垂直搜索引擎应用研究①

Research and Application of Vertical Searching Engine

孔祥春　李义杰　郑凯明　(辽宁工程技术大学 研究生学院 辽宁 葫芦岛 125105)

摘　要： 随着Internet和WWW的迅速发展，Internet上的资源日趋丰富，使得人们在浩如烟海的互联网中得到有用的信息日益困难，垂直搜索引擎应运而生。文章简要介绍了垂直搜索引擎的关键技术及其在政府门户中的应用价值，主要研究了垂直搜索引擎在中央人民政府门户的应用。

关键词： 垂直搜索 因特网 信息检索 搜索引擎

1　引言

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，是针对某一行业或组织，满足行业专业需求、或者组织某项业务需求的专业搜索引擎，是搜索引擎的细分和延伸，是对某类网页资源和结构化资源的深度整合，并为用户提供符合专业用户操作行为的信息服务方式[1]。其特点就是“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，垂直搜索引擎则更加专著、具体和深入。

搜索引擎的出现，整合了互联网上众多的网页资源，并提供信息导航和信息查询服务，使信息的价值得到了网民和厂商的普遍认可。但是，众多专业性网站、行业网站独立于互联网的成功，恰恰证明了互联网的发展格局应该是多方面的。通用搜索引擎的性质，决定了其不能满足特殊领域、特殊人群的精准化信息需求服务。市场需求多元化决定了搜索引擎的服务模式必将出现细分，针对不同行业提供更加精确的行业服务模式。可以说通用搜索引擎的发展为垂直搜索引擎的出现提供了良好的市场空间，势必将出现垂直搜索引擎在互联网中占据部分市场的趋势，也是搜索引擎行业细分化的必然趋势[2]。

2　垂直搜索引擎的关键技术

由于垂直搜索引擎服务具有其自身的特性，因此其技术要求特点上与一般互联网搜索引擎(水平搜索)有很多不同之处，下面通过和水平搜索的比较，列举出垂直搜索引擎的四大关键技术。

2.1 聚集、实时和可管理的网页采集技术

一般互联网搜索面向全网信息，采集的范围广、数量大，但往往由于更新周期的要求，采集的深度或说层级比较浅，采集动态网页优先级比较低，因而被称为水平搜索。而垂直搜索带有专业性或行业性的需求和目标，所以只对局部来源的网页进行采集，采集的网页数量适中。但其要求采集的网页全面，必须达到更深的层级，采集动态网页的优先级也相对较高。在实际应用中，垂直搜索的网页采集技术能够按需控制采集目标和范围、按需支持深度采集及按需支持复杂的动态网页采集，即采集技术要能达到更加聚焦、纵深和可管控的需求，并且网页信息更新周期也更短，获取信息更及时。

2.2 从非结构化内容到结构化数据的网页解析技术

水平搜索引擎仅能对网页的标题和正文进行解析和提取，但不提供其时间、来源、作者及其他元数据的解析和提取。由于垂直搜索引擎服务的特殊性，往往要求按需提供时间、来源、作者及其他元数据解析，包括对网页中特定内容的提取。比如：在论坛搜索、生活服务、订票服务、求职服务、风险信用、竞争情报、行业供需、产品比较等特定垂直搜索服务中，要求对于作者、主题、地区、机构名称、产品名称以及特定行业用语进行提取，才能进一步提供更有价值的搜索服务[3]。

① 收稿时间:2008-11-05

2.3 精、准、全的全文索引和联合检索技术

水平搜索引擎并不能提供精确和完整的检索结果，只是给出预估的数量和排在前面部分的结果信息[(TOPN)]，但响应速度是水平搜索引擎所追求的最重要因素；在文本索引方面，它也仅对部分网页中特定位置的文本而不是精确的网页正文全文进行索引，因而其最终检索结果是不完全的。

