个性化服务技术综述[1]

曾 春, 邢春晓, 周立柱

(清华大学 计算机科学与技术系,北京 100084)

E-mail: bobofu00@mails.tsinghua.edu.cn; {xingcs,dcszlz}@tsinghua.edu.cn

http://dbgroup.cs.tsinghua.edu.cn

摘要: 对个性化服务技术中用户描述文件的表达与更新、资源描述文件的表达、个性化推荐技术、个性化服务体系结构以及该领域的主要研究成果进行了综述.通过比较现有原型系统的实现方式,详细讨论了实现个性化服务的关键技术.此外,分析了 3 个具有代表性的个性化服务系统.最后对个性化服务技术进一步研究工作的方向进行了展望.

关 键 词: 个性化;基于规则的个性化;信息过滤技术;基于内容的过滤;协作过滤;推荐系统

中图法分类号: TP393 **文献标识码**: A

Web 已成为人们获取信息的一个重要途径,由于 Web 信息的日益增长,人们不得不花费大量的时间去搜索.浏览自己需要的信息.搜索引擎(search engine)是最普遍的辅助人们检索信息的工具,比如传统的搜索引擎 AltaVista,Yahoo 和新一代的搜索引擎 Google 等.信息检索技术满足了人们一定的需要,但由于其通用的性质,仍不能满足不同背景、不同目的和不同时期的查询请求.个性化服务技术就是针对这个问题而提出的,它为不同用户提供不同的服务,以满足不同的需求.个性化服务通过收集和分析用户信息来学习用户的兴趣和行为,从而实现主动推荐的目的.个性化服务技术能充分提高站点的服务质量和访问效率,从而吸引更多的访问者.

目前存在着许多个性化服务系统[1],它们提出了各种思路以实现个性化服务.个性化服务系统根据其所采用的推荐技术可以分为两种:基于规则的系统和信息过滤系统.信息过滤系统又可分为基于内容过滤的系统和协作过滤系统.

基于规则的系统如:IBM 的 WebSphere(www.ibm.com/websphere),BroadVision(www.broadvision.com),ILOG(www.ilog.com)等,它们允许系统管理员根据用户的静态特征和动态属性来制定规则,一个规则本质上是一个 If-Then 语句,规则决定了在不同的情况下如何提供不同的服务.基于规则的系统其优点是简单、直接,缺点是规则质量很难保证,而且不能动态更新,此外,随着规则的数量增多,系统将变得越来越难以管理.

基于内容过滤的系统如:Personal WebWatcher[2],Syskill & Webert[3],Letizia[4],CiteSeer[5],ifWeb[6],SIFTER[7],PVA[8],WebMate[9],WebACE[10],ELFI[11]和 WebPersonalizer[12]等,它们利用资源与用户兴趣的相似性来过滤信息.基于内容过滤的系统其优点是简单、有效,缺点是难以区分资源内容的品质和风格,而且不能为用户发现新的感兴趣的资源,只能发现和用户已有兴趣相似的资源.

协作过滤系统如:WebWatcher[13],Let's Browse[14],GroupLens[15],Firefly[16],SELECT[17],LikeMinds(www.macromedia.com),和 SiteSeer[18]等,它们利用用户之间的相似性来过滤信息.基于协作过滤系统的优点是能为用户发现新的感兴趣的信息,缺点是存在两个很难解决的问题,一个是稀疏性,亦即在系统使用初期,由于

[1] 收稿日期: 2001-12-09; 修改日期: 2002-07-01

基金项目: 国家重点基础研究发展规划 973 资助项目(G1999032704)

作者简介: 曾春(1976－),男,江西萍乡人,博士生,主要研究领域为数字图书馆,个性化服务技术;邢春晓(1967－),男,河南南阳人,博士,副教授,主要研究领域为海量信息处理及其在数字图书馆中的应用;周立柱(1947－),男,江苏连云港人,教授,博士生导师,主要研究领域为数据库,海量信息处理,Web 技术.

系统资源还未获得足够多的评价,系统很难利用这些评价来发现相似的用户.另一个是可扩展性,亦即随着系统用户和资源的增多,系统的性能会越来越低.

还有一些个性化服务系统如:WebSIFT[19],FAB[20],Anatagonomy[21]和 Dynamic Profiler[22]等,同时采用了基于内容过滤和协作过滤这两种技术.结合这两种过滤技术可以克服各自的一些缺点,为了克服协作过滤的稀疏性问题,可以利用用户浏览过的资源内容预期用户对其他资源的评价,这样可以增加资源评价的密度,利用这些评价再进行协作过滤,从而提高协作过滤的性能.

本文第 1 节讨论实现个性化服务的关键技术.第 2 节分析 3 个具有代表性的个性化服务系统.第 3 节总结全文并对进一步研究工作的方向进行展望.

1 个性化服务的实现

为了实现个性化服务,首先需要跟踪和学习用户的兴趣和行为,并设计一种合适的表达方式.为了把资源推荐给用户,必须组织好资源,选取资源的特征,并采用合适的推荐方式.此外,还必须考虑系统的体系结构,考虑在服务器端、客户端和代理端实现的利弊.下面,我们从用户描述文件的表达与更新、资源描述文件的表达、个性化推荐以及体系结构这 4 方面讨论个性化服务的实现.

1.1 用户描述文件

对个性化服务系统来说,最重要的是用户的参与,为了跟踪用户的兴趣与行为,有必要为每个用户建立一个用户描述文件(user profile).用户描述文件刻画用户的特征与用户之间的关系.在制定用户描述文件之前,需考虑下面几个问题:① 有没有现成的标准?② 收集什么数据?收集的数据用于什么目的?③ 如何收集数据?根据什么信息源来收集?④ 收集的数据如何组织?⑤ 用户信息能否自适应地更新?

