一种基于 UCL 的中文网页信息过滤方法

邢 玲[1,2],马建国[2],李幼平[3],刘志文[1]
(1. 北京理工大学电子工程系,北京 100081;2. 西南科技大学信息工程学院,四川绵阳 621010;
3. 中国工程物理研究院北京应用物理与计算数学研究所,北京 100088)

摘 要: 围绕如何在浩瀚的中文网页中找到用户感兴趣的内容,提出了基于 UCL (Uniform Content Locator) 的"二阶过滤法".它将媒体空间中的信息用 UCL 语义格 (Semantic Cases based on UCL, SCU) 表示,通过语义向量空间模型 (Semantic Vector Space Model, SVSM) 对网页的语义矩阵进行分析计算,粗略筛选出用户感兴趣的网页;再借助精细语义逐句解读其内容,提取用户所关注的信息.根据用户的阅读行为动态了解用户的兴趣变化,建立用户兴趣的本体模型,并分析和定义了用户兴趣度的度量.实验验证了上述过滤方法的有效性,其测试结果同向量空间模型 (Vector Space Model, VSM) 进行了比较,性能明显优于 VSM.

关键词: UCL; 信息过滤; UCL 语义格; 语义向量空间; 兴趣本体模型
中图分类号: TP391.1 **文献标识码**: A **文章编号**: 0372-2112 (2006) 10-1752-06

An Information Filtering Method for Chinese Web Pages Based on UCL

XING Ling[1,2], MA Jian-guo[2], LI You-ping[3], LIU Zhi-wen[1]
(1. *Department of Electronic Engineering, Beijing Institute of Technology, Beijing* 100081, *China*;
2. *School of Information Engineering, Southwest University of Science and Technology, Mianyang, Sichuan* 621010, *China*;
3. *Institute of Applied Physics and Computational Mathematics, China Academy of Engineering Physics, Beijing* 100088, *China*)

Abstract: The work focuses on filtering users' interested contents in Chinese web pages. Two-stage filtering method based on UCL is presented. SCU is brought forward to express the information of Medium Space. SVSM is introduced to filtrate cursorily web pages, and then contents of these pages are understood by virtue of some elaborate semantic characteristics, so the web pages which users are interested in can be extracted. At the same time, the users' interested changes are tracked dynamically according to the reading actions, and the interesting ontological profile is submitted, then the measure of interestingness is analyzed and calculated. Laboratory simulations demonstrate the arithmetic feasibility and validity.

Key words: UCL; information filtering; SCU; SVSM; interesting ontological profile

1 引言

20 世纪 90 年代以来,Internet 以惊人的速度发展起来,它容纳了各种类型的原始信息,包括文本信息、声音信息、图像信息等等.人们每天都要获取和处理大量的信息,面对浩瀚如海的信息,往往会导致"信息过载"和"信息迷航".因此,信息过滤 (Information Filtering, IF) 技术得到越来越广泛的关注.

对于网页的信息过滤,特别是针对中文网页的文本处理,许多研究者试图从分析网页的特定结构着手,研究网页的不同部位标识信息的主题表达能力,从而提出信息过滤的各种算法,如文献[1～3];文献[4]通过研究 Semantic Web 中的关键技术,利用本体论建立用户语义模型并提出基于此模型的过滤算法;文献[5～8]利用数据挖掘技术,采用不同的策略挖掘用户日志得到用户访问信息,从而了解用户的访问模式,得到用户感兴趣的话题.但是这些方法的不足之处在于:

1. 网络信息空间中的数据大多以半结构化和非结构化的形式存在,对信息资源的内容缺乏形式化的语义描述.即使研究网页不同的标识信息,也只能将部分的数据进行语义的扩展.能不能将网页内容的基本特征映射到语义空间中,通过形式化的语义描述将网页的内容再现?

2. 通过本体论建立的用户语义模型虽然从语义理解的角度诠释了用户模型,但是却不能动态地了解用户兴趣的变化,对信息过滤的结果有着直接的影响.通过什么方式可以动态了解用户兴趣的变化?又怎样度量用户的兴趣度呢?

3. 挖掘用户日志可以得到用户的访问信息,提取用户感兴趣的话题.但是在文献中基于既往浏览内容的个性化服务很难捕捉到新鲜的网页和用户兴趣的变化,同时数据挖掘的工具和数学模型比较复杂,实现起来也比较困难.能不能用相

收稿日期:2006-01-19;修回日期:2006-07-21
基金项目:国家 863 高技术研究发展计划(No. 2005AA121520);国家自然科学基金(No. 60272014)

对简单的方法通过自我学习收集和分析用户的信息需求,选取用户感兴趣的内容?

