网络信息分类

—— 原理与应用

施国良◎著

科学出版社
www.sciencep.com

网络信息分类

——原理与应用

施国良　著

科学出版社
北京

内 容 简 介

随着网络信息的迅猛发展，庞大的网络信息资源和人们特定的信息需求之间形成了巨大的矛盾。在此背景下，本书专门讨论了分面分类法在网络信息组织中应用的理论与实践问题。首先，本书分析了网络信息组织面临的问题；其次，专门论述了分面分类法的原理、特征和独特的性能及其用于网络信息组织的长处、必要性和可能性；再次，详细讨论了分面分类法应用于网络信息组织的基本技术特点、过程和方法，并结合实例说明其可操作性，使读者既对网络信息组织的特殊性有一个完整的了解，又对分面分类法的具体应用有了感性的认识；最后，将理论、技术与实践相结合，用实验的方法将分面分类法应用于网络信息组织，并结合一个案例做了具体的说明。本书集原理和应用于一体，语言深入浅出，通俗易懂，并配有必要的图表，具有较强的可读性。

本书适合作为高等院校图书情报类和信息管理类专业各层次学生的教学参考书和补充读物，也可作为各类信息资源管理部门（包括政府部门和科研机构）工作人员、众多网络公司工作人员以及广大计算机与网络爱好者的参考书。

图书在版编目(CIP)数据

网络信息分类：原理与应用/施国良著. —北京：科学出版社，2008
ISBN 978-7-03-023763-7

Ⅰ. 网… Ⅱ. 施… Ⅲ. 分面分类法-应用-计算机网络-信息管理
Ⅳ. G254.11 TP393

中国版本图书馆 CIP 数据核字（2008）第 206499 号

责任编辑：林 建 / 责任校对：陈玉凤
责任印制：张克忠 / 封面设计：耕者设计工作室

科 学 出 版 社出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

北京市文林印务有限公司印刷

科学出版社发行 各地新华书店经销

*

2008 年 12 月第 一 版 开本：B5（720×1000）
2008 年 12 月第一次印刷 印张：10
印数：1—2 000 字数：220 000

定价：26.00 元

（如有印装质量问题，我社负责调换〈文林〉）

前　言

随着网络信息的迅猛发展，庞大的网络信息资源和人们特定的信息需求之间形成了巨大的矛盾。在此背景下，本书专门讨论了分面分类法在网络信息组织中应用的理论与实践问题。信息量的迅速增长与人们对信息的特定需求是人们学习、工作和生活中的一对矛盾。只有研究网络信息资源的现状与特征，找到网络信息资源组织的规律，才能从根本上解决这一矛盾，为人们方便、快捷地利用网络信息提供一个理想的途径。

正是在这样的背景下，本书首先从搜索引擎的不足之处出发，对分类法、分面分类法、网络信息组织、数据库等相关领域进行文献调查；其次运用实验方法对实物、服务和网络文献三个类型的网络信息进行分面分析；再次选择其中的实物类型进行概念设计、逻辑设计和物理设计；最后建立一个简单的分面分类网络应用模型，并将这个模型初步应用于上海宝山钢铁股份有限公司（以下简称宝钢公司），旨在建立一个基于工艺和流程的企业知识组织系统。这个系统的底层正是分面数据库。设计系统模型的参数为：开源服务器 Apache 2.0.50，开源数据库 5.0，服务器端脚本为 PHP 5.2.4，浏览器端使用 AJAX 技术实现无刷新更新数据功能，从而基本上适应了分面分类法的应用原理。

本书的观点是：搜索引擎的局限性决定了仅仅依靠特性检索方式无法解决用户个性化的信息需求问题；枚举式分类法用于指导实物排架的功能特性决定了其无法根本解决网络信息激增所带来的网络信息组织问题；分面分类法并不能简单地等同于阮冈纳赞的冒号分类法，网络上应用分面分类法必须遵循严谨的分面分析过程；网络既为分面分类法的复苏提供了广阔的应用舞台，也为分面分类法再次兴起提供了强大的技术支持。