用户描述文件还没有一个统一的标准,如 W3C(www.w3c.org)有两个涉及用户描述文件的标准:PICS(platform for internet content selection)和 APPEL1.0(a P3P preference exchange language 1.0),PICS 是父母和老师用来控制孩子的浏览能力的,提供了过滤规则定义语言 PICSRules.APPEL1.0 可定义用户感兴趣的站点和过滤规则,这些规则大部分是在 PICSRules 的基础上发展起来的.此外,Netscape,Firefly 和 VeriSign 曾向 W3C 的 P3P(platform for privacy preferences)工作组提交了一个 OPS (open profiling standard)草案,由于目前 P3P 版本不打算考虑如何进行数据传输,因此该草案被搁置一边,OPS 描述了如何表示一个用户描述文件以及用户与 Web 站点交互的问题.

在收集用户的信息之前,首先需分析用户愿意提供什么信息,用户一般都很注意个人信息的保密性[23],www.cyberdialogue.com的调查显示,80%的用户愿意向 Web 站点提供自己的姓名、性别、年龄、教育背景和兴趣,但大多数用户不愿意提供私有、敏感的信息,比如个人收入和信用卡号等,该公司另一项调查显示,28%的用户愿意 Web 站点向其他 Web 站点共享自己的信息.为了规范 Web 用户信息的保密性,W3C 成立了 P3P 工作组来解决这个问题,它允许用户有选择地向 Web 站点提供自己的信息,从而达到保护用户信息的目的,目前已有一些站点和浏览器支持了 P3P,比如 www.w3c.org,www.microsoft.com,www.aol.com,www.att.com 等站点和 Microsoft/AT&T P3P 浏览器等等,但还处于试用阶段.

1.1.1 用户描述文件的表达

不同个性化服务系统的用户描述文件各有其特点,用户描述文件从内容上可以划分为基于兴趣的和基于行为的两种类型[24].基于兴趣的用户描述文件可以表示为加权矢量模型、类型层次结构模型、加权语义网模型、书签和目录结构等.基于行为的用户描述文件可以表示为用户浏览模式或访问模式.在具体实现时可以综合基于兴趣和基于行为这两种表达方式.

用户描述文件可以用文件来组织,也可以用关系数据库或其他数据库来组织.目前有一些系统采用基于 XML 的 RDF(resource definition framework)来表达用户描述文件,并利用支持 XML 的数据库系统来存储用户描述文件,这样,不仅利用了 XML 的优点,也保持了系统的性能.表 1 从用户描述文件的表达、学习的信息源两方面比较了几个典型的个性化服务系统.

Table 1 Comparison of several prototypes in user profile

表 1 各个原型系统在用户描述文件方面的对比

Prototype①	Representation of user profile②	Data source for learning③
BroadVision	用户静态信息	用户注册信息
Personal WebWatcher	基于加权关键词矢量,隐式创建与更新	利用指向文档的超链内容
Syskill & Webert	表示为兴趣类,基于加权关键词矢量,隐式创建,显式反馈更新	用户显式反馈的信息
Letizia	基于加权关键词矢量,隐式创建与更新	用户隐式反馈的信息,访问和标记某网页等行为
CiteSeer	一个文件的集合,集合中每个文件可以包含关键词、URLs、引用等,允许显式或隐式创建,允许显式或隐式更新	用户行为和对推荐文档的反应
ifWeb	基于加权语义网,表达关键词和它们之间的上下文关系,考虑用户感兴趣和不感兴趣的内容	用户显式和隐式反馈的信息
PVA	表示为个人视图,是一种类型层次结构,表达领域的知识,隐式创建和更新	Proxy 日志信息
WebPersonalizer	从 Web 访问日志和站点文件脱机产生的 URL 聚类	用户浏览行为
GroupLens	用户个性信息是放在数据库中,基于关键词矢量,显式创建、显式反馈或隐式更新	显式反馈信息和用户在某页所花的时间
SELECT	基于加权关键词矢量,显式创建、显式反馈或隐式更新	阅读的文档、阅读文档所花的时间和添加书签等行为
SiteSeer	用户书签和目录结构等信息,显式创建、显式更新.	书签、引用文件的内容、用户定义的目录类型
WebSIFT	用户浏览记录,隐式创建与更新	Web 访问日志
Anatagonomy	基于加权关键词矢量,显式创建、显式反馈或隐式更新	用户行为和显式反馈信息

①原型系统,②用户描述文件的表达,③学习的信息源.

1.1.2 用户信息的收集与更新

在用户第 1 次使用个性化服务系统的时候,系统可以要求用户注册自己的基本信息和感兴趣的内容,系统也可以隐式地收集用户信息.在定制好一个用户描述文件之后,系统可以让用户自主修改,也可以由系统自适应地修改,这样,系统就可以随用户兴趣的变化而变化.系统要自适应修改用户信息,必须根据学习的信息源分析当前用户的行为,从而调整用户兴趣的权重或调整用户兴趣层次结构.根据学习的信息源,用户跟踪的方法可分为两种:显式跟踪和隐式跟踪.显式跟踪是指系统要求用户对推荐的资源进行反馈和评价,从而达到学习的目的.隐式跟踪不要求用户提供什么信息,所有的跟踪都由系统自动完成,隐式跟踪又可分为行为跟踪和日志挖掘.

