● 焦玉英　刘伟成　李法运

网络环境下专题文献信息过滤模型及服务系统设计*

摘　要　基于专题文献的信息过滤系统的设计,主要解决以下关键问题:(1)针对学术用户信息需求特点,构建动态可适性用户模型。(2)建立结构合理、高效运行、容量高的本地数据库后台支撑体系。(3)利用智能信息推送技术,为专业用户提供经过过滤的有针对性的文献信息。图3。参考文献5。

关键词　信息过滤　专题文献　系统模型　智能代理

分类号　G354.4

ABSTRACT　The authors think that the design of a subject document information filtering and service system should include the following key issues: (1) Constructing a dynamic adaptive user model according to the information needs of academic users; (2) Establishing an efficient background support system; (3) Utilizing intelligent information push technology to provide filtered and customized information. 3 figs. 5 refs.

KEY WORDS　Information filtering. Subject document. System model. Intelligent agent.

CLASS NUMBER　G354.4

信息过滤系统的目的是对大量动态产生的信息进行分类并提供给可能满足其信息需求的用户[1]。最初的信息过滤方法是人工提醒服务,即将新信息告知研究图书馆或专门图书馆的用户,当时这一过程被称作 SDI(定题情报服务)。

本文的目的旨在运用已有的信息检索技术、信息过滤技术、机器学习技术和人工智能技术等设计一个能提供满足专题研究的个性化信息查找服务的高效的信息过滤系统。通过信息过滤机制,克服第一、二代搜索引擎所遗留的一些弊端,减少网络用户所面临的信息超载现象,在主动推送个性化定制信息的同时,尽量过滤掉无关信息,净化用户信息需求空间,提高从事专项科研的用户满意度[2]。

1　信息过滤系统的一般模型

一个信息过滤系统主要包含信息分析模块、过滤模块、过滤模板生成模块、学习模块、信息采集模块(可选)。信息分析模块主要是对海量信息进行分析,提取其中的特征信息,如将每条信息表示为空间向量或索引等。过滤模板生成模块是收集用户对信息的需求和喜好来生成过滤模板。过滤模块就是根据上述两个模块的结果来处理海量信息并将最终结果返回给用户从而实现过滤的目的。学习模块就是不断跟踪用户兴趣的变化来调整甚至更改过滤模板,达到正确过滤无用信息的目的[3]。过滤模块的匹配算法和用户需求模板的描述方法、信息的揭示方法是相互联系的,常用的匹配模型有布尔模型、向量空间模型、概率模型、聚类模型、基于知识的表示模型、混合模型等,主要任务是剔除不相关的信息、选取相关的信息并按相关性的大小提供给用户。为了提高信息过滤的效率,系统还根据用户对过滤结果的反应通过反馈机制作用于用户和用户需求模板,使用户逐渐清晰自己的信息需求,对用户需求模板的描述也会越来越明确、具体。在现有技术条件下,全自动的信息

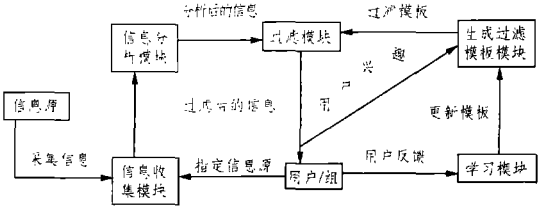

图1　网络信息过滤系统模型

* 本文系国家自然科学基金资助项目(70473067)研究成果之一。

过滤系统还处于试验阶段，为了提高实用性，往往会在这些关键部分进行必要的人工干预，如对动态的信息流先作预处理、人工修改用户需求模板等。

2 专题信息过滤系统的设计目的、结构原理和功能特点

本文在吸收各种信息过滤系统原理和方法的基础上，试图构建一个基于Web的信息过滤模拟实验系统，即一个相对优化的具有异构代理功能的信息聚类过滤系统。

代理，作为一种半智能化的计算机程序，正不断用于帮助处理重复性的或耗费时间的任务。本系统旨在引入代理的方式从各种分布式资源中发现用户感兴趣的信息，并以聚类和排序两种方式提供给用户。一般系统用户对过滤系统的反馈有两种形式：显式反馈和隐式反馈。本系统的设计尽量避免用户的显式反馈，而主要采用隐式反馈的方式获取用户的兴趣。当然，用户的信息检索和过滤行为是一种复杂的学习过程，不可能一蹴而就。因此系统也允许用户对所阅读的文献作出相关性判断，并把这种判断作为用于学习用户兴趣的重要手段。

