第16卷 第12期
2006年12月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol.16 No.12
Dec. 2006

一个基于MAS的搜索引擎模型

杨烁颖，白万民
（西安工业大学，陕西 西安 710032）

摘 要：在巨大的Internet/Web信息中很难积极地搜索到准确的信息，搜索引擎技术解决了用户检索Web信息困难的问题，而现有的搜索引擎返回的信息却并不总令用户满意。文中在对MAS理论调研的基础上，提出一个基于MAS的搜索引擎的模型，并将其与著名的Google搜索引擎作比较和分析。

关键词：Agent；MAS；搜索引擎
中图分类号：TP391.3　　**文献标识码**：A　　**文章编号**：1673－629X(2006)12－0195－04

A Search Engine Model Based on MAS

YANG Shuo-ying, BAI Wan-min
(Xi'an Technological University, Xi'an 710032, China)

Abstract: With the development of Internet/Web technology, people have submerged into a huge of Internet/Web information. The research on search engine technology becomes more and more important. After a brief description on the MAS theory, this article discusses a search engine model based on the MAS, and then compares it with Google search engine.

Key words: Agent; MAS; search engine

0 引 言

随着Internet/Web的迅猛发展，全球的网页已超过20亿。用户要在浩瀚的信息里寻找信息非常困难。搜索引擎正是为了解决这个“迷航”问题而出现的技术，它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，从而起到信息导航的目的。然而Web的非结构化信息广泛地分散在Internet中，即使有像Google这样的搜索引擎，对Web上大量的非结构化的和无管理的信息进行索引，也很难保证用户找到真正需要的信息，屏蔽不需要的信息。将Agent作为这些问题的求解方法已经成为现在的一个研究方向。文中将从MAS的角度出发，提出一个基于MAS的搜索引擎模型。该模型的特点是：提高搜索引擎智能性，以及检索信息的有效性。

1 搜索引擎的原理

1.1 搜索引擎的定义

搜索引擎是指因特网上专门提供查询服务的一类网站，这些网站通过网络搜索软件或网站登录等方式，收集因特网上大量网站的页面，经过手工分类或Spider软件Agent处理后建库，根据用户的查询请求，按照一定的算法从索引数据库中查找对应的信息并返回给用户。

1.2 搜索引擎的组成

一个搜索引擎由搜索器、索引器、检索器和用户接口等四个部分组成：

(1) 搜索器（如：自动搜索软件Spider）：在互联网中发现和搜集信息，通常是一个日夜不停运行的计算机程序。它要尽可能多、快地搜集各种类型的新信息。

(2) 索引器：分析搜索器所搜索的信息，从中抽取出索引项，用于表示文档以及生成文档库的索引表。

(3) 检索器：根据用户的查询在索引库中快速检出文档，进行文档与查询的相关度评价，对将要输出的结果进行排序，并实现某种用户相关性反馈机制。

(4) 用户接口：输入用户查询，显示查询结果，提供用户相关性反馈机制。

1.3 搜索引擎的原理

搜索引擎的原理，可以看做三步：从互联网上抓取网页→建立索引数据库→在索引数据库中搜索排序。如图1所示。

(1) 从互联网上抓取网页。

利用Spider系统程序，自动访问互联网，并沿着任何网页中的所有URL爬到其它网页，重复这过程，并把爬过的所有网页收集回来。

(2) 建立索引数据库。

由分析索引系统程序对收集回来的网页进行分析，提

收稿日期：2006－03－25
作者简介：杨烁颖（1978－），女，重庆人，硕士研究生，研究方向为计算机辅助技术；白万民，教授，计算机学院院长，研究方向为计算机辅助设计。