显式跟踪是简单而直接的做法,系统可以要求用户反馈自己对推荐资源的喜好程度.一般情况下,这种做法很难收到实效,因为很少有用户向系统主动表达自己的喜好.比较实际的做法是行为跟踪,因为用户的很多动作都能暗示用户的喜好.用户行为可以表现为查询、浏览页面和文章、标记书签、反馈信息、点击鼠标、拖动滚动条、前进、后退等等,文献[25]的研究表明,简单的动作(比如点击鼠标)不能有效地揭示用户的兴趣,而浏览页面和拖动滚动条所花的时间可以有效地揭示用户的兴趣.文献[4]的研究表明,用户查询、访问页面、标记书签能有效揭示用户的兴趣.

目前,基于 Web 日志的挖掘技术发展迅速[26~28],利用 Web 日志可以获得页面的点击次数、页面停留时间和页面访问顺序等信息.通过分析 Web 日志可以获得相关页面、相似用户群体和用户访问模式等信息,个性化服务系统可以利用这些信息创建或更新用户描述文件.Web 日志挖掘中最常使用的方法是根据网页的点击次数来评价用户对该网页的兴趣,其实这种方法是不完整的,而且经常是不正确的.但该方法可用于辅助其他日志分析技术.尽管 Web 日志的信息不够全面,但还是可以从中发现许多有意义的信息,比如通过收集用户顺序请求的日期和时间,可以分析出用户在每个资源上所花费的时间,从而可以推断用户对该资源感兴趣的程度;通过收集用户感兴趣的领域,有利于对用户感兴趣的内容进行分类;通过分析用户请求的顺序有利于预测用户将来可能的行为,从而推荐合适的信息.

一般 Web 日志挖掘可分为 4 步:① 首先清除 Web 日志中无关的信息,比如请求失败信息,页图片请求信息

等等,然后将剩下的数据存放到数据库中.② 将 URL、动作、资源的类型、大小、请求的时间、请求者域名、用户、服务器状态作为维变量构建数据立方体.③ 进行在线分析处理,通过对数据立方体的切块和切片,分析用户在不同域的分布情况,分析用户对资源的使用情况等等.④ 利用各种数据挖掘方法来预测、分类和发现有意义的关系,比如用户的行为模式、用户行为的变化、不同用户群在使用和行为上的相似性等等.

1.2 资源描述文件

个性化服务系统所应用的领域决定了它所处理的资源.Anatagonomy,SmartPush[29]应用的领域是报纸;GroupLens 应用的领域是 Usenet 新闻;CiteSeer 应用的领域是科技文档;FireFly 应用的领域是音乐和电影;Amazon.com,eBay 应用的领域是电子商务;还有一些个性化服务系统并不面向特定的领域,它们用于导航、推荐、帮助或搜索,不过它们所处理的资源不太相同,比如 Personal WebWatcher,WebWatcher,Letizia 处理 Web 页与链接;WebSIFT 处理 Web 访问日志;SiteSeer,PowerBookmark[30]处理 BookMark 和相关文档;Syskill & Webert,ProFusion[31]处理从其他搜索引擎返回的查询结果;还有一些处理 E-mail[32]等等.目前,个性化服务系统所处理的资源都属于文本范畴,FireFly 面向音乐和电影,其实现是通过用户评价喜欢的音乐家和电影来进行协作过滤的,所以仍然属于文本处理.

资源的描述与用户的描述密切相关,一般的做法是用同样的机制来表达用户和资源,资源描述文件可以用基于内容的方法和基于分类的方法来表示,下面从这两方面分析文档资源描述文件的表达.

1.2.1 基于内容的方法

基于内容的方法是从资源本身抽取信息来表示资源,使用最广泛的方法是用加权关键词矢量.对文档来说,关键的问题是特征选取,特征选取要达到两个目标:一是选取最好的词;二是选取的词最少.要抽取特征词条,需要对文档进行词的切分,在切分的同时,利用停用词列表(stop word)从文档特征集中除去停用词,在完成词切分后,接着除去文档集中出现次数过少和过多的词.经过这些处理后,特征数目一般还很大,还需对特征进行进一步的选取,以降低特征的维数.特征选取的方法很多,比较简单的做法就是计算每个特征的熵,选取具有最大熵值的若干个特征;也可以计算每个特征的信息增量(information gain),也就是计算每个特征在文档中出现前后的信息熵之差;还可以计算每个特征的互信息(mutual information),也就是计算每个特征和文档的相关性;还可使用 X^2 统计方法.文献[33]的对比研究表明,信息增量方法和 X^2 统计方法表现较好,但这两种方法的计算量比较大.

在完成文档特征的选取后,还得计算每个特征的权值,使用最广泛的是 TFIDF 方法,对某一特征,TF(term frequency)表示该特征在文档中出现的次数,IDF(inverse document frequency)表示 log(所有文档数/包含该特征的文档数).矢量模型的代价是比较大的,有时为了加快处理速度,可以只考虑 TF 一项,文献[34]对比研究了矢量模型在只考虑 TF,IDF 以及没有考虑 TF 和 IDF 等几种情况,该研究表明,单独考虑 TF 或 IDF 时都使效果显著下降.

1.2.2 基于分类的方法

基于分类的方法是利用类别来表示资源,对文档资源进行分类有利于将文档推荐给对该类文档感兴趣的用户.文本分类方法有多种,比如:朴素贝叶斯(Naï ve-Bayes),k 最近邻方法(KNN)和支持向量机(SVM)等.

资源的类别可以预先定义,也可以利用聚类技术自动产生.许多研究表明:聚类的精度非常依赖于文档的数量,而且由自动聚类产生的类型可能对用户来说是毫无意义的,因此可以先使用手工选定的类型来分类文档,在没有对应的候选类型或需要进一步划分某类型时,才使用聚类产生的类型.