2.1 系统的设计目的

设计一个针对专题文献信息服务的过滤系统，目的就是在更大程度上提高网络环境下图书情报机构为特定用户提供专题文献服务的精度；增强图书馆员在网络时代的信息导航职能；提高我国图书情报领域文献信息服务的水平；为国内各类型文献信息中心开发优化的信息过滤服务系统提供参考，从而改善和优化科研人员利用网络信息的环境。

在基于专题文献的信息过滤系统的设计过程中，本文主要解决以下关键问题：

(1)针对学术用户信息需求的特点，构建动态可适性用户模型。该模型将采用用户界面代理技术跟踪并捕获用户变化的兴趣，即避免基于内容过滤方式中仅使用关键词来表征用户兴趣的不精确性，又避免基于协作过滤机制中运行早期的“冷起动”问题。使系统能够快速平稳地调适到用户不断变化的需求。

(2)妥善解决本地系统与搜索引擎的互联互动关系，及时分析网页内容，强化对异构的搜索结果进行去重、排序的功能，建立结构合理、高效运行、容量高的本地数据库后台支撑体系。

(3)利用智能信息推送技术，及时、高效地为专业用户提供满足科研需求的、经过过滤的有针对性的文献信息。

2.2 系统的结构原理

为了解决以上几个关键问题，我们首先提出了基于异构多代理的Web信息过滤推荐系统的拓扑结构图(如图2)。系统由用户界面代理等7个部分组成。用户界面代理是直接面向用户的。信息过滤代理负责对从网络上获取的信息进行过滤查找。反馈更新代理负责接受用户界面上来自用户的各种反馈，采用有效的学习机制，更新、调适用户的兴趣文档。操作代理负责提供由用户点击的链接文献的组织和显示。搜索代理负责利用所选定的搜索引擎从网络上获取相关文献并进行去重、排序并显示，对中心索引数据库进行更新和查找。中心数据库保存从网络上下载的针对各个用户需求的相关文献的特征描述，也可以包括题目和摘要片段。新到信息提醒代理负责跟踪网络文献的变动情况，通过和用户所存检出文献对比，一旦发现有新文献到来，即通知用户。

图2 基于异构多代理的网络信息过滤系统结构

2.3 系统的特点和功能

本系统在小范围内进行了真实环境下的联网试验性运行，证明有以下的功能和特点。

(1)能够处理分布在异构资源中的非结构化信息，如通过http，nntp，ftp，gopher连接的信息来源，这里的“信息”意指来自WWW网页、gopher网站、ftp网站、基于WWW的信息服务、新闻组等各种来源的文献。

(2)系统的运行不需要用户过多地参与和关注。信息将含有关键词的摘要和网页片段两种形式提交给用户。系统通过考察热链接和浏览历史并允许用

户对阅读的文献评分,以及对显示的关键词进行增、删、改操作。通过这些方式获得用户的兴趣特征,跟踪用户的兴趣并调整搜索策略,调适到最佳的用户兴趣上来。

(3)本系统并不直接搜索 WWW 数据源,而是发出多个代理利用现有的搜索引擎并执行"元搜索"(本系统考虑到速度的需要,暂限定为同时可向 5 个搜索引擎发出请求),以发现用户感兴趣的信息。

(4)本系统将基于内容的过滤、基于协作的过滤、基于环境的过滤和基于经济学的过滤机制有机结合起来,以确保系统的易用性,解决"冷起动"问题,加速用户兴趣模型的进化,保证系统运行的速度。在用户必要的参与下,主要依靠系统的各种代理学习用户的兴趣,获得最佳的过滤精度并为用户发现、提供适用的信息服务。

