论点

浅谈垂直搜索引擎技术的应用

余楷鑫

摘　要：本文以垂直搜索为主题，介绍搜索引擎的发展及其现状，对比通用搜索引擎和垂直搜索引擎的优缺点，论述了垂直搜索引擎技术及其发展潜力。

关键词：搜索引擎；通用搜索；垂直搜索

Internet 的发展，给人类社会带来了翻天覆地的变化，将人类文明推向一个新的高度的同时，也给人类提供了无限的商机。它的到来，使信息技术成为当今世界各国发展的主题。Internet 的普及，网民数量的猛增，web 资源的指数增加，都激励着人们探索新的突破。以服务大众出名的通用搜索引擎为我们指引了方向。它们为无数的网民提供了从无底的 web 中寻找资源的机会。然而，随着 Internet 的发展，通用搜索引擎有时很难在庞大的信息库中搜索到准确的信息。它的缺陷，给垂直搜索引擎提供发展的空间，以及无限的潜力。垂直搜索的出现，便是对通用搜索引擎的补充，在未来的时间里，它将与通用搜索引擎相辅相成，服务人类的同时，共享新的金矿。

一、搜索引擎的涵义

所谓搜索引擎，是指在 Internet 下，网站根据用户输入的查询条件（关键字），自动从 web 资源里提取出与用户输入条件相关的信息的一类网站。搜索引擎以一定的策略在互联网中搜集、发现信息，对信息进行理解、提取、组织和处理，并为用户提供检索服务，从而达到信息导航的目的。

随着 Google（谷歌）、baidu（百度）等搜索引擎在 Internet 上经营的成功，越来越多的 IT 企业开始进军搜索市场，经过 IT 精英的不断开拓新领域，创造新价值。搜索引擎从广义上可以将其划分为通用搜索引擎和垂直搜索引擎。

二、通用搜索引擎与垂直搜索引擎的对比

随着网络的发展，它一方面让我们更容易获取到信息，另一方面，信息的爆炸发展，也彻头彻尾地使我们陷入了无边无际的信息海洋之中。在海量的信息页面之前，我们想要找到自己需要的信息简直就如“大海捞针”。搜索引擎的横空出世让我们有了探索信息海洋的指南针。

（1）通用搜索引擎的最大优点是，实现全文搜索，检索到的信息量大，信息覆盖范围广，同时引擎更新信息速度快。目前 Internet 上搜索引擎可索引到的网页数量已超过 110 亿页，由于通用搜索引擎搜索范围的广，导致搜索的匹配度低，命中率低，层次结构不清晰，而且重复连接较多，查询结果信息量大，用户很难在海量的链接结果中找到想要的信息。

（2）垂直搜索引擎是针对某一个特定行业的专业搜索引擎，是通用搜索引擎的细分和延伸，是对网页库中的某类专门的信息进行一次整合，定向分字段抽取出需要的数据进行处理后再以某种形式返回给用户。垂直搜索引擎是相对通用搜索引擎的信息量大、查询不准确、深度不够等提出来的新的搜索引擎服务模式，通过针对某一特定领域、某一特定人群或某一特定需求提供的有一定价值的信息和相关服务。其特点是“专、精、深”，且具有行业色彩，相比较通用搜索引擎的海量信息无序化，垂直搜索引擎则显得更加专注、具体和深入。

三、垂直搜索引擎的原理及组成

搜索引擎主要由搜索器、索引器、检索器组成。基本原理和主要功能组件方面，垂直搜索引擎与通用搜索引擎基本相同。两者主要的区别在于 Spider 爬行范围和网页信息处理深度两方面。通用搜索引擎 Spider 爬行的范围是面向几乎所有网页，而垂直搜索只爬行跟主题相关的网页。因此，垂直搜索引擎能够比通用搜索引擎更快速地找到相关主题的信息。

搜索器（Spider）：也称网络蜘蛛、网络机器人等，是搜索引擎的灵魂。它根据特定算法负责抓取网页，从抓取到的网页里采集信息，对信息进行分词，分词根据词语的特殊属性选择分词算法，并将信息与其关联的 URL 保存进服务器数据库。搜索器必须保证及时的发现新网页，定时的重新采集已有网页信息更新保存数据库数据。

索引器（Indexer）：根据搜索器，即网络蜘蛛采集后经过分词等处理后产生的关键字（keyword），建立从关