

一种基于智能 Agent 的科技文献快速协作推送机制

徐小龙 赵昌耀 耿卫健 程春玲

(南京邮电大学计算机学院 南京 210003)

摘要 针对目前网络科技文献共享平台上汇聚的海量的科技文献,提出了一种新的基于智能 Agent 的科技文献快速协作推送模型,通过引入由智能 Agent 技术构建的用户 Agent、数据传递 Agent、信息评估 Agent、信息分析 Agent、服务器监控 Agent 和数据挖掘 Agent,在客户机和服务器之间建立一种有效的协作关系,联合为每个用户主动提供最符合其需求的科技文献。提出了一种新的文献-读者契合度评估算法,算法综合考虑用户主动提供的订阅信息、用户的阅读历史情况、文献价值客观评价以及文献发表的先后时间,获取契合度值向量对文献与读者的匹配情况进行客观的评估,从而在准确获得用户真实兴趣和关注领域的情况下及时为用户提供最有价值、最符合用户需求的高质量科技文献信息。分析了目前的网络信息获取和传送方法,然后介绍了基于智能 Agent 的科技文献快速协作推送系统模型,最后详细探讨了文献-读者契合度评估算法及契合度向量的各个参数的计算方法。

关键词 科技文献,智能 Agent,信息推送,协作

中图法分类号 TP391 **文献标识码** A

Rapid Collaborative Scientific and Technical Literature Push Mechanism Based on Intelligent Agent

XU Xiao-long ZHAO Chang-yao GENG Wei-jian CHENG Chun-ling

(College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract The current scientific and technical literature network-sharing platforms have a mass of papers, which is difficult for users to get what they really need. We presented a new Rapid collaborative scientific and technical literature push model based on intelligent Agent. This model adopts Intelligent Agent technology to build the User Agent, Data Transfer Agent, Information Evaluation Agent, Information Analysis Agent, Server Monitor Agent and Data Mining Agent, to establish an effective collaborative relationship between clients and servers, provide scientific and technical literatures jointly for each users meeting their needs perfectly. We also proposed a Literature-Readers Matching Assessment Algorithm(LRMAA), which is to comprehensively consider the subscription information provided by user initiatively, the user's reading history, the objective assessment of literature as well as the literature publishing time, to get the matching vector, resulting in a true and accurate access to the user's research interests and focuses in time to provide users with the most valuable, suitable and high-quality scientific and technical literature information. This paper first analyzed the current network Information Pull and Push methods, and then introduced the Intelligent Agent-based rapid collaborative scientific and technical literature push model, at last, discussed LRMAA in detail about all parameters computing of the vector.

Keywords Scientific and technical literature, Intelligent Agent, Information push, Collaboration

1 引言

目前科技文献共享平台^[1-5](如国外的 DOAJ, J-STAGE, BMC, SciELO 和 PubMed Central 等,以及国内的中国科技文献在线、中国期刊网、万方数字化期刊和维普中文科技期刊数据库等)已经汇聚了海量的科技文献,关于同样或相似主题(尤其是热点主题)的论文常常也很多,并且每天均会增加大

量的新的论文(例如“中国科技文献在线”有效注册用户数万人,总访问量单以 IP 地址计算,已达百万人次,日均访问量达万人次,在线发表的论文数为每月数百篇,且每月发表论文篇数不断上升^[6]),但是这些论文的质量良莠不齐。查阅论文的科技工作者需常常主动登录网站,并在网站搜索引擎的协助下,半手工地搜索信息,这常常需要花费大量的时间,有时还找不到需要的论文,也不能及时追踪到最新、最有价值的论

到稿日期:2010-05-10 返修日期:2010-08-10 本文受教育部科技发展中心网络时代的科技论文快速共享专项研究资助课题(2009117),教育部博士点基金(20093223120001),江苏省科技支撑计划(BE2009158)以及江苏省高校自然科学基金项目(09KJB520010)资助。

