计算机科学2003Vol.30№.5

面向虚拟企业联盟的多级信息滤波体系设计与实现*)

田力威 尹朝万
(中国科学院沈阳自动化研究所 沈阳110016)

To Design and Implement the Multi-Level Information Filtering System for the Virtual Enterprise

TIAN Li-Wei YIN Cao-Wan
(The Shenyang Automatic Institute of the Chinese Academy of Sciences, Shenyang 110016)

Abstract This paper describes a multi-level filtering information based on the CORBA for the virtual enterprise. The distributed agent and artificial intelligent technology are used to solve the filtering system to adapt to the environment that the user's interest and the resource of information change frequently. This paper raises the function and inference algorithm for the any level of the filtering system. During creating the yellow page of the virtual enterprise, this paper takes the authoritative influence of the leader of an alliance into account. Moreover, the paper shows the better filtering result of the multi-level filtering system in the situation which has a lot of information in a dynamic distributed environment by comparing the filtering performance between other system and this system in the network of the virtual enterprise.

Keywords VE(Virtual Enterprise), Multiple agent, Information filtering, CORBA

1. 引言

虚拟企业联盟是一个基于广域网络的动态信息集成系统。它的特点是信息系统构成的动态性和可重构性，需要根据用户的需求和企业的成员情况，高效率、高精度地提供相应信息资料。然而，随着企业联盟的不断扩大，用户需求的不断增加，网上查询需求及相应的信息搜索结果成指数地膨胀。在这种情况下，根据用户的需求，在动态的信息搜索结果中，利用信息滤波器准确地提取用户感兴趣的信息，有效地屏蔽信息"噪声"则变得越来越重要。

目前，对于信息检索系统中的信息滤波功能，大多数系统都是用机器学习和人工智能方法实现的。其一般分为三种方式：即基于内容的滤波(content-based filtering)、协作滤波(collaborative filtering)和经济滤波(economic filtering)[1]。基于内容的信息滤波方法是首先假设每个用户是相互独立操作的，利用矢量空间模型和词频测量方法、相近语义索引等方法实现搜索关键词与搜索到文档相关性计算，进而实现滤波；协作信息滤波方法的思想是假设任何人的兴趣都不是孤立的，应属于某一群体。因此，它根据拥有相同或相近兴趣对相关文档的评价，向其他用户进行推荐，进而抑制不感兴趣文档。

然而基于内容的滤波、协作滤波都要依赖事先定义好的用户模板或公共模板计算匹配性来实现滤波，这样往往存在"信息噪音"。因此，如何使滤波工作更个性化则成为近年来信息滤波研究的重点。文[2]针对用户兴趣的可变性和文档动态变化所产生的不确定因素，提出了一种层次式的智能信息滤波模型。其基本思想是将整个计算任务分解成若干子任务，利用信息检索中已经使用的技术以及人工智能有教师学习实现信息滤波功能。然而，它的实现机制仍是基于内容的滤波和个性化信息交互技术的结合。

综上所述，如将上述滤波方法应用于虚拟企业联盟系统中，存在以下问题：

(1)虚拟企业内部是相互合作与信息相互共享的信息集成体。因此，基于内容的滤波方法无法保证成员间的协作和信息共享。

(2)虚拟企业内部，盟主对资源分类推荐的权威性和被其他用户的认同程度都远远大于一般用户，因此一般的协作滤波无法保证盟主的权威性和用户兴趣的一致性的统一。

(3)虚拟企业成员在不断地加入和退出，因此其信息资源也不断地变化。目前的滤波机制无法适应这一高动态环境。

(4)虚拟企业成员的兴趣一般会根据联盟的动态变化而发生变化，即其在某一时期内的兴趣域是一定的。目前的滤波机制无法保证用户个性化特征的时效性和变化性的统一。

随着分布式人工智能技术的不断发展，特别是CORBA体系的不断完善，利用CORBA技术实现对信息滤波系统的智能化封装，进而实现多级智能化分布式信息滤波，成为代替人们从事繁杂信息收集、过滤、聚类以及信息融合的有力工具。本文即是利用分布式人工智能技术和CORBA技术，提出了一个面向虚拟企业联盟的智能化滤波器，有效地解决了虚拟企业成员个性化信息检索对实效性与变化性的要求。

2. 多级滤波器的总体结构

多级滤波器的整体结构如图1所示。该系统从物理分布上可分为用户端和服务器端两部分。

2.1 用户端

用户端软件为用户通过网络下载到本地安装的应用软件，其功能涵盖用户层的全部职能。用户不但可以通过用户端界面向系统提出查询请求，还可对查询结果满意度和兴趣度提出评价形成感兴趣文本集合$S\{I_1, I_2, \cdots, I_n\}$和不感兴趣文