(5)在本系统中,产生一个由进化代理构成的人造生态环境。各种代理可以在一个有限的资源环境中实现相互合作和竞争,使代理的效率和整个系统的效率最佳。通过信息过滤类代理的原理是它能提供一个最有效地利用对问题可能的现有方案的方法(在我们的系统中,就是跟踪一个新的用户兴趣或调适该领域中的变化),以此实现系统的个性化并负责跟踪、调适用户的兴趣。信息发现类代理则继续利用突变、遗传等操作更新并培训处理代理的进化技术,对搜索空间进行考察以提供更好的方案。它主要负责信息资源处理、调适信息资源发现并取得用户感兴趣的实用信息。

3　专题文献信息过滤系统实现的关键技术

3.1　用户兴趣模型的获取技术

尽管信息过滤主要针对用户比较长期的兴趣,但也不可忽视用户需求的变动性。这使得用户模型的构建成为信息过滤系统构建过程中的一大难点。

用户需求信息的获取方式可粗略分为两种类型:

(1)显式知识获取:通过提问来获取知识。这种方法有一定的缺陷:首先它只能利用有限的提问来确定用户的偏好,要求用户主动填写预先设定的提问模式,所以系统不能主动跟踪用户的兴趣变化,故其兴趣文档没有随时间动态更新。其次,由于语言表达的问题和分类的模糊性与多样性,互联网一般用户在如何生成合适的关键词,如何选择相关的类别上还有一定的困难,往往不能将信息需求表达清楚,用户不精确的信息表达将影响信息查准率。

(2)隐式知识获取,也即用户兴趣的学习,可根据用户对浏览信息的选择,采取某种学习方法逐步明确用户兴趣所在。实质上它是一个机器学习的过程。对于用户的信息需求,可以通过反映用户信息需求的各种痕迹和线索来获得:①跟踪用户的热链、经常访问的站点或浏览历史,分析、记录用户的行为和选择倾向,隐性地获取对用户需求信息的描述,确定用户的兴趣和偏好,获取用户信息需求,自动产生个性化的知识库规则来指导过滤。但由于用户的兴趣时常变化,用户的行为信息所反映的用户的信息需求往往是多条线索混合在一起,这给识别信息需求带来了很大的困难。②相关性反馈,即用户对以前过滤结果和所访问的网页的反馈,所获得的知识可用来更新用户模型。

将显式知识获取与隐式知识获取两种方法结合使用可获得更好的效果[4]。通过对新用户的提问来获取用户的初步信息,并将其归入某一个用户原型;在用户交互过程中动态获取的规则不断地被用于修改其用户模型,使每个用户兴趣文档既要反映用户感兴趣的主题,又要通过不断地对用户兴趣的学习,了解用户的兴趣变动,不断更新用户兴趣文档。

有鉴于此,在本系统中,用户兴趣模型准备采取综合的方法:首先要求用户注册登记,通过填写表单的方式主动提供自己的兴趣,得到用户的初始模板;然后,通过对用户行为的跟踪分析,充分挖掘 Web 数据来扩充完善用户的个性化向量;最后,利用用户的反馈信息不断修正用户的个性化模板。

3.2　进化式信息过滤代理技术

本试验将进化的原理引入代理技术,通过多个代理的竞争和合作,经过一个类似于生物界中自然进化的过程,使系统性能达到最优。

代理的进化受两个因素控制:它们自身的健康和整个系统的健康。在整个代理当中仅有若干个排序最高的代理才被允许产生后代。一个代理的排序仅仅基于其健康状况。允许产生后代的代理的数量与因表现低劣(健康差)而被排除的代理的数量有关。进化的速度与整个系统的健康状况有关。如果系统的整体健康变差,那么为了寻找对用户新兴趣的适应性就有必要加快进化;如果系统的整体健康增强,进

化会保持一个稳定的可配置比率,允许系统以较慢的速度考察检索空间,获得更好的答案。

代理由基因型和显型两部分组成。代理的可进化部分叫做基因型。代理的其他部分,称作显型,含有不应当被进化的信息,通常是如何处理可进化部分的指令。就本系统来讲,基因型指的就是加权关键词向量。其显型含有该代理的不可进化的部分,如其健康状况、长期兴趣字段,当然还有使代理能相互交流并与系统交流的命令等。显型类似一个固定的模板,该模板由基因型信息填充然后被"执行"。图3对信息过滤代理的基因型和显型之间的关系进行了可视化描述。