本书的结构为：第 1 章，主要从网络信息资源的现状出发，分析目前网络信息组织与检索存在的问题，从而引出全书的研究问题；第 2 章，主要对目前网络信息组织尤其是网络信息分类的理论与实践进行评述，让读者对与本书有关的研究有一个简要的了解；第 3 章，主要在理论上探讨用分面分类法对网络信息进行组织的必要性；第 4 章，主要在理论上探讨用分面分类法对网络信息进行组织的可能性；第 5～7 章，主要通过实验的方法将分面分类法具体应用于网络信息组织；第 8 章，作为案例，对宝钢公司做了基于业务和流程的分面分类系统模型的介绍；第 9 章，主要讨论分面分类法网络应用的一些拓展问题；第 10 章，主要是概括全书的结论，并粗略地勾勒了未来的一些研究方向，主要起到抛砖引玉的

作用。

本书的特色包括：从网络信息自身特性出发，系统地探讨了分面分类法应用于网络信息组织与检索的必要性与可能性，从而为分面分类法的网络应用提供了理论依据；用实验的方法证明了分面分类法在网络信息组织中应用的可能性，在Apache服务器架构上用PHP脚本和AJAX技术实现单页面无刷新浏览，从而有可能让用户在网络上真正感受到事物多向成类的属性和分类思维方式的本质；为分面分类法在网络信息组织中的应用建立了三个方面的模型——实物类、服务类和文献资源类，也为今后的拓展提供了一个初步的框架；研究了分面分类法在界面设计和数据库设计方面的一些具体应用，也为后续研究提供了广阔的空间。当然，分面分类法也有自身的局限性。即便再接近人类的思维方式，分面分类法自身仍然存在一些问题。例如，如果作为第三方分类，就存在与用户需求相匹配的问题；又如，如果一个分类对象的属性太多，在网络实现时对于优先选取和展示哪一个面这一问题就需要进行调查研究，并根据调研结果进行统计分析，才能确定所展示的面的顺序。

本书的写作得到了多方面的帮助和支持。感谢我的导师沈固朝教授对我的研究给予的指导，还要感谢南京大学信息管理系叶继元教授、郑建明教授、朱庆华教授、华薇娜教授和谭华军教授给我提出了很好的建议。在本书初稿写作过程中，得到了国家留学基金项目（网络信息组织与传播，编号：2003832061）的资助。本书出版受“211工程”三期重点学科建设项目（技术经济与管理）及江苏省国家重点学科培育建设点（技术经济与管理）资助。感谢科学出版社的编辑，正是他们的辛勤工作才使本书得以付梓。在本书写作过程中，我的爱人欧阳宁兰女士给了我莫大的理解和支持，她承担了绝大部分家务和照顾孩子的责任，家庭是我工作、研究的可靠保障。

由于作者水平有限，恳请广大读者不吝赐教，指出书中的不足与错误。最后，向本书引用的所有参考文献的作者表示感谢。

施国良

2008年10月

目　录

第1章 绪 论

本章首先从网络信息发展的历史与现状着手，分析网络信息的特征与问题及其对信息组织的挑战；其次讨论搜索引擎的现状、特征、效果、问题及其形成原因，从而引出本书所讨论的主题，即分类，尤其是分面分类法能否作为网络信息组织的最佳方式以及在网络上如何实施分面分类法；最后概括全书的思路、结构及写作意义。

1.1 网络信息资源的现状与特征

网络的出现改变了传统的信息组织、存储与检索的格局，网络信息的复杂多样性、分散凌乱性，使得信息量的增长与人们个性化的需求之间的矛盾比之传统的纸张环境更加突出，问题也更加严重。为了解决这一矛盾，首先必须了解网络信息资源的现状与特征，分析目前网络信息资源组织的不足之处，从而找到问题的根源。

网络信息资源是以数字化形式记录的以多媒体形式表达的存储在计算机上并通过网络通信方式进行传递的信息内容的集合。

因特网的前身是美国高级研究计划局网络（Advanced Research Projects Agency Network，ARPANET），20世纪70年代由美国高级研究计划署开发。它最初应用于美国国防部内部传递情报，但后来被学术团体采纳成为用于信息交换的学术网络。由于采用客户服务器构架和以太网技术，到80年代后期，因特网开始得以对普通公众开放。到90年代中期，因特网已经由6万个网络构成，全世界有5000万个电子邮件用户，使用文件搜索和检索工具的用户每年增加10倍。因特网是自发和协作的网络，遵守公共的网络协议，无人被授权管理这个网络。万维网是因特网的多媒体部分，由超文本结构和浏览导航工具构成。它最早于1989年由位于瑞士的欧洲粒子物理实验室开发，主要用于共享文献，第一个万维网商业软件于1991年得到广泛使用。