*)本文得到沈阳先进制造基地基金课题"基于网络的设计制造管理技术研究与实现"(CX01-02-01)的资助。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。

本集合 $S\{P_1,P_2,\cdots,P_n\}$。此外，用户还可以根据系统提供的虚拟企业资源黄页和利用相关工具，对网上资源进行分类和重组，进而形成符合用户兴趣取向的个性化虚拟企业资源黄页和基于领域的兴趣集 $S\{I_1^w,I_2^w,\cdots,I_n^w\}$ 和非兴趣集合 $S\{P_1^w,P_2^w,\cdots P_n^w\}$。从而提高信息搜索的目的性，同时也有效地抑制不相关领域的信息干扰。用户的查询请求则利用 ORB 技术实现封装，并通过用户代理提交给 CORBA 总线层。此外，用户代理还将用户的兴趣方向提交给服务器端，以便形成符合用户需要的虚拟企业资源分类。

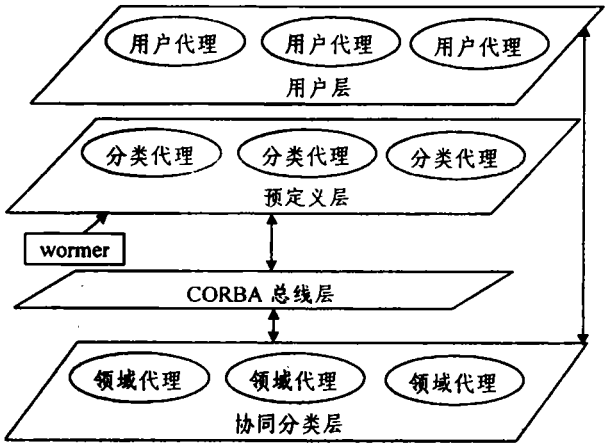

图1 多级滤波总体结构

2.2 服务器端

服务器端软件运行在分布式服务器上。(1)预定义层一方面利用"蠕虫"及时发现和更新每个成员企业的资源变化，以保证信息源的时效性；另一方面，虚拟企业内的信息资源虽然都以文档形式体现的，但由于成员类型不同，该层需将异构的文档进行统一；并通过 CORBA 协调机制利用用户代理实现基于内容的初步文档分类，然后按照其与某个领域的相似度形成面向领域的信息索引，并存入系统索引库。(2)协同分类层，在预定义层生成的预定义分类的基础上，根据用户在一定时期内的兴趣范畴，特别是要充分考虑盟主在决定文档类别时的权威性影响及联盟企业内其他用户的推荐，通过协调机制协调各个领域代理形成公共虚拟企业黄页。

3. 个性化用户代理

3.1 个性化用户代理基本功能

用户通过客户端登录系统向搜索引擎表明个人身份，并通过用户界面向搜索系统提出搜索请求、对虚拟企业黄页中领域分类的评价和对搜索结果的兴趣度评价。运行于客户端的个性化用户代理则提取这些信息，一方面它通过推理向系统提供用户某一时期的兴趣取向，并基于这些取向对搜索结果分类提出推荐意见；另一方面则将用户在某一时期的兴趣情况存入本地的用户兴趣索引库，以形成本地的个性化虚拟企业黄页。

3.2 个性化用户代理的结构

由于个性化代理不仅对环境的变化产生反映，还会受其它代理行为的影响。根据活动=f(状态，个性)[3]可得到个性代理的基本结构为：

IA={信念(belief)，意愿(intention)，能力(capability)，承诺(commitment)，行为(act)，个性(personality)}

其中个性(personality)=(PS,PO)，指的是对自己和其它代理的个性描述；行为(act)=(AS,AC)是指自己信念、意愿、承诺等方面的行为和与其他代理间的通讯行为。而在虚拟企业联盟中，盟主的行为对其它用户的影响是十分大的，而其它用户间的影响则较小，一般用户对盟主的用户代理的行为影响几乎为零。基于以上分析，结合文[3]中给出的3种不同的意图承诺机制，本文提出2种用户代理行为函数：

(1)盟主个性化用户代理行为函数：

$AS_{goal}:AS\times PS_{goal}\rightarrow interest(c)$

$AC_{goal}:AC\times PO_{goal}\rightarrow 0$

(2)普通用户个性化用户代理行为函数：

$AS_{goal}:AS\times PS_{goal}\rightarrow interest(c)$

$AC_{goal}:AC\times PO_{goal}\rightarrow interest(c)$

$AC_{goal}:AC\times H-PO_{goal}\rightarrow H-interest(C)$

其中 $H-PO_{goal}$ 为来自盟主的个性化描述；$H-interest(C)$ 为盟主的个性目标偏好；interest(C)根据文[4]提出根据个性心理学中个性特质，代理的个性倾向是指代理对不同个性目标偏好程度的定义。本文利用基于 CORBA 的概念化定义对象 $C=\langle D,W,R\rangle$，其中 D 为文档集合；W 为领域集合；R 为领域空间上的兴趣关系集合。特设个性倾向评价函数为 $INTEREST\in\{5,4,3,2,1,0,-1,-2,-3,-4,-5\}$。其中正数表示用户某一时期对某个和某类文本感兴趣的程度；负数表示用户某一时期对某个和某类文本不感兴趣的程度。

