

基于多智能体的个性化信息系统研究

王 勇, 陈 新, 陈新度

(广东工业大学计算机学院, 广州 510090)

摘 要: 提出了一种基于多智能体机制的个性化信息系统 (PNSS), 通过建立用户操作的动态特征模型, 运用基于语义及统计的文本匹配算法, 结合用户基本特征, 实现了信息的自动归类和排序, 使得用户能够对其最重要的信息迅速作出决策。

关键词: 多智能体; 文本分类; 个性化信息

Research of Personalized Information System Based on Multi-agent

WANG Yong, CHEN Xin, CHEN Xindu

(College of Computer, Guangdong University, Guangzhou 510090)

【Abstract】 This paper proposes a personalized information system based on multi-agent. It realizes automatic classification and arrangement of information by setting up the dynamic feature model of users' operations, using algorithm of text match based on semantic and statistics, and combining with users' basic feature. The users can rapidly make decision about their most important information.

【Key words】 Multi-agent; text classification; Personalized information

本文提出并建立了一个智能化的自适应模式的个性化新闻服务系统 (PNSS), 为用户提供基于个性需求特征的新闻服务。该系统通过建立用户操作的动态特征描述模型, 运用文本匹配算法, 过滤和排列恰当的新闻信息, 例如当用户浏览到他在注册时没有“申报”的行业时, 新闻信息服务系统能够补充该行业为用户新的兴趣增长点, 并为其送上该行业的新闻。同时, 新闻排列的顺序也反映了用户对不同行业的关注程度, 用户最感兴趣的行业或领域的新闻将被排列在新闻列表的最前面。多智能体机制^[2]作为分布式智能化信息系统环境的一种重要的应用模式, 在 PNSS 系统中被采用, 使得系统具有足够的灵活性、智能性和扩展性。

1 系统结构

通常来说, 一个信息系统包括了信息收集、存储、发布等几个重要功能, 在 PNSS 中, 新闻信息的发布功能与用户的动态操作行为相关联。考虑到对新闻信息的服务主要是对文本信息的处理, 要应用自然语言处理技术 (NLP), 系统要求较高的智能性, 且在互联网环境中应用, 需要系统具有灵活性和扩展性。因此, PNSS 系统采用了多智能体结构, 由多个智能体, 亦称智能代理, 通过分工协作来完成信息服务, 系统结构见图 1。

用户注册智能代理 (ULAgent) 负责记录建模用户特征描述, 并记录到用户基本特征数据库中。其中包含了用户的个人基本信息和感兴趣的行业选择信息。

用户操作监控代理 (UOCAgent) 负责对用户的在线操作进行跟踪、监测, 并记录用户访问操作的动态数据集, 保存信息到用户操作数据库。

用户特征抽取代理 (UFEAgent) 负责实时连接用户操作

数据库, 从中抽取动态操作特征数据, 建立用户特征模型, 并保存到用户动态特征库中。

新闻搜索代理 (NSearchAgent) 根据预先制定的新闻搜索和过滤策略, 实现对有关新闻网站或者特定的新闻信息源的基于互联网的搜索和搜集, 相应的信息被保存到新闻信息库中。

