

# 基于 Agent 的个性化信息过滤系统的设计与实现

费洪晓, 穆 珺, 巩艳玲, 黎 成  
(中南大学 信息科学与工程学院, 湖南 长沙 410075)

**摘 要:** 针对用户个性化服务的特定需求, 文中提出了一种基于 Agent 的个性化信息过滤系统的设计思想及其实现过程。采用基于主题的过滤和基于兴趣的过滤相结合的过滤方法对信息分两次过滤, 同时利用 Agent 跟踪用户的浏览行为, 从而提供隐式反馈。系统能够根据文本的内容自动判别文本所属主题分类, 并计算待过滤信息与用户兴趣之间的相关度, 最后利用用户的反馈对用户兴趣模型进行更新, 从而帮助用户准确获取有用信息。

**关键词:** Agent; 个性化信息过滤; 用户兴趣模型

**中图分类号:** TP18

**文献标识码:** A

**文章编号:** 1673-629X(2006)12-0001-03

## Design and Implementation of Agent - Based Personalized Information Filtering System

FEI Hong-xiao, MU Jun, GONG Yan-ling, LI Cheng

(Information Science and Engineering College of Central South University, Changsha 410075, China)

**Abstract:** With the requirement of user's specific information service, a framework of personalized information filtering system, based on Agent, is proposed and implemented in this paper. Combine the two methods to filter the information. Use the topic-based filtering method and then use the interest-based method. The system is able to judge the categories of the texts automatically according to the content of the texts and then calculate the similarities between the texts and user's needs. Finally update the user profile depend on the user's feedback, so as to help the user to obtain useful information accurately.

**Key words:** Agent; personalized information filtering; user profile

随着网络信息的迅猛发展, 网上数以亿计的网页给信息获取带来了极大的困难, 向人们提出了如何快速地从信息海洋中获取其所需信息的挑战<sup>[1]</sup>。自动搜索引擎虽然可以帮助人们寻找信息, 但因信息量过大而不能正确对信息进行分类和个性化的服务。对于万维网这种信息量巨大而结构杂乱无章的网络环境, 基于 Agent 的信息过滤技术成为处理万维网信息的有效途径<sup>[2]</sup>, 这对于提高 Web 信息检索的个性化水准和准确度非常有利。

## 1 信息过滤系统的设计

### 1.1 系统的设计思想

文中所设计的个性化信息过滤系统是一个智能 Agent 系统, 系统由用户接口 Agent、检索 Agent、过滤 Agent 三个子 Agent 组成<sup>[3]</sup>。由于单个 Agent 无法独立完成任务, 故采用多 Agent 协作来共同完成任务。信息检索 A-

gent 从网络上的各种信息源中发现并获取相关的信息; 过滤 Agent 解决中文信息的不一致性, 过滤掉无关或不需要的信息; 用户接口 Agent 能够动态知晓用户的兴趣并能跟随用户兴趣的变化。根据用户的兴趣, 这三种不同类型的 Agent 通过共同的通信语言和通信机制进行相互交流和协作来快速、准确地完成对中文网络信息的个性化过滤。

### 1.2 系统的体系结构

本系统由用户接口 Agent、检索 Agent、过滤 Agent、领域模型库、文档库、用户模型库六部分组成, 系统的体系结构如图 1 所示。

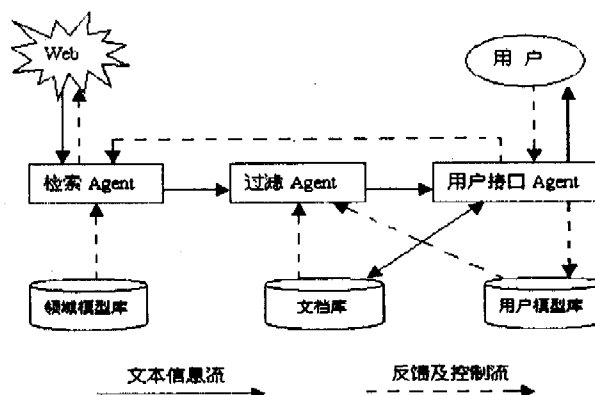


图1 系统体系结构图

收稿日期: 2006-03-20

基金项目: 国家自然科学基金资助(60173041); 湖南省自然科学基金资助(02JJY2094)

