个性化服务技术研究

吴辉娟,袁 方
(河北大学 数学与计算机学院,河北 保定 071002)

摘 要:对个性化服务技术中的用户识别、用户描述文件、个性化推荐技术、个性化服务系统的体系结构及目前的研究方向进行了概述。从实现角度详细讨论了3种个性化推荐技术。个性化服务具有针对性,它的目的就是为了使用户更好地找到需要的信息,通过从用户访问网站的历史记录中得到用户的个人信息,利用个性化推荐的方法将信息推荐给用户。个性化推荐避免用户陷入信息的海洋,提高用户查询效率,使得用户可以得到他们真正想得到的信息,避免繁多的人工搜索。

关键词:个性化; 内容过滤;Web日志;协作过滤
中图分类号:TP393.4 **文献标识码**:A **文章编号**:1005-3751(2006)02-0032-03

Research of Technologies on Personalized Information Service

WU Hui-juan, YUAN Fang
(College of Mathematics and Computer, Hebei University, Baoding 071002, China)

Abstract: Some technologies related to personalization are introduaced in this paper, which include the representation of user profile, the identification of users, the recommendation technologies, the frame of the system and some research directions for personalization. In addition, three recommendation technologies about how to implement personalization are discussed in detail. Personalization aims at everyone and is to help the users to find the information they needed; personalization gets users' information from the history of users' access, and commends the information to users by personalized recommendation technologies; personalized recommendation avoids users dropping into the large information, and improves the efficiency of finding information. The result is that users can get what they wanted, avoid more finding.

Key words: personalization; content filtering; Web log; collabarative-filtering

0 引 言

Web已成为人们获取信息的重要手段,但是随着Web信息日益增长,人们需要花费大量时间去搜索、浏览自己需要的信息。搜索引擎(search engine)是人们查找信息的重要工具,但是传统的搜索引擎是根据用户输入的查询关键词进行搜索并且提供信息的,所以不同的用户提供同一关键词,系统返回的信息是相同的,但事实上,不同用户所需要的信息是不同的。个性化服务是就这个问题提出的,采用个性化服务技术的搜索引擎能够为用户提供他们想要查找的、满足他们兴趣和爱好的信息,避免用户陷入信息的海洋。实现个性化服务需要准确的识别用户,跟踪用户的兴趣和行为,对用户的兴趣和行为进行描述,针对基于不同推荐技术的系统,可能需要对网络资源进行描述。

1 用户的识别

个性化服务系统首先要能准确地识别出用户,这是为用户建立模型和实现向用户推荐信息的前提。

对于信息服务系统来说,用户可以分为两类:注册用户和非注册用户。用户在系统中注册,注册时输入个人信息,包括姓名、性别、年龄、教育背景和兴趣。由于用户一般都很注意个人信息的保密性[1]。因此,这些信息不能代表用户的兴趣,往往用于用户身份的验证。系统为每个不同的注册用户赋予一个ID。

对于非注册用户,他通过一个浏览器访问一个或多个Web站点。实际上,由于本地Cache和代理服务器(Proxy)的存在,使得用户识别比较困难。例如,不同的用户使用同一个代理服务器,在日志文件中会形成相同的IP地址;同时由于代理服务器中的缓存功能,使得同一用户的访问被认为是不同的用户。用户可以用一个浏览器,也可以用多个浏览器;可以访问一个服务器,也可以访问多个服务器,因此,用户识别比较困难。在识别用户时,可以将Access Log,Refer Log和用户提交的注册信息结合起来。注册用户根据系统记录的ID容易辨别。未注册用户识别应遵循以下启发式原则[2]:

(1) 如果用户的IP地址不同则认为是不同的用户。

(2) 如果IP地址相同但浏览器软件或操作系统不同,则认为是不同的用户。

收稿日期:2005-05-14
作者简介:吴辉娟(1981—),女,河北栾城人,硕士研究生,研究方向为数据挖掘;袁 方,副教授,博士研究生,硕士生导师,研究方向为数据挖掘。

(3) 通过 Refer Log 和站点的拓扑结构图构建每个用户的访问路径,如果所请求的页面和以前访问的所有页面不存在直接的超链接关系,则认为具有相同 IP 地址的用户是不同的用户。

用户识别出来以后,可以为他赋予一个 ID。

2 用户描述文件

描述用户兴趣的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[3]。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。

收集用户信息的数据来源有下面的几种:用户注册信息、利用指向文档的超链接内容、用户显式反馈的信息、用户隐式反馈的信息、访问和标记某网页等行为、Proxy 日志信息、用户浏览行为、用户在某页所花的时间、阅读的文档、阅读文档所花的时间和添加书签等行为、引用文件的内容、用户定义的目录类型、Web 访问日志。

