面向联盟企业的智能化专业搜索体系设计与实现*)

A System of Intelligent Professional Search of Virtual Enterprise

田力威　尹朝万

(中国科学院沈阳自动化研究所　沈阳110016)

Abstract The existing Meta-search system finishes the parallel Web search services by the general multi-Agent. It's applications in dynamic VE(Virtual Enterprise)have been limited because it is eyeless while beginning to search and it is slowly when facing dynamic system . This paper designs a system of intelligent professional search for VE(Virtual Enterprise)with intelligent search technology and multiple agent techniques. Towards the request of the exact and efficient professional query in the VE, This paper supports the realizable method and effective mechanisms about distribute multi-engine cooperation basing on multi-agent, CORBA and the two-pole information filter system on server/client structure. It is very valuable to improve the speed of rebuilding of VE and the efficiency of using the sharing information of VE.

Keywords VE(Virtual Enterprise), Multiple agent, Intelligent search

1. 引言

联盟企业是指在日益激烈的市场竞争中具有不同、互补的技术和生产能力的企业间为响应市场需求而进行动态联合的方式,是一种随着市场需求变化的可重构、可重用、可扩展的动态组织形态[1]。其成员或网络用户是否通过协同网络在分布式组织结构中高效、快捷地获得用户需求、用户定单、产品市场前景和企业生产资源、加工能力、产品资源信息以及虚拟企业内各单元的生产状况、数据、文档等信息,将直接影响联盟企业的决策速度和决策准确性。因此为联盟企业内各种用户提供高效、准确的专业检索体系已成为目前实现虚拟企业面临的一个重大问题。

传统的互联网搜索体系如图1所示。网络蜘蛛周期性地到互联网上发现新的网页、刷新已发现网页,并将发现结果提交给存储管理模块;查询引擎则根据用户的查询需求,通过索引模块获得用户感兴趣的网页。在这种体系中,搜索引擎的起始点是一个随机的网页,其搜索过程初期有相对的盲目性,而搜索范围又根据其各自的搜索算法具有一定的局限性。在一般搜索体系下,虚拟企业中的用户就不得不使用各种搜索引擎界面来操作多个搜索引擎进行搜索。这不但加大了搜索操作的难度,同时由于大量无关、过时或重复搜索结果的存在大大降低了搜索的效率[2]。目前,许多研究者已经利用"元请求"技术设计了多引擎搜索工具,如 MetaCrawler、SavvySearch、Softbot、Amathaea 等。这些工具实际上是将大量通用搜索引擎(如:Yahoo、AltaVista 等)集成在一起,利用自身的分布协调功能来调度各个引擎。但如将其应用于虚拟企业中时则会出现以下问题:(1)非专业检索方式无法适应专业化用户的需求。现有大部分信息检索系统采用关键词输入方式进行检索,对任何用户都是一种模式,很容易让一般用户感到迷茫,同时使专业用户无法准确地找到自己感兴趣的对象。(2)用户与检索系统的交互方式比较单调。现有系统普遍采用相关反馈技术作为用户和系统进行交互的主要手段。针对不同需求的用户,提供不同的输入方式是目前现有系统所缺少的。(3)缺少分布式智能专业信息检索和适应异构信息源信息变化的能力。现有系统(如 WebWatcher,InfoFinder)主要通过学习用户的历史关联信息,在线引导用户检索感兴趣的信息。这种为用户导航的方式每次只能浏览一个站点,效率比较低,而且无法避免用户浏览以前已经浏览过而现在不需再看的文档或链接[3]。此外,由于没有有效地适应信息源信息变化的机制,不能及时为用户提供新的信息,因而无法为用户快速定位感兴趣的主题。目前的搜索机制一般为全局的、集中的搜索。它无法适应虚拟企业物理和逻辑上的异构性、分布性和可重构性。针对上述问题,本文设计了一个基于 Web、Java 和 CORBA 技术的、面向虚拟企业的专业化信息检索系统,它不仅为各种用户提供高效、准确的专业搜索引擎,同时也可帮助用户选择欲加盟的联盟企业或帮助盟主约请兴趣企业构建联盟企业。

