# <<网络信息采集与利用>>

## 图书基本信息

书名：<<网络信息采集与利用>>

13位ISBN编号：9787300119205

10位ISBN编号：7300119204

出版时间：2010-6

出版时间：贾朝辉 中国人民大学出版社 (2010-06出版)

作者：贾朝辉

页数：157

## 前言

随着科技的发展，信息的总量在迅速增长，网络信息采集方式也在进行着革命性的发展，对信息质量的要求不断提高。
根据第25次《中国互联网络发展状况统计报告》，截至2009年年底，中国网民数量已经达到3.8亿，互联网普及率稳步上升，这既给网络信息的采集与利用提出了更高的要求，也提供了现实基础。
本书从教学实践出发，理论和实践相结合，系统地阐述了与网络信息采集与利用的相关内容。
第一章为网络信息资源概论；第二章介绍了网络信息处理方式及关键技术；第三章介绍了搜索引擎及其使用；第四章介绍了其他网络信息资源及其使用；第五章介绍了联机检索技术及其应用；第六章介绍了网络学术数据库信息采集；第七章介绍了非万维网网络信息的采集；第八章介绍了网络信息编辑；第九章介绍了网络竞争情报采集与分析。
本书在编写过程中得到许多同行和北京第二外国语学院图书馆的大力支持，特别是中国人民大学出版社的大力支持，也参阅了大量的相关著作和网站，在此表示衷心的感谢！
本书在编写过程中，注重内容更新，紧跟现代检索技术的发展，然而作者能力、知识有限，错误、疏漏之处在所难免，请读者予以批评指正。

# <<网络信息采集与利用>>

## 内容概要

《网络信息采集与利用》从教学实践出发，理论和实践相结合，系统地阐述了与网络信息采集与利用的相关内容。
第一章为网络信息资源概论；第二章介绍了网络信息处理方式及关键技术；第三章介绍了搜索引擎及其使用；第四章介绍了其他网络信息资源及其使用；第五章介绍了联机检索技术及其应用；第六章介绍了网络学术数据库信息采集；第七章介绍了非万维网网络信息的采集；第八章介绍了网络信息编辑；第九章介绍了网络竞争情报采集与分析。

# <<网络信息采集与利用>>

## 书籍目录

## 章节摘录

插图：从一个网页到另一个网页，从一个网站到另一个网站采集网页资料。
为保证采集的资料最新，还会回访已抓取过的网页。
网络机器人采集的网页，还要经过其他程序进行分析，根据一定的相关度算法进行大量的计算建立网页索引，才能添加到索引数据库中。
我们平时看到的全文搜索引擎，实际上只是一个搜索引擎系统的检索界面，当你输入关键词进行查询时，搜索引擎会从庞大的数据库中找到符合关键词的所有相关网页的索引，并按一定的排名规则呈现给我们。
不同的搜索引擎，网页索引数据库不同，排名规则也不尽相同，所以，当我们以同一关键词用不同的搜索引擎查询时，搜索结果也就不尽相同。
大型全文搜索引擎的数据库储存了互联网上几亿至几十亿的网页索引，数据量高达几千G甚至几万G。
但即使最大的搜索引擎建立超过二十亿网页的索引数据库，也只占到互联网上普通网页的30%，不同搜索引擎之间的网页数据重叠率一般在70%以下。
我们使用不同搜索引擎的重要原因，就是因为它们能分别搜索到不同的内容。
而互联网上有更大量的内容，是搜索引擎无法抓取索引的，也是我们无法用搜索引擎搜索到的。
和全文搜索引擎一样，分类目录的整个工作过程也同样分为收集信息、分析信息和查询信息三部分，只不过分类目录的收集、分析信息两部分主要依靠人工完成。
分类目录一般都有专门的编辑人员，负责收集网站的信息。
随着收录站点的增多，现在一般都是由站点管理者递交自己的网站信息给分类目录的编辑，然后由编辑人员审核递交的信息，以决定是否收录该站点。
如果该站点审核通过，分类目录的编辑人员还需要分析该站点的内容，并将该站点放在相应的类别和目录中，所有这些收录的站点同样被存放在一个“索引数据库”中。
用户在查询信息时，可以选择按照关键词搜索，也可按分类目录逐层查找。
如以关键词搜索，返回的结果跟全文搜索引簿一样，也是根据信息关联程度排列网站。
需要注意的是，分类目录的关键词查询只能在网站的名称、网址、简介等内容中进行，它的查询结果也只是被收录网站首页的URL地址，而不是具体的页面。
分类目录就像一个电话号码簿一样，按照各个网站的性质，把其网址分门别类排在一起，大类下面套着小类，一直到各个网站的详细地址，一般还会提供各个网站的内容简介，用户不使用关键词也可进行查询，只要找到相关目录，就完全可以找到相关的网站（注意：是相关的网站，而不是这个网站上某个网页的内容，某一目录中网站的排名一般是按照标题字母的先后顺序或者收录的时间顺序决定的）。

## 编辑推荐

《网络信息采集与利用》：21世纪高职高专规划教材·新闻传播系列

# <<网络信息采集与利用>>

## 版权说明