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 definiton 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。

不同的系统有不同的算法来得到用户描述文件,不同的系统也有不同的用户描述文件的表达方式。用户描述文件的表达方式有以下几种:用户静态信息;基于加权关键词矢量,隐式创建或更新;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;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;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;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;从 Web 访问日志和站点文件脱机产生的 URL 聚类;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;用户浏览记录,隐式创建与更新;基于加权关键词矢量,显式创建,显式反馈或更新。

3 个性化推荐

目前存在许多个性化服务系统[4],根据其采用的推荐技术可分为两种: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。

3.1 基于内容过滤的技术

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。

基于内容过滤的系统的优点是简单、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。

描述用户兴趣和网络上的文本比较直接的做法是利用文本的特征。用户兴趣是多方面的,可以根据其浏览过的文档选取合适的主题词来表达用户兴趣。

实际中一般的方法都是基于一个向量空间模型来描述网页文档集中的内容的。首先利用停用词表将文本中的停用词去除,然后将主题词表中只在一个文本中出现的词去掉,然后根据主题词表与文本进行匹配,找出可代表文本特征的关键词,这些关键词代表了文挡集的特征,即用户的兴趣[5]。文档的集合可表示为一个向量,代表每个单词的维用一个权值来表示。关键词的权重可用 TFIDF 方法来表示,一个项 ti 的 TFIDF 计算公式定义如下:

$$TFIDF(ti) = TF(ti) * \log(n/DF(ti))$$

TF(ti)为项 ti 在所有文档中出现的次数,DF(ti)为出现 ti 的文档数,如果 ti 在越多的文档中出现,说明 ti 区分这些文档的能力越差,n 为文档的总数。用户兴趣可表示为一个关键词权重的向量。$\boldsymbol{U}=\{w_1,w_2,w_3,\cdots,w_i\}$,$w_i$ 表示第 i 个主题词的权重。向量的维数 n 是固定的,这就保证了文档和用户兴趣之间相似性计算的精度。向量 $\boldsymbol{d}=\{d_1,d_2,d_3,\cdots,d_i\}$,其中,$d_i$ 表示关键词 ti 在文档 $\boldsymbol{d}$ 中出现的次数。

用户 $\boldsymbol{U}$ 和文档 $\boldsymbol{d}$ 的相似度[6]可表示为:

$$Sim(\boldsymbol{u},\boldsymbol{d}) = \boldsymbol{u}\cdot\boldsymbol{d}/||\boldsymbol{u}||\cdot||\boldsymbol{d}||$$

计算用户兴趣和网页文本的相似度,根据相似度的降序排列将网页推荐给用户。但是,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无关。

3.2 基于协作过滤的技术

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。它是根据相似用户来推荐出新的感兴趣的内容。

用户描述文件由用户浏览过的网页 URL 来描述。用户浏览过的网页 URL 可以从近一段时间内 Web 日志中得到。假设近期内日志中用户的会话为 n 次,每次会话(Session)都由若干 URL 组成,则用户的描述文件可表示为 $\{URL_1,URL_2,URL_3,\cdots,URL_m\}$,其中,$URL_i$ = N_i *(用户会话总数/出现 URL_i 的会话总数)[7],N_i 为 URL_i 在所有会话中出现的次数,在网站的日志中找到与用户 A 最相似的 k 个用户,将这些用户访问的 URL 集推荐给用户 A。寻找与用户相似的其它用户的方法只适合于用户数量较少的系统,对于用户数量成千上万的系统,可将用户聚类,每类用户同样用 URL 的向量组成,即 $\{URL_1,URL_2,URL_3,\cdots,URL_i\}$。然后,判断用户属于哪一类,将此类用户访问的网页按 URL_i 的降序排列推荐给用户。

3.3 基于用户兴趣关联规则的推荐技术

基于用户兴趣关联规则的技术用于基于规则的个性化服务系统。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实现[8],利用规则来推荐信息依赖于规

则的质量和数量,基于规则的技术的缺点是随着规则的数量增多,系统将变的难以管理。

关联规则生成可用于找出在某次服务器会话中最经常一起出现的网页,在 Web 使用挖掘中,发现的关联规则往往是指支持度超过预设阈值的一组网页。这些网页之间可能并没有超链接直接互相连接。例如,Apriori 算法[9]发现关联规则可能会发现访问电子产品网页的用户会访问体育用品网页。在个性化信息服务中,用户关联规则的挖掘有助于网站设计者重新组织和设计网站结构。