本文围绕以上三个问题,运用 UCL 技术实现基于个人语义的智能信息过滤. UCL 技术是解决网络信息语义理解和主动服务的一种独特方法. 文章从分析信息的媒体空间着手,将网页信息映射到 UCL 语义空间中,通过对网页进行粗略筛选和精细的语义过滤,将用户感兴趣的网页直接推送给用户. 同时根据用户的阅读行为和阅读评价综合度量用户的兴趣度,从而动态了解用户的兴趣变化,准确地刻画用户兴趣的本体模型. 最后构建了基于 UCL 的个性化信息服务模型(Personalizing Information Service Model,PISM),实现了主动提供个性化的信息服务.

2　语义向量空间模型

在向量空间模型(Vector Space Model,VSM)中,将文档看作是由相互独立的词条组(T_1, T_2, …, T_n)构成,对于每一词条 T_i,都根据其在文档中的重要程度赋予一定的权值 ω_i,并将 T_1, T_2, …, T_n 看成一个 n 维坐标系中的坐标轴,ω_1, ω_2, …, ω_n 为对应的坐标值. 这样由(T_1, T_2, …, T_n)分解而得到的正交词条向量组就扩张成了一个文档向量空间,文档则映像为空间中的一个点.

由于在 VSM 中仅仅是将词条在坐标系中按照其对文档的贡献程度进行一定的赋值,并没有考虑词条本身的语义信息,以及整个信息空间的语义结构. 因此我们将向量空间扩展为语义向量空间,通过 UCL 对信息进行标引和分类,从而更有效地表示出信息空间的语义特性,形成 UCL 语义空间. 关于 UCL 的定义以及标引的详细过程请参见文献[9～11].

定义 1　信息空间中的每一条信息都包括信息的语义和信息的载体两种特征.

定义 2　Internet 上所有的电子化文档构成了信息的电子载体集合,简称媒体空间,记为 Ω.

$$\Omega = \{A, P\} \tag{1}$$

其中:$A = \{a_i \mid a_i \in$电子信息资源集合$\}$

$P = \{p(a_i) \mid p(a_i) \in$电子信息资源出现的概率$\}$

定义 3　信息空间的语义结构是指可将 Ω 按照语义划分成众多的、有层次结构和相互联系的子空间,每一个子空间都代表了特定的语义.

设媒体空间 Ω 的基于某种内容属性的分类 g,即映射 $g_i: A \to C_i$. 记:

$$g_i(A) = \{c \mid c_i \in C_i: \exists a \in A; y_i = g_i(a)\} \tag{2}$$

定义 4　映射 g_i 下信息集 A 的一个像 $g_i(A)$ 称为信源集 A 在 g_i 关系下的一个内容描述. 称各内容描述子空间的直积

$$C_w = C_{w1} \times C_{w2} \times C_{w3} \times \cdots \times C_{wn} \tag{3}$$

为基于网页的信源空间 Ω 的 UCL 内容分类所张的网页信息 UCL 分类描述空间,简称 UCL 语义空间[12]. 称有序组(u_{w1}, u_{w2}, …, u_{wn})为信源空间 Ω 的基于 UCL 的内容分类的一个内容描述向量或 UCL 语义向量,记为 U_w. 即

$$U_w = (u_{w1}, u_{w2}, \cdots, u_{wn}) \tag{4}$$

式中 $u_{wi} \in C_{wi}, i \in N; n$ 是 U_w 的分量数

在基于 UCL 的语义向量空间模型(Semantic Vector Space Model,SVSM)中,网页信息与用户 Profile 均需要表示为语义矩阵. 下面我们来具体说明如何将网络信息进行 UCL 映射.

定义 5　在网页信息中,除了描述的内容以外,还需要使用其它的辅助项从不同的角度来进一步描述网页信息. 将各种辅助项归纳为一组语义功能,称之为语义格(Semantic Cases)[13].

我们将网页媒体空间中提取的 UCL 字段称为 UCL 语义格(Semantic Cases based on UCL,SCU).