取相关网页信息，根据一定的相关度算法进行大量复杂计算，得到每一个网页针对页面内容中和超链中每一个关键词的相关度，然后用这些相关信息建立网页索引数据库。

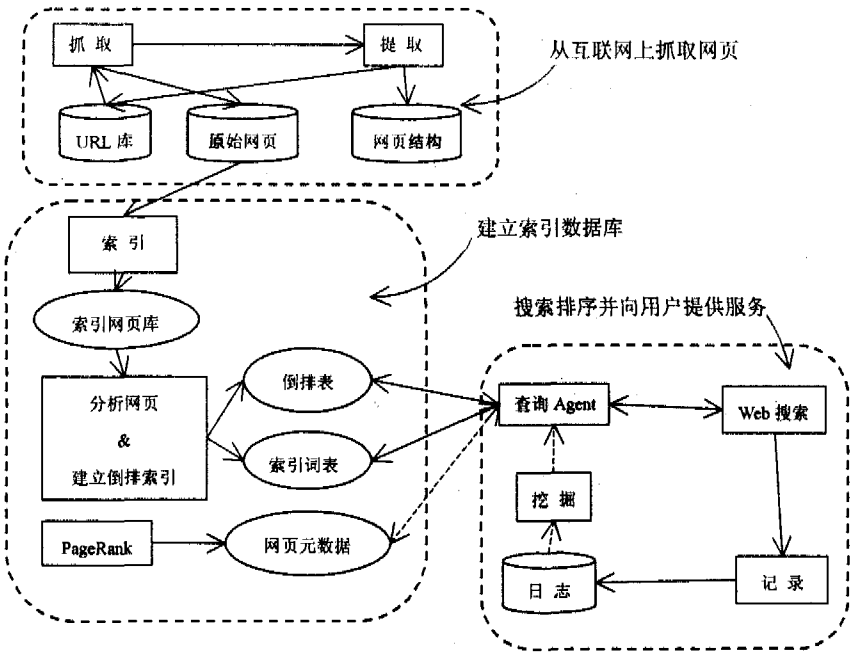

图1 搜索引擎原理图

(3) 在索引数据库中搜索排序。

当用户输入关键词搜索后，由搜索系统程序从网页索引数据库中找到符合该关键词的所有相关网页。因为所有相关网页针对该关键词的相关度早已算好，所以只需按照现成的相关度权值排序，相关度权值越高，排名越靠前。最后，由页面生成系统将搜索结果的链接地址和页面内容摘要等内容组织起来返回给用户。

著名的两种排序算法[1]：

* HITS(Hypertext - Induced Topic Search)算法。

由康奈尔大学 Jon Kleinberg 博士于 1998 年首先提出。Kleinberg 认为搜索开始于用户的检索提问，将用户检索提问分为三种：特指主题检索提问；泛指主题检索提问；相似网页检索提问。HITS 算法专注于改善泛指主题检索的结果，它的目标是通过一定的迭代计算方法得到针对某个检索提问的最具价值的网页。

* PageRank 算法。

PageRank 算法的基本思想是：如果一个页面被许多其他页面引用，则这个页面很可能是重要页面；一个页面尽管没有被多次引用，但被一个重要页面引用，那么这个页面很可能也是重要页面；一个页面的重要性被均分并传递到它所引用的页面。页面的重要性用 PageRank 度量。最后搜索引擎按照页面的 PageRank 值对搜索结果进行排序，将 PageRank 值高的重要页面放在前面。

1.4 搜索引擎的分类

按照信息搜集方法和服务提供方式的不同，搜索引擎系统可以分为三大类：

(1) 目录式搜索引擎。

以人工方式或半自动方式搜集信息。因为加入了人的智能，所以信息准确、导航质量高，缺点是需要人工介入、维护量大、信息量少、信息更新不及时。这类搜索引擎的代表是 Yahoo。

(2) 机器人搜索引擎。

由一个称为蜘蛛(Spider)的机器人程序以某种策略自动地在互联网中搜集和发现信息，由索引器为搜集到的信息建立索引，由检索器根据用户的查询输入检索索引库，并将查询结果返回给用户。该类搜索引擎的优点是信息量大、更新及时、毋需人工干预，缺点是返回信息过多，有很多无关信息，用户必须从结果中进行筛选。这类搜索引擎的代表是 AltaVista, Infoseek, Google, 天网等。

(3) 元搜索引擎。

元搜索引擎是基于 GSE(通用搜索引擎)框架建立的一种搜索机制。用户只需递交一次检索请求，由元搜索引擎负责转换处理后提交给多个预先选定的独立搜索引擎，并将所有查询结果集中起来以整体统一的格式呈现到用户面前。这类搜索引擎的优点是返回结果的信息量更大、更全，缺点是不能够充分使用所使用搜索引擎的功能，用户需要做更多的筛选。这类搜索引擎的代表是 Web Crawler, Info Market 等。

2 基于 MAS 的一个搜索引擎模型

2.1 MAS 介绍

MAS(多 Agent 系统)是由多个可以相互交互的、称为 Agent 的计算单元所组成的系统。Agent 作为计算机系统具有两种重要的能力。首先，每个 Agent 至少在某种程度上可以自治行动，由它们自己决定需要采取什么行动实现其设计目标。其次，每个 Agent 可以与其他 Agent 进行交互，这种交互不是简单的交换数据，而是参与某种社会行为，就像人们在每天的生活中发生的那样：合作、协作和协商等。