1.3 个性化推荐

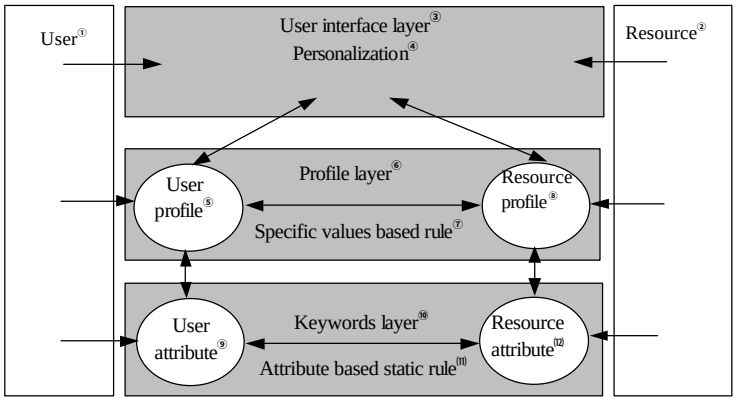
个性化推荐可以采用基于规则的技术、基于内容过滤的技术和协作过滤技术,前面已经提到支持这些技术的个性化服务系统,现在从实现角度分析这几种技术.

1.3.1 基于规则的技术

规则可以由用户定制,也可以利用基于关联规则的挖掘技术来发现[35],利用规则来推荐信息依赖于规则的质量和数量,基于规则的技术其缺点是随着规则的数量增多,系统将变得越来越难以管理.一个规则本质上是一

个 If-Then 语句,规则可以利用用户静态属性来建立,也可以利用用户动态信息来建立.为了利用规则来推荐资源,用户描述文件和资源描述文件需用相同的关键词集合来进行描述.信息推荐时的工作过程是这样的:首先根据当前用户阅读过的感兴趣的内容,通过规则推算出用户还没有阅读过的感兴趣的内容,然后根据规则的支持度(或重要程度),对这些内容排序并展现给用户.

基于规则的系统一般分为 3 部分(如图 1 所示):关键词层、描述层和用户接口层.关键词层提供上层描述所需的关键词,并定义关键词间的依赖关系,在该层可以定义静态属性的个性化规则.描述层定义用户描述和资源描述,由于描述层是针对具体的用户和资源,所以描述层的个性化规则是动态变化的.用户接口层提供个性化服务,根据下面两层定义的个性化规则将满足规则的资源推荐给用户.



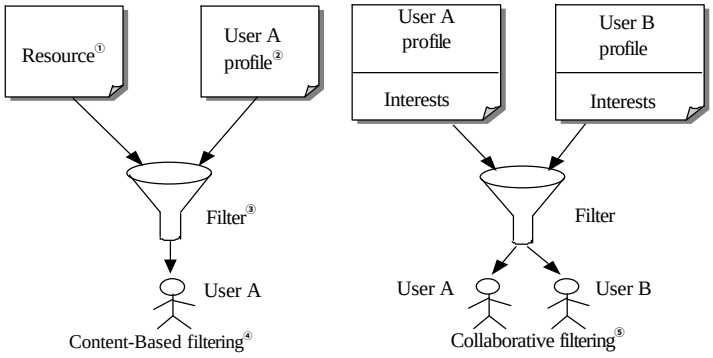
①用户,②资源,③用户接口层,④个性化,⑤用户描述文件,⑥描述层,⑦基于特定值的规则,⑧资源描述文件,⑨用户属性,⑩关键词层,⑪基于属性的静态规则,⑫资源属性.

Fig. 1 Rule-Based technology

图 1 基于规则的技术

1.3.2 信息过滤技术

信息过滤技术可分为基于内容过滤的技术(content-based filtering)和协作过滤技术(collaborative filtering),如图 2 所示.基于内容过滤的技术是通过比较资源与用户描述文件来推荐资源.它的关键问题是相似度计算,对于矢量空间模型来说,通常采用的方法是余弦度量.如果用户的描述文件没有正确描述用户的兴趣和行为,那么该方法推荐的数据可能和用户真正的兴趣根本不相关.基于内容过滤的系统其优点是简单、有效,缺点是难以区分资源内容的品质和风格,而且不能为用户发现新的感兴趣的资源,只能发现和用户已有兴趣相似的资源.



①资源,②描述文件,③过滤器,④基于内容的过滤,⑤协作过滤.

Fig.2 Information filtering technology

图 2 信息过滤技术

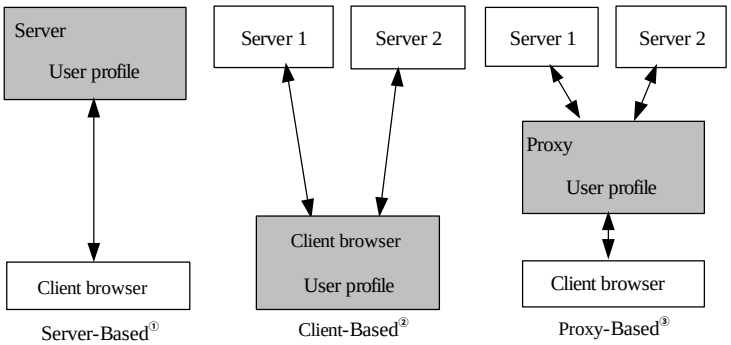
协作过滤是根据用户的相似性来推荐资源.它与基于内容的过滤技术不同,它比较的是用户描述文件,而不

是资源与用户描述文件.其关键问题是用户聚类.由于它是根据相似用户来推荐资源的,所以有可能为用户推荐出新的感兴趣的内容.

