智能化个性服务在WEB上的应用研究

王向星 （中央财经大学信息系2000研究生 100081）

摘要：智能化个性服务可以为用户提供准确全面的信息，给出了两种流程结构；为了达到智能化个性服务Agents必须具有良好的接口，利用用户信息表对搜索引擎提供的信息进行再过滤，进而提供相关信息的推荐；以及跟踪用户的需求偏好，动态更新用户信息表。

关键词：智能化个性服务　搜索引擎　Agents　用户信息表

1 发展历程

(1) 传统的网络上网页的样式和层次结构都是固定的，信息分布在不同层次的页面中。但是，用户并不是对所有的信息都感兴趣，而且不同用户所感兴趣的内容也不相同。因而用户就不得不到不同的页面中查找不同栏目的信息，用户连接服务器的时间变长，即服务器对用户的平均处理时间变长。如果服务器能够知道用户的需求，直接提供给用户所需的信息，就可以准确、快速地为用户服务，从而减少服务器的负荷，这就出现了搜索引擎。

(2) 搜索引擎通过自动浏览程序搜索各网站的资源，并将各资源的关键词、摘要、资源地址等信息提取出来，形成一个自己的索引库。当用户向搜索引擎提交关键词进行查询时，搜索引擎就在该索引库中查找匹配的信息，提供相应的资源地址供用户选择。但是，用户提交的关键词也只能片面地反映用户的需求，产生的结果仍然存在冗余。

(3) 智能服务根据能够较为全面地反映用户需求的用户信息表动态地、及时地为用户提供准确的信息。智能服务的优点有：

① 可以为用户提供更为准确的信息，减少用户自身的搜索时间。

② 根据信息关联规则进行信息的预取，放入快速缓存中以加快信息的访问速度[1]；或者进行相关信息的推荐。

③ 可以利用数据挖掘技术从访问Log文件中提取用户的访问模式，用于市场决策推荐服务[2]。

④ 采用聚类用户访问模式方法，预测用户未来的访问行为，进行相应的信息推荐[3]。本文将在下面中阐述这些实现方法。

2 智能化个性服务的流程分析

2.1 智能化个性服务的流程结构

2.1.1 WEB服务器提供智能化个性服务如图1

(1) 用户通过客户浏览器向WEB服务器发出查询请求。

(2) WEB SERVER向搜索引擎传递请求。

(3) 搜索引擎根据用户的请求提示对所有数据进行组织分类提取初步的数据并将结果传递给Agents。

(4) Agents用已有的用户信息表对(3)中的结果进行数据提取，过滤掉不合用户需求的信息；同时，Agents根据新的需求信息和数据挖掘的结果更新用户信息表。

(5) 将符合用户需求的信息结果传递回Web服务器。

(6) 由Web服务器将结果返回客户浏览器。

智能网站的个性服务就采用了这种方式。但它具有缺点：如果有多个用户同时发出请求，WEB服务器就需要生成多个Agents，用户信息表的处理及相应的数据过滤是一件费时的工作，给WEB服务器造成严重负担。为了避免这个问题，Amalthaca[5]建议应将智能化的个性服务适当地分散到用户端。按照这个思想，我们提出另外一个流程。

