

·学术方阵·

## 智能搜索引擎信息过滤机制研究

张 帆 林 建 ( 华中师范大学信息管理系 湖北武汉 430079)

摘 要: 智能搜索引擎是人工智能技术和传统搜索引擎技术相结合的产物。面对信息无时无刻不在进行更替的网络环境, 智能搜索引擎具有自然语言过滤智能化、多文档处理智能化、用户服务智能化等信息处理机制。为促进智能搜索引擎发展, 应重视用户建模技术研究, 加强基于多 Agent 智能搜索引擎系统的研制与实践, 加大智能搜索引擎关键技术研究力度。

关键词: 智能搜索引擎 信息过滤 自然语言理解 人工智能

中图分类号: TP391.3

文献标识码: A

文章编号: 1003-6938(2007)04-0052-05

### Research on Filtering Mechanism in Intelligent Search Engine

Zhang Fan Lin Jian ( Department of Information Management, HuaZhong Normal University, Wuhan, Hubei, 430079)

Abstract: Intelligent search engine is a product that combines the traditional search engine technology and artificial intelligence technology. Facing the incessantly information changing of environmental network, the intelligent search engine can solve the problem by the mechanism that intelligent filters of natural language, intelligent multi-document processing and intelligent customer services. To promote the development of intelligent search engines, we should pay more attention on user modeling technology, enhance search engine research based on Multi-Agent System and practice intensified research in key technologies of smart search engine.

Key words: intelligent search engine; information filtering; natural language understanding; artificial intelligence

CLC number: TP391.3

Document code: A

Article ID: 1003-6938(2007)04-0052-05

20 世纪 80 年代以来, 国内外种类繁多的搜索引擎, 如 Google、Alta vista、Sohu 等在为用户提供浏览和查询信息、拦截与过滤不良信息和无用信息方面起到了一定的作用, 成为广大网络用户获取网络信息的首选工具。但是, 随着网络信息的爆炸性增长及用户信息需求的个性化发展, 搜索引擎简单的过滤网络信息状况已难以满足用户精确查询信息的需要。杜亚军等人曾对 Google 中文、百度、天网三大中文搜索引擎的智能性进行过测试,<sup>[1]</sup> 测试结果表明, 基于关键词的搜索引擎在“容错性”(用户检索结果集与其真正需要的匹配程度)、“适语性”(查询的结果与查询概念书面用语的耦合程度)及“个性化”(针对不同用户提供针对性信息)等方面的智能较差, 并指明上述三个引擎均未能满足“适应性”及“个别性”要求, 惟有百度对用户的误输入有一定的辨别能力。

由此可见, 传统的基于 Web 搜索引擎虽然在索引库构建上不完全一致, 但其缺陷大致相同。其一, 查询效率低下, 主要体现在“大海捞针”和“资源漏检”两个方面。笔者最近做了一个简单的实验, 利用 Google 引擎查询有关“基因”的研究信息, 点击后系统反馈有 25, 100, 000 个网页。假设一秒钟浏览一个网页, 则需要 6962 个小时查阅完这些结果信息。在网络信息爆炸性增长的今天, 不可能有用户会花费这么多的时间与精力来浏览搜索到的每一个网页, 何况大部分网页内容和查询意图并不相关, 因此, 要获得真正需要的信息宛如大海捞针。“资源漏检”是指传统搜索引擎由于不能理解和联想用户的检索需要而致使信息丢失的现象。笔者使用“红薯”一词进行检索时, 虽然获得了数量巨大的结果网页, 但是仍然丢失了以白薯、地瓜、红苕、番薯等同义概念和近意概念为标引词的

基金项目: 本文系国家社科基金项目(06BTQ024)研究成果之一。

收稿日期: 2006-12-29; 责任编辑: 王景发

信息源。其二,在服务功能方面,传统搜索引擎只能通过匹配方式为用户提供符合检索条件的信息,难以利用匹配算法排除不符合需求的信息,主动推送符合用户实际需要的信息。

上述问题存在的主要原因是由于传统搜索引擎在 Internet 中的搜索过程缺少系统对信息进行检索与筛选的智能行为,其关键是对用户查询条件与目的以及网络资源缺乏理解和认识。因此,有关智能搜索引擎的理念与实践主要源于人的大脑分析与查询信息的高度智能化过程。智能搜索引擎采用机器学习的方法研究文本信息的自动搜集、抽取与分类处理,解决网络信息的主动推送,实现网络信息服务个性化。文章拟就新一代智能搜索引擎信息过滤机制与发展策略进行初步探讨。