基于近邻用户的协作过滤技术应用比较普遍,它的核心问题是为当前用户寻找 k 个最相似的邻居来预测当前用户的兴趣.该方法在实践过程中遇到两个很难解决的问题,一个是稀疏性,也就是指在系统使用初期,由于系统资源还未获得足够多的评价,该方法很难利用这些评价来发现相似的用户.另一个是可扩展性,也就是指随着系统用户和资源的增多,该方法性能会越来越低.对第 1 个问题,文献[36]提出了基于内容的协作过滤方法,也就是利用用户浏览过的资源内容来预期用户对其他资源的评价,这样可以增加资源评价的密度,并利用这些评价再进行协作过滤,从而提高协作过滤的性能.文献[37]提出了 LSI(latent semantic indexing)方法来降低维空间,增加数据的密度,从而更容易发现用户间的相似性.对第 2 个问题,人们提出了基于规则[38]、聚类方法、贝叶斯网[39]、Horting 图[40]、基于近邻资源的协作过滤方法[41]等,它们通过预先建立一些反映相关性或相似性的模型,从而提高系统在预测和推荐时的性能.

1.4 个性化服务体系结构

基于 Web 的个性化服务体系结构和用户描述文件分布的位置有很大的关系.用户描述文件可以存放在服务器端、客户端、代理端,如图 3 所示.大部分个性化服务系统的用户描述文件都存放在服务器端,比如 Syskill & Webert,Letizia,GroupLens,Anatagonomy 等等,它的优点是可以避免用户描述文件的传输,除了支持基于内容的过滤,还可以支持协作过滤.缺点是用户描述文件不能在不同的 Web 应用之间共享.也有一些系统的用户描述文件是存储在客户端的,比如 PointCast Network(www.pointcast.com),这种体系的个性化服务可以在服务器端实现,也可以在客户端实现,它的优点是用户描述文件可以在不同的应用之间共享,缺点是只能进行基于内容的过滤.还有一些系统的用户描述文件是存储在代理上的,比如 Personal WebWatcher,PVA 等,这种体系的个性化服务可以在服务器端实现,也可以在代理上实现,它的优点是不仅可以支持基于内容的过滤和协作过滤,还支持用户描述文件在不同 Web 应用之间的共享,缺点是可能需要传输用户描述文件.



①基于服务器端,②基于客户端,③基于代理端.

Fig. 3 Architecture of personalization

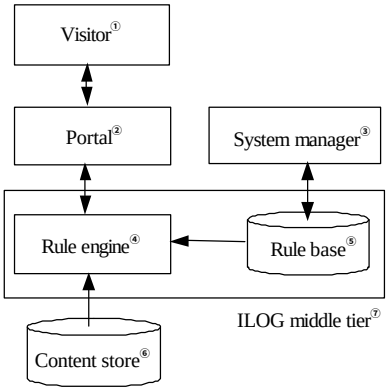
图 3 个性化服务体系结构

2 个性化服务系统

每个个性化服务系统都有自己的特点,下面分析一下具有代表性的 3 个系统:基于规则的系统 ILOG、基于内容过滤的系统 Personal WebWatcher 和基于协作过滤的系统 GroupLens.

2.1 ILOG

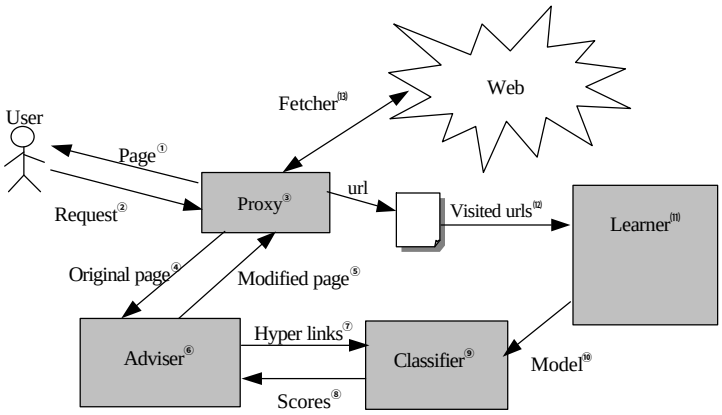
ILOG 的个性化服务是基于规则的(如图 4 所示).系统管理员只需定义业务规则.系统的核心是规则引擎,它用于解释规则,并为站点的访问者产生符合其兴趣的动态内容.ILOG 是作为一个中间件形式提供的,提供 Rules(C++)和 JRules(java)两种组件用于第 2 次开发.此外,ILOG 还提供了一种业务规则定义语言.



①访问者,②入口,③系统管理员,④规则引擎,⑤规则库,⑥内容库⑦ILOG 中间层.

Fig.4 Architecture of BroadVision

图 4 BroadVision 体系结构



①网页,②请求,③代理,④原始网页,⑤修改后网页,⑥建议器,⑦超链,⑧得分,⑨分类器,⑩模型,⑪学习器⑫访问过的 urls,⑬收集器.

Fig.5 Architecture of personal WebWatcher

图 5 Personal WebWatcher 体系结构

2.2 Personal WebWatcher

Personal WebWatcher 的个性化服务是在服务器端提供的.它主要由两个部分组成:代理服务器(proxy server)和学习器(learner),代理服务器是用户 Web 浏览器与 Web 之间的桥梁,它保存了所有访问过的 URL 地址,学习器主要是为系统提供用户模型,整个系统用 Perl 语言和 C++语言编写.