2.1.2 用户端提供智能化个性服务如图2

(1) 客户浏览器向Agents提交请求。

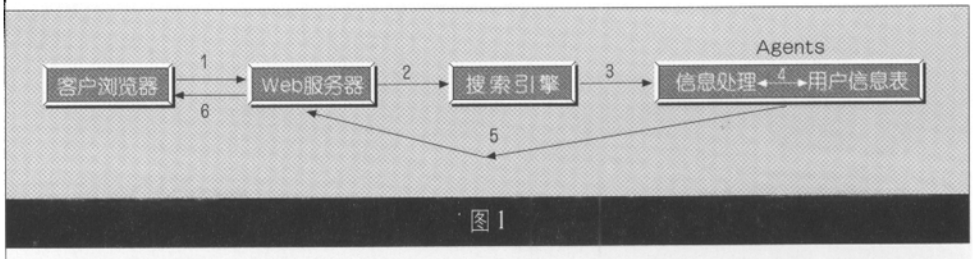


图1

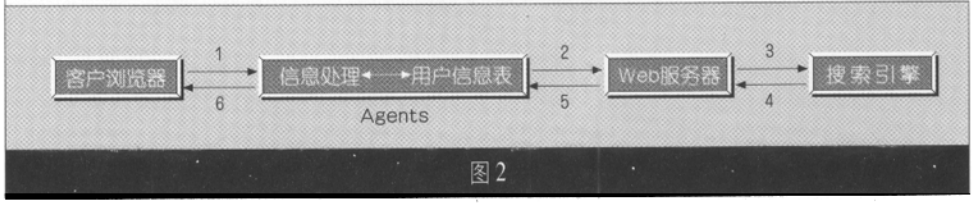


图2

(2) Agents解析用户请求，转化成Web服务器可辨认的请求。

(3) WEB服务器向搜索引擎传递请求。

(4) 搜索引擎根据用户的请求提示对所有数据进行组织分类提取初步的数据并将结果返回Web服务器。

(5) Web服务器将结果返回Agents。

(6) Agents处理如1中(4)的步骤，并将结果返回客户浏览器。

这种方式下，WEB服务器提供传统的搜索引擎功能，不增加服务器负担。其中，搜索引擎查询的唯一依据是经Agents转化过的用户请求。只要提交的请求一样，搜索引擎的查询结果就一样，而不管是不是同一个用户提交的请求。再通过客户端的Agents处理，实现智能化个性服务。

显然，Agents必须具备用户与机器的接口转化问题。引用1990年Nielsen [4] 提出了9条可用性原则：

① 人机对话简明、自然；

② 使用用户的自然语言；

③ 减少用户的记忆负担；

④ 促进一致性实现；

⑤ 提供返回信息；

⑥ 提供清楚的出口标记；

⑦ 提供热键；

⑧ 提供有效的出错处理信息；

⑨ 能够防止出错。

3 Agents的信息处理

这是实现智能化个性服务的环节所在。用户提交的请求只能反映用户某一方面的需求，而不是用户全面、准确的需求，这就意味着以此作为唯一查询根据的搜索引擎所产生的结果也是片面的，必然包含了一些用户并不需要的信息结果。由于用户信息表能更全面、准确地反映用户需求，所以用它作为过滤标准处理这些信息，就可以更好地过滤掉冗余信息，提供更为准确的信息给用户，这就是Agents信息处理的作用。

那么，Agents是如何利用用户信息表来进行信息处理？笔者提供以下一些方法：

(1) 信息对用户信息表的相关性绝对值。首先，给此相关性绝对值下个定义：假设用户信息表中的关键词类为{K1,K2,...,Kn},对应的重要性权值为{W1，W2，.....,Wn};同时假设一条信息中包含K1,K2,...,Kn的个数为N1，N2，.....,Nn;那么这条信息对用户信息表的相关性绝对值即可表示为：Σ(Ni*Wi),其中I=1,...,n。这样就可以根据相关性绝对值的大小来判断信息对用户的重要性，并决定返回推荐信息的优先级。举例说明，有一个用户信息表如下表所示：