万维网主要用超文本标记语言将文献组织成不同的信息页面，每一个网页有一个唯一的地址，称为统一资源定位符（Uniform Resource Locator，URL），每一个网页可以与别的网页链接，每一个网页的内部信息也可以与别的网页链接，文献可以通过交互式界面允许用户浏览和导航，这种交互式界面被称为浏览器。浏览器和服务器之间的通信通过一种叫做超文本传输协议的公共语言进行。

超文本传输协议负责解释页面的超文本标记，从而使页面能够正常显示并能传输文件（Ellis，Vasconcelos，1999）。

需要注意的是，网络信息与传统的以纸张为载体的信息有很大的区别，这突出表现在两个方面：

（1）网络信息资源呈分散混乱的状态。网络中信息资源分散无序，经常处于更替状态。例如，有时网页内容变动和更换频繁，有时因为服务器及其他原因造成网页地址变化。网络信息的动态性让用户无法判断网上信息有多少是自己需要的，这使得网络信息资源的组织显得尤为重要。另外，因特网是一个多网络、无中心主管的分散型互联结构，网上信息资源具有分布式的特点，缺乏一个主管机构进行集中领导和组织。因此，整个网络信息资源的分布呈现一种混乱无序的状态。

（2）网络信息资源呈非对称性分布。网上信息资源在内容上没有一个完善的体系和结构，更新周期不一，信息质量得不到保证，冗余信息泛滥。另外，网上信息资源在不同学科专业领域、不同行业、不同地理位置和技术水平上呈现较大的差异性，各个网站产生和传播的信息量在数量上差别很大。

如果我们将传统文献与网络信息做一个对比，就会发现二者的差异可以用表1-1来说明。

表 1-1 传统文献与网络信息对比

项目	传统文献	网络信息
形态	实物	数字化
数量	庞大的实体	海量信息
内容	相对规范	杂乱无章
种类	图书、报刊、胶片、手稿等	文本、音频、图形、图像、视频、动画等
存取方式	线性排架	超链接
用户	数量稳定	数量激增
相对成本	较低	较高

表1-1中存取方式的差异非常关键，因为它将会影响到信息组织和检索的方式。

随着计算机和通信技术的发展和普及，因特网和万维网带来一个非常严峻的问题，即网络信息资源不仅数量巨大，而且增长迅速。可以说，在因特网上，每分钟都在产生大量的信息，这些信息的产生、发布、检索和使用背后没有规则。不仅如此，目前既没有关于全部网络信息的记录，也没有用于存储和检索这些信息的公认的分类与描述框架。网络信息具有格式全（文本、图形、音频和视频）、

类型多（电子期刊、营销信息、商业报告、图书馆目录）、主题范围广（上至天文、下至地理）的特点。曾有机构尝试计算网络信息的总量与增量。例如，美国加利福尼亚大学伯克利分校（Berkeley）信息管理与系统学院曾于2000年和2003年两度对因特网信息总量和增长的速度进行了研究，发现如下：

（1）纸张、胶片、磁介质和光学介质4类载体2002年新增的信息量为5艾[①]，92%的新增信息储存在硬盘中。

（2）2000～2002年3年中，新信息量翻了一番。

（3）2002年电子渠道（电话、广播、电视和互联网）的信息流拥有8艾，大约是硬盘信息流的3.5倍。其中，固定电话和无线电话的语音和数据占总量的98%（表1-2）。

表1-2 全世界信息量的增长（1999～2003）

存储介质	2002年（已解压）/太字节	2002年（未解压）/太字节	1999～2000年（已解压）/太字节	1999～2000年（未解压）/太字节	（已解压）变化/%
纸张	1634	327	1200	240	36
胶片	420254	7669	431690	58209	−3
磁介质	5187130	3416230	2779760	2073760	87
光学介质	103	51	81	29	28
合计	5609121	3424277	3212731	2132238	74.5

资料来源：How Much Information 2003. http://www2.sims.berkeley.edu/research/projects/how-much-info-2003. 2008-10-04