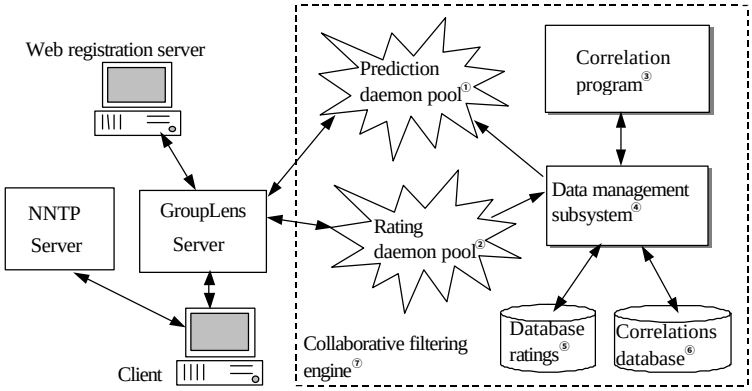
代理服务器主要由 3 部分组成:代理(proxy)、建议器(adviser)和分类器(classifer).当代理接到一个请求时,先下载请求的文档,如果该文档是 HTML 格式,将会加上一些建议并将结果发给用户,增加建议的过程是这样的:代理将下载的文档发给建议器,建议器先从文档抽取超链,接着将结果发给分类器,分类器利用学习器产生的用户模型推荐满足某一阈值的超链返回给用户.图 5 是 Personal WebWatcher 的体系结构.

学习器有两种版本:从头开始创建一个新模型的学习器和更新一个已存在模型的学习器.它们之间的差别是:前一个不得不定义好领域信息,同时从一个空模型开始学习;后一个可以利用已定义好的领域信息,同时修改已存在的模型.系统假定被用户访问过的文档是揭示用户兴趣的,也就是正面的例子,所有其他被忽略的文档就是负面的例子,这个假定可以省却用户参与系统学习的过程.用户的兴趣模型很简单,系统是根据指向文档的超链来预测文档的兴趣度,而不是真正文档的内容,因为检索文档是一个非常耗时的过程,在系统空闲的时候(比如晚间),也可以根据文档的内容来预测文档的兴趣度,这样会更准确一些.此外,也可以利用超链来预测文档内容,然后利用这些文档内容来预测真正文档的兴趣度.

2.3 GroupLens

GroupLens 是一个应用于 Usenet 新闻的协作过滤系统,它的目标是让用户一起协作,从大量的 Usenet 新闻中发现他们感兴趣的内容.系统分为两部分:客户端和服务器端.客户端是一个新闻阅读器 NewsReader,服务器端提供协作过滤.NewsReader 一般连接到本地 NNTP 服务器,同时也连接到 GroupLens 服务器共享过滤信息,只要用户下载一篇文档,NewsReader 都会向 GroupLens 服务器发送消息请求对该文档内容的预报,也就是其他用户对该文档的评价.此外,用户也可以评价文档,NewsReader 会将该用户评价发送到 GroupLens 服务器上进行处理,以提供给其他用户浏览,GroupLens 会利用这些信息调整该用户和其他用户的相关性.

GroupLens 服务器端分为请求代理和过滤引擎.请求代理将用户的请求分派到合适的进程来处理.过滤引擎分为 4 部分(如图 6 所示):预报模块、评价模块、相关性计算模块、数据管理模块.预报进程为客户提供精确的文档内容预报,文档的预报处理需要两类信息:用户间的相似性和用户对文档的评价;评价进程接收用户的评价并将其安全存储;用户相关性计算程序计算用户间的相似性;数据管理子系统管理用户的评价和用户相似性等数据.



①预报进程池,②评价进程池,③相关性计算,④数据管理子系统,⑤评价库,⑥相关性库,⑦协作过滤引擎.

Fig.6 Architecture of GroupLens

图 6 GroupLens 体系结构

3 总结与展望

个性化服务技术是目前非常流行的一种技术,本文分析了各种具有代表性的个性化服务系统,并在此基础上详细描述了建立个性化服务的关键技术.面对日益增长的 Web 信息,要满足不同背景、不同目的和不同时期的查询请求,必须针对不同用户提供不同的服务才能真正解决这个问题.

目前已经存在很多个性化服务系统,不过大部分都只是研究原型,也有一些系统已经推向了市场,比如 GroupLens.随着电子商务的不断发展,个性化服务显得越来越重要,它能将电子商务网站的浏览者转变为购买者、提高电子商务网站的交叉销售能力、提高客户对电子商务网站的忠诚度.尽管已经存在许多个性化服务系统,但个性化服务技术仍有很多值得研究和探讨的领域,归纳起来有以下几个方向:

(1) 用户兴趣和行为的表达.由于用户兴趣是多方面的,动态变化的,跟踪、学习和表达用户兴趣是一个最基本和难以解决的问题,这也是进一步研究的方向.

(2) 分类和聚类技术.分类和聚类技术是个性化服务的基本技术,不过有一些新的特点,比如能处理属于多个类的数据,类可以互相重叠;能进行增量的处理;能处理高维和大数据量,具有良好的可扩展性.

(3) 个性化推荐技术.现有的个性化推荐技术都存在一些缺点,如何克服这些缺点也是进一步研究的方向.

(4) 安全技术.现有的大部分个性化服务系统都忽略了如何保护用户的隐私.为了规范 Web 用户信息的保密性,W3C 成立了 P3P 工作组来解决这个问题,但还处于试用阶段.个性化服务技术要发挥作用,必须提出一个有效的保护用户隐私的机制,只有先保障系统的安全,才能顺利实现个性化服务.