作者简介: 费洪晓(1967-), 男, 浙江嵊县人, 副教授, 研究方向为网络管理与网络安全。

(1) 用户接口 Agent。为用户提供操作和查看过滤结果的界面,将过滤得到的网页呈现给用户,并接受用户反馈。通过跟踪用户行为来分析用户的兴趣,自动地对用户兴趣模型进行更新和修正,使之能够快速地对适应用户兴趣和环境的变化<sup>[4]</sup>。

(2) 检索 Agent。其任务就是不断地从网络中搜寻与相应领域相关的网页。通过调用搜索引擎获取相关信息的 URL,并下载提交给过滤 Agent 进行分析。文中的检索 Agent 通过集成 Yahoo 和 Lycos 搜索引擎来实现对网络信息大范围的搜索,使检索到的结果具有更高的广泛性和准确性,满足用户快速、准确的查询需求。

(3) 过滤 Agent。其主要功能是对网页进行分析,抽取用户感兴趣的有用的信息。过滤 Agent 首先对网页进行特征抽取,形成结构化的网页属性。分析的手段主要是统计主题概念库中关键词在网页中出现的频度并计算该网页与主题概念的隶属度。然后,根据领域主题库和用户模型库,计算目标文档特征向量同用户兴趣模型特征向量的相似度,对文档进行过滤,并将过滤后的网页的结构化信息提交给用户 Agent。

(4) 领域模型库。领域模型库用来描述该知识领域中主题概念间的关系及与各概念相关的关键词。与各概念相关的关键词可以通过领域训练样本学习获得,主题概念和关键词可在系统运行过程中自适应调整。用户可通过选择领域的主题概念作为其基本的兴趣点,形成用户模型的框架基础;主题概念模型库是网页进行分类和过滤的基础,过滤 Agent 用主题概念库中的信息,计算网页与各主题概念的隶属程度和用户兴趣的相关度,从而对网页进行个性化的过滤。

(5) 文档库。记录所收集的网页、相应文档的结构化信息及与之相关性最大的若干主题概念的隶属度。保存用户感兴趣的文档资料,为用户提供快速的获取相关网页的手段。

(6) 用户模型库。记录反映各用户兴趣点的模型。其作用是用来保存用户 Agent 从过滤 Agent 所提供的与领域相关的文档中提取具体用户最感兴趣的文档。

## 2 用户接口 Agent

### 2.1 用户接口工作流程

用户接口 Agent 主要由注册界面,登陆界面、用户界面、兴趣跟踪和分析器四部分组成,并与用户文档库、用户模型库协同工作。

用户接口 Agent 的工作流程如下:用户注册,提交用户信息和感兴趣的资料,建立一个初步用户兴趣模型;用户登陆,确认用户名和口令;输入查询关键字,送入信息检索 Agent;用户仔细阅读其感兴趣的搜索结果,接收用户反馈、兴趣跟踪和分析器,对当前用户的兴趣进行跟踪和自适应地调整,并及时地更新和修正用户模型;将用户感兴趣的结果信息保存入文档库。

### 2.2 兴趣模型的更新机制

用户兴趣模型的更新机制的基本思想是在相关反馈基础上,通过用户提供的初步兴趣模型和跟踪用户的行为来分析某一方向兴趣的文本,经过文本映射和文本结构分析,获得文本的特征向量表示,然后使用权值更新的算法使用户的兴趣模型得到更新和修正。

用户在阅读当前搜索过滤的结果时,会对结果进行某些操作<sup>[5]</sup>(如:打开浏览,将页面保存、添加收藏等),用户接口 Agent 会进行分析和判断该信息是否符合自己的要求和需要。对于符合要求的将该页面的 IP 地址记下来并将文本保存到文档库中,同时将该文档的特征项提取出来,并计算出各特征项的权值,存入用户模型中。如果某特征项在用户模型中已经存在,那么它将获得一个新的权值。其新权值计算公式<sup>[4]</sup>如下:

$$\text{Value} = \frac{r * o * k}{n * D} + \text{Oldvalue} \quad (1)$$

其中,Value 为用户兴趣模型中的关键字的新权值, $r$  为关键字所在结果的相关度, $o$  是用户对关键字所在结果文档采取的操作类型的对应权值, $k$  是关键字所在结果的关键字序列序号对应的权重, $n$  是本次任务所返回的所有文档数,Oldvalue 是用户兴趣模型中的旧权值, $D$  是一个调节常量。

由公式(1)就可以得到一些新的兴趣特征向量和旧特征向量的新的权值,按照新权值的大小排序,取排在前面的特征值对原有的用户兴趣模型进行修正和优化。这样设计的优点是可以随用户浏览行为中表现出来的用户兴趣的改变而改变用户兴趣模型,及时更新了用户兴趣信息,对信息的过滤更具导向性,同时,大大提高了用户查找信息的准确度和效率。

