2014/2020)。虽然谷歌公司秉承“不作恶”的理念，但是这种广告模式也隐藏了风险。有些赞助链接提供了虚假信息，从而误导了用户，甚至造成了惨剧。2011年，戴维·惠特克(David Whitaker)事件[1]曝光了谷歌为假药投放广告，而后谷歌被美国联邦政府处以5亿美元的罚款。人们希望这些互联网企业在引导着技术革命前进的同时，更应该担负起相应的社会责任。

人们在使用语言的过程中不断地思考语言的结构、功能、历史演变以及语言对于人类生存与进化的影响，于是逐渐形成了一门研究语言的科学——语言学。同时，人们在改进语言通信工具，发明语言信息检索技术与系统的过程中，又自觉地运用到语言单位(词、短语)、词类(虚词、停用词)等基础的、初级的语言学知识。可以想象，为了进一步提升信息检索的效率与精度，必然要尝试利用语义(同义表达)等更加精细的知识，从而开启语义搜索的研究与系统开发。

1 Jake Pearson，2013，How a career con man led a federal sting that cost Google $ 500 million，https://www.wired.com/2013/05/google-pharma-whitaker-sting/(2022年4月5日获取)。

3.4 基于关键词匹配方法的局限性

前面介绍了搜索引擎的工作原理和体系结构，讲解了背后的检索模型和返回结果的排序技术。下面，我们结合例子来梳理搜索引擎的工作流程，并借此说明基于关键词匹配的搜索引擎技术的局限和不足。

搜索引擎的工作流程包括“搜集网页”“预处理”和“查询服务”这三个部分。基于关键词匹配的搜索系统建立在对词的处理、分析和统计的基础上，这是一种“词袋”（bag of words）技术。在一个搜索引擎的系统中，搜集到的原始网页是一个集合S，预处理后得到的是S的一个子集，其中的元素具有某种内部表示，包括原始网页文档、URL和标题、编号和关键词集合等。这个关键词的集合和文档的编号构成倒排文件结构，这样在系统得到一个关键词的输入后，就可以迅速给出相关文档编号的集合输出。那么，系统如何实现查询输入与文档集合的匹配呢？从用户的角度来看，一般需要提供一个查询短语，例如“诚信的故事”。对于中文来说，由于词与词之间没有界限，系统需要对原始查询进行切分，即分词。例如上例的分词结果为“诚信、的、故事”。分词后，还要删除那些没有查询意义或者在每篇文档中都会出现的词，如“的”“地”“得”等，这些词也叫作停用词。在上例中，系统需要删除“的”。在这之后，就得到了一个参与匹配的查询词表，即q={诚信，故事}。接着，系统要找到这个查询词表中的每一个元素，并对应到倒排文件中的倒排表，即文档编号的集合，这些倒排表集合的交集就是对应的查询结果的文档集合。比如，“诚信”的倒排表集合是{网页文档1、网页文档3、网页文档5，……}，“故事”的倒排表集合是{ 网页文档3、网页文档5、网页文档7，……}，取交集后，结果为{网页文档3、网页文档5}。这样，就得到了既包含“诚信”，也包含“故事”的网页，即实现了查询需求和网页文档的匹配。

如前面图3.1所示，排名前几位的返回结果跟查询词语比较接近，但

是并不完全相同，没有出现停用词“的”。如果综合分析查询结果，可以归纳出如下三种类型（袁毓林 2008a）：

题为《诚信的故事》《一个关于诚信的故事》等文章或帖子，因为这些文章（或帖子）的标题跟查询串完全一样，或包含着跟查询串一样的短语；

题为《诚信故事》《诚信小故事》《共铸诚信小故事》《中国古代诚信小故事汇编》等帖子或网页，因为这些帖子（或网页）的标题跟查询串比较接近，或包含着跟查询串比较近似的短语；

题为《天大举办“诚信故事”晚会》《诚信通故事：悄悄地“成交”》等跟查询目标无关的帖子，因为这些帖子的标题中包含着跟查询串比较近似的短语。

可见，这种基于词（关键词）匹配的搜索引擎技术在检索词语与目标文档中的词语一致时，检索结果比较理想。但是，如果二者不一致，如同义不同形或者同形不同义时，都会出现问题。

第一，同义不同形。不管是查询词语还是网页文档中的文本内容都是以自然语言为载体的。自然文本中有大量的同义词，如“奥运标志”“奥运五环”“奥林匹克标志”“奥运会会徽”，这些词都指由五个不同颜色的圆环构成的奥运会的标志。如果按照上面严格的基于关键词的搜索引擎技术，就找不到包括“五环”或“会徽”的文本，而实际上文本的内容是相同的。

第二，同形不同义。自然语言还有大量的同形词，如“苹果”，既可以指一种水果，也可以指苹果公司。同形词一般不会导致人与人之间的交际障碍，这是因为文本中有大量的上下文信息，并且同一个词的不同意义会在不同的语境中使用。例如“苹果好吃”中的苹果自然指的是第一种意义，但是，如果只根据词形是无法进行区分的。虽然“苹果”表示一家科技公司或手机的意义跟水果相比是新产生的意义，但是在搜索引擎中，返回的结果中只有很少的条目是表示水果义的。再如，“鲁迅的文章”这种