Apriori 算法利用了一个层次顺序搜索的循环方法来完成频繁项集的挖掘工作。这一循环方法就是利用 k -项集来产生$(k+1)$ -项集。具体做法就是:首先找出频繁 1-项集,记为 L1;然后利用 L1 来挖掘 L2,即频繁 2-项集;不断如此循环下去直到无法发现更多的频繁 k -项集为止。此算法的性质:一个频繁项集中任一子集也应是频繁项集。

算法利用层次循环发现频繁项集。

输入:Web 日志中用户访问网页的 URL,记作 D。

输出:D 中的频繁项集。

●流程:

(1)L1 = find_frequent_1_itemset(D);//发现 1-项集

(2) For(k=2;Lk-1! =空集;k++)

(3)Ck = aprior_gen(Lk-1,min_sup);//根据频繁(k-1)项集产生候选 k-项集

(4)for each t∈D{//扫描数据库,以确定每个候选项集的支持频度

(5)Ct = subset(Ck,t);//获得 t 所包含的候选项集

(6)for each c∈Ct c.count++;

(7)}

(8) Lk = {c∈Ck | c.count > min_sup}

(9)return L = UkLk

●procedure apriori_gen(Lk-1,min_sup)

(1)for each l1 ∈ Lk-1

(2) for each l2 ∈ Lk-1

(3)if((l1[1] = l2[1] ∧ … ∧ …(l1[k-2] = l2[k-2 ∧](l1[k-1] < l2[k-1]}

c = l1⊕l2;//将两个项连接到一起

(4)if has_infrequent_itemset(c,Lk-1)

(5)detete c;//除去不可能产生频繁项集的候选

(6)else Ck = Ck∪ {c}

(7)}

(8)return Ck

●procedure_ infrequent_subset(c,Lk-1)

for each(k-1) subset s of c

(1) if s ! ∈ Lk-1 return TRUE;

else return FALSE

在 D 中挖掘出所有的频繁项集后,就可以较为容易获得相应的关联规则,也就是要产生满足最小支持度和最小信任度的强关联规则。

可以利用公式(1)来计算所获关联规则的信任度。这里的条件概率是利用项集的支持频度来计算的。

$$\text{Confidence}(A=>B)=P(B|A)=(\text{Support\_count}(A\cup B))/(\text{Support\_count}(A)) \quad (1)$$

其中,Support_count(A∪B)为包含项集 A ∪B 的会话记录数目;Support_count(A)为包含项集 A 的会话记录数目。具体产生关联规则的操作如下:

① 对于每个频繁项集 l,产生 l 的所有非空子集。

② 对于每个 l 的非空子集 s 如果((support_ count(1))/(support_count(s)) ≥min_conf,则产生一个关联规则 s=>(l-s)。其中 min_conf 为最小信任度阈值。

4 个性化服务系统的体系结构

根据用户描述文件的存放位置不同,个性化服务系统有 3 种体系结构。用户描述文件可存放在服务器端、客户端、代理端。

4.1 基于服务器端

在这种体系结构中,用户描述文件可存入服务器端。这种体系结构的优点是可以避免用户描述文件的传输;缺点在于用户的描述文件不能在不同的 Web 应用之间共享。

4.2 基于客户端

在这种体系结构中,用户描述文件可存入用户浏览器端。这种系统的个性化定制服务既可在服务器端实现,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 Web 应用之间共享。

4.3 基于代理端

这种体系结构中,用户描述文件存入代理服务器端。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。

5 总 结

个性化服务技术是目前非常流行的一种技术,能有效解决"知识过载"和"迷航"问题,因此,它必将受到用户的青睐。目前已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继续研究的领域:

(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,是进一步研究的方向。

(2)分类和聚类技术。这两种技术有一些新的特点,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;能处理高维和大量数据,具有良好的可扩展性。

(3)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的研究方向。

(下转第 37 页)

页,然后打开"数据库"面板。

(2)单击该面板上的加号(+)按钮,然后从弹出式菜单中选择"数据源名称(DSN)"。出现"数据源名称(DSN)"对话框,输入连接名称。

(3)DreamweaverMX2004 会创建名为相关的数据源连接,它指向 SQL Server 服务器的学生成绩数据库。如果连接失败,请执行以下操作:复查 DSN;核对数据库的用户名和密码。

(4)单击"确定"。新连接出现在"数据库"面板上。

4)相关模块和页面的实现:DreamweaverMX2004 提供了大量实用的服务器行为,并自动生成代码,根据具体任务可直接调用,因此功能模块实现很简单。在这里重点介绍该系统的核心模块——成绩查询模块的实现过程。其它模块可参照实现。