## 3 信息过滤 Agent

检索 Agent 将搜索返回结果给过滤 Agent,进行信息的分类和筛选,将那些不属于用户需求信息领域范围和不符合用户兴趣的信息过滤掉。文中对这些检索出来的信息进行两次过滤:基于主题的过滤和个性化的过滤。基于主题的过滤的作用是将那些不属于用户所需信息主题的信息滤掉,个性化过滤则是将主题过滤后存留下来的信息中与用户兴趣无关的信息过滤。经过两次过滤,这样得到的信息的准确度就大大提高了。

信息过滤 Agent 它有自己的学习和过滤机制,必须先给予它充足的学习和训练,方能实现过滤的自动化、智能化。知识训练机制的功能就是通过对训练文本的训练为基于主题的过滤提供主题概念和关键词向量。系统初始化时,这些训练文本都是精心挑选出来的样本,它们是能够表现各个主题和类型的词的组合;在用户不断和系统交互的过程中,将以一定频率从用户文档库中选取文档更新训练库,从而使主题概念模型自适应地跟随用户的变化。

信息过滤的工作流程是:首先,对采集到的 Web 页面

进行预处理,将 HTML 页面里的文本提取出来,然后使用中文分词技术将 Web 文本切分成单个的中文词语并进行词频统计,根据统计的结果从得到的中文词向量中提取出能够表达出该文本主题的特征向量。运用贝叶斯分类算法对该文本进行分类,如果该文本信息不属于用户需求的类型则将它滤掉。留下来的结果,采用向量空间模型法实现个性化过滤,从而使用户得到相对准确的网络信息。信息过滤 Agent 的结构及工作流程如图 2 所示。

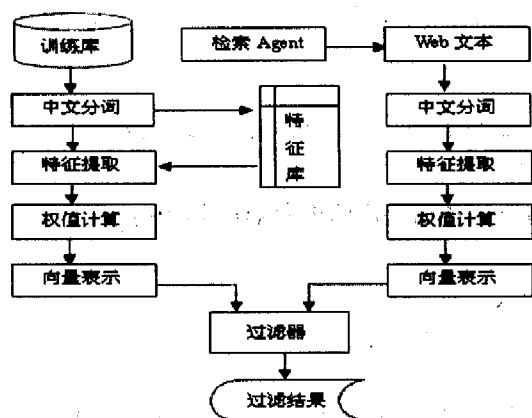


图 2 信息过滤 Agent

### 3.1 基于主题的分类过滤

文中采用的文本分类过滤算法为朴素贝叶斯分类算法。该算法的基本思路是计算文本属于类别的概率,文本属于类别的几率等于文本中每个词属于类别的几率的综合表达式。

朴素贝叶斯分类算法的具体算法步骤如下:

1) 计算特征词属于每个类别的几率向量  $(w_1, w_2, w_3, \dots, w_n)$ ,

$$\text{其中, } w_k = P(W_k | C_j) = \frac{1 + \sum_{i=1}^{|D|} N(W_k, d_i)}{|V| + \sum_{j=1}^{|V|} \sum_{i=1}^{|D|} N(W_j, d_i)} \quad (2)$$

2) 在新文本到达时,根据特征词分词,然后按下面的公式计算该文本  $d_i$  属于类  $C_j$  的几率:

$$p(C_j | d_i; \theta) = \frac{P(C_j | \theta) \prod_{k=1}^n P(W_k | C_j; \theta)^{N(W_k, d_i)}}{\sum_{r=1}^{|C|} P(C_r | \theta) \prod_{k=1}^n P(W_k | C_r; \theta)^{N(W_k, d_i)}} \quad (3)$$

其中,  $P(C_j | \theta) = \frac{C_j \text{ 训练文档数}}{\text{总训练文档数}}$ ,  $P(C_r | \theta)$  为相似含义,  $|C|$  为类的总数,  $N(W_k, d_i)$  为  $W_k$  在  $d_i$  中的词频,  $n$  为特征词总数。