图 2 描述了一个标准的多 Agent 系统结构。系统包含一些 Agent，它们通过通信互相交互。这些 Agent 可以在环境中动作，不同的 Agent 有不同的“作用范围”，表示它们可以控制、影响环境的不同部分。在有些情况下，影响的范围可能会重叠，而影响范围重叠的事实会产生 Agent 之间的依赖关系[2]。

2.2 一个基于 MAS 的搜索引擎模型

根据已有的 MAS 的标准结构和搜索引擎原理，笔者提出了一个基于 MAS 的搜索引擎模型，如图 3 所示。

(1) 图 3 中最底层是 Internet 环境，是个人和公司的主页。

(2) 搜索器 Agent 的功能是:发现和搜集互联网中的信息,每一个搜索器 Agent 都有自己的作用范围,如:搜索器 Agent 1 遍历北京地区的网页,搜索器 Agent 2 遍历上海地区的网页等等,它们都是相关作用范围的专家。

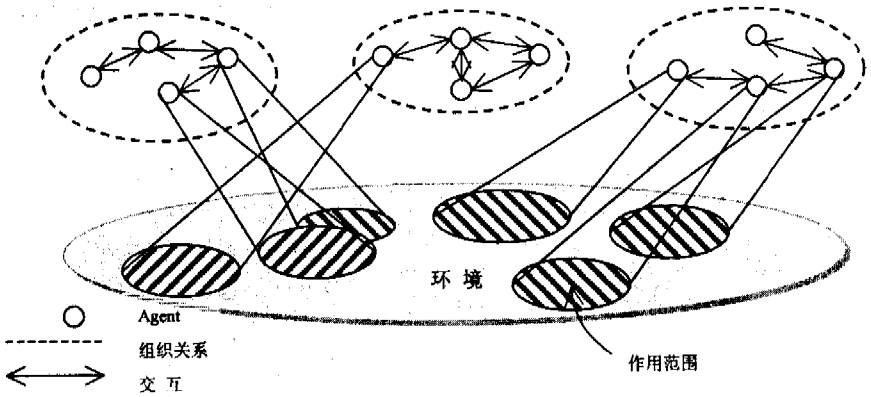

图 2　MAS 的标准结构

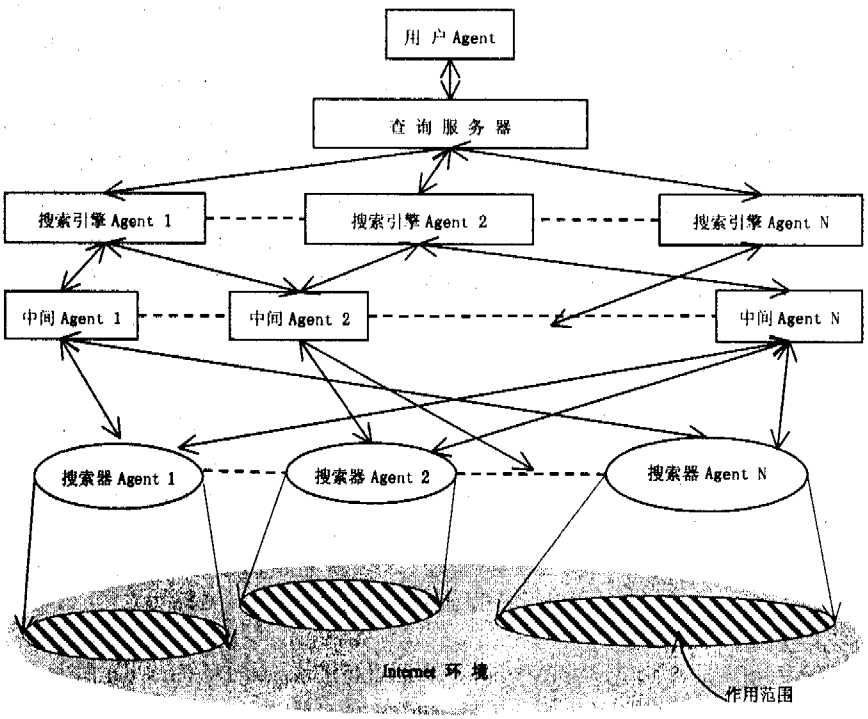

图 3　一个基于 MAS 的搜索引擎模型

(3) 中间 Agent 可以访问多个搜索器,它的功能是:提取相关网页信息,建立索引数据库,计算所有与关键词相关网页的相关度,并对其权值排序。

(4) 搜索引擎 Agent 可以访问多个中间 Agent 来查找信息,搜索引擎 Agent 可以通过使用 Agent 通信语言与中间 Agent 通信,告知它们自己需要的信息。