关键词类	关键词集合	权值
信息	信息、资讯	3
安全	安全、保密	3
通信	传输、通信、传送	3

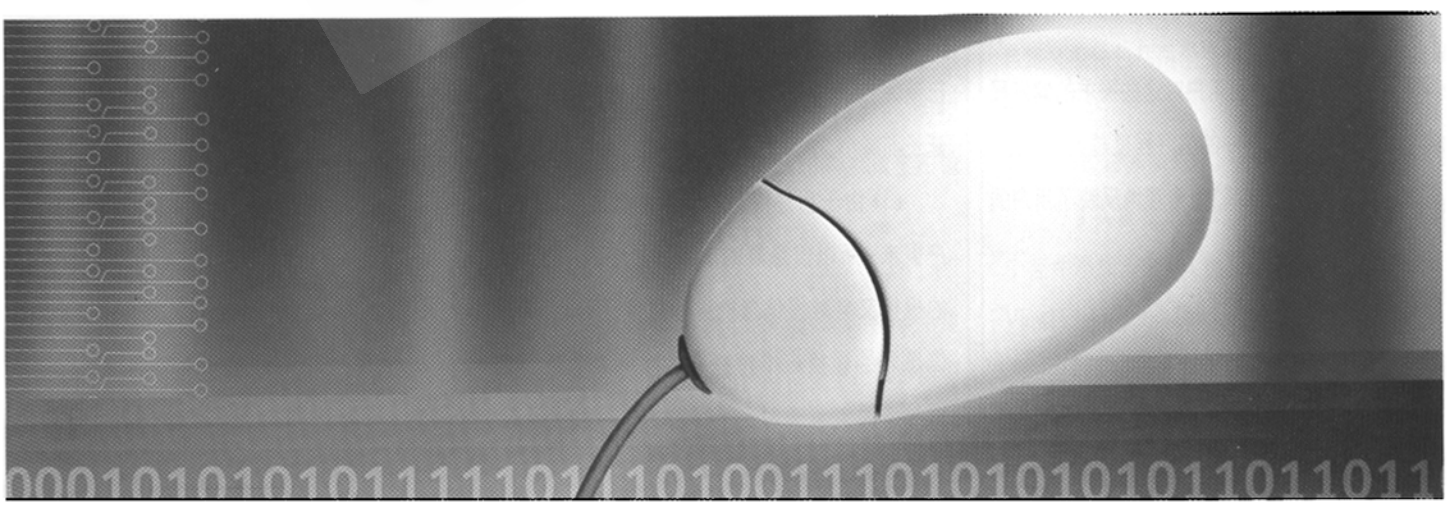
有一条信息包含关键词的个数如下：信息50个，资讯0个、安全40个，保密20个，传输30个、通信36个、传送0个。那么，该信息的相关性绝对值为：

(50+0) *3+ (40+20) *3+ (30+36) *3=528

(2) 信息之间的相关性。传统上，信息的分类和链接一般只是按单关键词的相关性进行。但是，单关键词只能片面反映信息内容，所进行的信息链接就会存在冗余。因此，我们有必要改进，采用多关键词来判断信息间的相关性。假设每条信息I分别由n个相互独立的关键词类Ki组成，即I1={K11，K12，.....，K1n},I2={ K21，K22，.....,K2n},其中Kij表示第i条信息的第j个关键词类；取两条信息的关键词类的交集I1 ∩ I2={K1，K2，.....，Kt}，Kj(j=1,2,...,t)在I1中的权值为W1j,在I2中的权值为W2j。这样，信息I就可用相应的关键词类的权值向量V [Ij] 表示：V [I1] = {W11，W12，.....，W1t},V [I2] ={W21，W22,.....,W2t}。计算出V [I1] 与V [I2] 向量间的夹角即可得出I1与I2的相关性程度：

$$S(I1, I2)=\cos(V[I1],V[I2])= \sum_{i=1}^{t} W1iW2i \Big/ \sqrt{\sum_{i=1}^{t} W1i^2}\sqrt{\sum_{i=1}^{t} W2i^2}$$

值越小表示信息间的相关性越大。这样，我们以此为依据进行信息的分类以及高相关性信息间的相互链接。



(3) 用户间的相关性应用。现实生活中，某些人的习惯之间存在相似性，分析其中几个人的行为可以推出其他人的行为。我们将此思想应用到用户在网络上的行为：通过分析用户信息，将相似程度高的人群划为同类组；接着，就可以将同类组中一些用户需求的信息推荐给该组成员。

那么，我们如何计算用户间的相关性？这里采用概率论的方法来定量计算用户U1和U2在过去获取信息的相似程度：

$$S(U1,U2)=count(I(U1)\cap I(U2))/(count(I(U1))+count(I(U2)))$$

其中，I (Uj) 表示用户Uj访问过的信息集合，count()是记数函数，该公式表示了用户U1与U2过去所访问过的信息中相同部分的概率，反映了用户间的兴趣相似程度。S (U1, U2) 越大，相似程度也越大。