智能搜索引擎是以自然语言理解技术为基础的新一代搜索引擎。所谓智能是指该搜索引擎所具有的一种综合能力,包括对网络信息环境及用户信息的感知能力;对感知到的环境与用户信息进行记忆与存储的能力;通过学习实现某一目标的知识获取与过滤能力等。目前虽然对于智能搜索引擎的研究尚处于概念层次的讨论,<sup>[2]</sup>但是关于如何提高搜索引擎的智能性探索近十年来一直都在进行。<sup>[3]</sup>

## 1 智能搜索引擎信息过滤特点

### 1.1 面向动态数据流

智能搜索引擎面对的是半结构化和非结构化的数据,为用户长期的信息需求提供服务,而传统搜索引擎面向的是用户短期的实时查询。智能搜索引擎注重向个人或一组具有相同或相近兴趣的用户提供信息,用户访问的是动态数据流,而不是静态数据库。

### 1.2 语义理解

智能搜索引擎是一种基于内容的信息搜索工具,能够实现对以自然语言形式的用户请求内容和文档内容的理解。其语义理解体现在两个方面:一是理解用户的搜索请求;二是分析信息内容。由于智能引擎对知识具有一定的理解和处理能力,可实现分词技术、同义词技术、短语识别、机器翻译和支持自然语言查询等,因此可将目前的基于分类浏览与简单关键词查询提高到基于概念和知识层面的检索,从而为用户提供更方便、更确切的搜索服务。上述功能的实现主要基于所采用的语义网络等智能技术。中文智能引擎通过汉语分词、句法分析以及统计理论可有效地理解用户查询请求。

### 1.3 动态获取用户兴趣、构建需求模板(Profile)

智能搜索引擎可动态观察和记录用户行为,不断获取用户长期的相对固定的兴趣与爱好,并通过不断的训练学习,增长获取用户兴趣的智能。对返回的信息进行及时评价,不断分

析用户请求,了解用户的真正需要以便调整搜索策略。与传统搜索引擎相比,智能引擎具有用户数据登记和兴趣自动识别机制,这是构建个性化信息需求模块的基础,也是实施有效信息过滤的关键。

### 1.4 个性化信息服务

个性化信息服务的实质就是针对性服务,即针对不同个体采用不同的服务策略,提供不同的服务内容。如前所述,与现有搜索引擎相比,智能引擎不仅可以根据不同用户的信息需求建立需求模板(Profile)并进行自我学习,以便动态地获取不断变化着的用户兴趣,而且可以主动将切合需要的有关信息推送给用户,为用户提供具有针对性的、个性化的信息服务。智能引擎个性化服务的核心就是通过跟踪分析用户的搜索行为,发现其某段时间内的高频检索词,了解用户关心的内容,然后由引擎主动地将与高频检索词相关的信息进行针对性地推送,以提高用户的搜索效率。

### 1.5 智能化信息过滤

智能搜索引擎是一个网页信息的智能获取与处理工具,其智能性首先体现在智能搜索器的使用方面:搜索器通过对特定站点或者遍历 Internet 不断寻找可利用的知识,自动过滤掉非需求信息,完成在线信息索引,再通过启发式学习、类比学习、归纳学习或发现学习等调整搜寻策略。智能搜索是信息过滤技术中的关键技术,智能浏览器则是智能搜索引擎基于机器学习理论设计的智能系统。智能索引数据库或采用客户推送式(由客户数据操作启动信息推送)或采用服务器推送式(由数据库中的触发器启动信息推送)将符合需要的信息推送(过滤)给需要者。将智能代理应用于客户端和服务端可起到自动的不断过滤信息的作用。智能搜索引擎信息过滤的运行结构如图 1 所示。

## 2 智能搜索引擎信息过滤机制

智能搜索引擎是一种基于智能代理的信息过滤和个性化信息服务系统。如图 1 所示,其工作原理是通过智能代理自动获得的资源模型(如 Web 知识、领域资源等)与用户模型进行匹配,并智能化地主动将信息推荐给特定用户,智能代理具有不断学习和不断适应信息资源与用户兴趣动态变化的能力,从而提供个性化的信息服务。智能代理既可以在客户端进行,也可以在服务器运行,其智能机制主要体现在以下方面。