徐小龙(1977—),男,博士,副教授,硕士生导师,主要研究方向为网络计算、智能 Agent、信息安全技术等,E-mail: xuxl@njupt.edu.cn; 赵昌耀(1987—),男,学士,主要研究方向为基于网络的计算机软件技术等; 耿卫健(1987—),男,学士,主要研究方向为基于网络的计算机软件技术等; 程春玲(1972—),女,硕士,副教授,硕士生导师,主要研究方向为计算机应用、信息安全技术等。

文。信息主动推送技术由信源主动、及时地向用户推送不断更新的动态信息,但是现有系统推送的内容难以满足各种用户不同的个性化需求,这是当前需要重点解决的问题。

针对这种现状,本文的贡献在于:提出利用智能 Agent 技术来实现个性化快速协作推送机制,使得提供的论文推送服务内容不再是千篇一律,而是各得其所。本文还提出了一种文献-读者契合度评估算法,算法综合考虑了用户主动提供的订阅信息、用户的阅读历史情况、文献价值客观评价以及文献发表的时间,从而获取契合度值向量对文献与读者的匹配情况进行客观的评估。结合文献-读者契合度评估算法和快速协作推送机制的科技文献主动推送系统不仅能按照用户明确的要求提供信息,而且在用户并未有意识参与的情况下能自动跟踪用户访问各种信息留下的痕迹并进行分析,从而在准确获得用户真实兴趣和关注领域的情况下及时为用户提供最有价值、最符合用户需求的高质量科技文献信息。

2 相关工作

目前的基于 Internet 的信息获取和传送方法及技术从“信源”与“用户”的关系来看可分为两种模式^[7,8]:

(1)信息拉取模式(Information Pull),由用户主动地、针对自己的需求、有目的地从信源查询、搜索所需的信息,信息系统只是被动地接受查询,提供用户所需的部分信息。

(2)信息推送模式(Information Push),由信源主动、及时地向用户推送不断更新的动态信息,用户被动地接受服务器主动发来的新信息。

信息拉取模式的主要缺陷是及时性差,当信源中信息更新变化时,用户难以及时拉取新的动态信息;对用户要求高,要求用户掌握信息查询甚至高级检索技术。而信息推送模式一般是利用频道式的网络广播软件、电子邮件系统、网页信息推送软件等向用户主动地发布、推送各种信息。用户像收听电台广播或看电视那样,可选择自己感兴趣的频道或节目,有目的地获取信息。

但是,目前的信息推送软件主要缺点是推送的信息内容常常缺乏针对性,没有满足用户的个性要求,而且信源服务器任务重,信源系统要主动地、快速地、不断地将大量信息(很多信息是用户不需要的)推送给用户。个性化信息服务首先是能够满足用户的个体信息需求的一种服务,其次还应该是一种帮助个体培养个性、发现个性、引导需求以完成用户特定任务,同时提高检索效率的服务,主要包括服务时空的个性化、服务方式的个性化、服务内容的个性化。实现个性化信息服务需要采用先进的智能化技术。

3 基于智能 Agent 的科技文献快速协作推送系统模型

智能 Agent^[9,10]是一种能够完成委托任务的计算机系统,具有自治性、社会性、反映性、能动性智能特性;采用 Agent 技术可将一个静态的系统转变成一个有效的由用户驱动的动态信息平台。开放分布式网络环境下的 Agent 系统中,每个节点上都有相对独立的智能 Agent。多个智能 Agent 之间彼此在逻辑上相互独立,通过共享知识、任务和中间结果,协同在工作中形成问题的解决方案。分散意味着控制、运算数据、知识在逻辑上或者地理上是分布式的,从而可以分担

负载;而松散耦合表明每个智能 Agent 均将更多的时间用于自主计算或者知识处理,在网络上传递的仅仅是由具有移动性的 Agent 所携带的最终结果,而不是中间内容。智能 Agent 技术应用于网络科技文献共享系统可以获得更高的性能。

基于智能 Agent 的科技文献协作推送系统模型的运行平台分为 3 个区域:用户域、门户服务器、数据服务器域(一般采用集群(Cluster)进行数据分布式存储和处理),其上运行着相同的 Agent 运行支撑环境和不同功能的 Agent:

(1)用户 Agent(UA, User Agent) 驻留于用户客户机上,负责管理用户订阅信息、用户阅读领域和其它个人信息,代替用户向服务器提交能尽量反映用户真实需要的论文信息需要请求,向服务器主动报告用户是否在线等情况,将服务器推送给用户的文献信息按对论文的客观评价进行先后排序;

(2)数据传递 Agent(DTA, Data Transfer Agent) 负责在用户客户机和服务器之间以及服务器直接传递各类数据信息;

(3)信息评估 Agent(IEA, Information Evaluation Agent) 驻留于服务器端,负责根据指定的策略来对文献价值根据其引用次数等参数进行客观评价,评价结果存储于服务器端;

(4)信息分析 Agent(IAA, Information Analysis Agent) 驻留于门户服务器上,将由驻留于数据服务器上的服务器监控 Agent 派来的 DTA 所携带的数据,再进行分类整理,并按照一定的方式存储至索引数据库中,或是更新索引数据库中的内容;

(5)服务器监控 Agent(SMA, Server Monitor Agent) 驻留于数据服务器上,负责监视论文更新情况,并将相关信息(尤其是新增论文信息)通过 DTA 主动报告给门户服务器上的 IAA,将信息变动情况即时通知 IEA;

(6)数据挖掘 Agent(DMA, DMA Mining Agent) 驻留于门户服务器上,负责根据由 UA 派来的 DTA 所携带的用户信息以及论文信息,通过内在的算法从数据库中挖掘出用户真正需要的论文信息,并通过 DTA 推送给用户。

4 文献-读者契合度评估算法 LRMAA

科技文献快速协作推送系统中应用智能 Agent 技术可将用户订阅信息、用户阅读情况、文献价值客观评价等进行综合考量,以向用户提供最符合用户需求的文献。

(1)用户订阅信息

用户首先需要向系统提供自己的研究领域、研究方向、研究兴趣和目前所关注的研究热点,这表明系统首先尊重用户的主观意愿,可以有效缩小检索的范围,提高搜索的效率,也使得最后推送给用户的信息更加具有针对性。

(2)用户阅读情况

系统不但要接收用户的订阅信息,还要记录用户的历史阅读信息,并从中挖掘出用户的真正需求,从而能够将用户需要的科技文献及时、准确地发送给用户。

(3)文献价值客观评价

由于同样研究主题的论文,其价值本身有较大的差异,一般价值较高的论文将获得较高的阅读率和引用率以及评价,在将论文或摘要信息推送给用户时,应该在排序过程中获得优先的位阶。目前对于科技期刊的评价已经有一些研究成

果,比较广泛的是采用影响因子等方法来评价科技期刊的价值,但是针对某一文献的评价机制则研究得还不够。

本文提出一种文献-读者契合度评估算法(Literature-Readers Matching Assessment Algorithm, LRMAA),采用契合度值对某一读者的匹配情况进行评估。契合度值为向量 M :

$$M \rightarrow \langle S, H, E, T \rangle \quad (1)$$

式中, S 为用户主动提供的订阅信息(subscription)与文献的匹配度; H 为从用户的阅读历史情况(History)挖掘出的与文献的匹配度; E 为文献价值客观评价价值(Evaluation); T 为文献发表的时间(Time)。由于用户有指定搜寻某个时间段信息的需求,文献发表的时间 T 值可由用户自行指定,在推送的结果中由系统简单地判断即可,因此下面将重点描述 S, H, E 的确定方法。

4.1 S 值的确定

为了使向用户主动推送的文献信息能够体现用户的主观意愿,首先需要用户在初始时主动提供需求。用户可以在如表 1 所列的项目中进行选择和设置。

表 1 用户指定文献范围列表

项目	说明
技术领域	读者可以设置与自己研究方向相关的技术领域。可以设置技术领域 1、技术领域 2、技术领域 3、……,如读者可设置(计算机网络、分布式计算、信息系统)。
发表时间	读者可以设置希望获取的论文的发表时间段。由于技术进步和更新速度越来越快,一般读者希望获取到较新的文献资料。如读者可设置文献发表时间为“2009 年至目前”。
期刊类型	针对科技文献而言,有几种典型的划分方法,例如可以将文献划分为 SCI 来源期刊、EI 来源期刊、核心期刊等。
文献作者	读者可以设置希望获取的论文的作者身份层级(学历、职称等)。
成果区域	国内、国外(可指定是哪个国家)以及所用的语种。

