●焦玉英 王 娜

数字图书馆中主动信息过滤系统的构建研究*

摘 要 设计了一个结合使用协作过滤和基于内容过滤的主动信息过滤的实验系统。其结构框架的主要部分有:智能代理、检索服务器、用户需求文档数据库、过滤服务器、结果处理器和推送服务器。它采用机器学习的机制来预测用户新的兴趣。图 1。参考文献 7。

关键词 数字图书馆 信息过滤 基于内容的过滤 协作过滤
分类号 G250.76

ABSTRACT The authors design an experimental information filtering system based on collaborative filtering and content-based filtering. Its major components include intelligent agents, retrieval servers, user requirement document databases, filtering servers, result processors and push servers. It can predict new interest of users by machine-learning mechanisms. 1 fig. 7 refs.

KEY WORDS Digital library. Information filtering. Content-based filtering. Collaborative filtering.
CLASS NUMBER G250.76

信息过滤技术是一种系统化方法,可以根据用户的个性化需求,从动态的信息流中提取和用户需求相关的信息。在数字图书馆使用它,某种程度上可以解决信息过载问题,并更好地为用户提供个性化服务。主动信息过滤是信息过滤的一种,这种方法可以主动为用户寻找相关信息。随着用户个性化需求的发展,将主动信息过滤技术应用于数字图书馆,可以基于用户的需求整合信息资源,并能够主动为用户提供个性化的信息服务。

1 主动信息过滤系统及代表

主动信息过滤系统是一种可以帮助用户从特定类型的信息资源数据库中搜寻信息的应用系统。它可以根据被提供的用户兴趣文档,利用各种数据分析技术,为用户检索相关信息并将信息及时准确地提供给用户。它还可以根据用户对推送内容的反馈来改善过滤的效果。Malone 曾把过滤分为两种方法:基于内容的过滤方法和基于协作的过滤方法。目前,基于这两种方法的过滤模型都被广泛应用于主动过滤系统中。随着对主动信息过滤系统的深入研究,许多实验系统已经陆续出现,而且一些系统还被用于数字图书馆。其中,SIFI,Tapestry 和 Fab 被认为是主动信息过滤系统最初发展阶段中最为典型的 3 个系统。许多后续的研究都是建立在这 3 个系统的基础上。分析这 3 个系统对我们研究数字图书馆中的主动信息过滤系统有非常重要的意义。

1.1 SIFT 系统

SIFT(Stanford Information Filter Tool)是美国斯坦福大学 1994 年开发的一个主动信息过滤系统[1]。这是一个典型的基于内容的信息过滤系统。它支持布尔模型和向量空间模型。每当系统获得新的文献列表,过滤服务器就开始逐个地将文献与已存储的用户需求进行匹配。与需求相匹配的文献摘要将被存入用户的目录中。到了一定的时间,SIFT 系统将通过电子邮件或是个人的主页将文献推送给用户。

SIFT 系统的结构包括电子邮件/互联网提问处理器、过滤引擎和提醒器。提问处理器处理被用户提交的电子邮件和填写的表格内容。过滤引擎负责对比文献和用户的需求文档。提醒器则排列过滤的结果,并将文献的摘要逐一提交给每个用户。使用该系统,用户需要设定一个相关性阈值,这个阈值可以表明用户所希望的结果相关性的程度。如果用户没有给出这个阈值,系统可以使用一个缺省值。如果用户对周期性的通知结果感到不满意,也可以修改提问或阈值。相关反馈可以改善整个系统的效果。

SIFT 系统的工作原理和结构都很简单,但是从它的增长率来看,它对用户是非常有效的。例如,1996 年的 4 月,它的增长率是每个月大约 680 名用户

* 本文系国家自然科学基金项目(70473067)的研究成果之一。

和1500个需求文档。当然,该系统也有一些缺陷:用户的文档没有有效期;系统是根据关键词的权重或阈值来判断匹配条件的。

1.2 Tapestry

Tapestry系统是由Goldberg等人在1992年设计和建造的一个支持协作过滤的实验性的邮件系统。它允许用户对读过的文献给出注解。用户通过关键词检索文献时,可以根据其他用户的注解来决定阅读哪些文献。这些注解不仅包括接收与拒绝的建议,还包括文本信息。