从表1-2可以看出，2000～2002年，在纸张、胶片、磁介质和光学介质的对比中，磁介质信息增长速度最快，大约是纸张信息量增长速度的两倍，是光学介质信息量增长速度的2.86倍，而胶片信息量则呈负增长。

实际上，以磁介质信息和光学介质信息为主要形式的电子信息的总量和增长速度一直是人们关注的问题，但其测量工程不仅庞大而且复杂。2003年以后，Berkeley中断了这项研究工作。令人欣慰的是，2008年6月，加利福尼亚大学圣迭戈（San Diego）分校与几家著名的公司（AT&T、Cisco、IBM、Oracle等）一起发起了一项长达三年的研究，继续对网络信息总量和增长速度进行测

① 中国物理学会曾将艾作为10^{18}的词头名称（exa）——国际单位制叫“艾可萨”，简称“艾”，符号E，中文为穰

量，预计首次研究成果将在2008年底公布①。

1.2 网络信息资源组织的问题

不难看出，目前网络信息庞大的总量与人们日益增长的个性化的需求形成了矛盾，这种矛盾的解决途径在于网络信息的组织。通过优化网络信息的组织结构，可以最大限度地为人们存取信息带来方便。这就如同某个人家里衣服很多，需要用衣橱分门别类地放置。如果将衣服胡乱地堆放，日后势必很难在短时间里找到他所要的衣服。

所谓信息组织，也称为信息资源组织，是为了方便人们检索、获取信息而将庞杂、无序的信息进行系统化和有序化处理的过程。它是根据信息资源检索的需要，以文本及各种类型的信息源为对象，通过对其内容和外形特征的分析、选择、处理，使其成为有序化集合的活动。信息组织是采用各种方法和手段依照事物之间的同一性、包容性、交叉性、排斥性等关系使信息有序化的过程。它通过揭示信息间内在的逻辑关系，对信息进行加工、整理，使之系统化，从而达到便于信息传递和交流的目的。传统的信息组织主要是指对纸质文献的组织，包括分类、标引、描述等，形成的结果包括目录、索引、文摘等。随着互联网广泛而深入的发展，网络信息在社会信息量中的比重呈日益上升的趋势。今天我们从事的信息工作很重要的一部分就是将传统信息组织上网，或将网上信息下载供人们使用，而其中的关键就是网络信息组织。由于在网络环境下，信息资源多以磁介质为载体，以电子文献（如数据库、超文本文件、图形、声频、视频等）的形式存在，信息组织的对象逐渐多样化，信息组织范围随之扩大，信息组织的形式出现了更新，信息组织的技术也随之更新。信息资源不再停留在对文献特征的描述，而是深入到知识单元，它不再仅仅依赖于机械的方式，还依赖于一些网络化的工具，因而网络信息组织与传统的信息组织有较大的区别。

标引是信息组织的正式工具。标引系统通常能反映人类依靠思维来组织和使用知识的方式。标引系统将人类思维的组织能力拓展到人工的信息存储与通信系统。标引系统将存储与通信系统转变为信息检索系统，而这种检索系统在很大程度上又是人类思维的信息检索系统的模拟。标引系统最初的功能是指明文献的内容和外形特征。所以，标引一直与知识组织的理论相联系，它们共享下述前提（Anderson，1985）：

---

① Jagoda B. Groundbreaking University of California，San Diego Research Study to Measure "How Much Information?" is in the World. http：//ucsdnews. ucsd. edu/newsrel/international/06-08InfoInTheWorld. asp. 2008-11-08

（1）人类的知识可以通过概念框架来组织，知识本质上是概念与概念间关系的结构物。人类自古以来的全部知识形成了一个庞大的知识体系，其中既包括原来的传统载体的知识，也包括新媒体中的知识。

（2）这种概念框架是类和子类的等级结构。概念框架实际上反映了人们思维的逻辑关系，这种关系通过等级结构来反映知识体系，其中既包括从属关系也包括平行关系。

（3）类的形成过程是将共享相同的属性与特征的事物归在一起，而将其他事物区别开来。一个类中至少有一个属性为同一个类中所有成员共享，而其他类中也至少有一个属性特征为本类所没有①。