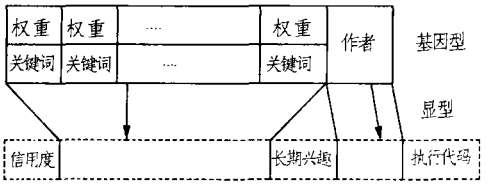

图3 信息过滤代理基因型和显型

新的代理通过遗传或突变(或两者同时进行)来产生。遗传算子运用于代理的可进化部分,该算子随机选择关键词向量中的两点,改变位于其中间的两个父代理的所有字段,产生两个新代理。新的基因型继承了其父代理的关键词向量的一部分。突变是产生下一代代理的另外一种方式。突变算子采用一个代理的基因型作为自变量产生一个新代理,这个新代理是其父代理的一个随机修改版。突变关键词的权重被随机修改,而新的被"突变"关键词是从另一个簇的代理中随机选择出来的关键词,依次类推。

3.3 基于权重向量的文献表征技术

只有将文献表示成计算机可以理解的形式,文献才能自动地和用户模板进行匹配比较,为用户提供个性化的信息服务。文献的表征技术基本上决定了信息过滤中的匹配机制。

本系统中,文献的表征采用的是基于权重向量的标准信息检索技术。在执行了WWW搜索引擎检索后,由"加权关键词向量产生器"(WKVG)对文献进行处理,把所获文献转换成加权关键词向量。为减轻系统运行的负担,本系统选择仅对WWW查找返回的标题和摘要片段进行关键词抽取。由于这一步操作是在WWW检索之后进行的,故不会对过滤精度产生太大的影响。第一步,通过HTML解析器对HTML文献进行解析,把HTML文献转换成文本格式。HTML文献中包含的链接可以用HTML URL检索器检索出来并作为一种特定类型的关键词来处理,只保留一些重要的HTML特征如头标识等,用于产生加权关键词向量。然后将新产生的文本文件通过词干处理程序排除单词的后缀,只保留词根。然后从词干处理程序的输入结果中排除所有经常使用的英语单词(如the,it,for,will等)。对中文文献将采用遗传算法和PAT树相结合的方法而不采用词典法进行关键词抽取[5]。最后,利用一种标准的信息检索加重机制,即tifdf,为每一个关键词计算权重:

$$W_N = H_c T_f idf_k$$

这里,T_f为关键词在当前文献中的词频,H_c为头常数,idf_k按下式定义:

$$idf_k = \log\left(\frac{N}{df_k}\right)$$

N为已由系统检出的文献总数,df_k为K个词语的文献频率。词语idf_k为整个文献集的文献频率。在这种情况下,文献集是所有加权关键词向量的集合,后者为检出文献的内部表征。如果在文献的正文文本中发现关键词,头常数即等于1.0。若相反,关键词为题目的部分,那么其权重将乘以一个常数。这样,题目关键词比一般字体的关键词拥有更大的权重。

通过使用诸如文献的URL、服务器名称以及作者等字段,来增大加权向量,完成向量产生的过程。

3.4 基于向量空间模型的语义聚类技术

聚类是信息过滤中的核心技术。聚类与分类不同:分类是一种监督学习(supervised learning),其类别是根据应用的需要事先确定的,根据表示事物特征的数据可以识别其类别;聚类是一种非监督学习(unsupervised learning),其类别不是人为指定的而是分析数据的结果,聚类完全由计算机自动进行,不需要人工干预。文献聚类能动态地维护类目结构,有助于个性化服务的实现。

文本聚类主要有两个步骤:首先提取文档特征矢量;其次,对文档特征矢量进行聚类,即将文档按特征矢量的不同,分为有限个数的类。当前使用的文献聚类技术可分为两大类:层次聚类技术和分割聚类技