3) 比较新文本属于所有类的几率,将文本分到几率最大的那个类别中。同时,将那些不属于用户所要信息主题范围的信息摒弃。

### 3.2 基于兴趣的个性化过滤

朴素贝叶斯分类过滤后,得到的是一些该主题类的

Web 文本,但仍会含有部分无用的信息,在此引进了个性化的信息服务进行再一次的过滤,其思想是通过用户对用户提交的初步兴趣模型的分析 and 用户行为的跟踪,学习用户的操作习惯,来获得用户的兴趣<sup>[5,6]</sup>。随着用户对本系统的使用,系统通过对该用户进行跟踪分析,不断地对该用户的兴趣模型进行更新和优化。然后通过个性化的处理,得到最可能接近用户需要的结果。

具体操作的步骤为:

第一步:新文本到来后,分词,将文本表示为特征向量。

第二步:从用户模型中得到当前用户的当前兴趣向量。

第三步:计算新文本特征向量和当前兴趣向量间的相似度,公式为:

$$\text{Sim}(d_i, d_j) = \frac{\sum_{k=1}^M W_{ik} \times W_{jk}}{\sqrt{(\sum_{k=1}^M W_{ik}^2)(\sum_{k=1}^M W_{jk}^2)}} \quad (4)$$

其中,  $d_i$  为新文本的特征向量,  $d_j$  为第  $j$  类的中心向量,  $M$  为特征向量的维数,  $W_k$  为向量的第  $k$  维。

第四步:比较每个新文本与兴趣向量的相似度,按照相似度的大小,按某一阈值进行过滤。

## 4 多 Agent 的通信和协作

### 4.1 Agent 间的通信

多 Agent 通信机制直接影响了多 Agent 系统的协同工作能力和系统的性能。本结构下的多个 Agent 通过 ACL 通信语言和 FIPA 消息传递机制进行通信和协同工作。在 FIPA 规范下的 Agent 平台有三个主要的子系统:索引器、AMS (Agent Manage System, Agent 管理系统) 和 MTS (Message Transport Service, 消息传输服务)。Agent 间消息的传送通过 MTS 系统中的 ACC (Agent Communication Channel) 提供的消息传送服务以信封 + 消息的格式传输来实现。根据信封中定义的地址等信息,在 MTS 为 Agent 平台上的 Agent 找到合适的消息传送通道 ACC 后,ACC 进行通信协议的判定和消息的传输。当消息抵达目的地后,再通过 ACC, Agent 就可以解读 ACL 所要传达的信息,多 Agent 之间的沟通也随即完成。

### 4.2 Agent 间的协作

多 Agent 协作是一种集体行为,指一个 Agent 在采取行动或做决定时,要受到别的 Agent 的存在或知识的影响。正是协作使得几个智能 Agent 能将它们各自的努力组合起来完成信息的检索和个性化的过滤。

本系统各 Agent 协同工作具体的工作流程为:首先用户向检索 Agent 发出请求,请求系统搜索出用户需要的信息,信息检索 Agent 监测到用户 Agent 的请求,马上响应并开始搜索信息,同时将搜索到的结果返回并提交给信息

(下转第 6 页)

表 1 不完备信息系统(U, AT)

AT \ U	A	B	C	D	E
1	0	1	1	0	1
2	1	0	0	*	0
3	0	1	*	0	1
4	*	0	0	1	1
5	1	0	1	0	0
6	0	*	0	1	0
7	0	1	0	*	0
8	1	*	0	0	0
9	1	1	0	0	1
10	0	0	1	0	1

取  $\alpha = 0.6$ , 有:

$$S_A^{\alpha}(1) = \{1, 3, 7, 10\}, S_A^{\alpha}(2) = \{2, 4, 5, 6, 8, 9\},$$

$$S_A^{\alpha}(3) = \{1, 3, 4, 7, 8, 10\}, S_A^{\alpha}(4) = \{2, 3, 4, 6, 8\}$$

$$S_A^{\alpha}(5) = \{2, 5, 8\}, S_A^{\alpha}(6) = \{2, 4, 6, 7, 8\},$$

$$S_A^{\alpha}(7) = \{1, 3, 6, 7, 8, 9\}, S_A^{\alpha}(8) = \{2, 3, 4, 5, 6, 7, 8, 9\},$$

$$S_A^{\alpha}(9) = \{2, 7, 8, 9\}, S_A^{\alpha}(10) = \{1, 3, 10\},$$

$$\text{设 } X = \{2, 4, 5, 6, 7, 8, 9\}, R^{\alpha}(X) = \{2, 5, 6, 9\},$$

$$\bar{R}^{\alpha}(X) = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$$

显然,  $R^{\alpha}(X) \subseteq X \subseteq \bar{R}^{\alpha}(X)$  成立。

$$R^{\alpha}(R^{\alpha}(X)) = \emptyset, \bar{R}^{\alpha}(\bar{R}^{\alpha}(X)) = \{2, 4, 5, 6, 7, 8, 9\}$$