References:

[1] Pretschner, A. Ontology based personalized search [MS. Thesis]. Lawrence, KS: University of Kansas, 1999.
[2] Mladenic, D. Machine learning for better Web browsing. In: Rogers, S., Iba, W., eds. AAAI 2000 Spring Symposium Technical Reports on Adaptive User Interfaces. Menlo Park, CA: AAAI Press, 2000. 82~84.
[3] Pazzani, M.J., Muramatsu, J., Billsus, D. Syskill & Webert: identifying interesting Web sites. In: Weld, D., Clancey, B., eds. Proceedings of the 13th National Conference on Artificial Intelligence and 8th Innovative Applications of Artificial Intelligence Conference. Menlo Park, CA: AAAI Press, 1996. 54~61.
[4] Lieberman, H. Letizia: an agent that assists web browsing. In: Burke, R., ed. Proceedings of the International Joint Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 1995. 924~929.
[5] Bollacker, K.D., Lawrence, S., Giles, C.L. Discovering relevant scientific literature on the Web. IEEE Intelligent Systems, 2000,15(2):42~47.
[6] Asnicar, F., Tasso, C. ifWeb: a prototype of user modelbased intelligent agent for documentation filtering and navigation in the World Wide Web. In: Tasso, C., Jameson, A., Paris, C.L., eds. Proceedings of the UM 1997 Workshop on Adaptive Systems and User Modeling on the World Wide Web. West Newton, MA: User Modeling Inc., 1997. 3~12.
[7] Mostafa, J., Lam, S.W., Palakal, M. A multilevel approach to intelligent information filtering: model, system, and evaluation. ACM Transactions on Information Systems, 1997,15(4):368~399.
[8] Chen, C.C., Chen, M.C., Sun, Y.S. PVA: a self-adaptive personal view agent system. In: Schkolnick, M., Provost, F., Srikant, R., eds. Proceedings of the 7th ACM SIGKDD International Conference on Knowledge discovery and data mining. New York: ACM Press, 2001. 257~262.

[9] Chen, L., Sycara, K. WebMate: a personal agent for browsing and searching. In: Sycara, K.P., Wooldridge, M., eds. Proceedings of the 2nd International Conference on Autonomous Agents. New York: ACM Press, 1998. 132~139.

[10] Han, E.H., Boley, D., Gini, M., *et al.* WebACE: a web agent for document categorization and exploration. In: Sycara, K.P., Wooldridge, M., eds. Proceedings of the 2nd International Conference on Autonomous Agents. New York: ACM Press, 1998. 408~415.

[11] Schwab, I., Pohl, W., Koychev, I. Learning to recommend from positive evidence. In: Riecken, D., Benyon, D., Lieberman, H., eds. Proceedings of the International Conference on Intelligent User Interfaces. New York: ACM Press, 2000. 241~247.

[12] Mobasher, B., Cooley, R., Srivastava, J. Automatic personalization based on Web usage mining. Communications of the ACM, 2000,43(8):142~151.

[13] Joachims, T., Freitag, D., Mitchell, T. WebWatcher: a tour guide for the World Wide Web. In: Georgeff, M.P., Pollack, E.M., eds. Proceedings of the International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers, 1997. 770~777.

[14] Lieberman, H., Dyke, N.V., Vivacqua, A. Let's browse: a collaborative web browsing agent. In: Maybury, M., Szekely, P., Thomas, C.G., eds. Proceedings of the International Conference on Intelligent User Interfaces. Los Angeles, CA: ACM Press, 1999. 65~68.

[15] Konstan, J., Miller, B., Maltz, D., *et al*. GroupLens: applying collaborative filtering to usenet news. Communications of the ACM, 1997,40(3):77~87.

[16] Shardanand, U., Maes, P. Social information filtering: algorithms for automating word of mouth. In: Roberts, T., Robertson, S., eds. Proceedings of the ACM CHI'95 Conference on Human Factors in Computing Systems. New York: ACM Press, 1995. 210~217.

[17] Alton-Scheidl, R., Ekhall, J., Geloven, O.V., *et al.* SELECT: social and collaborative filtering of web documents and news. In: Kobsa, A., Stephanidis, C., eds. Proceedings of the 5th ERCIM Workshop on User Interfaces for All: User-Tailored Information Environments. 1999. 23~37.

[18] Rucker, J., Polanco, M.J. Siteseer: personalized navigation for the web. Communications of the ACM, 1997,40(3):73~75.

[19] Srivastava, J., Cooley, R., Deshpande, M., *et al.* Web usage mining: discovery and applications of usage patterns from Web data. In: Fayyad, U., ed. Proceedings of the ACM SIGKDD Explorations. New York: ACM Press, 2000,1(2):12~23.

[20] Balabanovic, M. An adaptive Web page recommendation service. In: Johnson, W.L., Hayes-Roth, B., eds. Proceedings of the 1st International Conference on Autonomous Agents. New York: ACM Press, 1997. 378~385.

[21] Sakagami, H., Kamba, T., Sugiura, A., *et al.* Effective personalization of push-type systems——visualizing information freshness. Computer Networks and ISDN Systems, 1998,30(1~7):53~63.

[22] Wu, K.L., Aggarwal, C.C., Yu, P.S. Personalization with dynamic profiler. In: Wu, K., Datta, A., eds. Proceedings of the 3rd International Workshop on Advanced Issues of E-Commerce and Web-Based Information Systems. Los Alamitos, CA: IEEE CS Press, 2001. 12~20.

[23] Volokh, E. Personalization and privacy. Communications of the ACM, 2000,43(8):84~88.