上述信息由用户 Agent 来维护,主要负责存储和更新等。我们假定所有的技术领域集合为 $ResearchField$,其中包含了若干领域:

$$ResearchField = rf_1 \cup rf_2 \cup rf_3 \cup \dots \cup rf_i \cup \dots \quad (2)$$

$$rf_i = \{k_1, k_2, k_3, \dots, k_r, \dots\} \quad (3)$$

式中, rf_i 为某一技术领域,模型中为一个集合。集合中的 k_r 元素是包含于 rf_i 集合的词条,即该技术领域可能出现的关键词,这由客户机上的知识库提供,并由系统初始化 k_r 元素所属的集合,在后续过程中,将随着用户的使用情况和服务器的内容进行更新。 k_r 在 rf_i 中的权值为 $w_r(rf_i)$,权值大小的差异取决于 k_r 在各个技术领域中分布的程度:如果 k_r 仅存在于其中某一个 rf_i 中,就表明由 k_r 可完全确定该领域 rf_i ,权值自然最大;如果 k_r 分布于每个领域集合,则 k_r 不具备确认任何领域的的能力,权值自然最小。文献对于 rf_i 就可分为相关文献和无关文献^[1]。

UA 通过与用户提交的研究领域信息,获取用户选定的领域集合 $Focus$,由 m 个单类领域集合组成,代表该用户感兴趣的 m 个研究领域:

$$Focus = rf_1 \cup rf_2 \cup rf_3 \cup \dots \cup rf_m \quad (4)$$

式中, $Focus \subset ResearchField$ 。UA 根据用户的研究领域集合 $Focus$,利用概率统计模型,可得 k_r 在 rf_i 中的 $w_r(rf_i)$ 的值^[11]:

$$w_r(rf_i) = \frac{y/(Y-y)}{(n-y)/((N-n)-(Y-y))} \quad (5)$$

式中, N 是文献数据库中与领域子集合 $Focus$ 相关的文献数量, n 是文献数据库中与用户设定的领域集合 $Focus$ 相关且包含词条 k_r 的文献的数量, Y 是与 rf_i ($rf_i \subset Focus$) 相关的文献数目, y 是与 rf_i 相关而且包含 k_r 的 rf_i 数目。

S 值的选取可以有效缩小检索的范围,即服务器可根据用户设定的技术领域,将推送给用户的文献限制在符合用户指定的技术领域内;同时,再根据用户设定的“发表时间”、“期刊类型”、“文献作者”、“成果区域”等进一步缩小文献的范围,从而减轻服务器和网络通信设施的负载,提高搜索的效率。但是由于研究点的快速更新,选项无法设置得很详细。为了使推送给用户的信息更加具有针对性,就必须根据用户的历史阅读信息进一步挖掘出用户的真正需求。下面讨论反映用户阅读历史情况的 H 值的计算方法。

4.2 H 值的确定

用户阅读历史情况可以表现为用户当前重点检索和查阅的某一研究热点及其相关领域技术内容的情况。如用户当前重点关注“云计算”领域的相关内容,在查阅文献时查询关键词通常是“云计算”或是其中相关的关键技术,如计算机集群技术、分布式数据挖掘、聚类与分类技术等,而且这些检索关键词出现的频次会比较高。但是用户阅读历史会表现为短期历史和长期历史。长期历史会表现为一个大范围的概念,一般应该与用户的总的研究方向相吻合。例如该读者长期从事网络和分布式计算领域的研究,但可能以前研究的领域为“网格计算”或“对等计算”,近期研究“云计算”,即用户的阅读重点(即研究重点)发生了迁移,因此应重点关注用户的短期历史情况,方法是设置相应的“时间窗口”。可以把长期阅读历史看成是由一段段的短期阅读历史组成的,通常相邻的短期阅读历史具有一脉相承的延续性特征,如图 1 所示。