通过对 Web 信息的分析,可以得到网页的几个很重要的媒体特征,这些 UCL 字段从根本上表示了网页的语义范畴. 表 1 列出了 UCL 语义格的主要类别,其中的项都可以通过对 Web 文档的提取而得到,而其中的类别是可以扩展的.

表 1　网页 UCL 语义格的主要类别

类别(Sort)	栏目(Column)
标题(Title)	时间(Time)
作者(Author)	摘要(Abstract)
字符数(Number)	段落(Paragraph)
地址(Address)	其他(Undefined)

对任一网页 D,在经过句法和语义结构分析之后,其每一关键词都有一组数值表示其在语句层次上于各类语义格的相关概率. 称此数值为"语义格加权值". 非关键词则不予考虑. 关键词的定义和提取参见文献[14].

假定总共有 n 个语义格类型,那么对应于关键词 T 在网页中的第 i 次出现,其语义格加权值即可表示为一个向量 P_i,即:

$$P_i = (\omega_{i1}, \omega_{i2}, \cdots, \omega_{in}) \tag{5}$$

若关键词 T 在网页中出现的频率为 k,则有 k 个这样的向量,它们可以方便地表示为 $k \times n$ 的矩阵 M_T,即:

$$M_T = \begin{bmatrix} \omega_{11} & \omega_{12} & \cdots & \omega_{1n} \\ \omega_{21} & \omega_{22} & \cdots & \omega_{2n} \\ \cdots & \cdots & & \cdots \\ \omega_{k1} & \omega_{k2} & \cdots & \omega_{kn} \end{bmatrix} \tag{6}$$

其每一行对应于关键词 T 在网页 D 中的一次出现,每一列对应于该关键词相对于某一语义格类型在网页中的分布. 矩阵 M_T 的行向量的线性组合定义了一个新的向量 P_T, P_T 描述关键词 T 在网页 D 中的整体语义格分布,称之为语义格向量(Case Vector).

$$P_T = (r_1, r_2, \cdots, r_n) \tag{7}$$

式中 $r_j = \frac{1}{k} \sum_{i=1}^{k} \omega_{ij}, j \in (1, 2, \cdots, n)$

虽然此向量描述了关键词的语义结构,限定了它在网页中的语义内涵,但是作为对网页内容的标引还不够完整. 因此我们将语义格向量加以扩充,加入关键词本身的加权值 ω,其计算表达式为:

$$\omega_i(d) = tf_i(d) \times \log\left(\frac{N}{n_i} + 0.01\right) \tag{8}$$

$\omega_i(d)$ 表示第 i 个关键词的加权值,tf_i 对应于关键词 T_i 在网页 D 中出现的频率,N 为所有的网页数目,n_i 为包含 T_i 的网页数目.

现在,假定有 m 个关键词用于标引网页 D,而对网页中的每一个关键词 T_i,有如下的向量定义:

$$U_i = (\omega_i, r_{i1}, r_{i2}, \cdots, r_{in}) \quad (9)$$

其中,ω_i 为关键词 T_i 本身的加权值,r_{ij} 为关键词相对于第 j 个语义格类型的语义格加权值,$i \in (1, 2, \cdots, m)$,$j \in (1, 2, \cdots, n)$. 而这 m 个关键词语义向量的叉积则又定义了一个 $m \times (n+1)$ 矩阵 M,矩阵 M 表示一文档的语义内涵,称为语义矩阵 (Semantic Matrix),即:

$$M = \begin{bmatrix} \omega_1 & r_{11} & r_{12} & \cdots & r_{1n} \\ \omega_2 & r_{21} & r_{22} & \cdots & r_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \omega_m & r_{m1} & r_{m2} & \cdots & r_{mn} \end{bmatrix} \quad (10)$$

在网页信息和用户 Profile 都表示为语义矩阵后,就可以通过矩阵相似值来预测网页和用户需求信息之间的相关度. 矩阵的相似程度可用向量之间的夹角余弦来度量.

定理 1 若两个语义矩阵 M_D 和 M_Q 中的语义向量为 U_D 和 U_Q,那么向量 U_D 和 U_Q 的夹角越小说明向量的相似度越高,此定理称为语义矩阵余弦定理.

$$Sim(Q, D) = Cos(U_Q, U_D) = (U_Q \cdot U_D) / (|U_Q| \times |U_D|) \quad (11)$$

3 用户兴趣模型

用户兴趣模型 (User Interesting Profile, UIP) 是指描述用户感兴趣的信息范围的模型,它是个性化信息服务模型中的重要组成部分. 建立用户兴趣模型的目的就在于通过统计、分析用户行为和用户反馈意见建立一个反映用户基本兴趣和信息需求的信息模型,并将该模型用于帮助用户更好、更快地获取新的信息.