按照 Anderson（1985）的理论，标引系统通常由以下组成部分构成：

（1）文献集合（如万维网的网页）；

（2）独立的能够表征文献的替代的集合（如网站或网页的标题、摘要和作者等）；

（3）独立的索引，这些索引能够将文献集合与替代品联系起来。

从总体上看，网络信息组织可以分为族性和特性两大类。族性检索是对具有某种共同性质或特征的众多事物、概念的检索，其中分类途径是族性检索的重要工具（华薇娜，2008）。分类检索既适合查询具有同一特征的多个目标，也适合主题范围广、概念宽泛的问题。虽然，有学者认为元数据和受控词汇也可以看做是网络信息组织（Schwartz，2001），但元数据实际上是能够表征文献的替代的集合，而受控词汇的功能也不过如此②。与此相关的另一个重要概念在计算机界比较热门，那就是本体（Gruber，1993）。本体实际上是对客观事物进行概念化和抽象化所得到的模型，并通过计算机处理成某个领域公认的概念集。从某种意义上看，元数据是计算机标识信息对象时所规定的一般的类型或者格式，如作者、篇名、时间、主题、载体等。而本体则是某个领域中可以用计算机来表达的概念的集合，这些概念可以用元数据来标引，也可以用其他的通用和特殊的方式来标引。从本质上看无论是元数据还是本体，都不构成信息组织的全部，因为它们并没有涉及如何组织从特定领域中抽象出来的概念。这种组织方式才真正反映了人们的思维方式，对于网络而言，就是网络用户的思维方式，并最终由界面来实现。所以网络信息组织可以归结为两类，即特性组织方式和族性组织方式，前

① 实际上，人们在日常生活中一直在潜意识里对事物进行着分类，这类活动伴随着人们认识世界和改造世界的全过程。很多学科（如语言学、数学、计算机等）都从不同角度研究分类，这一点也可以说明分类及分类研究的普遍性

② 本书从信息组织与检索的角度来考察族性检索与特性检索，前者主要指分类途径，而后者主要指搜索引擎，这主要是为了使研究对象更加明确，因此不去深究对信息组织与检索的不同方式的划分及其细微的差别，后面谈到族性检索时也主要指一般意义的分类法，目的是与搜索引擎相对应

者可以通过主题标引来实现，而后者则可通过分类标引来实现。组织信息是为了方便检索和查找，而信息检索中出现的问题反过来又反映了信息组织的问题。目前，网络信息检索实践中出现了很多不尽如人意的地方，所以在研究网络信息组织之前，应当首先来研究一下搜索引擎的现状与问题。

1.3 搜索引擎的现状与问题

互联网出现后，搜索引擎是一种重要的网络信息检索工具，它让用户键入关键词，然后到数据库中去匹配。不同于学科目录，搜索引擎不用人工索引员来编制索引，而是通过软件来自动生成包含网页的数据库。概括说来，搜索引擎的工作原理包括三个部分：一是用某种程序，如“网络爬虫”（crawler），来收集众多的网页内容；二是以某种利于高效检索的方式（如标引）组织这些网页，形成数据库；三是接受查询，并用某种排序软件进行排序，并输出结果。

第一，爬虫负责跟踪网络，以广度优先或深度优先的方法从 Web 上下载页面，按照链接从一个网站到另一个网站。不同的搜索引擎有不同的爬虫，有的遍历所有网站，有的则根据自己的标准选择一些流行的网站遍历。前者返回的结果量很大，而后者会返回更加相关的结果，速度也更快。

第二，爬虫返回的每一个页面都存放在一个数据库中，对下载页面的内容进行分析以用于索引，具体包括分词、过滤、转换等工作；然后将文档表示为一种便于检索的方式并存储在索引数据库中，一般采用的方法有矢量空间模型、倒排文档、概率模型等；并通过自动抽词和字顺排列编制好索引。索引是每一个有效词的列表，并有一个相应的指针指向它在数据库中的位置。

第三，实现用户查询关键词和目标文档匹配度的计算，根据计算结果所有符合查询要求的页面 URL 按照相关度递减的顺序排列，并返回给用户；用户接口为用户提供一个输入查询请求，定制查询结果的 Web 页面并将查询结果格式化后返回给浏览器。同样，不同的搜索引擎会遵循不同的原则：有的标引爬虫返回页面中每一个单个的词，有的只标引标题或短语。搜索引擎的第三个要素是搜索软件。该软件将用户键入的提问关键词与索引进行比较，发现匹配的结果并按照相关度进行排序。相关度排序标准依不同的搜索引擎而不同。