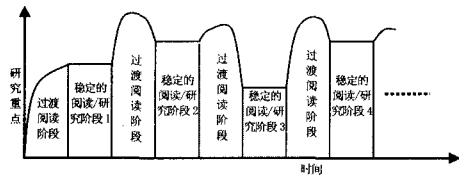


图 1 研究重点迁移示意图

经过一段不确定的带有试探性的文献查询和阅读后,用户逐渐确立自己的研究重点和研究内容,选择阅读的文献会稳定集中于某些确定的研究点,与此相关的重要词汇在查阅关键词和阅读文献中出现的频次常常会比较,并且具有较高的区分度。

设在时间段 Δ 期间(即某一稳定的阅读/研究阶段),用户经常使用的查询关键词及其使用频次为一个集合:

$$S_{\Delta} = \{(t_1, \omega(t_1)), (t_2, \omega(t_2)), \dots, (t_n, \omega(t_n))\} \quad (6)$$

式中, t_i ($i=1, 2, \dots, n$) 为一列独立词条项, $\omega(t_i)$ 为 t_i 在时间 Δ 期间被用户使用的频次,显然被用户反复使用的词条代表了用户关注的重点。因此设定某一阈值 α ,要求 $\omega(t_i) \geq \alpha$ 时, t_i 才能作为反映用户近期需求的稳定关键词。UA 负责维护用户的短期和长期阅读情况。

用户很可能同时关注不同的领域,如用户如果作为高校的科研工作者承担着某一技术领域的科研任务,同时承担某一门课程的教学任务并正在进行教学研究工作,那么他所使用的关键词一般都分属于当前关注的这几个领域。但是知识库的信息可能难以及时涵盖许多崭新的信息,不加以遴选的

新信息都添加到知识库中也会使知识库变得过于庞大。如何对本阶段所使用的关键词进行分类呢? 本文采用时间相关性的归类方式, 流程为:

步骤 1 当用户采用某一关键词 t_i 时, 首先判断 t_i 是否唯一属于用户选定的领域集合 $Focus$ 中的领域 rf_i 。如果唯一属于, 则直接确认并转步骤 4, 否则进入步骤 2。

步骤 2 如果 t_i 并非唯一属于用户选定的领域集合 $Focus$ 中的某一领域, 例如同时属于 rf_i 和 rf_k , 则由系统设定一个时间段, 这意味着将当前使用的关键词与他刚刚使用或马上将使用的关键词归并在一起, 即认为有较大概率的关键词 t_i 与临近时间使用的关键词 (设为 t_h) 相关并同属于一个领域。如果 t_h 属于 rf_i , 则将 t_i 和 t_h 归并在一起; 如果 t_i 不属于用户选定的领域集合 $Focus$ 中的任一领域, 则进入步骤 3。

步骤 3 由于当前使用的关键词 t_i 不属于用户选定的领域集合 $Focus$, 也同样由系统设定一个时间段, 将当前使用的关键词与他刚刚使用或马上将使用的关键词归并在一起, 即认为有较大概率的关键词 t_i 与临近时间使用的关键词相关。

步骤 4 由以上可将 S_Δ 按领域和所用关键词相关性分为若干当前关注子集:

$$\begin{cases} S_\Delta = s_1 \cup s_2 \cup s_3 \cup \dots \cup s_i \cup \dots \\ s_i = \{(t_x, \omega(t_x)), (t_y, \omega(t_y)), \dots\} \end{cases} \quad (7)$$