由于 Web 信息的特殊性,在媒体空间中的任一相互联系的概念,因每个人对知识的理解不同,对它感兴趣的程度也不同,故有着不同的表现. 在媒体空间分类与具体网页的映射上,每个用户可能会把相同的文档赋予不同的概念. 如,一个用户可能认为一篇网页属于“体育”类,而另一个用户可能认为这篇网页属于“新闻”类. 同样,对于同一个概念,不同的用户对其感兴趣的程度也不同. 因此,在构建用户兴趣模型时要进行媒体空间概念与具体网页/文档之间的映射,这也就是我们所说的本体概念,文献[14]中详细说明了媒体分类和本体分类的概念. 本文建立的用户兴趣模型是基于本体分类的兴趣模型,即用户的兴趣本体模型 (Interesting Ontological Profile, IOP).

3.1 兴趣本体模型的构建

用户信息需求和兴趣可以用 user_profile 文件来表示. 用户向系统提供了一次 user_profile 文件之后,系统就可以及时的把与之相关的数据送给用户,而无需向系统反复地提出同样的查询. 这种自动的信息流使用户可以与不断更新变化的信息保持同步. 实际上可以将 profile 文件看成是一种持续执行的查询命令[15]. 另外,基于用户的兴趣是随时间而变化的事实,系统需要及时修改文件以反映用户信息需求的变化. 其 user_profile 文件格式为:

```
user {
    keyword1 ,sort1 ,column1 ,threshold;
    keyword2 ,sort2 ,column2 ,threshold;
    ……
    keywordn ,sortn ,columnn ,threshold;}
```

其中 keyword 表示用户输入的关注词句即用户希望获得的信息;sort 表示本体分类中的类别名称;column 表示本体分类的子集如网站栏目分类;threshold 表示用户对某本体分类的兴趣度,即用户阈值,用于判断用户对某一本体分类感兴趣的程度,通常在 0 和 1 之间,阈值越大说明感兴趣的程度越高. 并有以下的规则:

规则 1 对本体分类 d 与用户 u,若用户 u 对 d 的感兴趣程度 $\mu_u(d) >= user_u : threshold$,则称 d 是用户 u 所需要的分类.

另外,每个用户根据自己的需要,用户阈值即 $threshold$ 值可各不相同. 用户阀值是指在过滤网页时,对网页进行匹配计算所得值的最低要求. 它既可以由用户代理通过观察、记录、学习用户的行为来自动地维护,以便适应用户不断变化的信息需求,也可以由用户自己动手调整.

3.2 用户兴趣度的度量

在个性化信息服务中,用户对分类的兴趣度是反映用户个性化信息的重要指标. 兴趣度量化的主要意义在于用来对多个分类的兴趣进行比较,兴趣度的计算构成了模型学习算法中最重要的一环. 它涉及到许多问题:兴趣是如何变化的,如何判断用户对一个网页是否感兴趣,计算兴趣度应该考虑哪些因素,兴趣度如何量化等等.

用户在浏览的过程中,会不断地接收到有关本体分类的新信息,因此用户对分类的兴趣度也会发生变化,其计算公式为:

$$\mu' = \mu + \Delta\mu \quad (12)$$

μ 为用户对某本体分类原来的兴趣度,μ' 为变化后的兴趣度,$\Delta\mu$ 为用户浏览一个新的网页之后对该网页所对应本体分类的兴趣度的变化值,也可理解为用户对该网页的兴趣度.

3.2.1 影响 $\Delta\mu$ 的因素

本文总结归纳了两个影响用户对某一网页兴趣度变化的重要因素:用户行为和用户对网页的显式评价.

可以根据用户的行为来表示用户对某一网页感兴趣的程度,即用户的行为体现了用户的兴趣程度[15]. 用户的行为可以是添加网页至兴趣库、下载网页、阅读网页、从兴趣库中删除网页等,这些行为体现了用户不同的兴趣,所以具有不同的意义. 表 2 指出了用户行为和其代表的不同意义.