[24] Wu, Y.H., Chen, Y.C., Chen, A.L.P. Enabling personalized recommendation on the web based on user interests and behaviors. In: Klas, W., ed. Proceedings of the 11th International Workshop on Research Issues in Data Engineering. Los Alamitos, CA: IEEE CS Press, 2001. 17~24.

[25] Claypool, M., Le, P., Waseda, M., *et al*. Implicit interest indicators. In: Campbell, M., ed. Proceedings of the ACM Intelligent User Interfaces Conference (IUI). New York: ACM Press, 2001. 14~17.

[26] Zaiane, O.R., Xin, M., Han, J. Discovering Web access patterns and trends by applying OLAP and DATA mining technology on Web logs. In: Howe, S.E., Smith, T.R., eds. Proceedings of the IEEE International Forum on Research and Technology Advances in Digital Libraries. Los Alamitos, CA: IEEE CS Press, 1998. 19~29.

[27] Paliouras, G., Papatheodorou, C., Karkaletsis, V., *et al.* Clustering the users of large Web sites into communities. In: Danyluk, A., ed. Proceedings of the 17th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 2000. 719~726.

[28] Nanopoulos, A., Manolopoulos, Y. Mining patterns from graph traversals. Data and Knowledge Engineering, 2001,37(3):243~266.

[29] Kurki, T., Jokela, S., Sulonen, R., *et al*. Agents in delivering personalized content based on semantic metadata. In: Murugesan, S., Macarthur, S., O'Leary, D.E., eds. Proceedings of the AAAI Spring Symposium Workshop on Intelligent Agents in Cyberspace. Menlo Park, CA: AAAI Press, 1999. 84~93.

[30] Li, W.S., Vu, Q., Agrawal, D., *et al*. PowerBookmarks: a system for personalizable web information organization, sharing, and management. Computer Networks, 1999,31(11~16):1375~1389.

[31] Fan, Y., Gauch, S. Adaptive agents for information gathering from multiple, distributed information sources. In: Murugesan, S., Macarthur, S., O'Leary, D.E., eds. Proceedings of the AAAI Symposium on Intelligent Agents in Cyberspace. Menlo Park, CA: AAAI Press, 1999. 40~46.

[32] Mock, K. Dynamic email organization via relevance categories. In: Bastani, F., ed. Proceedings of the 11th International Conference on Tools with Artificial Intelligence. Los Alamitos, CA: IEEE CS Press, 1999. 399~405.

[33] Yang, Y., Pedersen, J.O. A comparative study on feature selection in text categorization. In: Danyluk A., ed. Proceedings of the 14th International Conference of Machine Learning. San Francisco: Morgan Kaufmann Publishers, 1997. 412~420.

[34] Lee, D.L., Chuang, H., Seamons, K. Document ranking and the vector-space model. IEEE Software, 1997,14(2):67~75.

[35] Adomavicius, G., Tuzhilin, A. User profiling in personalization applications through rule discovery and validation. In: Lee, D., Schkolnick, M., Provost, F., *et al*., eds. Proceedings of the 5th International Conference on Data Mining and Knowledge Discovery. New York: ACM Press, 1999. 377~381.

[36] Balabanovic, M., Shoham, Y. Fab: content-based, collaborative recommendation. Communications of the ACM, 1997,40(3):66~72.

[37] Sarwar, B.M., Karypis, G., Konstan, J.A., *et al*. Application of dimensionality reduction in recommender system——a case study. In: Jhingran, A., Mason, J.M., Tygar, D., eds. Proceedings of the ACM WebKDD Workshop on Web Mining for E-Commerce. New York: ACM Press, 2000.

[38] Sarwar, B.M., Karypis, G., Konstan, J.A., *et al*. Analysis of recommendation algorithms for e-commerce. In: Proceedings of the ACM Conference on Electronic Commerce. New York: ACM Press, 2000. 158~167.

[39] Breese, J.S., Heckerman, D., Kadie, C. Empirical analysis of predictive algorithms for collaborative filtering. In: Cooper, G.F., Moral, S., eds. Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers, 1998. 43~52.

[40] Aggarwal, C.C., Wolf, J.L., Wu, K., *et al*. Horting hatches an egg: a new raph-theoretic approach to collaborative filtering. In: Chaudhuri, S., Madigan, D., Fayyad, U., eds. Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 1999. 201~212.

[41] Sarwar, B., Karypis, G., Konstan, J., *et al*. Item-Based collaborative filtering recommendation algorithms. In: Shen, V.Y., Saito, N., eds. Proceedings of the 10th International World Wide Web Conference (WWW10). 2001. 285~295.

A Survey of Personalization Technology[1]

ZENG Chun, XING Chun-xiao, ZHOU Li-zhu

(*Department of Computer Science and Technology, Tsinghua University, Beijing* 100084, *China*)

E-mail: bobofu00@mails.tsinghua.edu.cn; {xingcx,dcszlz}@tsinghua.edu.cn

http://dbgroup.cs.tsinghua.edu.cn

Abstract: The crucial technologies related to personalization are introduced in this paper, which include the representation and modification of user profile, the representation of resource, the recommendation technology, and the architecture of personalization. By comparing with some existing prototype systems, the key technologies about how to implement personalization are discussed in detail. In addition, three representative personalization systems are analyzed. At last, some research directions for personalization are presented.

Key words: personalization; rule-based personalization; information filtering technology; content-based filtering; collaborative filtering; recommendation system

[1] Received December 9, 2001; accepted July 1, 2002
Supported by the National Grand Fundamental Research 973 Program of China under Grant No.G1999032704