如上, 设 s_i 与用户选择的领域 rf_i 相对应, 然后 UA 会将二元组 $(t_x t_y, \dots, rf_i)$ 通过 DTA 传递给服务器。服务器首先根据 rf_i 来将检索范围缩小为某一领域。基于向量空间模型^[12,13], 可将某一篇文章映射为一个特征向量:

$$V(d) = (q_1, \mu_1(d); \dots; q_n, \mu_n(d)) \quad (8)$$

式中, $q_i (i=1, 2, \dots, n)$ 为一列独立词条项, $\mu_i(d)$ 为 q_i 在 d 中的权值, 一般定义为 q_i 在文献 d 中的词频 f_{id} 的函数。由 TF-IDF 函数可得^[12,13]:

$$\mu_i(d) = \frac{f_{id} \log(\frac{N}{n_i})}{\sqrt{\sum_{k=1}^n (f_{kd})^2 \times \log^2(\frac{N}{n_k})}} \quad (9)$$

式中, N 为文献库中文献的数目, n_i 为含有词条 q_i 的文献数目。关键词查询串 $t_x t_y, \dots$ 与文献 d 之间的匹配程度 S 值为:

$$S(t_x t_y, \dots, d) = \frac{\sum_{i=1}^n \mu_i(d) \times t_i}{\sqrt{(\sum_{i=1}^n \mu_i^2(d)) (\sum_{i=1}^n t_i^2)}} \quad (10)$$

式中, 若关键词 t_i 出现在页面 d 中, 则值为 1, 否则为 0。

4.3 E 的确定

在服务器选取用户当前关注的相关科技领域涉及的文献后, 对文献的价值进行合理评价, 将最有价值的论文优先展现在用户面前, 即通过 E 值的确定来实现对检索结果的排序。但是, 从读者的角度来看, 通过网络共享的文献的价值^[14]可以从论文评审专家意见及评分值、论文读者的评分值、论文的关注度、论文发表期刊的影响因子 (Impact Factor) 和论文收录情况、论文的研究支撑等方面来进行评估。

不同的读者对于这几个因素的重视程度并不一样, 例如有些作者注重论文评审专家意见, 有些则重视论文发表期刊的影响因子, 有些则认为论文的关注度是最重要的因素。

假设某读者认为论文的关注度是最值得关注的因素, 则在推送结果排序时, 需要根据论文被引频次及网络读者的点击浏览率和下载率来计算优先值。设在当前时间段内, 系统

中所有论文总的浏览次数和下载次数累计为 A , 文献 d 的浏览次数和下载次数累计为 $C(d)$, 则容易得到论文的浏览率和下载率为 $C(d)/A$ 。显然, 该值越大, 则表明该论文受到的关注度越高。

而论文被引频次代表的论文价值程度则可以参考 Page Rank^[15,16]的基本思想: 一篇文献被作为参考文献多次引用, 则这篇文献很可能是重要的; 一篇文献尽管没有被多次引用, 但被一篇重要的文献引用, 则这篇文献很可能是重要的; 一篇文献的重要性被均分并传递到它所引用的文献。设 d 是一篇文献, $F(d)$ 是 d 引用的文献集合 ($|F(d)|$ 为集合 $F(d)$ 的大小), $B(d)$ 是引用 d 的文献集合, 由此确定表示文献 d 客观的基于被引频次的论文价值为:

$$R(d) = \sum_{d' \in B(d)} (R(d') / |F(d)|) \quad (11)$$

由此可以综合计算出基于论文关注度的文献评价价值:

$$E(d) = \alpha \cdot C(d)/A + \beta \cdot R(d) \quad (12)$$

式中, α, β 为调节点击浏览率/下载率和论文被引频次两者对于论文关注度计算时的重要程度, 这两个参数应该通过实验来不断地优化确定。按照 $E(d)$ 值的大小来对按照前面 S 和 H 值及用户设定的时间范围从文献数据库中提取出来的资料排序, 然后推送到用户的客户机上。