(上接第 3 页)

过滤 Agent 进行主题的过滤、个性化过滤;最后将得到的结构返回给用户 Agent,并显示出来。信息过滤系统 Agent 的协同工作的流程如图 3 所示。

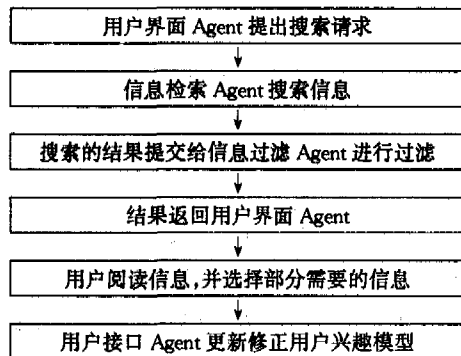


图 3 信息过滤系统 Agent 的操作流程图

## 5 小结

文中提出了一种基于 Agent 的个性化信息过滤系统结构,适用于一般万维网上的智能信息搜索系统。与传统搜索引擎相比,该结构下的信息搜索可以帮助使用者根据本人的兴趣和偏爱获得较高的匹配。同时,该结构下系统

$$R^{\alpha}(\bar{R}^{\alpha}(X)) = \{2, 4, 5, 6, 7, 8, 9\}$$

$$\bar{R}^{\alpha}(R^{\alpha}(X)) = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$$

显然,  $R^{\alpha}(R^{\alpha}(X)) \subseteq R^{\alpha}(X) \subseteq \bar{R}^{\alpha}(\bar{R}^{\alpha}(X)) \subseteq X \subseteq \bar{R}^{\alpha}(\bar{R}^{\alpha}(X))$  成立。

## 4 结论

集对分析自 1989 年提出来以后,在许多领域得到了应用,但仍有许多认识上和理论上的问题值得研究<sup>[1,2]</sup>。文中在文献[2]基础之上,考虑用集对分析方法进一步刻画不完备信息系统,把此理论与粗糙集理论有机地结合在一起,提出了一种新的集对粗糙集理论,定义了一种不完备信息系统的上、下近似运算,得到了一些性质,并且通过一个简单的例子说明了上述方法的可行性,在粗糙集用于研究不完备信息系统方面做了一定推广。

## 参考文献:

- [1] 赵克勤. 集对分析及其初步应用[M]. 杭州: 浙江科学出版社, 2000.
- [2] 张 鹏, 王光远. 新集对论[J]. 哈尔滨建筑大学学报, 2000, 33(3): 1-5.
- [3] 张文修, 吴伟志, 梁吉业, 等. 粗糙集理论与方法[M]. 北京: 科学出版社, 2001.
- [4] 黄 兵, 周献中. 基于集对分析的不完备信息系统粗糙集模型[J]. 计算机科学, 2002, 29(9): 1-3.
- [5] 黄 兵, 钟 斌. 改进集对粗糙集模型[J]. 计算机工程与应用, 2004, 40(2): 82-84.

的实现面临着一些难点问题,突出表现为用户兴趣更新问题、Agent 间的通信问题、概念的标准化定义问题和多系统的集成问题,这些研究的进展将为多 Agent 系统的进一步实用化奠定基础。

## 参考文献:

- [1] Mladenic D. Personal WebWatcher: design and implementation [EB/OL]. 1996. <http://ranger.uta.edu/alp/ix/readings/mladenic96personalWeWatcher.pdf>.
- [2] 冯 翱, 刘 斌, 卢增翔, 路海明, 等. Open Bookmark—基于 Agent 的信息过滤系统[J]. 清华大学学报: 自然科学版, 2001, 41(3): 85-88.
- [3] 费洪晓, 巩艳玲, 谢文彪, 等. 基于混合学习策略的多 Agent 信息过滤系统[J]. 计算机应用, 2006, 26(2): 267-269.
- [4] 李 俊, 张灵玲, 周文辉, 等. 一个智能用户接口 Agent 的设计与实现[J]. 软件学报, 1999, 10(8): 23-27.
- [5] 白丽君. 基于智能 Agent 的用户兴趣发现和更新[J]. 计算机工程, 2003, 21(2): 70-72.
- [6] 张国印, 陈 先, 皮 鹏. 基于词频统计的个性化信息过滤技术[J]. 哈尔滨工业大学学报, 2003, 24(1): 63-67.