### 2.1 网络“蜘蛛”智能化

网络蜘蛛的概念起源于 1990 年,目前能在各种搜索引擎中运行的蜘蛛程序约有 30 多个,<sup>[4]</sup>其中著名的网络蜘蛛是:AOL Search NetFind、WebCrawler 等。智能引擎的网络蜘蛛面对信息更替无时无刻不在进行的网络环境(如文档常被增加

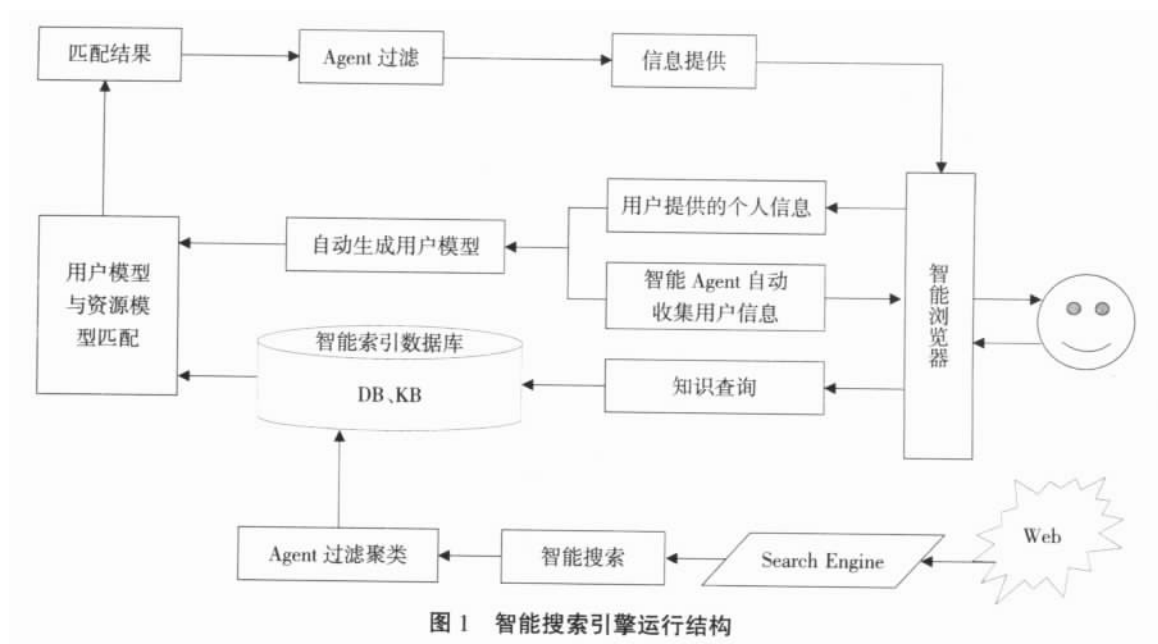


图1 智能搜索引擎运行结构

或删除、改变或添加), 采用启发式或类比式学习方法以及最有效的搜索策略, 选择最佳时机, 从 Internet 上抓取信息并自动完成在线信息的索引, 其可以遍历 Internet 和 Intranet 的任何地方, 并将尽可能挖掘获取的信息进行索引, 其中包括获取特定论点的信息。为了提高搜集速度, 智能系统可同时启动多个引擎进行并行工作, 然后将各引擎的搜索结果加以整合后再存放于索引数据库。

## 2.2 多文档处理智能化

智能引擎为了使文本信息处理的精度得到提高, 降低向量空间的维数, 通常采用基于统计、模式识别、禁用词表或奇异值分解等方式对文本进行预处理, 过滤掉一些无关属性, 以减少无关信息对文本信息处理过程的干扰。与此同时, 智能搜索引擎还具有跨平台工作以及处理多种文档结构的能力。如对网络上的各类文档进行智能化处理。

## 2.3 自然语言过滤智能化

智能搜索引擎支持直接采用自然语言中的字、词或整个自然语句作为过滤检索式, 如使用“超媒体技术向纵深方向发展”作为过滤检索式等, 这对于一般用户更适合。此外, 利用自然语言中的管道过滤也有助于检准率的提高, 所谓管道过滤即使用管道符“—”连接若干检索词, 智能系统首先对第 1 个词进行过滤和检索, 然后在其结果信息的基础上, 对后一个词所涉及的信息进行检索和过滤, 依次类推, 以达到逐步缩小过滤范围, 提高检准率的目的。此外智能搜索引擎的语种过滤以及相关性排序等功能也较强。