它的主要部分包括:索引器、文献存储器、注解存储器、过滤器、小存储箱、回邮器、评价器和阅读器/浏览器。索引器的主要责任是分析文献并建立索引。文献存储器负责为所有的文献和文献索引提供长期存储。注解存储器是为与文献相关的注解提供存储的部分。过滤器会反复地比较用户提问和文献,并将与用户提问相匹配的文献放置于提问者的小存储箱中[2],每个用户都有一个小存储箱。回邮器可以将小存储箱中的内容,通过周期性的电子邮件提交给每个用户。评价器将会根据用户个性化的需求来排列文献。阅读器/浏览器将为用户提供系统界面。

在Tapestry系统中,使用客户端/服务器的模式进行过滤的过程可以分为两个部分。在客户端,根据用户结构文档中复杂的规则确定用户的最终需求文件。在服务器端,通过使用一些简单的规则,新的文献被过滤,而且用户感兴趣的文献也将被确定。系统中的注解主要是用来帮助用户决定需要阅读哪些文献。注解可以将新的和用户感兴趣的文件推荐给用户。这样,系统会更有效,用户也会越来越多。Tapestry系统也有不足之处,因为愿意阅读所有文献并作注解的用户始终缺乏。

1.3 Fab

Fab是斯坦福大学图书馆项目的一个部分。它是一个混合型的过滤系统,该系统可以通过内容分析来构建和维护用户的需求文档,并可以直接将用户的需求文档进行比较,为协作推荐来确定相似的用户群[3]。Fab可以通过结合协作过滤和基于内容的过滤两种方法,减少基于一种过滤方法所造成的一些缺陷。

Fab的设计原则是:每位用户都有一个用户需求文档,这些需求文档由用户评价文献所留下的一些特殊特征构成。需求文档和文献可以用向量来表示,用户需求文档和文献之间的相似度也就可以通过向量间的计算获得。和用户需求文档相似度高的文献将会被推荐给用户。另外,两个用户间的相似度,可以通过对比他们的需求文档来得到。这样,系统可以推荐与上述两个用户有相似兴趣的其他用户的意见。向某一用户推荐的最终结果可以通过整合两种推荐结果来获得。

Fab的主要组成部分有3个:收集代理、选择代理和中部路由器。收集代理可以为某一主题寻找文献并形成一个数据库或是索引。收集代理的需求文档代表了一个动态变化的用户组的兴趣主题。选择代理会为某一用户寻找文献,一个选择代理的需求文档代表了一个用户的兴趣。中部路由器可以将收集代理发现的,与用户需求相匹配的文献推送给用户。

2 代表系统的分析及结论

上述3个系统可以称之为信息过滤系统的基石。SIFT系统是一个典型的基于内容的信息过滤系统,Tapestry是推荐系统中协作过滤应用的先例,而Fab是首个结合基于内容的过滤和协作过滤的实验系统。

2.1 信息过滤的方法

基于内容的过滤,其优势是简单而有效的,尤其是对新的信息资源而言,例如SIFT。但是劣势在于很难区分资源的特点和模式,并且只能发现与用户感兴趣的资源相似的资源,而不是新的、感兴趣的资源。因为这种信息过滤方法只考虑预定的文献内容,故而可能会提供较差的推荐。如果有两种文献有同样的内容特征,它们将被预测为有同样的相关性等级[4]。

协作过滤的优势在于能够发现新的、用户感兴趣的资源但也同时存在两种劣势。就Tapestry系统而言,在系统使用初期对资源的评价还不是很充分,对提供给用户的新文献排序也很困难。还有一个问题是,随着用户和资源的增长,系统的容量也会逐渐减少。这种过滤方法对于先前不知道信息需求,或者是很难表达需求的用户而言,是非常有效的。

通过Fab系统,我们可以发现:协作过滤可以克服基于内容过滤的一些缺陷,这是因为它的用户需求文档是从各个方面对用户特征化的,这样的用户文档更为全面。另外,在混合型系统中,存在着两种过滤方法无法单独提供的优势。因此,结合了基于内容的过滤和协作过滤两种方法的混合型过滤系统,是信息过滤系统发展的一个趋势。