不同的搜索引擎所使用的爬虫和标引方法会导致不同的结果。这就是为什么在不同的搜索引擎中键入相同的关键词会得到不同的结果。而且，由于这样的操作方式，使得搜索引擎会返回较分类目录更全面更专业的结果。这是因为它们搜索整个网页而非网站的顶层页面，使用的索引也是自动生成的，而非向分类目录那样使用预先生成的索引。

搜索引擎经过了多年的发展之后，还出现了一些新的成果，功能越来越强

大，提供的服务也越来越全面，其目标是发展成用户首选的 Internet 入口站点，而不仅仅是提供单纯的查询功能。具体表现为以下几个方面：

（1）个性化和多样化的服务。现在绝大多数搜索引擎都提供多样化的服务，以吸引更多的用户，商业搜索引擎尤其注重这一点。以 Yahoo！为例，用户可以从它的首页中查看新闻、金融证券信息、天气预报、黄页，可以进行网上购物、拍卖、找人，或者使用免费 Email 和网上寻呼等服务。许多搜索引擎也开始提供个性化的服务，如 Google 的学术搜索、邮件列表搜索；Yahoo！的“My yahoo!”、邮件列表搜索；Infoseek[①] 的“Personalized start page”；Lycos[②] 的“My Lycos”等，它们允许用户为自己定制起始页面，并选择感兴趣的内容和经常使用的服务放在该页面中。

（2）查询功能更加强大。与最早的搜索引擎相比，现在的搜索引擎在查询功能方面已有了很大的改进。除了简单的 AND、OR 和 NOT 布尔逻辑外，不少搜索引擎还支持相似查询、截词功能。例如，AltaVista[③]、Northern Light[④]、Lycos 等支持短语查询，AltaVista 的高级搜索功能支持 NEAR 逻辑等。域搜索也是一项很实用的功能，它允许用户把查询范围限制在网页的某个域中，如标题、URL、图像标记或链接等，AltaVista、Northern Light 和 Infoseek 等搜索引擎都支持对网页的多种域进行搜索。

（3）分类目录和搜索引擎相互结合。由于分类目录和基于 Robot 的搜索引擎有各自的优缺点，目前它们谁也无法完全取代谁，于是很多搜索站点都同时提供这两种类型的服务。例如，Yahoo！主要是一个分类目录，但它也从有名的搜索引擎服务商 Inktomi[⑤] 那里获取网页搜索结果，当用户查询一个关键词时，Yahoo！首先返回从目录中查到的匹配项，如果用户对结果不满意，或者目录中没有匹配项，那么用户还可以继续查找与关键词匹配的网页。国内两个有名的中文搜索引擎搜狐和 Yeah 也都是这种模式。

下面的问题是，既然已经有了很多搜索引擎可以对网络文献进行主题搜索，为什么还要研究网络信息分类呢？答案很简单，目前的搜索引擎不尽如人意（高广太，2001）。主要表现为：服务器响应经常很慢、重复信息很多、常有垃圾和广告，很多网页是动态的，仅从搜索结果往往看不出是否与你的需求相关。另外，搜索引擎还有一个众所周知的缺陷，那就是搜索的结果往往庞大无比，用户需要花很多时间，才能找到所需信息。举例说明，在 Google 搜索引擎中输入

① GO. com-Official Home Page. http：//go. com/? pg=home. html&sv=a2. 2008-10-04

② Lycos. http：//www. lycos. com. 2008-10-04

③ Altavista. http：//www. altavista. com . 2008-10-04

④ Welcome to Northern Light. http：//www. northernlight. com. 2008-10-04

⑤ Inktomi Support Offerings. http：//support. inktomi. com/Offerings. 2008-10-04

“三聚氰胺”，在 0.46 秒内就返回了 36 900 000 条结果（图 1-1）。很多人对现在的搜索引擎不太满意，因为他们找不到在图书馆进行架位浏览的感觉。而且还有实验表明人们无法通过搜索引擎的方式找到自己真正想要的信息，即使是像电话号码或地址这样简单的信息也是如此（Rosati et al.，2004）。