## 2.4 多种语言信息过滤与检索智能化

智能搜索引擎可为用户提供多种语言查询, 包含有两层意思: 其一, 系统可以按用户指定的任何语种进行信息搜索,

并输出查询结果; 其二, 支持用户采用某一语言提交查询, 系统在多种语言的索引库中搜索并返回所有的结果文档, 再经过机器翻译把信息结果呈现给用户。

## 2.5 用户服务智能化

智能引擎可通过跟踪用户行为, 了解用户兴趣爱好, 根据用户每次返回的评价, 调查其查询行为, 并对搜索结果做出合理解释。智能服务一般包括以下内容: 其一, 根据用户需求的变化不断提供动态的信息服务, 提高用户获取特定信息的效率、减轻用户认知负担; 其二, 根据用户反馈不断调整搜索策略, 并选择最佳时机, 自动搜集与整理结果信息; 其三, 允许用户为自己定制起始页面, 以便将感兴趣的内容与经常使用的服务置于该页面, 供搜索引擎推送服务时参考; 第四, 将多个搜索引擎的结果文档进行整合, 并将其整合后的整体存放在索引数据库备用。

## 2.6 用户界面智能化

中文智能引擎采用诸如语义网络等智能技术, 通过汉语分词、句法分析以及统计理论等, 使用自然语言与用户交互, 力求最大程度地了解用户, 与用户实时交流。目前已有一些搜索引擎采用自然语言智能答询, 用户可以输入简单的疑问句, 搜索引擎在对提问结构进行分析和内容分析后, 或直接给出答案, 或引导用户从几个可选择的问题中进行再选择。多数智能搜索引擎可用人机对话的方式, 在专业、智能、多媒体搜索的基础上, 为用户提供即时、准确的所需信息。

## 3 国内外著名智能搜索引擎

目前, 人们已经研究出了一些具有一定智能性的搜索引擎。这类搜索引擎的代表有 Lotus Domino Extended search、

Webwatcher、Letizia、FSA、Lexxe、FAQ Finder、Imatch、GHunt 等。

### 3.1 Lotus Domino Extended Search

在该系统中,有关客户的信息保存在多个不同位置的关系数据库、Team Room 数据库、Notes 数据库、竞争对手的 Web 站点以及 Domino Doc 中,只要使用一个直观的界面,用户便可同时搜索所有这些不同类型的数据库,查找并接入相关客户信息。当用户使用 Notes 客户机点击某个文档时,Domino 能够立即引导你进入相关文档所在的 Notes 数据库并调出相应文档;如果所需文档不在 Notes 数据库中,系统将从 DBI 图形文件或电子表格 Domino Extended Search 中查找该文件并保存为 Notes 的附件,以方便用户转发等处理。

### 3.2 Imatch 高智能搜索引擎<sup>[5]</sup>

该系统可以根据用户搜索习惯和意图,智能匹配相关搜索结果,力求贴近用户的实际需要。采用全球领先的高智能匹配技术,率先打破了现有搜索引擎只能匹配搜索结果的不足。实现智能化的模糊匹配,如用户搜索“我要买电视”,“电视购买”,“电视”这些词时,Imatch 都会处理各有关词与“电视”进行匹配,以供用户查询。由于该引擎可智能地匹配关键词,可减少客户频繁调整关键词方案的麻烦。

### 3.3 GHunt<sup>[6]</sup>

智能搜索引擎 GHunt 是由慧聪国际软件与中国网联手共同研发的中文智能搜索引擎。该引擎首次将自动分类技术,中文内容分析技术及区域识别技术应用到大型搜索引擎当中,中文网页覆盖率已超过 2 亿,网页更新频率快,搜索功能较强。GHunt 不仅提高了系统对关键词的理解水平,而且针对用户应用较多的搜索要求,增加了广泛的地域和强大的 MP3 搜索,并可提供针对内容的相关性查询和符合汉语特性的模糊查询等。此外,该系统支持分布式的网络信息并行搜索与内容过滤;其次能识别文本类别并从概念层次上理解文本信息,包括文本分类、聚类,文本概念抽取;第三是实现了有效的、统一的基于概念语义空间的文本信息管理;第四可提供高效的基于语义的文本信息检索、基于内容的图像检索以及个性化的专题信息推送服务等。