表 2 用户行为和用户评价代表的不同意义

用户行为	意义	用户评价
添加兴趣	非常有兴趣	非常满意
下载网页	有很大兴趣	满意
阅读网页	有兴趣	一般
删除网页	没有兴趣	不满意

表 2 中也指出了,若用户浏览了一个网页之后,对该网页进行显式的评价:非常满意、满意、一般、不满意等.

用户浏览一个网页,对该网页对应的本体分类的兴趣度变化有如下的计算公式:

$$\Delta\mu = \omega_1 \times g_1(time, length) + \omega_2 \times g_2(evaluation) \quad (13)$$

在式 (13) 中,g_1,g_2 为两个函数,它们将所有的因素规格化为一个可以比较的量. ω_1,ω_2 为权重. 在公式中也没有考虑

用户兴趣随时间衰减的因素,只是考虑用户的兴趣在一段时间内比较相对稳定的兴趣.

$$g_1(time, length) = \frac{activity}{4.0} 10^{[-\lg(length)]/time} \quad (14)$$

$$g_2(evaluation) = \frac{evaluation}{3.0} \quad (15)$$

式(14)中 *time* 为用户浏览网页的时间,单位为秒. *length* 为网页的字符数. 函数 g_1 说明网页长度对兴趣度的影响起负作用,而用户阅读网页的时间起正作用, *length* 被取对数说明时间的作用比长度更强. *activity* 是用户浏览网页行为的权值,对应表 2 中的行为从上到下依次取为 4.0、3.6、3.6、0. 如果用户的行为不是阅读网页,则规定式(14)中的指数部分为 1. 式(15)中的 *evaluation* 是用户对网页文档的反馈,对应表 2 中的四个等级评价,即 *evaluation* 的值为 3.0、2.0、1.0、0,所以 g_2 的可能取值为 1.0、0.67、0.33、0. 因此由式(13)可知,$\Delta\mu$ 在[0,1]之间.

在表 2 中阅读网页即包括打开网页、浏览全文、关闭网页等用户行为. 如果用户打开网页之后,阅读网页的时间没超过规定的驻留时间,则认为用户对此网页不感兴趣,即 *activity* = 0;如果用户打开网页之后,阅读网页的时间超过规定的驻留时间,并且将该网页关闭,则认为用户正在阅读网页 *activity* = 3.6;如果用户下载了很多网页,在一定时期内,并没有阅读网页,则也认为用户对该网页不感兴趣 *activity* = 0. 对驻留时间有如下规则:

规则 2　如果用户在阅读某一网页的过程中,阅读此网页的时间小于 30 秒或大于 3 小时,则此网页被用户忽略,定义为不感兴趣的网页.

3.2.2　ω_1,ω_2 的取值

ω_1,ω_2 为用户行为和用户反馈这两个因素在影响兴趣度变化量中所占的权重. 它必须满足 $\omega_1 + \omega_2 = 1$ 的条件. ω_1,ω_2 决定了在计算 $\Delta\mu$ 的权重体系中各个因素的重要性是不同的. 通常用户反馈最能反映用户的兴趣度,用户行为次之,因而 ω_1,ω_2 的重要性从大到小依次为 ω_2,ω_1. 在本文中,它们的取值分别为 $\omega_1 = 0.4$,$\omega_2 = 0.6$.

由式(12)可得到如下规则:

规则 3　用户对某一网页感兴趣的程度越高,对该网页对应的本体分类的感兴趣程度也越高.

将上面得到的 μ 值与用户阈值进行比较后即可知道哪些本体分类是用户感兴趣的. 通过上面的算法就可以得到用户对本体分类的兴趣图谱,从而得到通行于各个兴趣图谱的个人用户兴趣模型,在文献[10]中有详细的讨论.

3.3　用户 profile 的维护

一定时期内经过用户端的信息流是一个信息集合,记为 Q;在 Q 中符合用户兴趣需求的子集记为 U;其他不属于用户兴趣范围的信息构成子集 M;显然有:

$$Q = U + M \quad (16)$$

从 Q 中依据表达用户需求的向量$\langle p_1, p_2, \cdots, p_n\rangle$而生成的信息集合记为 U!用户代理通过观察、记录、分析用户对 U 的行为,将 U 中用户不感兴趣(没有阅读等操作)的信息特征从向量$\langle p_1, p_2, \cdots, p_n\rangle$中去除或是修改其特征值;同时将新发现的用户感兴趣的信息特征加入到向量$\langle p_1, p_2, \cdots, p_n\rangle$中. 从而不断动态地调整、修正用户的兴趣 profile 文件,使其能更准确地表达用户兴趣的变化,使 U 能逐渐逼近 U.