(5) 查询服务器的功能是:由页面生成系统根据搜索引擎 Agent 反馈的信息权值按权值由高到低的顺序,将搜索结果的链接地址和页面内容摘要等内容组织起来返回给用户 Agent。

(6) 最顶层的用户 Agent 利用了数据挖掘技术来挖掘用户的兴趣关联规则,并且充分考虑用户当前的兴趣状况,可根据用户的兴趣习惯很好地预测用户即将发生的行为,进而从查询服务器返回给用户的页面与用户的兴趣关联显示出最有价值(用户最感兴趣)的页面,从而达到提供主动服务的目的[3]。

3　搜索引擎实例——Google 搜索引擎

最大的搜索引擎 Google 从 2002 年的 10 亿网页增加到现在的近 40 亿网页。据估计,整个互联网的网页数达到 100 多亿,而且每年还在快速增长。因此一个优秀的搜索引擎,需要不断地优化提升其性能。图 4 是 Google 的逻辑结构[4]。

Google 的搜索原理:

(1) 几个分布的爬行者(自动搜索软件)同时在网上爬行,URL 服务器负责向爬行者提供 URL 的列表。爬行者所找到的网页被送到存储服务器中。

(2) 存储服务器于是把这些网页压缩后存入一个知识库中。每个网页都有一个 doc ID,当一个新的 URL 从一个网页中解析出来时,就被分配一个 doc ID。

(3) 索引库和排序器负责建立索引,索引库从知识库中读取记录,将文档解压并进行解析。每个文档就转换成一组词的出现状况,称为 hits。hits 记录了词、词在文档中的位置、字体大小、大小写等。索引库把这些 hits 又分成一组"barrels",产生经过部分排序后的索引。索引库同时分析网页中所有的链接,并将重要信息存在 Anchors 文档中,该文档包含了足够信息,可以用来判断一个链接被链入或链出的结点信息。

(4) URL 分解器阅读 Anchors 文档,并把相对的 URLs 转换成绝对的 URLs,并生成 doc ID,它进一步为 Anchor 文本编制索引,并与 Anchor 所指向的 doc ID 建立关联。同时,它还产生由 doc ID 对所形成的链接数据库。