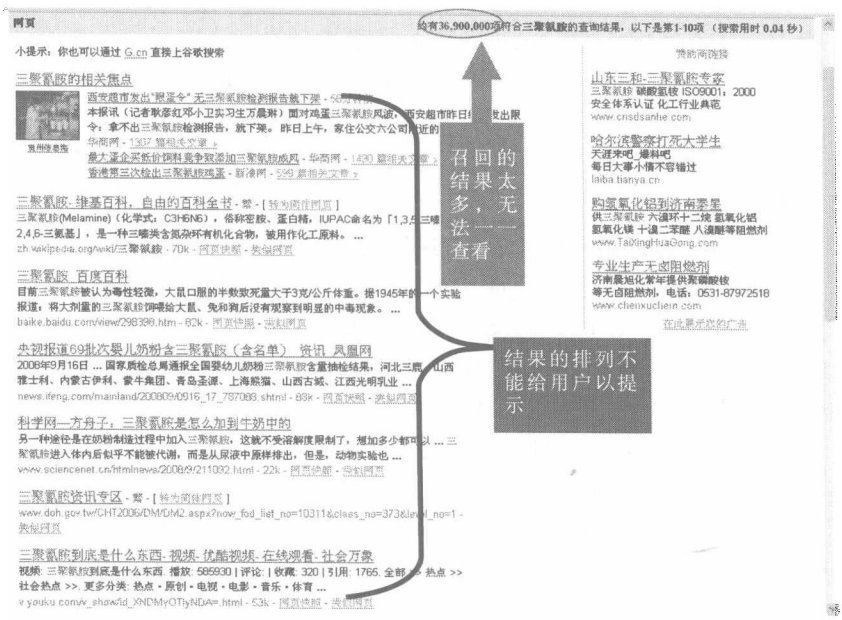

图 1-1 搜索引擎中输入“三聚氰胺”的返回页面

搜索引擎的不足主要表现在以下几个方面：

（1）无法感知在分类表中的位置，也无法知道某结果与其他结果之间的关系。搜索的结果是一条条地排列出来的，用户顶多知道这些结果按照某种顺序排列，如相关度或者时间先后顺序。搜索引擎的结果是孤立的。

（2）背景噪声大，会出现很多无关的结果。可能很多人都有类似的经历，输入一个关键词以后，会返回成千上万条结果，但实际上这其中有很多都是无关的，甚至还有很多类似广告的东西。某些专业的搜索引擎虽然可以部分地解决这一问题，但搜索结果往往又不如综合性的搜索引擎来得全面。

（3）某些信息对象并不包含用户的检索词，但实际上内容是相关的，搜索引擎对这部分信息对象无能为力。也就是说，搜索引擎并不能把你要查找的东西全部反馈给你，否则你所要做的只是花时间筛选而已。另外，互联网上还有相当一部分信息无法通过搜索引擎找到，如存放在后台的动态数据。

为什么会出现上述情况呢？这主要取决于搜索引擎所采用的网络信息的组织

方式，实际上也就是其标引数据库的方式。主要包括两个方面：即选词的原则和标引的深度（王栾生，2004）。例如，是标引链接，还是标题，抑或是正文内容？如果是正文内容，全文标引吗（Schwartz，2001）？

从标引的角度看，分类目录和搜索引擎所使用的方法并不新。分类目录使用的是概念方法，这种方法已经使用了几百年了，也是很多领域使用的分类表的基础（如图书馆分类法和生物分类法）。搜索引擎使用的时间并不长，它基于自动抽词编制索引，应用于20世纪60年代的第一代文本检索软件，并在基于倒排档文本检索系统中实现了。从本质上看，这就是万维网的两种主要的标引和搜索文献的方法——语词标引和概念标引。

语词标引方法自动输入待标文献中的语词，而不考虑其他加工，也不考虑原文献中精确的意义，直接用这些词来描述文献。这种方法基于计算机，不考虑语词的意义，因而标引过程经济且快速，目前使用也较多。万维网的搜索引擎就是使用网站的自动词汇标引。自动标引系统频繁地使用统计工具来为每个词分配权值，按照词频来决定文献中的词哪一个是最重要的。这种方法可以用来选择表征那些高频词。