垂直搜索由于在信息的专业性和使用价值方面有更高的要求，因此能够支持全文检索和精确检索，并按需提供多种结果排序方式，比如按内容相关度排序(与水平检索的 Page Rank 不同)或按时间、来源排序。另外，一些垂直搜索引擎还要求按需支持结构化和非结构化数据联合检索，比如结合作者、内容、分类进行组合检索等。

2.4 高度智能化的文本挖掘技术

垂直搜索与水平搜索的最大区别是它对网页信息进行了结构化信息抽取加工，也就是将网页的非结构化数据抽取成特定的结构化信息数据，好比网页搜索是以网页为最小单位，基于视觉的网页块分析是以网页块为最小单位，而垂直搜索是以结构化数据为最小单位。基于结构化数据和全文数据的结合，垂直搜索才能为用户提供更加到位、更有价值的服务。整个结构化信息提取贯穿从网页解析到网页加工处理的过程[4]。同时面对上述要求，垂直搜索还能够按需提供智能化处理功能，比如自动分类、 自动聚类、 自动标引、自动排重，文本挖掘等等。这部分是垂直搜索乃至信息处理的前沿技术，虽然尚不够成熟，但有很大的发展潜力和空间，并且目前在一些海量信息处理的场合已经能够起到很好的应用效果。

3 垂直搜索引擎在中央人民政府门户的应用

专业的政务垂直搜索引擎是整合政务资源的有效方式之一，中央人民政府门户搜索引擎实现了对全国省级以上政府网站、国务院公报等内容的搜索，是政务垂直搜索引擎的典范之作。

3.1 中央人民政府门户搜索引擎的功能

通过 http://sousuo.gov.cn 进入中央人民政府门户搜索引擎主页，搜索主页简洁，包含了本网站搜索、国务院公报搜索、政府网站搜索、图片搜索、文档搜索等搜索分类。本网站搜索是指对中央人民政府门户自身发布内容的搜索功能，国务院公报搜索是指对国务院公报内容进行搜索，政府网站搜索是指对各级政府网站的网页内容搜索，图片搜索是指对各级政府网站上的图片进行搜索，文档搜索是指对各级政府网站上的文档内容进行搜索，比如 WORD、PDF、EXCEL、PPT 等。通过这样的分类，可以方便公众有针对性地选择搜索目标。

对于每一种分类搜索，系统都提供“高级搜索”功能，在高级搜索界面上，用户可以根据来源、日期、标题、作者、正文等属性进行搜索，并且可以指定结果的排序方式是按照网页的时间排序还是按照内容的相关度进行排序。

图 1　中央人民政府门户搜索引擎主页

3.2 中央人民政府门户搜索引擎架构

中央人民政府门户搜索引擎总体架构实现了跨平台应用，使整个系统不受硬件平台的限制，具有良好的扩展性和可管理性，设计了集群和负载均衡，在负载增加和并发访问压力增大的情况下，具有扩展能力。中央人民政府门户搜索引擎由采集层、数据加工层、搜索层、系统管理层等部分组成。

3.3 中央人民政府门户搜索引擎特点

(1) 垂直专业搜索：整合政务网络信息

中央人民政府门户搜索引擎实现了对全国省级以上政府网站的内容和服务的采集；实现了包括按信息分类、条件组合、文件类型、图片、区域等多种检索方式，同时实现了对多语种、多文种的检索。

(2) 实时更新搜索信息：第一时间获取一手信息

中央人民政府门户搜索引擎所提供的搜索内容，必须能够及时反映政府网站的内容变化，各级政府网站上新发布的政务信息和办事指南应能及时搜索。目前各级网站发布的新网页一般在 30 分钟之内就可在中央人民政府门户搜索引擎中搜索到。