(5) 链接数据库用于计算所有文档的页面等级(PageRank 算法)。

(6) 排序器会读取 barrels,并根据词的 ID 号(word ID)列表来生成倒排文档。

(7) 一个名为 DumpLexicon 的程序则把上面的列表和由索引库产生的一个新的词表结合起来产生另一个新的词表供搜索器使用。

(8) 搜索器就是利用一个 Web 服务器,并使用由

DumpLexicon 所生成的词表,并利用上述倒排挡以及页面等级来回答用户的提问。

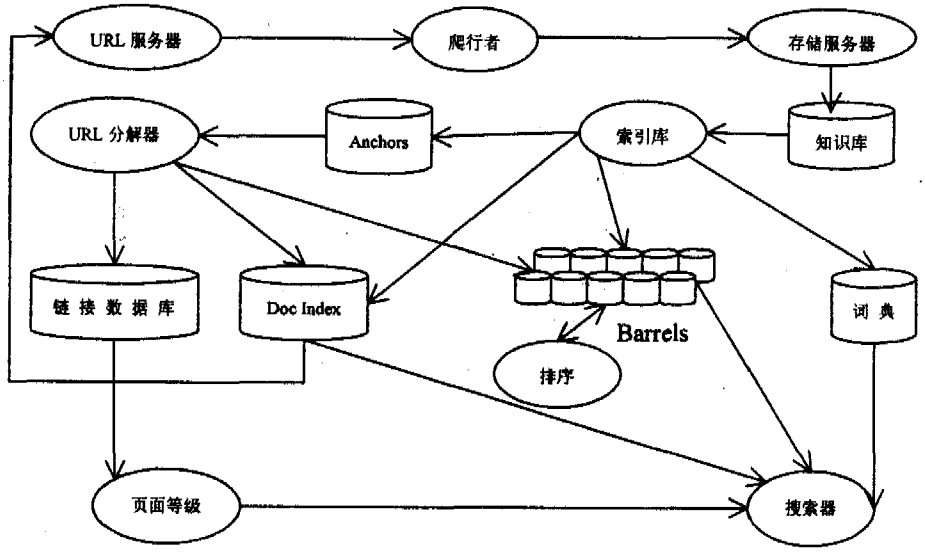

图 4 Google 的逻辑结构

4 基于 MAS 的搜索引擎模型与 Google 的比较

(1) 图 3 中的搜索器 Agent 与图 4Google 的逻辑结构中的爬行者的功能类似,是搜索互联网上的网页,而与图 4Google 的逻辑结构中的搜索器不同,Google 中的搜索器是利用一个 Web 服务器来回答用户提问的。图 3 模型中的搜索器 Agent 运用了类似垂直主题搜索的原理,假设它是按地区搜索的,也可以假设它按主题范围去搜索,如股票、天气、新闻等类的搜索器,垂直主题的搜索器更具有高度的目标性和专业性。

(2) 图 3 模型中每一个搜索引擎 Agent 都相当于一个现有的 Web 搜索引擎,比如 Google 就可以成为该模型中的一个搜索引擎 Agent。

(3) 图 3 模型中还用到了元搜索引擎的原理,设计了一个查询服务器,它可以并行查询许多搜索引擎,然后将这些搜索引擎返回的结果进行整理并递交给用户。这样它的覆盖面积就比一般的搜索引擎大很多。

(4) 图 3 模型中的用户 Agent,它是一个智能 Agent,观察用户的操作能力,通过 Cookie 机制能实现将用户的兴趣信息存放在本地 Agent 上,把查询服务器提交的页面中与用户兴趣一致的返回,从而提供了个性化服务[5]。

通过以上分析,可以看出文中提出的基于 MAS 的搜索引擎在信息分类搜索及返回满足用户查询信息方面更具优势。

5 结束语

目前的搜索引擎虽然能给用户提供海量的搜索结果,却很少有用户看 10 页以后的搜索结果。未来的搜索引擎有必要引入人工智能技术,尝试去理解用户的查询意图,并优先显示用户需要的结果。它使用自动获得的领域模型(如 Web 知识、信息处理、与用户兴趣相关的信息资源、领域组织结构)、用户模型(如用户背景、兴趣、行为、风格)知识进行信息搜集、索引、过滤(包括兴趣过滤和不良信息过滤),并自动地将用户感兴趣的、对用户有用的信息提交给用户。智能 Agent 具有不断学习、适应信息和用户兴趣动态变化的能力,从而提供个性化的服务。这种基于智能 Agent 的信息过滤和个性化服务的搜索引擎将有待人们去研究和实现。

参考文献:

[1] 何晓阳,吴 强,吴治蓉.HITS 算法与 PageRank 算法比较分析[J].情报杂志,2004(2):85-86.
[2] Wooldridge M. An Introduction to MultiAgent Systems[M]. 北京:电子工业出版社,2003.
[3] 张卫丰,徐宝文,许 蕾.利用 Agent 个性化搜索结果[J].小型微型计算机系统,2001,22(4):724-727.
[4] Brin S,Page L. The Anatomy of a Large-Scale Hypertextual Web Search Engine[J]. Computer Networks and ISDN Systems,1998,30(7):107-117.
[5] 徐宝文,张卫丰.搜索引擎与信息获取技术[M].北京:清华大学出版社,2003.

(上接第 194 页)

列工作的循环,实现三维模型在地表的真实感运动。

2 结 论

文中讨论了虚拟战场环境中军事目标三维模型与地形匹配的几个关键问题,并给出了相应的解决方法,重点解决了四点匹配过程中支撑平面的确定和悬空点的判断,提出了一种有效的姿态匹配算法。三维模型与地形的无缝匹配对于保证虚拟环境的真实性有至关重要的意义,如何根据系统的具体要求,在保证表现真实性的前提下进一步提高算法的效率,是以后需要做的工作。

参考文献:

[1] 朱竞夫.DTM 在坦克视景仿真中的应用[J].装甲兵工程学院学报,1998,12(2):53-57.
[2] 郭齐胜.车辆驾驶仿真中地形匹配的数学模型[J].装甲兵工程学院学报,1999,13(1):56-59.
[3] 张景赛.DVENET 中坦克装甲车辆机动性仿真的研究[J].系统仿真学报,2000,12(5):315-318.
[4] 齐 敏.虚拟环境中运动车辆行为仿真的程序方法研究[J].数据采集与处理,2000,15(4):500-503.
[5] 宋汉辰.三维对象模型与地形的匹配方法研究[J].计算机辅助设计与图形学学报,2003,15(9):1167-1171.