5 原型系统构建与性能分析

5.1 原型系统

下面介绍应用本文提出的基于智能 Agent 的科技文献快速协作推送机制的实验原型系统的构建方法。本系统由于是分布式的网络应用系统, 因此可以部署在基于 Intranet 的网络平台上。该原型系统可采用 Java 语言和 Eclipse 开发工具来构建。系统可基于 IKV++ 的 Grasshopper^[17] 来构建系统中的各个 Agent。Grasshopper 中的 Agent 分为静态 Agent (stationary Agent) 和移动 Agent (Mobile Agent) 两类。静态 Agent 不能移动但是可以与各种 Agent 通信, 从而实现本文提出的系统中存在的 6 种 Agent。静态 Agent 和移动 Agent 均运行于 Agency 中; 一个 Agency 由 core agency 和一个或多个 place 组成。Core agency 为 Agent 的运行提供通信服务、注册服务、管理功能、迁移功能、安全功能、持久化功能等。

基于 Grasshopper, 科技文献快速协作推送机制的实验原型系统的参考界面如图 2 和图 3 所示。系统分为 3 个部分: 用户客户端、门户服务器端和文献服务器端。用户客户端部分界面如图 2 所示, 用户首先需要向系统提供自己的研究领域和设定的论文发表时间以及限定的论文来源等, 从而使系统可以根据用户的主观意愿来有效缩小检索的范围, 提高搜索的效率, 也使得最后推送给用户的文献更加具有针对性。而对用户的阅读历史情况的记录和挖掘则在系统后台进行。

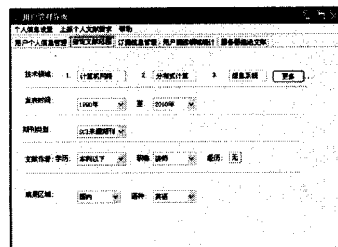


图 2 用户客户端系统部分界面

服务器端部分界面如图3所示。门户服务器端管理用户的信息和文献的相关信息,包括用户的信息和当前状态,这对移动 Agent 的迁移路由等尤其重要;同时需要维护索引数据库,并对文献的价值进行合理评价,得到一个综合评分值,以此作为用户获取所需文献信息列表的排序依据。文献服务器端主要管理具体的文献数据库,并负责对文献数据库进行更新。

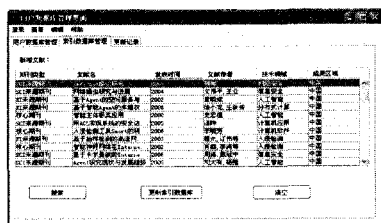


图3 服务器端系统部分界面

5.2 性能分析

本文提出的这种新型的基于智能 Agent 的科技文献快速协作推送机制改变了传统的获取科研文献的方式,在准确性、及时性等方面具有明显的性能优势。下面作具体的分析。

个性化——本文提出的基于智能 Agent 的科技文献快速协作推送机制是在节点与节点之间建立一种协作关系,利用智能 Agent 技术来实现个性化快速协作推送机制,并利用文献-读者契合度评估算法,综合考虑了用户主动提供的订阅信息、用户的阅读历史情况、文献价值客观评价价值以及文献发表的时间,从而获取契合度值向量对文献与读者的匹配情况并进行客观的评估,使得提供的论文推送服务内容不再是千篇一律,而是各得其所。

便捷性——结合文献-读者契合度评估算法和快速协作推送机制的科技文献主动推送系统能自动跟踪用户访问各种信息留下的痕迹并进行分析,从而在准确获得用户真实兴趣和关注领域的情况下及时为用户提供最有价值、最符合用户需求的高质量科技文献信息。以上的工作除了在用户加入系统时要求用户明确地提供自己的相关需求信息,以后则不再需要用户有意识地再次反复提交相关需求,系统就可以自主地、动态地反映用户当前的真实需求,用户的使用体验将更加便捷。

及时性——基于智能 Agent 的科技文献快速协作推送机制与传统的获取科研文献的方式最大的不同在于,用户可以改变以往的信息获取模式,将繁琐的信息追踪、获取、遴选和评判工作交由系统来完成,用户可以更多地精力放在对知识本身的学习、研究和探索上。只要用户在线,当系统中有新的符合用户需求的资源时,就可以立刻推送到用户面前,用户不必担心自己重复已有的研究工作,避免了智力资源的浪费。