4　基于 UCL 的信息过滤算法

在建立了基于 UCL 的用户兴趣模型和标引库之后,用户就可以进行按语义的查询、过滤. 文中介绍的"二阶过滤法",首先通过粗略语义(网站、时段、类别和栏目),在"全球网"(World Wide Web)中,筛选出有限的网页进入用户的信息库,形成"我的网"(my Web);其后,对存入 my Web 的文件,借助精细语义(个人自由定义的关注词句)逐句解读,提取含有该关注词的句子,连同适量的上下文字,供用户判读,并建立起用户兴趣模型.

为此我们将信息过滤分为两个步骤:

Step 1:分析、抽取新 Web 文档的特征并与标引库进行比较,确定该网页的 UCL 标引向量,并将其映射到语义向量空间中,得到用户感兴趣的类别信息并存入信息库中,也即粗略过滤;

Step 2:解读新网页并与用户 Profile 进行比较,最终确定该新网页是否为用户所需的信息,并提供给用户,即实现了精细过滤.

图 1 示出了基于 UCL 的信息过滤系统框图. 在图中 Web 信息通过 UCL 标引后,经过过滤器进入本地信息库,本地信息库是用户大容量的兴趣仓储. 而其中的文件管理主要是对标引后的信息进行有组织地分类、存储管理. 用户代理根据用户兴趣模型,从 Web 信息中截取与用户兴趣相匹配的信息,经过滤器后将结果送给用户. 而用户也可以通过指定关注词句对信息进行检索. 同时,用户要对得到的信息做出评价,以便使用户兴趣模型更贴近用户的真实兴趣. 这样,经过一定时间的积累,用户智能代理就会跟踪到用户的兴趣,代替用户搜寻和过滤信息并主动地显示给用户.

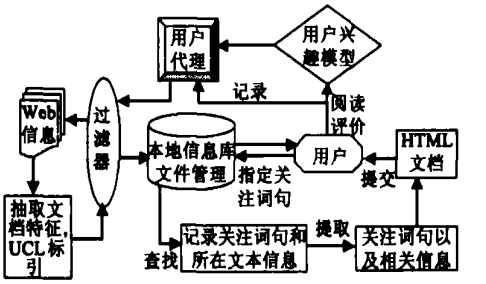

图 1　基于 UCL 的信息过滤系统

在过滤系统中,将 Web 页中的句子信息详细表示出来并存于本体信息库中. 将该信息与用户输入的关注词句进行全文匹配,提取出含有该关注词的句子以及该句子所在文档信息. 按照一定的顺序以 HTML 文档的形式提交给用户. 如果提交的结果过多,用户还可以对其进行加权处理,使大于兴趣阈值的文档提交给用户. 通过预处理,用户可决定是否有必要详细地阅读此网页,直接排除用户的疑虑. 用户也可对网页进行评价,然后通过兴趣学习方法使检索到的信息更加符合用户

的兴趣.

5 基于 UCL 的个性化信息服务模型

个性化信息是指由人类个性对信息需求的决定关系而产生的一系列对个体有用的信息. 个性化信息服务(Personalizing Information Service ,PIS)的主要目的就是要为用户提供一种满足个体信息需求的服务,即根据用户的要求提供信息服务,或者通过对用户个性、习惯的分析而主动地向用户提供其可能需要的信息服务.

本文构建的个性化信息服务模型如图 2 所示. 在模型中, Web 信息经过 UCL 标引后,进入 UCL 标引库分级、分类存储,并进入信息过滤模块和用户兴趣模型. 用户可以通过 UCL 智能代理或者自己指定的内容检索得到大量 Web 信息中自己最感兴趣的信息. 用户也可以显式地对信息做出评价,或者通过代理跟踪用户的阅读行为而获得隐式评价,从而维护用户 Profile,使用户兴趣模型更加贴近用户的需求,实现主动地提供个性化信息服务.