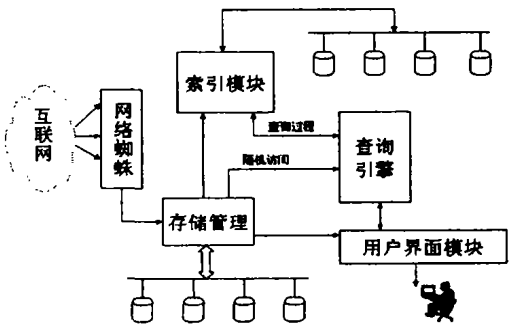

图1　一般搜索体系结构

2. 面向虚拟企业的智能搜索体系

面向虚拟企业的搜索体系如图2所示。它基于分布式 Agent 与 CORBA 技术。对于具有异构信息资源的可重构联盟企业的建立与运行,实现在互联网上的高效查询与检索。系统中用户通过统一的用户界面,输入自己搜索请求。用户代理将

*)本文得到国家自然学基金重点项目(79931000),国家863/CIMS 项目863-511-030-007-1资助。田力威　博士生,研究方向:企业信息集成及分布式处理,智能搜索技术,网络管理。尹朝万　博士生导师,研究方向:企业信息集成及分布式处理,智能搜索技术,计算机应用技术。

这一请求提交给 CORBA 总线，并根据 MARK 数据库中有关用户的个性化数据推理出用户对某一方向的兴趣度或对几个方向兴趣度大小的排序。服务代理则在 CORBA 系统的调度下，调用某一专业搜索引擎或几个搜索引擎来完成对用户请求的搜索。搜索结果通过信息滤波器的过滤和分类后提供给用户。该系统能够满足虚拟企业用户在信息检索时的个性化要求，它从用户的角度出发，为了满足不同用户个性化搜索的需求，采用基于 CORBA 规范的多代理技术和多用户个性化模式的协作信息滤波算法，过滤掉了大量不相关搜索结果，有效地消除了用户迷茫问题。

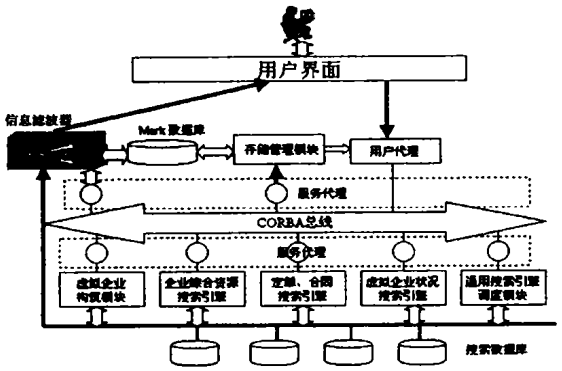

图2　面向虚拟企业的搜索体系

2.1　用户界面

本系统为用户提供了统一的定制界面，用户不但可输入欲搜索的关键词还可以输入对已搜索到网页的满意度 SL，共10个级别(－4～＋5)，用户代理根据 SL，通过反馈学习来确定用户对某关键词的满意度。系统会自动将用户感兴趣的网页及其满意度和感兴趣的关键词及其满意度，存入 Mark 数据库中，以便信息滤波器进行反馈学习。

2.2　多代理

系统中代理可分为三类：(1)用户代理：根据用户请求和对网页的满意度评价，创建请求元和推理出用户对关键词的满意度和讨厌度。(2)中介调度代理：它主要是调度各个服务代理和 CORBA 的公共设施间的交互，其作用相当于一个搜索领域的专家，是一个中介搜索引擎。在 CORBA 中，中介调度代理这一角色由具有协调能力的 Agent 扮演。它在分布式环境下传播用户代理与服务 Agent 间的请求和服务，并根据用户兴趣度动态地建立用户代理与服务 Agent 间的连接；从而实现针对某个用户的兴趣选择适合他的某一领域专业搜索引擎或各个引擎并行搜索的结果的按用户兴趣排序；同时它还为系统中的其它 Agent 提供一系列管理服务，如生命周期、安全性等。(3)服务代理：其主要功能是为各种专业引擎提供基于 CORBA 规范的封装。服务 Agent 既可以单独向用户代理提供服务，也可以进行合作以完成复杂的应用请求。

