网上搜索引擎的几个理论问题

李广健 张 蕾

(北京师范大学信息技术与管理系，北京 100875)

摘 要 本文概述了搜索引擎及其作用，搜索引擎的类型、搜索引擎的信息检索模型、建立搜索引擎的关键技术、搜索引擎的评价标准等理论问题并分析了目前搜索引擎的局限性。

关键词 搜索引擎 理论研究

Some Theoretical Issues about Search Engine on Internet

Li Guangjian Zhang Lei

(Dept. of Information Technology and Management, BNU, Bejing100875)

Abstract The paper discusses in brief some theoretical issues about search engine on Internet including: search engine and its functions, type of search engine, information retrieval models of search engine, some key technologies to build search engine and the evaluation of search engine, and points out the limitations of existing search engines.

Keywords Search engine Theory study

1 搜索引擎及其作用

搜索引擎（Search Engine）是一种在因特网上查找信息的工具。用户在搜索引擎的各种程序中键入要查找的关键词，引擎就会在自己的数据库中找出与该词相匹配的 URL 并将结果显示给用户，用户可根据显示的结果选择并访问相关站点。

搜索引擎主要的作用是网络导航，帮助用户快速的查找所需的站点。搜索引擎的最主要功能是迅速地从网上纷繁复杂的信息中筛选出符合用户需求的信息。据统计，网络上 90%的用户是通过搜索引擎来获得自己所需信息的。

搜索引擎还能为用户提供多种其他服务，如广告、免费的电子邮件、聊天室、地图等等。一般的搜索引擎主要靠广告来维持自身的生存和发展，甚至通过广告营利。

2 搜索引擎的类型

2.1 根据组织信息方式分类

(1) 目录式分类搜索引擎（网站级）。目录式分类搜索引擎（Directory）将信息系统地加以归类，按传统的信息分类方式来组织信息。用户按类查找信息。这种搜索引擎特别适合那些希望了解某一方面或范围内信息但又没有明确搜索目的的用户使用。最具代表性的目录式分类搜索引擎是 Yahoo!。目录式搜索引擎的特点是查准率高，但其查全率低，搜索范围较小。

(2) 全文搜索引擎（网页级）。全文检索（Full—Text Search）搜索引擎是指能够对各网站的每个网页中的每个词进行搜索的引擎。最典型的全文搜索引擎是 Digital 公司的 AltaVista。全文搜索引擎的特点是查全率高，查准率低，搜索范围较大，提供的信息多而全，缺乏清晰的层次结构，

收稿日期：1999 年 4 月 6 日

作者简介：李广健，男，副教授。

查询结果中重复链接较多。

(3) 分类全文搜索引擎。分类全文搜索引擎是针对全文搜索引擎和分类搜索引擎的缺点而设计的，通常是在分类的基础上再进一步进行全文检索。现在大多数的搜索引擎都朝这个方向发展。

(4) 智能搜索引擎。这种搜索引擎具备符合用户实际需要的知识库，搜索时，引擎根据已有的知识库来理解检索词的意义并以此产生联想，从而找出相关的网站或网页。同时，智能搜索引擎还具有一定的推理能力，它能根据知识库的知识，运用人工智能方法进行推理。这样就大大提高了查全率和查准率。

目前比较成功的智能搜索引擎有 FSA、Eloise 和 FAQFinder。FSA 和 Eloise 专门用于搜索美国证券交易委员会的 Edgar 商业数据库。这两个系统中均内嵌了特定领域中的商业知识，并使用推新——证明式的自然语言理解技术。芝加哥大学人工智能实验室开发的 FAQFinder，则是一个具有回答式界面的智能搜索引擎。它在获知用户问题后，查询 FAQ 文件，然后给出适当的结果。

2.2 根据搜索范围分类

(1) 独立搜索引擎。这种搜索引擎建有自己的数据库，搜索时通常只检索自己的数据库，并根据数据库的内容反馈出相应的查询信息或链接站点。目前常见的搜索引擎如 yahho!、1yocs、infoseek、AltaVista 等均属独立搜索引擎。