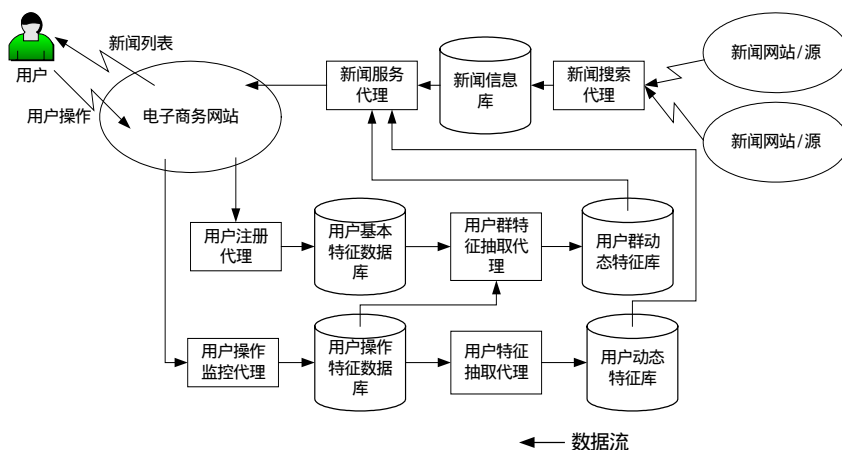


图 1 PNSS 系统框架结构

为了平衡用户个人兴趣与社会群体关注的内容, 系统建立了基于用户分群/段 (Segment) 的分布统计概率的用户群动态特征库, 由用户群特征抽取代理 (UCFEAgent) 结合用户基本特征数据和用户操作特征数据而成, 使得用户不会遗漏用户群体普遍感兴趣的重要新闻。

在上述代理的协同工作基础上, 当用户登录网站浏览新

基金项目: 国家“863”计划基金资助项目 (2003AA414023)

作者简介: 王 勇(1968—), 男, 讲师、博士生, 主研方向: 人工智能与知识工程, 网格计算和网络工程; 陈 新, 博士、教授、博导; 陈新度, 博士、副教授

收稿日期: 2004-07-13

E-mail: ygwang68@hotmail.com

闻信息时,新闻服务代理(NServiceAgent)通过把用户动态特征库中获取的当前用户的特征数据模型与文本匹配算法进行绑定,动态建构过滤器,对新闻信息库中的新闻进行匹配、检索和排序,形成用户个性化新闻列表,通过网站传给用户。

2 多智能体机制

2.1 智能体结构

智能体又称智能代理,是人工智能研究的新成果,它根据用户的需要和有关的专家知识,代替用户进行各种复杂的工作,如信息查询、筛选及管理,并能推测用户的意图,自主制定、调整和执行工作计划。智能体具有自治性、社会性、反应性、主动行为等特性,以及可移动性、合理性等特点,是可进行高级、复杂的自动处理的代理软件。智能代理可应用于广泛的领域,是人工智能领域近年来研究的一个热点,应用于信息检索领域之后,成为开发智能化、个性化信息检索的重要技术之一。

在智能体的设计中,主要有主动及被动反应两类模式。根据信息服务的特点,PNSS系统中的ULAgent、UOAgent、UFEAgent、UCFEAgent和NServiceAgent等智能体采用操作—反应的被动模式来设计各个智能体,其结构见图2。

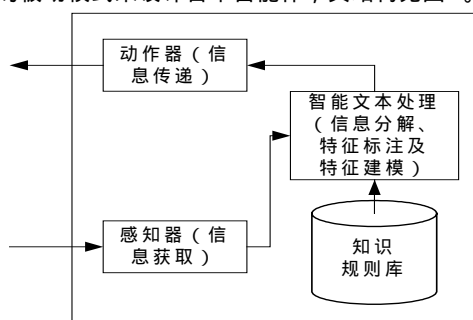


图2 PNSS中被动—反应模式的智能体结构

被动—反应式智能体包含了3个部分:感知器负责监视外界的相关工作信号,如启动信号量,并在启动工作程序时获取信息;智能处理模块负责智能体的推理工作,包括对获取信息的分解、理解、标注及建立特征模型,在该模块中,需要运用多种知识及推理规则,完成诸如自然语言处理、产生式规则推理等智能工作,因此,需要知识规则库的支持;动作器,也可以叫反应器,包括了信息传递和发布,建立与用户的回馈交流,以及与其它智能体的协作。