2.3　信息滤波

随着虚拟企业的不断重构，企业个体资源的不断增长，虚拟企业内部供应链组织形式的不断变化。使搜索结果中存在大量过时或与用户兴趣不相关信息，形成了信息洪水[4]。它大大地降低了搜索的效率，信息滤波器则是通过对搜索结果分析，过滤掉大量无关信息并将按兴趣度排序的结果呈现给用户。文[5]介绍了信息滤波的3种实现方式：即基于内容的滤波、社会滤波和经济滤波。社会滤波方法与分布式信息检索最为紧密，其思想是允许用户根据搜索结果的内容或其他用户对该结果的评价，通过用户代理给出其对搜索结果满意度。并结合矢量空间模型和词频测量方法、相近语义索引等技术来判断出新结果的兴趣度。其信息滤波完全驻留在服务器端。

本滤波系统则分两部分：SERVER 部分驻留在服务器端。它通过服务代理积累的用户们对某一网页的满意度评价，来简单决定网页符合某类用户的兴趣度。CLIENT 部分则安装在每个用户的终端上，根据社会滤波方法精确的推理结果实现对搜索结果符合某一用户兴趣的精确评价。

鉴于以上结构，本系统具备以下特点：(1)多个专业领域搜索引擎同时运行，大大地提高了系统搜索效率；(2)各个基于领域的引擎，实现了浏览信息检索的专业化；(3)能动态地适应不同用户的多种检索需求。

3　用户界面

用以对用户输入的关键词，进行预处理，对待西文，要去掉间歇词汇，然后将各种变形的词汇还原，如 rebuilding，rebuilt 还原为 rebuild；对中文要进行切词。

```
The search interface()
  {
    struct Keyword 定义关键词结构
    {char input;
     string term;
    }
    struct SL 定义满意度结构
    { char value
    long SL
    }
interface() 定义界面函数
{
string Domain; 定义搜索域
unsigned long Webnum; 定义相关专业引擎
boolean DataBase;
long num_keyword;
Add (Keyword,SL,add); 记录已输入的关键词及其兴趣度
Remove (Keyword,SL,remove); 删除已输入的关键词及其兴趣度
GetKeyword(long index); 增加新输入的关键词到系统内
GetSL(long input); 增加新输入的兴趣度到系统内
ExcuteSearch(); 执行搜索并得到相应信息
AbortSearch(); 结束这个搜索
  }
}
```

4. 用户代理

用户代理是与用户打交道，提供特定领域应用，实现用户的专业需求[6]。其结构如图3所示，其各部分功能为：

(1)接口层：用户代理与外部环境、用户代理与其它代理间的通讯中介。

(2)安全层：提供对用户代理内部的安全保护。

(3)任务推理机：是一个利用前馈人工神经网络推理机实现用户兴趣度推理功能的模块。我们将用户输入内容用 n 维矢量(W_1，…，W_n)进行表示。该词在该网页中出现的次数为 TF，而 IWF 代表含有该词的网页被返回的频率。由此可得用户对某一关键词的满意度为：

$$Interest_{v_i}(L) = Interest_{v_i}(L-1) + d \times SL \times TF_i \times IWF_i \quad L(1,2,3,\cdots,n)$$

上式中 d 为推理机学习阻尼；SL 为用户手工输入对搜索结果的满意度；对于搜索过程中关键词出现的次数 L；对于从未搜索过的关键词 $Interest_v$ 为0或在系统安装时用户初始化所赋予的值。通过系统自身的不断学习，当用户对某一关键词的满意度累计超过一定的值时，用户代理则将它提交给服务代理，以用来协助选择专业搜索引擎。