2.2 信息过滤系统的反馈机制

在信息过滤系统中的用户需求文档可以通过用

户的反馈来更新，用户的反馈也可以提高过滤效果。在SIFT系统中，对于向量空间模型而言，相关反馈技术可以用于修改用户的提问。用户只需要提供给系统自己感兴趣的文献。检测完这些文献，SIFT系统的服务器可以在用户的提问中调整关键词的权重。用户所提供的相关性的阈值也可以通过相关性反馈来修正。作为一个协作过滤系统，Tapestry使用了隐式反馈来调试或修正过滤提问。隐式反馈包括对传送给用户的文献的答复、与文献相关的注解等等。Fab系统则是一个混合型的系统，它的反馈机制同SIFT和Tapestry系统不同。当一个用户向系统提问时，同用户个人的需求文档最为匹配的文献将被选出。然后用户会对这些文献给出相关性等级的评定，这些评定将被用来作为反馈。系统的选择代理会用反馈来更新用户的个人需求文档。

相关性反馈技术被广泛地用于上述系统。通过修改用户的文档，相关性反馈可以被用于提高信息过滤系统的有效性。但是这种反馈机制是基于用户的行为。为了减轻用户负担，相关性反馈应该被更好地发展。

2.3 结果的处理

对过滤后的结果进行处理，可以减少用户的浏览时间，而且处理后的结果也更便于用户选择。SIFT系统允许用户设定一个相关性阈值，并可以使用缺省值来帮助无法提供阈值的用户。过滤系统将根据一系列的兴趣关键词和阈值来过滤文献。如果文献和需求文档间的相似性高于阈值，文献将被传递给用户。在Tapestry系统的评价器可以应用个性化的分类，自动地排序和分类文献。用户可以手动地改变文献的优先级别。Tapestry系统分两个步骤来过滤文献。首先是通过过滤的提问，将用户满意的文献放入用户的小存储箱中。第二步由评价器来完成，评价器在小存储箱中浏览文献并将结果提供给用户。在Fab系统中，被收集代理发现的文献将被推送至中部路由器。然后中部路由器推送这些文献给那些需求文档与文献匹配的用户。接下来，用户则必须将文献进行相关性排序。然后，系统将会根据用户的相关性排序排列用户收到的文件，也就是用户的偏好等级。

信息过滤的目的就是帮助用户发现符合需求的信息。SIFT系统基本不对结果进行处理，这样，用户必须花费一定时间来选择更为相符的结果。在Tapestry和Fab系统中有一些改进过的措施，但是这些措施都是基于用户的主动行为。为了从根本上达到过滤目的，一些自动措施应该被信息过滤系统采用。

3 主动信息过滤实验系统模型

鉴于上述分析，作者试图设计一个新的主动信息过滤系统来继承经典系统的优势，例如混合模型，并通过采用一些新的技术来克服它们的某些缺陷。本文设计了一个结合使用协作过滤和基于内容过滤的主动信息过滤的实验系统。图1显示了实验系统结构框架的主要部分：智能代理、检索服务器、用户需求文档数据库、过滤服务器、结果处理器和推送服务器。

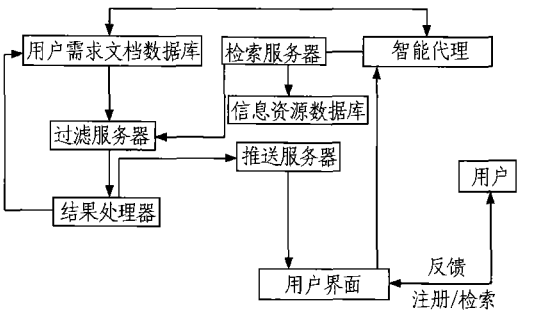

图1 主动信息过滤实验系统结构

当一个用户使用该系统时，他可以通过用户界面将信息输入系统。如果他是新用户，则须提供相关的注册信息给系统。注册后，用户可以登录系统并提交检索提问。已经使用过系统的用户则可直接登录系统，并上传提问或给出上次过滤结果的反馈信息。

所有信息将会从用户界面直接传送给智能代理。智能代理则具备5个功能。第一，会分析新用户的注册信息并建立起用户需求文档。第二，会根据相似的兴趣将用户需求文档划归为某一用户组，并且会根据用户组的兴趣来预测用户更多的兴趣。第三，通过该模块将用户的提问传递至检索服务器。第四，本模块可以根据用户的反馈修改用户的需求文档，并承担适应性学习的任务。第五，根据用户所预定的提醒时间，可以从用户的需求文档数据库中提取用户的文档，并按照文档中的时间将用户提问输入检索服务器。