而新闻搜索代理(NSearchAgent)则采取主动搜索模式,其结构如图3。

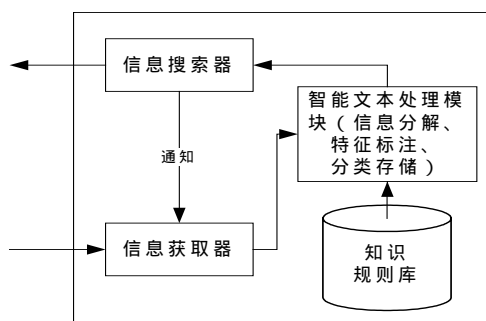


图3 PNSS中主动搜索模式的智能体结构

在主动搜索模式中,信息搜索器定期浏览相关的新闻信息网站,并判断是否有新的信息,如果有新的信息,则通知信息获取器下载信息内容,并交由智能文本处理模块来分析、标识特征及分类存储。

2.2 协调机制

由于PNSS系统中存在多个代理程序,必须建立合适的协调机制。综合考虑系统开发效率和系统运行性能,本系统采用了基于“黑板(blackboard)”结构的协调模型。在黑板结构中定义了用户注册信息和特征描述结构。

UOAgent通过操作—反应(action-reaction)模式来完成用户操作,并在把用户操作数据保存到数据库的同时,把用户操作数据记录到黑板中,并设置特征抽取代理启动信号量。UFEAgent获得黑板中的特征抽取启动信号量后,即综合用户操作数据库中的数据与当前用户操作数据,抽取和构造实时的用户特征数据模型,并记录到用户动态特征库中。UCFEAgent也同时启动,对当前用户操作进行统计分析,建立实时的用户群动态特征模型,保存到用户群动态特征库。

NSearchAgent则通过定时检索有关网站,获取新闻信息,进行文本挖掘工作,包括对文本进行分解、分析和分类,以标识新闻类别和特征,并存放结构化到新闻信息库中备阅。NServiceAgent运用操作—反应模式,把用户动态特征库中的特征模型与匹配算法绑定,动态构成新闻过滤器,并对新闻信息库中的新闻进行实时动态的排序和列表输出。

这种基于黑板结构的多代理协调机制使得各个代理之间相对独立,又通过信号量进行调度协作,同时,又依托全局黑板,实现了共享参数的传递。系统具有实现简洁、耦合度低、扩展性好的特点。

3 关键技术

3.1 动态特征结构

对用户的动态特征进行描述和记录的结构是PNSS系统的关键结构,下面给出有关结构的定义。

定义1 用户基本特征描述向量用以标识和记录注册用户的基本特征,其结构定义为 $\langle u_i, c_{i1}, c_{i2}, \dots, c_{im} \rangle \in U$,其中, U 为全体注册用户的集合, u_i 唯一标识第 i 个用户, c_{ij} 为用户 u_i 的第 j 个特征分量。需要注意的是,在用户基本特征向量的若干分量中,存在一个分量 c_{ik} 标识用户 u_i 注册时所选择的感兴趣和关注的行业或领域的特征描述集合 $\{\langle u_i, f_{ij}, r_{ij} \rangle | u_i \in U, f_{ij} \in F, j=1,2,\dots,m\}$,其中, F 为预先定义的新闻信息所涉及的全体领域的集合, m 为 F 集合中所包含的元素个数, f_{ij} 为用户 u_i 关注的第 j 个领域, r_{ij} 为用户 u_i 对 f_{ij} 领域的关注度,初始可设置用户 u_i 对感兴趣的行业的关注度值都相同,此值为静态数据。

定义2 用户操作向量是用以记录用户的操作行为及相关数据的特征向量,表示为一个四元组 $\langle u, o, d, t \rangle$,其中, u 为用户索引,以连接用户基本特征描述向量,唯一标识用户, o 为操作动作代码, d 为 o 操作所关联的数据集即保存现场数据,以便随后分析用户操作的相关数据领域, t 为执行 o 操作的时间戳。

定义3 用户操作特征集为 $O: \{ \langle u_i, o_{ij}, d_{ij}, t_{ij} \rangle | u_i \in U, j=1,2,\dots,p_i \}$,其中 p_i 为用户 u_i 所进行的所有操作的总次数, j 代表第 j 次操作。