(4)个性化数据库：用来保存用户个性的查询相关推理规则和信息，以帮助确定用户对查询结果的兴趣度。

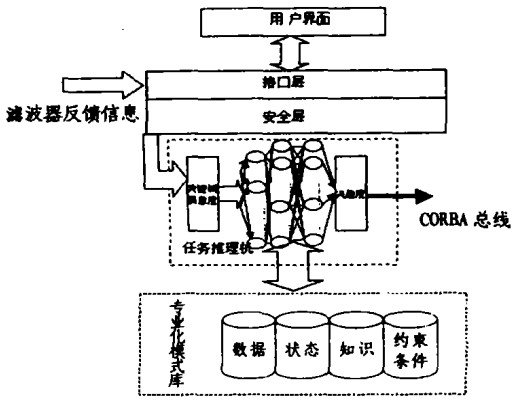

图3 用户代理

5 服务代理

在基于 CORBA 封装的体系下，尽管服务代理有能力向外界提供服务，但它并不关心到底向谁提供服务，只有当出现请求时给予响应。在本系统中服务代理共有三个主要任务：

(1)实现对 CORBA 各种公共设施的调度与使用，即“公共服务代理”；这里以查询请求为例定义如下。

```
module QueryService{
  interface query: active
  {
    while(){
    QueryableCollection.evaluate();
    create_iterator();
    retrieve_element_at();
    next();
     }
    }
  }
module CreateAgent:AGENT(QueryServiceAgent,〈ServiceAgent $
QueryServiceAgent〉)
  {
    integer now, timecontant,time;
    CAPABILITIES:=(〈动作〉〈精神条件〉);
    INITIALBELIEFS:=〈断言〉;
    COMMITMENTRULES:=〈承诺规则〉;
    〈动作〉::=(RUN〈time〉〈query()〉)|
      (INFORM〈time〉〈AGENT〉〈断言〉)|
      (REQUEST〈time〉〈AGENT〉〈动作〉)|
      (UNREQUEST〈time〉〈AGENT〉〈动作〉)|
      (REFRAIN〈动作〉)|(IF〈精神条件〉〈动作〉)
    〈精神条件〉::=〈精神条件〉OR〈精神模式〉|〈精神条件〉AND
                  〈精神模式〉|〈精神模式〉|ture;
    〈承诺规则〉::=(COMMMIT〈消息条件〉〈精神条件〉(〈agent〉
                  〈动作〉));
    〈消息条件〉:=〈消息条件〉OR〈消息模式〉|〈消息条件〉AND〈消
                  息模式〉|〈消息模式〉|ture
    〈断言〉:=(〈time〉(〈谓语〉〈参数〉))|〈变量〉;
  }
```

(2)实现对各种专业搜索引擎的 CORBA 封装，即“专业服务代理”；

(3)根据用户对关键词和网页的兴趣度，针对特定用户选择适合的专业搜索引擎或多个引擎优先级调整“调度服务代理”。其具体操作形式如图4所示。当服务代理接到搜索请求时，首先根据用户对关键词的满意度推理出用户感兴趣的领域，进而选择相应的专业领域搜索引擎进行搜索。当关键词涉及到多个领域时，服务代理则根据该词在专业引擎中所赋的权重，依次启动各搜索引擎并在其查询结果中加入相应的兴趣度权值。以关键词“订货”为例，如图4所示：其在“定单、合同搜索引擎”中的权重为2，所以先开始搜索；而在“企业综合资源搜索引擎”中权重为1，因此后启动；而其它引擎与其不相关，因此不启动。其主程序为：