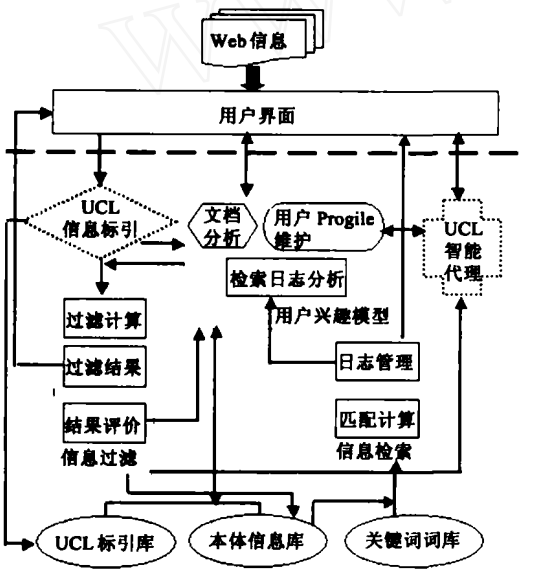

图 2 基于 UCL 的个性化信息服务模型

6 实验仿真

为了评价系统的性能,本文设计了信息过滤仿真实验,并对测试的结果进行了分析. 根据 CNNIC 最新统计的数据,我们选取了其中网民经常访问的四大中文网站:新浪、搜狐、人民网、新华网的 800 篇文档,涉及 4 个本体分类(体育、财经、教育、健康),每个本体分类对应 200 篇文档,将本体分类作为用户的需求领域,分别进行了测试,如图 3 所示.

信息检索领域通常采用查准率(Precision)和查全率(Recall)来衡量系统性能. 查准率定义为过滤到的目标类的样例集中所包含的属于过滤正确的样例所占的比例大小. 查全率定义为在一个过滤结果中所包含的过滤正确的对象数目占实际存在的满足查询要求的对象数目的比例大小.

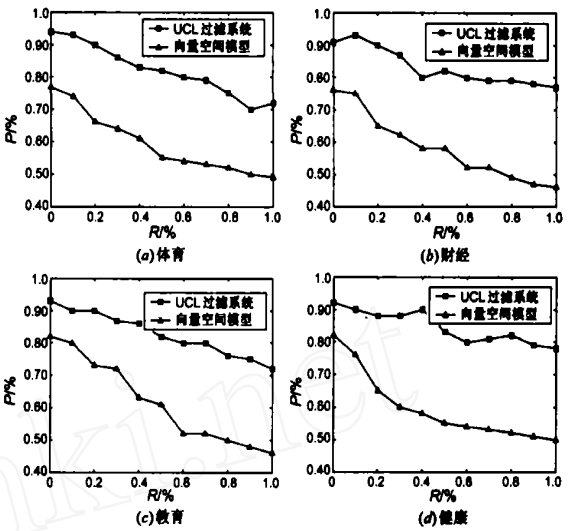

图 3 两种模型的性能比较

为了进行性能比较,选择目前运用较为广泛的 VSM 模型作为过滤对比方案. 在图 3 中,我们将区间[0,1]分成 10 等分,11 个边界点值作为 R,计算每个 R 下的 P,即在 R 区间范围内最大的 P 值. 由图可见,我们分别对涉及的 4 个本体分类进行了性能对比,基于 UCL 的信息过滤系统的性能要明显优于 VSM 模型.

7 结论及进一步工作

文章详细介绍了基于 UCL 的信息过滤技术,提出了"二阶过滤法"和 UCL 语义格的概念,建立了基于 UCL 的语义向量空间模型,并由此构建用户的兴趣本体模型. 其独特之处为:

1. 用 UCL 对网页信息进行标引,过滤时不再需要提取其它的文档特征,大大地提高了过滤的效率.

2. 采用 UCL 语义格表示媒体信息,通过语义向量空间模型映射网页,提高了系统的查准率.

3. 用户的兴趣本体模型运用基于用户阈值的类别、栏目及关注词来描述用户需求的信息,能够更加贴近用户的兴趣.

4. "二阶过滤法"使类别和栏目都与用户兴趣不符合的网页不进入信息库,过滤掉用户不关注的信息,节省了用户的精力. 同时通过先后两次语义过滤能够在信息库中精确地提取用户真正感兴趣的网页.

在未来的工作中,我们将集中以下几个方面进行研究:如何设定用户兴趣随时间的衰减指数;如何根据用户的行为动态确定兴趣阈值,如何运用用户兴趣图谱建立主动服务机制,如何分析用户行为的自相关性以及如何利用个体用户的兴趣图谱分析网络的小世界效应;如何根据网络信息的行为特性分析网络的无尺度效应和网络的动力学特性等等.