定义4 用户动态特征结构 $\langle u, f, r \rangle$ 是从用户持续的操作行为中分析和抽取而形成的与用户对新闻领域的关注程度相关的特征数据结构, u 为用户索引,用法与用户操作向量结构中的 u 相同,以唯一标识用户, f 为用户 u 关注的领域,

r 为用户 u 对 f 领域的关注度。不同于用户基本特征描述向量中的领域特征描述表, 此中的 r 是建立在对用户操作行为的分析和推理上, 是动态变化的。

定义 5 用户动态特征集合为 $D: \{ \langle u_i, f_{ij}, r_{ij} \rangle | u_i \in U, f_{ij} \in F, j = 1, 2, \dots, m \}$, 参数定义与用户基本特征描述向量中的相应参数定义相同。

定义 6 为了平衡用户的个人爱好与社会的关注热点, 建立了基于用户的特征分段的用户群动态特征向量 $\langle s, f, r \rangle$, s 为用户群特征分段标识, 即按照用户的有关特征的值, 可以把用户划分若干用户分段, 如根据年龄划分不同年龄段的用户集合, 根据性别划分男女两个用户分段等, 然后, 对各用户分段内的用户进行统计分析, 建立用户群分段相对所有行业的关注度的概率分布模型, f 为行业标识, r 为 s 用户群分段对 f 行业的平均关注度, 此值将跟随用户群体的操作而动态变化。

定义 7 用户群特征分段的集合为 $S: \{s_i, i = 1, 2, \dots, q | s_i \subset U, s_i \cap s_j = \emptyset (i \neq j), s_1 \cup s_2 \cup \dots \cup s_q = U\}$ 。

定义 8 用户群动态特征集合为 $C: \{ \langle s_i, f_{ij}, r_{ij} \rangle | s_i \in S, f_{ij} \in F, j = 1, 2, \dots, m \}$ 。

定义 9 用户信息输出特征集合为 $I: \{ \langle u_i, f_{ij}, r_{ij} \rangle | u_i \in U, f_{ij} \in F, j = 1, 2, \dots, m \}$, 其中 r_{ij} 值为行业 f_{ij} 信息输出的排序值, 信息输出按照 r_{ij} 的值由高至低排列。

3.2 相关算法

(1) 动态特征抽取算法

动态特征抽取主要是从用户操作特征向量中, 通过分析其特征提取而实现的, 可表示为从用户操作特征向量集合 O 到动态特征向量集合 D 的映射, 即:

$$\varphi: O \rightarrow D$$

其中, 函数 φ 为由用户操作 o_{ij} 的数据对象 d_{ij} 所对应的行业 f_{ij} 在时间轴上的平均概率分布, 其值即为 D 中的 r_{ik} 分量的值, 计算公式如下:

$$r_{ik} = \frac{\sum p(d_{ij})}{\sum q(o_{ij})}$$

其中,

$$p(d_{ij}) = \begin{cases} 1, & \text{当 } d_{ij} \text{ 与 } f_{ik} \text{ 相关时} \\ 0, & \text{当 } d_{ij} \text{ 与 } f_{ik} \text{ 不相关时} \end{cases}$$

$$q(o_{ij}) = 1, \text{ 操作 } o_{ij} \text{ 只统计对数据的操作}$$

然后对用户 u_i 的所有行业的 r_{ik} 值 ($k=1, 2, \dots, m$) 进行归一处理, 得到相应的 r_{ik} , 即为用户 u_i 对第 k 个行业的关注度值, 并可表示为一个用户动态特征行业关注矩阵。