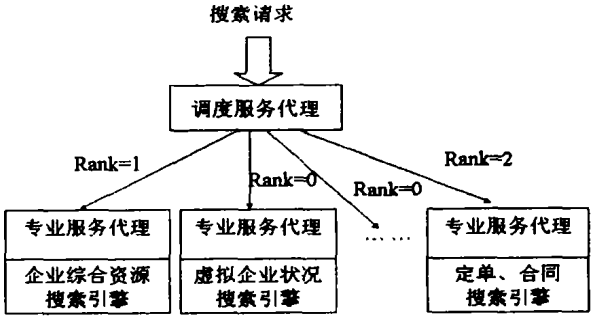

图4 多搜索引擎调度

```
main (query,InterestV)
{
 for Service_agent(i=1..Max_Searchengine)
  {
  Initialize(Service_agent(i),query);
  GetUsrProfile(Service_agent(i), InterestV);
  Service_agent.Rank(i)= PROFILE
  }
 Search_Process()
   {
    Whie(visited〈 Max_Searchengine)
    {
     Order(Service_agent.Rank(visited));
    }
    Whie(visited〈 Max_Searchengine)
    {
  if (Service_agent.Rank(visited))=INIT_PROFILE)
    {
      start{
       CORBA_Orbix_is_ready(Search(Service_agent.Rank(visit-
          ed)); }
       Catch(CORBA_SystemException&SysEx){
       ..... }
  }
}}}
```

6. 信息滤波

信息滤波的目的主要有两点，即过滤掉不符合用户兴趣或已查阅过的网页[7]；按其与用户兴趣的相关性进行排序，并将相关度最高的十个网页放在第一页。其具体主要步骤如下：

步骤1：服务器端为已检索到的网页进行初步分类。其分类公式为：

$$Result_c(d)=\log Pr(c)/n+\Sigma Pr(\omega_i|d)\log(Pr(\omega_i|c)/Pr(\omega_i|d)$$

其中 n 是文档 d 中的单词数；T 是句子的长度；ω_i 是在句子中的第 I 个词；$Pr(\omega_i|c)$是随机地从 C 类文档中抽取词 ω_i 的数学期望；$Pr(\omega_i|d)$是指在文档 D 中单词 ω_i 出现的期望。由此可以得到文档 D 属于 C 类的水平。

步骤2：客户端下载与用户兴趣相符的类别网页到客户机上。

步骤3：将结果按其后标识的用户兴趣度权重进行排序。

步骤4：根据 Mark 数据库中的反馈文档，剔除已查阅过的网页。

步骤5：提取 Mark 数据库中的用户感兴趣关键词，生成主题域词矢量。

步骤6：进行满意度计算，并按满意度大小进行排序。

通过信息滤波，与用户个性化模式不相关的文档或用户不感兴趣网页被过滤掉了，反馈的结果都是用户感兴趣的网页，并按满意度大小排序，从而提高了检索的精度。

7. 性能评价

本文从搜索效率和搜索适应性两个方面对本系统和其它基于多代理机制的搜索引擎[8]进行了性能对比。其具体过程

为：

(1)搜索效率：它是评价一个搜索引擎的关键性指标，也是体现搜索机制优越性的重要因素。搜索总时间由以下几部分组成[9]：①完成请求、服务代理到服务提供者过程规划所需时间S；②代理之间通讯所需时间C；③提供服务所需时间T；④等待提供服务所需时间Q；则其总时间R＝D+Q；其中D＝T+S+C。在此，R是服务提供数和请求产生周期的函数，如果请求产生的增长速度大于系统的吞吐量，即：P<D/N；那么系统将会崩溃而搜索时间将无限地增长。而只有在CORBA体系中，系统能准确地平衡请求和服务提供者间的供求，大大地抑制了"请求洪水"的现象，保证了系统在大负载情况下，搜索时间的稳定。其计算公式为：