负载均衡——从系统模型的构成可以明显看出,本系统是一个协同工作的分布式信息系统。系统模型的运行平台包含了用户域、门户服务器、数据服务器域这3个区域,其上运行着不同功能的 Agent。系统对于信息的推送工作是在不同的 Agent 所建立的协作关系下共同完成的。信息的处理工作和原来集中于检索服务器上的模块功能是由各个节点联合完成和承担的,因此可以避免所有的计算都由服务器承担,造成服务器负担过重而成为性能的瓶颈。

结束语 对于广大的科技工作者来说,能够及时获得最新、最有价值的文献,对于进一步顺利开展科研工作具有十分积极的意义。鉴于这种需求和目前网络科技文献共享系统中存在的可以进一步改进的空间,本文基于智能 Agent 的科技文献协作推送机制实质上是在服务器和客户机之间利用智能 Agent 建立了一种良好的协作关系,并提出了一种新的文献-读者契合度综合评估算法 LRMAA,以共同为用户提供最符合其真实需要的文献资料。

进一步的研究工作思路包括:

(1)构建真正具有实用价值的大规模网络科技文献共享系统,并将本文提出的基于智能 Agent 的科技文献协作推送机制应用到该系统中,以真实观察机制的运行情况。

(2)由于包含作者在线书刊检索模块的图书馆藏系统事实上也常常具有相同的需求,即读者希望能及时了解自己关注领域的书刊出版情况,因此本文提出的基于智能 Agent 的科技文献协作推送机制中的思想和方法也同样可以作为该系统设计与实现的有益参考。但具体如何裁剪使之适用,还需要进一步研究和实践。

参考文献

- [1] 互联网学术资源搜索[EB/OL]. http://www.paper789.com/paper_nr0759/, 2010-4-29
- [2] 吴贤奇. 现代文献信息检索[M]. 南京:东南大学出版社, 2007
- [3] 王园春,李瑞斌. 科技信息检索与利用[M]. 北京:石油工业出版社, 2006
- [4] 李瞳. 信息检索与利用[M]. 南京:南京大学出版社, 2006
- [5] 程维红,任胜利,王应宽,等. 国外科技期刊开放存取网络平台[J]. 中国科技期刊研究, 2009, 20(1): 36-44
- [6] 李志民,万猛. 网络时代的科技论文快速共享与多维表达研究[EB/OL]. http://www.paper.edu.cn/lizhimin_jiaoyu_wan-gluo.php, 2010-4-29
- [7] 智能信息推拉技术[EB/OL]. <http://baike.baidu.com/view/2638844.html?fromTaglist>, 2010-4-29
- [8] 陈鑫,常致全. 智能化搜索引擎原理及实现[J]. 计算机应用, 2003, (23): 191-193
- [9] FIPA. FIPA abstract architecture specification [EB/OL]. <http://www.fipa.org/specs/fipa00001/SC00001L.html>, 2010-04-29
- [10] 王汝传,徐小龙,黄海平. 智能 Agent 及其在信息网络中应用[M]. 北京:北京邮电大学出版社, 2006
- [11] Savoy J. Searching information in legal hypertext systems[J]. Artif. Intell. Law, 1994(2): 205-232
- [12] Salton G, Buckley S C. Term-weighting approaches in automatic text retrieval[J]. Inf. Process. Manage, 1988, 24(5): 513-523
- [13] 徐小龙,王汝传. 基于智能 Agent 的多维值信息检索模型[J]. 电子与信息学报, 2008, 30(2): 482-485
- [14] 葛新权,叶茂林. 科技论文学术水平评价方法研究[J]. 北京机械工业学院学报, 2005, 20(3): 49-51
- [15] 秦拯,张玲,李娜. 改进的 PageRank 在 Web 信息搜集中的应用[J]. 计算机研究与发展, 2006, 43(6): 1044-1049
- [16] 林泓,刘朋,李晶晶,等. 基于概率的 PageRank 改进算法[J]. 武汉理工大学学报, 2009, 31(3): 81-83
- [17] IKV+. Grasshopper [EB/OL]. <http://www.grasshopper.de/>, 2010-01-16