(2) 排序匹配算法

信息输出序列主要基于用户基本特征数据、用户动态特征数据及用户群动态特征数据, 即存在函数映射 $\psi: U \times D \times C \rightarrow I$ 。其中, 函数 ψ 为用户基本特征数据中的静态的行业特征关注度值, 记为 ru_{ik} , 用户动态特征中的行业关注度值记为 rd_{ik} , 用户所属的用户群分段所对应的动态特征行业关注度值记为 rc_{ik} 的符合简单加权平均值, 即 I 中的 r_{ik} 分量的计算公式如下:

$$\overline{r_{ik}} = w_u \times ru_{ik} + w_d \times rd_{ik} + w_c \times rc_{ik}$$

其中, w_u, w_d, w_c 分别为用户基本特征数据、动态特征数据、群体分段特征数据相应的权重。当然, 也可定义更复杂的权重计算公式, 如可由特征值按照概率分布计算而得。

然后, 再对每个用户的所有 $\overline{r_{ik}}$ 值 ($k=1, 2, \dots, m$) 进行归一处理, 得到相应的 r_{ik} 值, 即用户 u_i 的第 k 个行业的综合关注度值, 也就是用户 u_i 的个性化新闻信息输出队列的排序依据, 新闻信息列表依照 r_{ik} 值由高向低倒序输出。

4 实现与应用

PNSS 系统建立了基于浏览器/Web 服务器/中间件/数据库的多层架构的服务模式^[3], 把商业应用逻辑封装到 JavaBean 中, 前台接口由 JSP 程序提供服务, 后台数据库采用了 MS SQL Server 2000。多层结构的服务模式提高了系统开发效率, 提升了系统的整体性能, 增强了系统的灵活性和活性, 是多智能体结构的良好支撑环境。

在国家“863”计划项目及广东省科技创新重点项目“虚拟制造协作网”的“制造业新闻中心”功能板块中建立了 PNSS 系统, 作为网站客户关系管理的一个重要组成单元。当用户注册时, 除了登记用户的基本个人资料外, 还请用户选择他感兴趣的领域, 以便新闻信息的传送。在后续的用户操作当中, 通过追踪用户访问的页面数据, 判断与用户访问页面相关的行业特征值, 掌握用户感兴趣的内容和行业, 动态调整用户操作的特征模型。当用户浏览新闻时, 根据排序匹配算法, 把新闻信息列表输出给用户。用户感觉到大多数情况下, 总是可以在新闻列表的前面得到他最感兴趣的内容。经过测试, 系统的准确率达 75%, 用户获得了一种新的个性化的信息服务模式, 增强了满意度。同时, PNSS 系统的关键思路及技术也在被应用到其它的一些网站, 特别是在证券交易网站中进行的实验, 把用户的投资行为与有关新闻信息动态链接, 使得用户能够及时获取最感兴趣或者最重要的新闻信息, 及时决策, 这是一个有意义的应用领域。

5 结论

PNSS 系统是个性化信息服务模式和多智能机制系统结构的集成应用, 下一步的工作包括研究新的技术和算法, 全面跟踪用户的操作行为, 构造准确的用户动态特征模型, 准确抽取用户动态需求特征, 改进基于统计分布策略的排序算法, 提高新闻信息队列的准确性, 并加强网络搜索和文本挖掘能力, 获取高质量的新闻信息, 提升系统服务的整体性能。

个性化新闻服务系统是一个基于文本挖掘相关人工智能技术的有效的客户关系管理新模式。多智能代理结构的引入, 增强了 PNSS 系统的智能性和扩展性, 拓展了应用空间。它能够根据用户的动态行为特征, 提供有针对性的个性化信息, 可以被推广应用到多个领域, 例如网上购物系统、百货公司的电子商务系统、金融证券投资的信息服务等。总之, 个性化信息服务是信息时代的一个重要的特点和功能需求, 正在成为电子商务领域的研究热点之一。

参考文献

- 1 Turban. Electronic Commerce: A Managerial Perspective. Prentice Hall, 2000
- 2 何炎祥, 陈莘萌. Agent 和多 Agent 系统的设计与应用. 武汉: 武汉大学出版社, 2001
- 3 Wu Qiqing. JSP Website Design. Metallurgy Industry Publisher, 2001