$$R=D/(1-D/(P*N)) \quad (1)$$

其结果见图5，本系统搜索总时间与请求发生周期的关系为图5中实线所示。

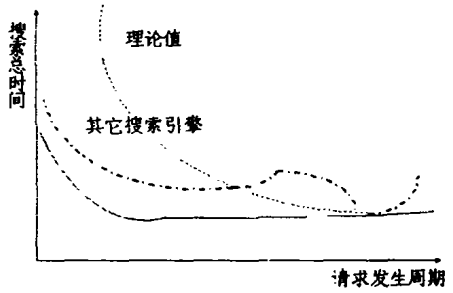

图5　基于CORBA的搜索引擎与其他智能搜索引擎搜索时间对比图

(2)搜索适应性：由于互联网上网站建设大都处于建设中，其服务内容和方式都在不断地变化，特别是在虚拟企业中，由于提供各种资源和服务的网站的动态性和可重构性，尤其是在原有联盟解体新的联盟形成时，服务的提供者的数量和类型将发生巨大的变化。而CORBA体系良好的系统可重构性恰恰满足了这一要求。假设当一些服务提供者退出，同时又有一些新的提供者加入时，系统的最大响应时间为R；而R是N的函数，此时，由于系统瞬间的请求过量，使得式(1)已无法满足。我们设这一阶段最大时间消耗t时间内的系统容量为：

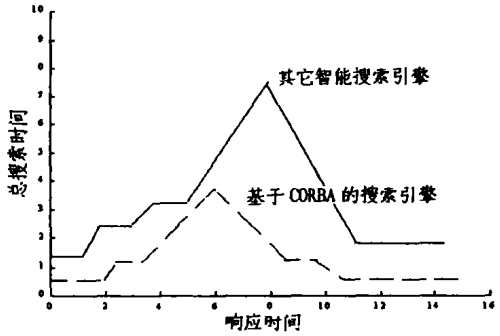

图6　搜索适应性对比图

$$E_x(t)=N(t)/D(t)-1/P(t) \quad (2)$$

则系统最大等待队列长度为：

$$Queue(t)=\max(0,Queue(t-1)-E_x(t)) \quad (3)$$

表明在CORBA系统中系统能迅速均衡请求的供求关系，使其快速稳定下来。而在其他系统中，这一过程不但很慢，而且系统会因等待队列会不断地加长而崩溃。其对比结果见图6所示。

结语　本文给出了一个面向虚拟企业的专业搜索体系，其对于流程企业用户同样适用。因为本系统采用了专业引擎与分布多引擎技术大大提高系统的信息搜索速度，又由于设计了服务器/客户端两极滤波器，不但保证了搜索结果的高准确度还大大提高了滤波效率。智能搜索与专业Agent的结合，有效地解决了虚拟企业中异构网络环境和动态信息源间的资源交互问题。本系统已在网络制造系统中被广泛地应用，其良好的搜索结果有力地证明了上述理论的分析。通过应用可知，本系统相对于现有采用普通代理和服务器单端滤波技术的搜索系统在对变化环境的高适应性和对网站变化的敏感性等方面有着明显的优势。同时，它成功地为用户提供了一个定制的搜索界面，操作简单方便，适用于各类非专业用户。是实现虚拟企业信息的检索和共享的有力工具。

参考文献

1　O'leary D E, et al. Artificial Intelligence and Virtual Organizations [J]. Communications of the ACM, 1997, 40(1)

2　Joachims T, et al. Web Watcher: A Tour Guide for the WWW [DB/OL]. http://www.cs.cmu.edu

3　汪晓岩，胡庆生，李斌，庄镇. 面向Internet的个性化智能信息检索[J]. 计算机研究与发展，1999，36(9)

4　邹涛. 信息的采集、文档的识别与分类[N]. 计算机世界日报，1999(5)

5　Green S, Cunningham P, Somers F. Agent Mediated Collaborative Web Page Filtering[DB/OL]. http://Citeseer.nj.nec.com

6　焦文品，史忠植. 构造MAS的动态体系结构的模型[J]. 软件学报，2000，23(7)

7　Viola M P. The AGENT0 Manual: [Technical Report STAN-CS-91-1389]. Department of Computer Science, Stanford University, Stanford. CA. USA. 1991

8　Egyhazy C J, Plunkett Jr T K, David M. Thompson. Intelligent Web Search Agents[DB/OL]. http://Citeseer.nj.nec.com

9　Decher K, Sycara K, Williamson M. Middle-Agents for Internet. [DB/OL]. http://Citeseer.nj.nec.com