### 3.4 Blinkx 中文版<sup>[7]</sup>

该引擎具有类似“模糊搜索”或“语文检索”的功能,该系统经过“学习”积累一定的“经验”之后,可以回答用户类似“最便宜的摄像机名称是什么”这样的搜索需求。这一点与传统的基于关键字的搜索方式截然不同。在搜索视频和音频资料方面具有非常强大的功能,在技术上已经能够准确定位到某一帧,能够准确地搜索到用户需要的声音视频资料,不仅可以搜索文章内容,还可以搜索互联网、本机和局域网上的内容,以及各种不同文本格式的内容。如 Text 等各种数据库中的数据

格式。

### 3.5 Lexxe<sup>[8]</sup>

其特征是实现了语言计算,利用人工智能识别不同类型语句,通过语法分析判断用户意图。Lexxe 引擎可将用户输入的文字当作语言来处理,而不是作为符号,其计算对象是语言,而不是一般的符号,因为它具备了语言的理解能力。但是由于语言技术上的局限,Lexxe 目前尚无法进行同义词自动代替。因此回答问题还局限于语言统计方法上,要求用户提问最好限在 10 个字以内,可达到最佳搜索效果,该系统目前尚不提供中文查询。

### 3.6 Letizia<sup>[9]</sup>

2003 年 8 月,麻省理工学院 Henry Lieberman 在国际人工智能联合大会上提出了个性化导航智能体 Letizia。Letizia 的特点是综合使用信息过滤与信息选择的策略,通过收集用户浏览习惯方面的信息,分析用户的兴趣爱好,并使用各种启发式策略对现有数据进行推理,从而实现一个有限资源的 Web 智能搜索。

## 4 促进智能搜索引擎发展策略

当前,如何向用户提供质量精良、数量适中的检索结果成为搜索引擎技术发展的一个新台阶。智能搜索引擎由于具有更多的“智力”,按照用户需求参与到整个信息过滤与检索的过程,使得搜索结果信息满足用户的需求而受到用户推崇。促进智能引擎技术的不断改进与完善,应注意采用一些有效的策略。

### 4.1 不断加强对自然语言理解及其应用的研究<sup>[10][11]</sup>

智能引擎必须要能理解用户的搜索请求和对网页内容进行分析,前者要求引擎对用户的查询意图和兴趣方向进行推理预测,后者要求系统能为用户提供有效的答案,以确保智能系统既不“漏检”,也不“泛检”。从人工智能的观点看,自然语言理解的研究任务就是要建立一种计算机模型,该模型具有类似人那样理解、分析与回答自然语言的能力。当前自然语言理解系统的研究,已经进入到文字识别和语音识别系统相配合,对书面语言与有声语言的识别与理解的新阶段,以自然语言理解技术为基础的智能搜索引擎比较有名的系统是 LUNAR、SAM 等。当前业界应不断加强对智能引擎的知识理解与分析能力研究,以便不断研发出具有较高知识理解与信息处理能力的软件和系统,以促进智能引擎的开发与利用。

### 4.2 加大智能引擎关键技术研究的力度

智能搜索引擎是人工智能技术和传统搜索引擎相结合的产物,搜索引擎智能化涉及到多方面的原理和技术,如知识理



解、专家系统、神经网络、知识挖掘、数据仓库、机器学习、智能代理等。加强这些技术的研究与开发应用是研制智能搜索引擎的基础工作。当前国内特别要加大对汉语分词技术、短语识别技术、同义词处理技术、知识库与推理机应用技术、文档信息压缩技术和人机对话智能技术等研究力度,在已有成果的基础上,寻求新的突破,使智能引擎的性能不断优化,提高过滤与检索网络信息的效率。

#### 4.3 重视用户建模的研究

智能搜索引擎是面向个性差异的用户,主动排斥或推送其不需或所需信息的智能工具,其信息过滤过程是基于用户模型进行的,它利用用户模型中存储的用户兴趣,判定资源信息是否与用户需求相关,从而确定取舍。但是由于用户兴趣与需求常常是模糊的,还可能是不断变换的,因此,采用科学的算法构造用户模型,描述用户模糊变化的信息需求是研制智能引擎信息过滤系统的关键,应给以重视。