所有的用户信息将被传递至用户的需求文档数据库，并被保存。数据库中存储的数据包括：用户的兴趣文档、用户组的信息、用户的个性化定制信息、最后一次过滤结果和需要提醒服务的用户的过滤时间。智能代理执行提醒服务时，将会提取用户的需求文

档、上一次的检索结果和检索时间,并将这些信息都输入过滤服务器。

将用户的信息输入用户需求文档数据库后,智能代理将把检索提问传递至检索服务器中。检索服务器会根据用户需求,负责在数字图书馆的信息资源数据库中检索,然后将检索结果输入过滤服务器。过滤服务器会将用户的需求文档和检索结果进行匹配。过滤服务器也将根据用户组的推荐来搜寻文献。然后,匹配后的结果将被提供给结果处理器。结果处理器会进一步处理结果并将其排序。最后,推送服务器将根据用户需求,将排序后的结果传送给用户。最终推送给用户的内容包括:最终的检索时间、新文献的数量、新文献的题目。本次检索的结果和时间将也被输入用户的需求文档数据库,用于系统对用户的提醒服务。

除了过滤信息,提醒服务也是该系统提供给用户的一种新的功能。用户如果想要使用提醒服务,只需通过系统界面来定制提醒的条件即可。这些条件可以通过智能代理传递给用户的需求文档数据库。这样,当提问被提交给检索服务器时,检索服务器会将检索结果与用户的需求文档和上次的检索结果相比较。如果有用户定制的新信息,推送服务器会将结果按照用户指定的时间和方式推送给用户。

4　实验系统模型的实现研究

为提高主动信息过滤系统的有效性,并减轻用户负担,可以考虑将一些较为成熟的技术用于该新模型的实现中。这些技术可以提高系统自动化和动态化的程度,并改善系统的过滤效果。

4.1　智能代理模块

在该系统中,智能代理模块可以处理各种信息,并从用户反馈中进行自我学习。可以在该模块中使用两种类型的代理:用户代理和提醒代理。用户代理可以通过用户的注册信息或跟踪用户行为,了解用户的兴趣和偏好;可以根据用户的反馈进行学习和训练[5]。然后,用户代理可以不断地修正用户的需求文档,以使需求文档更接近用户的需求。提醒代理可以从用户的需求文档数据库中,提取用户的提问、上一次的检索结果和检索时间。提醒代理将会把用户提问输入检索服务器,并将上一次的检索结果和时间传递给过滤服务器。

在该系统的智能代理中,基因算法可以被采用作为自动学习的方法。自然选择造成了适者生存,基因模式可以通过一代一代的个体继承下来[6]。在该系统中,通过使用向量空间模型,文献被表示为向量。在本模型中,一个基因可以被表述成一个条款,一个个体可以被表示为向量空间中的文献,而一个组可以被表述为用户需求文档。该方法的原则是:是否根据生存进程来更新用户的需求文档。使用这种方法,相关性反馈可以同基因算法互相作用。相关性反馈可以通过修改每一个关键词的权重影响以后的检索。基因算法的部分可以增加一些词到用户的需求文档中,并修改逆文献的频率值。

4.2　结果处理器模块

结果处理器模块可以为用户进一步地过滤结果,也是该系统的核心成分。在该模块中,聚类技术可以用于结果排序。使用聚类技术不仅可以保存分类的结构,也可以帮助用户更容易地做出选择。在聚类技术中的分类不是人工规定的,而是分析数据所得出的结果。文本聚类可以分为两个步骤:首先从文本中提取特征向量;然后根据不同的文本特征向量,文本可以被聚类为有限的类别。

K-means 聚类算法是最为简单和被普遍使用的采用平方差标准的一种算法[7]。在数据数量不大的情况下,它可以显示出更好的聚类效果。当算法被用于大量数据的聚类时,它的时间复杂度是 O(n)。在该系统是对过滤结果采用聚类技术,而结果的数据量并不是很大,因此为了提高响应周期,K-means 算法可以被用于该系统。

K-means 算法的原理是,将 K 作为一个参数,然后将 N 个对象分为 K 个簇。在簇内有更高的相似度,而在簇间则有更低的相似度。详细步骤是:首先,K 个对象应该是被任意选择的,这 K 个对象则被作为 N 个对象的簇的中心。然后,其他的 N—K 个对象将被分配到最相近的簇中心。接着,所有的簇中心都应该通过使用当前的簇成员间关系,进行重新计算。最后,如果不符合聚合的标准,则重复后两个步骤。典型的聚合标准是:对于新的聚类中心,没有(或只有最小的)模式再分配。这种算法的核心是一种反复的方法,就是将每个文献都视为一个对象,而且利用文献间的相似性和簇中心来聚类文献。