3.3 个性化代理功能实现机制

个性化代理功能实现的核心是如何根据环境的变化和用户需求来改变思维状态，产生意图和实现目标，因此个性化代理实现机制的核心是如何实现个性代理承诺某一规划以实现某个任务。因为用户代理的承诺刻画了代理对其未来行为的某种决策。从行为角度上看，承诺概念较意愿概念具有更强的约束，因为代理的意愿仅仅体现了用户代理对其未来行为的某种合理选择。具体实现为：

```
main(query,Interest)
{
  if User_agent(user)
  {
    Initialize(User_agent(user),query);
    GetUsrProfile(User_agent(user), Interestv);
    User_agent. Rank(user)= PROFILE
  }
Act_Process()
  {
    creategoal(query,Interest)=goal
    while(commite(plan,implement)= true)
    {
 selectplan(goal)=plan
 if(implement(plan)=goal)
  {
     start{
   CORBA_Orbix_is_ready(Search(user_agent. Rank(visited)); }
   Catch(CORBA_SystemException&SysEx){
        …… }
   reselectplan(goal)
   continue;
    }
    break;
   }
   recommite(plan,implement)
   return;
   }
   }
```

4. 文档的预分类

由于企业联盟的成员不断加入或退出，虚拟企业联盟内部的信息源是一个动态变化的分布式系统。因此，为提高虚拟企业黄页的建设效率，系统将对网络蜘蛛或网络蠕虫周期性发现的静态网页和文档进行初步的分类(在虚拟企业内，资源

的表达方式一般是以静态网页或文本形式提供给搜索系统的)。要解决文档是如何在不同领域间分布的问题,一般通过确定文档属于某一领域的概率有多大来判断。该功能是通过服务器端的预定义层实现的,其具体实现机制如下。

静态的网页或文档在不同领域中符合 $Zipf$ 分布,即 $q_i=k/i^{(1-x)}$,其中 q_i 是第 i 个文档出现的频率,i 为根据访问频率降序排列的索引,k 为一个常数,$0\leqslant x\leqslant 1$为参数。由于无论是静态网页还是文档,其标题均代表文档的关键内容,而且由于标题较短,易于识别。因此根据文档标题对文档进行分类,既有代表性又减少计算量。由上述分析可知文本在 n 个不同领域中的分布均采用 $Zipf$ 分布。现设:$T_i=\{T_1,T_2,\cdots,T_i\}$为文档标题集合,$W_n=\{W_1,W_2,\cdots,W_n\}$为预定义的领域分类集合。利用余弦定理可得到类别与文本相似度为:

$$SIM(T_i,W_n)=\sum_{i=1}^{i_{max}}\sum_{n=1}^{n_{max}}T_i\cdot W_n/\sqrt{\sum_{i=1}^{i_{max}}(T_i)^2\cdot\sum_{n=1}^{n_{max}}(W_n)^2}$$

由上式可得某文档与某一类领域的相似度较大,而与另一类领域的相似度则较小。由此可得类别相似度区分评价:$L_i=\sum_{k=1}^{n}(SIM(T_i,W_k)-SIM(T_i,W_{k+1}))$,其中 L_i 的绝对值越大说明文档属于两个领域的相关性就越小;而 L_i 的绝对值越小说明文档属于两个领域的相关性就越大。当 $\alpha\leqslant L_i\leqslant\beta$ 则文档同时属于两个领域,其中 α,β 分别为系统预设的上下阈值;当 L_i 值不在相关区间内则文档属于相似度大的那个领域。通过上述计算可形成虚拟企业信息资源的预定义分类。

5. 公共虚拟企业黄页的建立

公共虚拟企业黄页系统是系统提供给用户的信息资源公共分类集合。它是充分考虑用户和盟主的兴趣取向,在预定义分类的基础上,利用分布式领域代理间协调机制形成的资源分类。在每一个分类中,文档按其兴趣度进行排序,此外由于其是基于移动代码机制设计的,用户可利用用户端软件对公共虚拟黄页的结构进行调整,形成符合个性需要的用户虚拟企业黄页。其实现机制为:

设时间 t 被用户查询的特征向集为 $G=\{G_1,G_2,\cdots,G_n\}$,预分类定义为 $W_n=\{W_1,W_2,\cdots,W_n\}$,项 G_i 在 W_k 中出现的频率为 f_k,由此可得用户感兴趣类别在观察时间 t 内的利用率为 $U_{ik}=\sum_{i=1}^{n}f_i/\Sigma t$,其中 Σt 表示观测时间的和。按 U_{ik}的大小进行排序,以得到某个用户的兴趣取向索引。结合用户对某一文档在一定时期内的兴趣度评价,可得到用户对文档 d 在一定时期内的兴趣推荐:$W'_{di}=\alpha\times U_{ik}+\beta(\sum_{i=1}^{n}SIMM(W_i,T_i)/\Sigma t)$,其中 T_i 为推荐文档的标题特征项集合,α 和 β 为比例系数。按 W'_{di}的大小进行排序可得到用户对某一文档兴趣度的评价。将这一结果存入虚拟黄页分类检索数据库。由以上可得到用户对文档及其分类的建议。而在虚拟企业中盟主的权威性表现为它对资源分类和兴趣度的评价具有一定的代表性。即它的意见将被其他用户所采纳,并成为构建公共虚拟企业黄页的核心因素。因此考虑盟主的影响力可得到对于任意文档的分类和兴趣度的协同评价:

$CL(d)=\omega_0+\sum_{m=1}^{m_{max}}\sum_{i=1}^{n}k_mW'_{di}$,其中 ω_0为盟主对文档 d 的评价,k_m 为第 m 个用户评价的比例系数。通过以上过程可得到既符合要求又符合所有用户建议的资源评价集合 CL。基于此集合可构建公共虚拟企业联盟黄页系统。

6. 实验分析

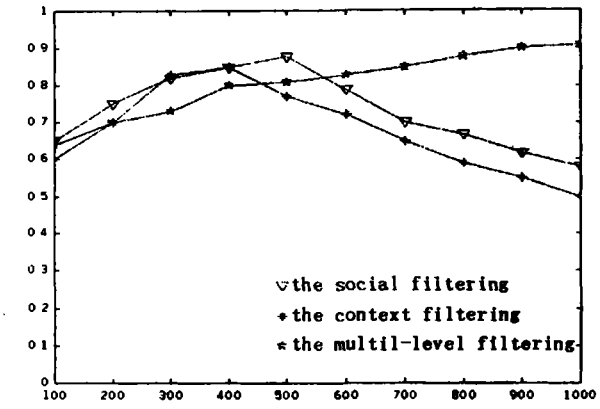

图2 三种滤波方法性能对比

本文将基于内容的滤波系统、社会滤波系统和多级信息滤波系统分别应用于虚拟企业联盟实验环境中,其性能对比情况见图2。从实验结果中可以发现,在小信息量、少用户需求的情况下,社会滤波系统滤波效果较好。但随着系统信息量的增加,社会滤波系统和内容滤波系统的滤波效果则明显下降,而多级滤波系统的效率则明显提高,并稳定在一个较高的滤波准确度水平。其原因是当信息量少时,多级滤波系统学习推理过程中的学习样本较少,因此误差相对大;而其他两种滤波方式则利用预先定义好的个人模板和公共模板进行滤波,因此准确性较高。随着多级滤波系统学习量的不断增加则其滤波准确性也随之提高,特别是系统对用户和盟主的兴趣取向有了明确的认识则系统滤波效果将进入相对稳定的高准确阶段;其它两种滤波方式因不能动态地适应用户和信息源的变化情况,其准确度将明显下降。

结语 本文在深入讨论现有滤波技术的基础上,根据虚拟企业联盟的特殊需求,提出了一个面向虚拟企业联盟的信息滤波系统。由于其充分考虑了用户个性化需求和盟主在虚拟企业黄页建立过程中的权威性影响力,使本系统能更好地适用于虚拟企业联盟中的动态信息滤波需要。本文给出了不同信息滤波系统在实验系统中运行情况的对比,验证了多级滤波系统在大信息量、动态分布式信息源环境下的卓越滤波效果。但由于其学习推理算法的限制,其小信息量、相对稳定环境下的滤波效果还不够理想。

参考文献

1 汪晓岩,等. 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)

2 Mostafa J,et al. A multilevel approach to intelligent information filtering: Model, system, and evaluation[J]. ACM Trans. on Information Systems 1997, 15(4): 368~399

3 Rao A S,Georgeff M P. Modeling rational agents within a BDI-architecture. in: allwn J, Fikes R, Sandewall W eds. principles of Knowledge Representation and Reasoning. California: [J]Morgan Kaufmann Publishers,1991. 473~484

4 徐晋晖,等. 一个具有个性的 Agent 实现机制[J]. 计算机发展与研究,2000,38(6)