#### 4.4 加强基于多 Agent 的智能搜索系统的研制与实践<sup>[12]</sup>

智能代理 Agent 具有利用相关反馈学习算法和基于多用户个性化模型的层次智能信息滤波算法,自动适应用户兴趣和信源的变化,高效并行地检索与过滤信息的特性。因此在现有 Agent 技术的基础上,加强利用 Agent 的特性,研制个性化的基于多 Agents 技术的智能信息过滤系统,以便从智能性、主动性、扩充性、易维护性等方面改善现有搜索引擎智能过滤的不足,提高检索速度和准确性。

#### 4.5 注重对多媒体网络信息过滤的研究

Web 页面是非结构化的(或称半结构化的),而 Web 页面中包含的多媒体数据更是复杂类型的非结构化数据。目前大多数实际应用的过滤软件主要是通过对 Web 页面的文本信息截取和拦阻而实现信息过滤的,对于将文本信息嵌入到图像文件中或直接以声像文件的形式出现的网络信息,常常可以轻易地躲避过滤系统的监测。也就是说,传统的搜索引擎对多媒体的搜索技术仅停留在对这些文档中文本识别的搜索,对多媒体文档的内容和情节还不能搜索。因此,研究与开发声像信息过滤技术具有重大的应用价值。

#### 4.6 实现 P2P 对等网络的信息过滤

计算机对等联网(Peer-To-Peer,简称 P2P)是目前流行的一种新兴的网络模型,其优势是通过直接信息交换,共享计算机的资源和服务。将 P2P 技术应用到网页的信息过滤当中,可以使系统共享所有硬盘上的文件、目录。用户在使用 P2P 技术进行搜索时无需通过 Web 服务器,不受信息格式的限制,其搜索深度将远远超过传统的搜索引擎,这将使 Internet 上信息的价值得到极大的提升。

#### 4.7 促进领域智能搜索引擎的优先开发应用<sup>[13][14]</sup>

一般来说,通用智能搜索引擎面对的数据信息覆盖面广,信息量大,数据不稳定,冗余度大,其研制相对比较困难,也不易现实。由于大多数用户查询对象局限于某一知识领域,因此出现了一种结合领域知识与智能技术来研究智能搜索引擎的新趋势,这种基于领域的智能搜索引擎(Domain-Based Intelligent Search Engine,简称 DBISE)近年来得到广泛应用。如 Deadliner 主要提供会议文献检索。利用 Web 资源及其传递本身所具有的特点,结合现有机器学习方法,优先研制并推广应用一些基于领域知识的智能搜索引擎是推动智能搜索引擎发展的重要举措。

参考文献:

- [1] 杜亚军等.中文智能搜索引擎的探讨[J].计算机应用,2004,(4):29-31.
- [2] 杜亚军.智能搜索引擎行为的研究与实现[D].博士学位论文,2005.
- [3] 陈鑫.中文智能搜索引擎[D].硕士学位论文,2004.
- [4] 杜亚军等.爬虫算法设计与程序实现[J].计算机应用,2004,(1):33-35.
- [5][7] 新浪网 EB/OL[2006-7-21].http://www.sina.com.cn.
- [6] 慧聪网 EB/OL[2006-12-20].http://www.huicong.com.
- [8] Pazzanim, J.A. Framework for Collaborative, Content-based and Demographic Filtering[J]. Artificial Intelligence Review, 1999, 13: 393-408.
- [9] Benarous, J. D. Etal. Filtering and Control with Information Increasing[J]. Methodology and Computing in probability, 2002,(2):123-35.
- [10] 李志义.搜索引擎发展中的问题与对策[J].情报科学,2002,20(5):556-558.
- [11] C.C.Yang, J.Yen, H.Chen. Intelligent Internet Searching Engine Base on Hybrid Simulated Annealing[J]. Decision Support System, 2000, 28(3):269-287.
- [12] 王伟平等.Web 智能搜索引擎多 Agent 系统结构及相关技术[J].计算机工程,2002,(3):38-40.
- [13] B.Thomas. Web spidering with all methods[J]. Christian wolff, 2001,(12):1-40.
- [14] 陈治平.智能门户搜索引擎技术[J].计算机工程,2004,(5):16-18.

作者简介:张帆(1944-),女,华中师范大学信息管理系教授,研究方向:信息组织与存取;林建(1981-),男,华中师范大学信息管理系情报学专业硕士研究生,研究方向:信息组织与检索。