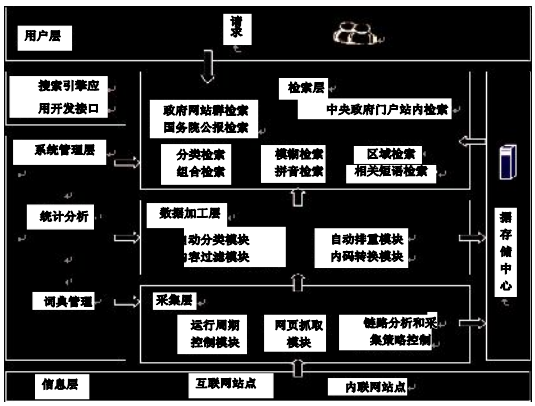

图 2　中央人民政府门户搜索引擎架构

(3) 与政府网站标准化建设紧密结合：实现更好的搜索结果

如果搜索引擎能够更好地“理解”网页内容，那么相信会提供更好的搜索服务。中央人民政府门户搜索引擎对于符合政府网站内容格式标准的网页，能够按标准进行内容分析，提高网页分析的准确性，实现更好的搜索效果。

(4) 分类搜索：方便用户的搜索过程

中央人民政府门户搜索引擎对公众提供了方便的政务信息、办事指南搜索，合理的分类可以方便用户的搜索过程。中央人民政府门户搜索引擎按照服务的类型将搜索内容细分为站内搜索、国务院公报搜索、图片搜索、文档搜索、政府网站搜索等几种类型。中央人民政府门户搜索引擎在采集到的搜索各个环节都需要进行细致的分类工作。

(5) 深度精准搜索：获得互联网搜索不能提供的满意体验

中央人民政府门户搜索引擎是对各级政府网站的全部网页内容进行采集和索引；中央人民政府门户搜索引擎对网页的全部内容建立索引，可以检索；中央人民政府门户搜索引擎能够为用户提供完整的搜索结果集，搜索到的所有网页都是可浏览的，采用的是精确检索技术。

4　垂直搜索引擎在政府门户中的应用价值

整合政务资源，有效提升政务资源价值。门户技术中的“网站群技术”和“全文检索技术”有效地整合了行政领导关系紧密的部门内部的信息资源；垂直搜索技术则有效整合了行政领导关系比较松散的机构间的信息资源，使得政务信息资源的聚拢和整合得以最大化，政务资源的可挖掘能力得到提高，从而有效地提升了政务资源的价值[5]。

一站式检索和导航服务，提高政府门户的公众服务水平。面对多如繁星的政府门户网站，用户查询信息和网上办事时往往无所适从，政务垂直搜索引擎的建设恰恰解决了这个问题，用户可以通过搜索引擎的各种检索方式，方便地获取过去需要访问多个网站才能查全的信息。同时也可以通过检索获取网上办事的入口。一站式检索和导航服务，大大方便了市民和企业，提高了政府门户网站的服务水平。

政治体制改革环境下，“凝聚”组织机构的有效手段。目前，市场格局的变化，按照“大社会、小政府”的思维模式，政府介入微观经济领域越来越少，国家各个行业的部分机构大都由事业型转为企业。这样部委和下面机构之间就没有行政领导职能，但业务上还存在千丝万缕地联系和业务指导关系[6]。垂直搜索引擎的出现将两者有效地“凝聚在一起”，通过“信息的关联”把大家联系在一起，有利于行业内信息的交流和协作。

5　结束语

可以预见，随着信息技术和因特网的发展，垂直搜索引擎在网络信息资源检索中的地位日渐重要。垂直搜索引擎将会更加流行，同时对人们网络生活的方方面面也将产生更为深刻的影响。

参考文献

1 Yang Sok Kim, Byeong Ho Kang, Paul Compton. Search Engine Retrieval of Changing Information. WWW, 2007.

2 Tan QZ, Zhuang ZM, Mitra P, Giles CL. Designing Efficient Sampling Techniques to Detect Webpage Updates. WWW, 2007.

3 俊英.垂直搜索引擎的研究与实现.哈尔滨:哈尔滨工业大学, 2004.

4 肖冬梅.垂直搜索引擎研究.图书馆学研究,2003,(2):90－93.

5 陈新颜.垂直搜索引擎辨析.现代情报, 2004,(9):128－156.

6 罗丽姗.垂直搜索引擎发展概述.图书馆研究, 2006,(12):20－25.