(2) 集搜索引擎。这种搜索引擎是一种调用其他独立搜索引擎的引擎。搜索时，它用用户的查询词同时去查询若干其它搜索引擎，作出相关度排序后，将查询结果显示给用户。用户利用这种引擎能够获得更多、更全面的网址。但其缺点是查询时间长。

集搜索引擎又可分为两类：串行处理引擎和并行处理引擎。所谓并行处理就是同时将查询词传送给几个独立引擎并进行搜索，而串行则是依次将查询词传送给几个独立引擎并进行搜索。集搜索引擎主要有 Metasearch、Digisearch、Fusion、Cyber411、Metacrawler、SavvySearch、Profusion、Mamma、Ask Jeeves、Highway61、Dogpile 等。

3 搜索引擎的信息检索模型

搜索引擎所使用的信息检索模型主要有布尔逻辑模型、模糊逻辑模型、向量空间模型以及概率模型等。

(1) 布尔逻辑模型。布尔型信息检索是最简单的信息检索模型，用户利用布尔逻辑关系构造查询并提交，搜索引擎根据事先建立的倒排文件确定查询结果。标准布尔逻辑模型为二元逻辑，并可用逻辑符“and”、“or”、“not”）来组织关键词表达式。布尔型信息检索模型的查全率高，查准率低。目前大多数搜索引擎均使用布尔逻辑检索模型，查询结果一般不进行相关性排序。

(2) 模糊逻辑模型。这种模型在查询结果处理中加入模糊逻辑运算，将所检索的数据库文档信息与用户的查询要求进行模糊逻辑比较，按照相关的优先次序排列查询结果。模糊逻辑模型可以克服布尔型信息检索模型查询中结果具有无序性的问题。例如，查询“搜索引擎”，则出现关键词“搜索引擎”多的文档将排列在较前的位置上。

(3) 向量空间模型。向量空间模型用检索项的向量空间来表示用户的查询要求和数据库文档信息。查询结果是根据向量空间的相似性而排列的。向量空间模型可方便地产生有效的查询结果，能提供相关文档的文摘，并对查询结果进行分类，为用户提供准确的信息。

(4) 概率模型。基于贝叶斯概率论原理的概率模型利用相关反馈的归纳学习方法，获取匹配函数，这是一种较复杂的检索模型。

目前，商用信息检索系统主要以布尔模糊逻辑加向量空间模型为主，辅以部分自然语言处理技术来构造自己的检索算法。

4　建立搜索引擎的关键技术

(1) 信息收集和存储技术。网上信息收集和存储一般分为人工和自动两种方式。

人工方式采用传统信息收集、分类、存储、组织和检索的方法。研究人员对网站进行调查筛选、分类、存贮。由专业人员手工建立关键字索引，再将索引信息存入计算机相应的数据库中。

自动方式通常是由网络机器人来完成的。“网络机器人”(Network Robot) 是一种自动运行的软件，其功能是搜索因特网上的网站或网页。这种软件定期在因特网上漫游，通过网页间链接顺序地搜索新的地址，当遇到新的网页时，就给该页上的某些字或全部字做上索引并把它加到搜索引擎的数据库中，由此，搜索引擎的数据库得以定期更新。

一般来说，人工方式收集信息的准确性要远优于“网络机器人”，但其收集信息的效率及全面性低于“网络机器人”。

(2) 信息预处理技术。信息预处理包括信息格式支持与转换以及信息过滤。目前，因特网上的信息发布格式多种多样，这就要求搜索引擎支持多种文件格式。从实际情况看，所有的搜索引擎都支持 HTML 格式，而对于其他文件格式的支持则不同的搜索引擎有不同的规定，最多的能支持 200 多种文件格式。一般地说，一个企业级的公用 Web 站点起码应该支持 40～60 种文件格式。同时搜索引擎还应具备信息格式转换功能，以保证不同格式的数据均能在网络流通。信息过滤也是搜索引擎的一项重要技术。在因特网中，存在有大量的无用信息，一个好的搜索引擎应当尽量减少垃圾站点的数量，这是信息过滤要着重解决的问题。