术。前者的代表是凝聚聚类技术,后者的代表是K-均值聚类技术。随着人工智能、神经网络等领域的交叉融合,聚类领域又产生了许多新技术,如模糊聚类、概念聚类、神经网络聚类等。

本系统采用的是基于文献信息关键词的语义聚类模型,即对产生的关键词向量进行语义扩展,是一种基于向量空间模型的语义扩展聚类方法。在系统设计中,我们采用概念扩充的方法和潜语义索引(Latent Semantic Index)技术来建立一个语义或概念空间,实现专题文献的语义聚类。并将过滤结果形成按关键词聚类的列表树,这样用户可以方便地获得自己所需的文献,起到聚类过滤的作用。

基于向量空间模型的文献聚类算法有分割算法中的k-means算法和层次算法中的凝聚层次算法。k-means算法在数据量较小时,有较好的聚类效果,当处理大规模数据时,时间复杂度是$O(n)$,但聚类效果较差;凝聚的层次算法方法简单,聚类效果一般比k-means算法要好,但时间复杂度却是$O(n_2)$。考虑到这里仅对过滤结果进行聚类,数据量不是很大,为了提高系统的相应时间,本系统采用k-means算法进行聚类。

k-means算法以k为参数,把n个对象分为k个簇,使簇内具有较高的相似度,而簇间的相似度较低。(1)在N个对象中随机地选取K个对象作为初始的聚类中心;(2)把其余N-K个对象归到距离最近的聚类中;(3)重新计算每一个聚类的中心;(4)重复(2)和(3),直到每一聚类的中心不再改变。这种算法实质是一种多次迭代的方法,把每篇文档看成一个对象,利用文档与聚类中心之间的相似度来进行聚类。

3.5　主动推送技术

在本系统中,采用电子邮件和屏幕两种信息推送方式。推送的参数由用户自行设定或采用缺省值。参数设置区位于过滤系统的主界面,由用户界面agent来维护和更新。其中包括搜索引擎的选择、新到信息提醒方式、更新周期、系统响应时间等参数的设置。关于搜索引擎的选择,目前系统最大值为5个,缺省搜索引擎为1个(google)。用户可以在该区中选择、增加、删除或修改所用的搜索引擎。搜索引擎的增加可能会增加系统应答的时间并影响过滤的精度。新到信息提醒方式为电子邮件和屏幕推送两种方式(默认是用户的桌面)。更新周期为用户设定的系统自动联网检索相关文献,更新用户兴趣文档的时间。根据专业学术信息数据库更新周期较长的特点,我们将它设置为7~30天,用户在这个范围内可以任意选择。系统响应时间为用户设定的允许系统在一次过滤操作中所耗费的时间。

推送功能是通过新到信息提醒agent实现的。系统对用户每一次提交过滤的结果进行存贮,并在用户兴趣文档中保留了文献获取时间。系统会根据用户所递交的提问,在系统的闲暇时间,对用户所指定的搜索引擎发出查询并进行聚类过滤,与原有文献比较,若发现有新到文献,即通知用户浏览查看。信息提醒内容包括:用户最后一次查询的日期和新到文献提醒日期、新到文献篇数等信息。

参考文献

1　Doug Oard. Information Filtering. 1995, Dec, 12. http://www.glue.umd.edu/~oard

2　李法运.基于Web的信息过滤模型优化及系统实现研究.武汉大学博士学位论文,2004

3　徐小琳,阙喜戎,程时端.信息过滤技术和个性化信息服务.计算机工程与应用,2003(9)

4　Wai Lam, Javed Mostafa. Modeling User Interest Shift Using a Bayesian Approach. Journal of The American Society for Information Science and Technology, 52(5):416-429, 2001

5　Jorng-Tzong Horog, Ching-Chang Yeh. Applying genetic algorithms to query optimization in document retrieval. Information Processing and Management 36(2000):737-759

焦玉英　武汉大学信息管理学院教授,博士生导师。通信地址:武汉市。邮编430072。

刘伟成　武汉大学信息管理学院2003级博士研究生。通信地址同上。

李法运　福州大学管理学院副教授,博士。通信地址:福州市。邮编350002。

(来稿时间:2005-06-07)