与基于语词的方法相反，基于概念的方法需要确定文献中语词代表的概念，而非仅仅从文献中抽出来即可。基于概念的方法由人工标引员确定文献中使用的概念，以确保所选择的词最能代表该文献。然而，概念是事物的观念，而不是事物的名称。因此，能够代表文献的标引词（概念或观念）常常不同于文献中实际使用的词，因为同一个概念可以用不同的词来表达（叶继元，1993）。分类标引中概念的产生有点类似于思维中概念的产生，即体验、知觉与表象。这种产生过程是基于一种联系，即将对象、观念、过程和其他实体与头脑中预先存在的数据联系起来（汪丁丁，2001）。基于概念的方法需要人类思维的介入，分析文献中的概念，选择最能反映该文献的词。这种方法引入了解释这一要素，耗时多，不如基于词语的方法经济，但是它关注每一个词的意义，万维网的分类目录拥有基于概念的组织，使用不同网站的人工标引。基于概念的标引结构性很强，用以表征不同概念间的关系，要么通过参照系统，要么通过聚类方法。由于基于概念的标引并不是直接从文献中抽词，而是区分语词不同的上下文环境中的意义，为每一个概念选择一个关键词或表达式。这些关键词或表达式将形成分类目录中的分类索引。

语词标引方法速度快而且经济，因而成了基于海量数据的搜索引擎的组织方式。但正是由于不考虑语词的意义，因而出现了很多问题。这些问题在检索中产生，因为系统使用语词标引方法无法区分同一个词表达的不同意义或不同的词表达相同的意义。而且不同的语词表达相同的概念也无法通过搜索引擎表现出来，除非你能输入该概念的每一个可能的语词，或者用参照系统。例如：“杜鹃”可

以指花，也可以指鸟；“食盐”和“氯化钠”都可以指代同一个物质，即 NaCl。建立参照系统可以解决概念与语词之间多值对应问题。有些搜索引擎使用非常深奥的计算机程序，基于概率统计和人工智能来模拟人类建立概念的过程与方式。可以用这些程序分析与同一个学科相关的不同语词，以确定该文献是属于什么学科的。

网页的超文本结构意味着可以通过浏览和导航，借助于不同的网页之间的链接就可以实现查找过程。由于每一个网页都可以有许多链接，因此要找到同一个信息，就可能有多种途径。通过链接来实现搜索既有优点也有缺点。主要优点是可以帮助人们找到没有意识到的信息，只需点击链接即可。主要缺点有两个：第一，为了找到特定的信息必须点击不同层次的链接；第二，由于没有公认的组织网络信息的框架，没有一个集中式的目录，要想找到特定的信息几乎不可能，而搜索引擎恰恰弥补了这个不足（柳远，2006）。

搜索引擎先收集众多的网页及其链接，然后处理查询。而处理查询的方式是根据用户输入的关键词来匹配数据库中的关键词。这中间的过程就是标引，并制成倒排档。具有这方面知识的人都知道其具体操作：先将所有页面的词汇集起来，然后对每一个词标出其原文出处，当然这其中会省略一些虚词，也会用到通配符。搜索过程发生时，引擎机根据用户输入的关键词进行匹配，即利用倒排档将符合要求的信息找出来，再按照一定的规则进行排序，如所匹配的目标词的数目、所匹配的目标词在原文中的位置、原文的时效性等，然后用一个复杂的算法进行权衡，得出最终的结果。

搜索网络信息资源存在上述问题的主要原因是：信息的设计者或创造者与潜在用户完全分离，用户的特征与信息需求距离数据库设计者或标引员目标越远，从数据库中检索相关信息资源所面临的问题越大。这一点在综合性搜索引擎上表现得最为明显，希望满足的用户越多，所包容的源网页也越多，结果集越大，对于特定用户而言，无关信息也就越多。主要原因是搜索者与信息源的距离太远，无法知道哪一条信息是相关的，为未知用户标引的问题颠覆了很多当代的标引理论基础和搜索算法。事实上，对于同一个搜索提问，不同的搜索引擎会得出完全不同的排序结果。虽然这一事实是不符合当代信息检索研究与实践的前提假设的，但这并不是说当代信息检索理论没有用处，只是说其不适合在网络信息环境中使用。

如果上述因素还算是客观因素，那么还有一些众所周知的主观因素。例如，可能会受到政治、宗教、利益集团的影响，又如网页过滤、屏蔽某些特定的 IP 地址等会影响到用户的搜索结果。搜索引擎提供商会受到广告客户及网络承包商和设备提供商等的影响而干预搜索结果的排序。另外，搜索引擎还存在付费收录与竞争排名的现象。