(3) 信息索引技术。信息索引就是创建文档信息的特征记录，以使用户能够快速地检索到所需信息。建立索引主要涉及到几个以下问题：①信息语词切分和语词词法分析。语词是信息表达的最小单位，由于语词切分中存在切分歧异，切分需要利用各种上下文知识。语词词法分析是指识别出各个语词的词干，以便根据词干建立信息索引。②进行词性标注及相关的自然语言处理。词性标注是指利用基于规则和统计（马尔科夫链）的数学方法对语词进行标注，基于马尔科夫链随机过程的 n 元语法统计分析方法在词性标注中能达到较高的精度。可利用多种语法规则识别出重要的短语结构。自然语言处理是指自然语言理解在信息检索中应用，可以提高信息检索的精度和相关性。③建立检索项索引。使用倒排文件的方式建立检索项索引，一般包括“检索项”、“检索项所在文件位置信息”以及“检索项权重”。④检索结果处理技术。搜索引擎的检索结果通常包含大量文件，用户不可能一一浏览。搜索引擎一 般应按与查询的相关程度对检索结果进行排列，最相关的文件通常排在最前面。搜索引擎确定相关性的方法有概率方法、位置方法、摘要方法、分类或聚类方法等。

概率方法根据关键词在文中出现的频率来判定文件的相关性。这种方法对关键词出现的次数进行统计，关键词出现的次数越多，该文件与查询的相关程度就越高。

位置方法根据关键词在文中出现的位置来判定文件的相关性。关键词在文件中出现得越早，文件的相关程度就越高。

摘要方法是指搜索引擎自动地为每个文件生成一份摘要，让用户自己判断结果的相关性，以便用户进行选择。

分类或聚类方法是指搜索引擎采用分类或聚类技术，自动把查询结果归入到不同的类别中。

5 搜索引擎的评价标准

(1) 查询速度的快慢。对于搜索引擎的海量数据库来说，检索速度是至关重要的。如果检索速度太慢，系统的实用性就会大打折扣。

(2) 查询结果的准确性。搜索引擎检索到的信息要准确，既不能“漏查”，也不能“误查”，也就是说，既要有较高的查全率，也要有较高的查准率。

(3) 系统的维护和更新。搜索引擎数据库中的信息应时常更新，以适应网络上信息的变化。

(4) 系统的安全性。搜索引擎应具有完整的容错、备份、崩溃修复机制。

6 目前搜索引擎的局限性

(1) 覆盖面有限。世界上至今没有任何一种搜索引擎可以覆盖全球的 Web 页。《科学》杂志最近一份研究报告表明，即使功能最完善的搜索引擎，也只能找到 Web 上大约三分之一的网页。

NEC 研究所对几种主要搜索引擎的 Web 网页搜索覆盖率作了统计，结果如下：

站点	覆盖率
HotBot	34%
AltaVista	28%
Northem light	20%
Excite	14%
Infoseek	10%
Lycos	3%

由此可见，查找重要的信息时，不应仅局限于单个搜索站点，而应使用多种搜索引擎对 Web 进行全面的搜索。

(2) 误查率、漏查率高。虽然利用搜索引擎可以获得大量的信息，但这些信息的准确性往往并不高。误查现象和漏查现象相当普遍。要解决误查和漏查问题，最根本的途径是让搜索引擎具备认知能力和推理能力。目前人工智能搜索引擎还处于研究开发阶段。

(3) 提供的信息滞后。网络上的信息总是处于不断更新、变化的状态中，每年都有数以百万计的网页添加到网络中来，同时又有许多站点变更位置或者消亡。这样，搜索引擎的检索结果中就不可避免地存在信息滞后的问题。

(4) 搜索速度不理想。搜索引擎一般对海量数据库进行搜索，检索结果中往往又带简要说明，从而导致搜索速度不理想。目前，为了提高效率，人们开始倾向于开发较小的专用搜索引擎，通过集中地执行特定任务，专用的搜索引擎在其运行领域中会表现出更大的灵活性。

参考文献

1 www.searchenginewatch.com
2 www.yahoo.com
3 www.excite.com
4 www.altavista.digital.com
5 www.sohoo.com.cn
6 gbchinese.yahoo.com

(责任编辑：赵立军)