4.3　推送服务器模块

在实验系统中,推送服务器是负责通过电子邮件和屏幕界面的方式,将过滤结果推送给用户。推送服务器可以采用用户指定的参数或是缺省值来将结果推送给用户。参数是可以通过智能代理来修改的,推

送周期也应该由用户指定和存储在用户的需求文档中。

推送技术是一种软件，也可以成为广播技术。这种软件可以根据用户限定的标准，来收集用户感兴趣的信息。而且它可以在用户指定的时间，将用户定制的信息传送到用户指定的地方。该实验系统采用的是一种周期性推送的方法。用户通过智能代理可以指定和修改推送的周期。鉴于数字图书馆使用的是分布式信息系统，推送技术对于该系统来讲是一个好的选择。在本实验系统中给用户推送的内容是经过过滤的文献。按照这种方法，推送的内容可以被处理和组织，以使用户不会感到被推送的结果是些过载的信息。

参考文献

1 Tak W. Yah, Hector Garcia-Molina. The SIFT Information Dissemination System. ACM Transactions on Database Systems, 12(1999)529-565

2 Goldberg D, et al. Using Collaborative Filtering to Weave an Information Tapestry. Communication of the ACM, 12 (1992)

3 Balabanovic M, Shoham Y. Fab: Content-based, Collaborative Recommendation. Communications of the ACM, 3 (1997)

4 Kim, BD, Kim, SO. A new recommender system to combine content-based and collaborativefiltering systems. Journal of Database Marketing, 6(2001)

5 Alexander Serenko, Brian Detlor. Intelligent agents as innovations. AI & Soc. 18(2004)

6 Uri Hanani, Bracha Shapira, Peretz Shoval. Information Filtering: Overview of Issues, Research and Systems. User Modeling and User-Adapted Interaction, 11(2001)203-259

7 Jain, A. K, Murty, M. N. , Flynn. Data clusting: A view. ACM Computing Surveys, P. J. 31(1999)

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

王 娜 武汉大学信息管理学院2005级情报学博士生。通讯地址同上。 （来稿时间：2006-09-27）

（上接第21页）制与惩罚措施执行中的政府角色及政府问责制度等。

2.3 亟待探索的政策领域

我国政府信息资源公益性开发服务的政策建议，一方面是要以开放的精神对现有法律与法规进行审查，尽快出台政府信息开放的专门法律制度；另一方面是探索过去从未涉及的一些政策领域。主要涉及：(1)明确政府部门在信息资源公益开发服务中的主体地位与具体任务，对企业、个人和其他组织进入和退出政府信息资源公益性开发服务领域的具体政策进行细化。(2)明确政府信息资源公益性开发服务的监管部门及其职责。(3)制定政府信息资源公益性开发服务的专项基金政策与管理制度，改革并完善对有关制度公益人的财政拨款制度，明确对参与政府信息资源公益性开发服务的营利组织的各种优惠政策。(4)制定政府信息资源公益性开发服务项目"外包"的运作与管理政策。(5)制定政府信息资源公益性开发服务的质量责任制度，出台有关服务质量评价标准。(6)有效保证政府信息资源产品得以全社会共享的技术标准与管理标准。

参考文献

1 王素芳. 我国信息资源开发利用政策法规初探. 情报学报，2004(3)

2 陶传进. 社会公益供给：NPO、公共部门与市场. 北京：清华大学出版社，2005：80.

3 周汉华. 政府信息公开条例(专家建议稿). 北京：中国法制出版社，2003：115.

4 李丹. 试论我国信息政策建设存在问题及应对策略. 晋图学刊，2004(2)

5 周毅. 我国情报政府效应偏差分析. 图书情报工作，1994(2)

6 周毅. 政府信息开放与开发的社会化和商业化：趋势、领域与问题. 中国图书馆学报，2005(6)

7 陈能华等. 美国信息资源共享市场的发展及启示. 中国图书馆学报，2006(5)

8 王正兴等. 英国的信息自由法与政府信息共享. 科学学研究，2006(5)

周 毅 苏州大学教授，博士。通讯地址：苏州大学东校区533信箱。邮编215021。 （来稿时间：2006-11-17）