成绩查询模块由 3 个页面组成,分别为:查询页面(search.asp)、判断页面(process.asp)、结果页面(result.asp)。具体实现如下:

①查询页面(search.asp):该页面的作用是接收用户查询条件并将它传送给判断页面。首先在页面中加入一个表单(包括一个文本域、提交按钮和重置按钮),并将表单 Action 属性设置为 process.asp。

②判断页面(process.asp):它不接受信息也不显示信息,主要是判断查询条件在数据表中是否存在,如果存在用 session 对象保存查询条件传到结果页面;如果不存在转回查询页面。因而,它没有任何可视化元素,必须加入判断代码。首先建立学生记录集(recordset1)并在代码视图中加入如下代码。

```
<% if not recordset1.eof then
      session(id) = request.form(xh)
      response.redirect("result.asp")
  else
      response.redirect("search.asp")
  end if
%>
```

③结果页面(result.asp):主要是显示查询结果。它是通过对记录集按照查询条件进行筛选并将结果显示出来。首先建立一个表格再建立一个结果记录集,然后在表格内加入动态元素。这里关键是记录集的筛选条件的建立。在记录集 Filter 属性中设置为 session("id")[6~9]。

5 结束语

编码实现的具体细节可参见文献[9]。另外,本系统还可实现从 ASP 平台向 ASP.net 平台转换。具体方法是:首先安装.NET 框架。可从 Microsoft Web 站点(网址为 http://asp.net/download.aspx)下载并按照安装说明进行安装;再将 DreamweaverMX2004 中 WEB 应用程序设置成 ASP.NET 应用程序服务器。为了提高系统效率可同时下载和安装 Microsoft Data Access Components (MDAC) 2.7 软件包(可从网址为 www.microsoft.com/data/download.htm 免费下载)。

参考文献:

[1] Tanenbaum A S. 计算机网络[M]. 熊桂喜,王小虎 译. 北京:清华大学出版社,1998.

[2] Comer D E. 用 TCP/IP 进行网际互联(I,II)[M]. 赵 刚,林 瑶,蒋 慧,等译. 北京:电子工业出版社,2001.

[3] 谢希仁. 计算机网络(第3版)[M]. 大连:大连理工大学出版社,2000.

[4] 陆 刚. 计算机网络操作系统[M]. 成都:电子科技大学出版社,2002.

[5] 王恩波. 网络数据库实用教程——SQL Server2000[M]. 北京:高等教育出版社,2004.23-24.

[6] 俞俊平,余安萍,俞俊军. Dreamwaver UltraDev4 网站开发实务[M]. 北京:电子工业出版社,2001.

[7] Siyan K. Windows2000 TCP/IP 实用全书[M]. 张 锦,彭宗仁,等译. 北京:电子工业出版社,2001.

[8] 曾清国. Windows2000 + ASP + SQL Server 案例教程[M]. 北京:中科多媒体电子出版社,2001.

[9] 飞思科技产品研发中心. Dreamwaver UltraDev4 网站设计与实现[M]. 北京:电子工业出版社,2001.

(上接第 34 页)

参考文献:

[1] Volokh E. Personlization and privacy[J]. Communications of the ACM,2000,43(8):84-88.

[2] 王熙照,王丽娟,袁 方,等. Web 用户访问模式挖掘[J]. 河北大学学报(自然科学版),2002,22(4):404-405.

[3] Wu Y H,Chen Y C,Chen AL P. Enabling personalized recommendation on the web based on user interests and behaviors [A]. In:Klas W. Proceedings of the 11th International Workshop on Research Issues in Data Engineering[C]. Los Alamitors,CA:IEEE CS Press,2002.17-24.

[4] Pretschner A. Ontology based personalized search[D]. Lawrence,KS:University of Kansas,1999.

[5] 曾 春,刑春晓,周立柱. 个性化服务技术综述[J]. 软件学报,2002,13(10):1952-1961.

[6] 曾 春,刑春晓,周立柱. 基于内容过滤的个性化搜索算法[J]. 软件学报,2003,14(5):1001-1002.

[7] 石 晶,龚震宇,袭杭萍. 基于 Web 挖掘的个性化服务技术[J]. 计算机科学,2002,29(8):168-169.

[8] Adomavicius G,Tuzhilin A. User profiling in preso-nalization applications through rule discovery and validation[A]. In:Lee D,Schkolnich M,Provost F,et al. Proceeding of the 5th International Conference on Data Mining and Knowledge Discovery [C]. New York:ACM Press,1999.377-381.

[9] 朱 明. 数据挖掘[M]. 合肥:中国科学技术大学出版社,2002.