4 用户信息表

4.1 定义：笔者认为可分为如下三部分，其中：

关键词类	关键词集合	权值

(1) 关键词类反映了用户的需求内容。

(2) 关键词集合是为了更好地满足用户使用自然语言的习惯。如同是表达"information"的意思，有些人习惯用"信息"，而有些人则习惯用"资讯"，所以在进行信息过滤时，就应认为两者等价，是属同一类的。

(3) 权值则反映了用户对这些需求的偏好程度，权值越大说明用户的偏好程度越大，反之，亦反。这样，整个表就较为客观地反映了用户的需求偏好。

4.2 初始化

(1) 最简单有效的办法是用户自己定制信息并赋予相应的权值，如和讯个性网站的定制方法。

(2) 由于用户处理过的文档或其他信息可以从一定程度上反映用户过去的需求偏好，所以我们可以通过这些信息的数据挖掘，提取出用户信息表。

4.3 动态更新

由于用户的需求偏好会发生变化，所以我们必须及时地根据用户的行为挖掘出用户的需求变化，并将它反映到用户信息表中，形成一个动态的更新过程。

那么如何根据用户的行为挖掘出用户的需求呢？

(1) 当用户浏览页面时，判断用户是否对该页感兴趣。最简单的办法是：假设用户浏览单位长度的页面信息所需的平均时长为T0，则用户浏览K个单位长度的页面信息所需的平均时长就为K*T0。分两种情况：

① 当用户访问某页面的时长大于或等于K*T0时，我们就认为用户对该页感兴趣。接着，从这些页面信息中提取出关键词反映到用户信息表中：若该关键词不存在于用户信息表中，则添加进去，并赋予相应的初始权值；否则，增加该词的权值。

② 当用户访问某页面的时长小于K*T0时，我们就认为用户对该页不感兴趣。若这些页面中的关键词存在于用户信息表中，则减少相应的权值。

(2) 通常，若用户在一段时间内不用某些关键词，则说明用户对这些关键词的需求在下降或已消失，反映到权值上即是：权值应随着时间的延长而减小。

(3) 由于用户信息表要较为全面地反映用户需求，所以信息表的规模要足够大；但过大的规模又会减慢信息的过滤速度。因此，用户信息表的规模应维持在适量水平线附近，在增加新关键词类时，要删去那些权值最小的关键词类。

5 结束语

尽管智能化的个性服务在WEB上的应用还只是处于尝试阶段，但其相关的思想及技术在数据挖掘及文档处理等方面的应用已取得了实际成效。所以，我们有理由相信，随着研究的进一步深入，该技术的WEB应用是可以预见的。

参考文献

1 Schechter S, Krishnan M, Smith M D. Using Path Profiles to predict HTTP request(C). Proceedings of 7th InternationalWorld Wide Web Conference,Brisbane,Australia,1998。

2 Cooley R, Mobasher B, Srivastava J. Data Preparation formining World WideWeb browsing patterns [J]. Journal of Knowledge and Information Systems,1999,1(1)。

3 Nasraoui O, Frigui H, Joshi A et al. Mining Web access logsusing relationalcompetitive fuzzy clustering (C). To appear in the Proceedings of the Eight International Fuzzy Systems Association World longress,1999。

4 E J Derrick, O Balci. A Visual Simulation Support Environment [J]. Journal of System Software。

5 A Moukas Amalthaca. Information Discovery and Filtering Using a Multiagent Evolving Ecosystem [Z]。

6 李煊等，基于关联规则挖掘的个性化智能推荐服务，计算机工程与应用，2002年第11期。

7 殷信义等，智能网站Agents的研究，计算机应用研究，2002年第1期。

8 冯永杰等，Agent在智能信息检索中的应用研究，计算机应用研究，2002年第2期。

9 魏子忠等，一种基于Agent的因特网信息获取系统，计算机工程与设计，2001年4月第2期。