参考文献:

[1] 王晓宇,熊方,凌波,周傲英. 一种基于相似度分析的主题提取和发现算法[J]. 软件学报,2003,14(9):1578 - 1585.
WANG Xiao-Yu, XIONG Fang, LING Bo, ZHOU Ao-Ying. A similarity-based algorithm for topic exploration and distillation

[J].Journal of Software,2003,14(9):1578 - 1585(in Chinese)

[2] 胡健,陆一鸣,马范援. 基于 HTML 文档结构的向量空间模型的改进[J]. 情报学报,2005,24(4):433 - 437.
Hu Jian,Lu Yiming,Ma Fanguan. Vector space model based on HTML document structure[J].Journal of the China Society for Scientific and Technical Information,2005,24(4):433 - 437. (in Chinese)

[3] 侯汉清,章成志,郑红. Web 概念挖掘中标引源加权方案初探[J]. 情报学报,2005,24(1):87 - 92.
Hou Hanqing, Zhang Chengzhi, Zheng Hong. Research on the weighting of indexing sources for web concept mining[J].Journal of the China Society for Scientific and Technical Information,2005,24(1):87 - 92. (in Chinese)

[4] 张玉峰,艾丹祥,金燕. 基于 Semantic Web 的个性化网络导航机制[J]. 情报学报,2005,24(4):438 - 444.
Zhang Yufeng,Ai Danxiang,Jin Yan. Personalized web navigation mechanism based on semantic web[J].Journal of the China Society for Scientific and Technical Information,2005,24(4):438 - 444. (in Chinese)

[5] Spiliopoulou M. The laborious way from data mining to web mining[J]. International Journal of Computer System Science and Engineer Special Issue on Semantics of the Web,1999,3(1):105 - 113.

[6] Cooley R,Mobasher B. Data preparation for mining world wide web browsing patterns[J]. Knowledge and Information Systems,1999,1(1):17 - 24.

[7] Buchner A G,Mulvenna M D. Discovering internet marketing intelligence through on-line analytical web usage mining[J]. SIGMOD Record,1998,27(4):54 - 61.

[8] 郭岩,白硕,杨志峰,张凯. 网络日志规模分析和用户兴趣挖掘[J]. 计算机学报,2005,28(9):1483 - 1496.
GUO Yan,BAI Shuo, YANG Zhi-Feng, ZHANG Kai. Analyzing scale of web logs and mining users' interests[J]. Chinese Journal of Computers,2005,28(9):1483 - 1496. (in Chinese)

[9] 马建国,邢玲,李幼平,李在铭. 数据广播中的 UCL 标引与传输机制[J]. 电子学报,2004,32(10):1621 - 1624.
MA Jian-guo, XING Ling,LI You-ping,LI Zai-ming. UCL indexing and transmission scheme in data broadcasting[J]. Acta Electronica Sinica,2004,32(10):1621 - 1624. (in Chinese)

[10] 马建国,邢玲,李幼平,文丽. 广播型网格的用户兴趣图谱[J]. 电子学报,2005,33(1):142 - 146.
MA Jian-guo, XING Ling, LI You-ping, WEN Li. User interest spectrum of broadcasting grid[J]. Acta Electronica Sinica, 2005,33(1):142 - 146. (in Chinese)

[11] 邢玲,史杏荣. 基于 UCL 的网页自动标引技术[J]. 计算机工程与应用,2004,40:148 - 151.
Xing Ling, Shi Xingrong. An automatic indexing method for web pages based on UCL[J]. Computer Engineering and Applications,2004,40:148 - 151. (in Chinese)

[12] 马建国. 基于内容标引的信息共享技术[D]. 四川成都:电子科技大学,2004,6.
Ma Jianguo. Information Sharing Technology with content Indexing[D]. Doctoral dissertation. University of Electronics Science and Technology of China,2004,6. (in Chinese)

[13] Geoffrey Z Liu. 语义矢量空间模型及其试验评价[J]. 情报学报,1996,15(6):402 - 413.
Geoffrey Z Liu. Experimental evaluation of the semantic vector space model[J]. Journal of the China Society for Scientific and Technical Information,1996,15(6):402 - 413. (in Chinese)

[14] 邢玲. 基于本体结构的网页信息自动标引技术[D]. 安徽合肥:中国科学技术大学,2005,6.
Xing Ling. Automatic indexing technology for web pages information based on ontology framework[D]. Master dissertation. University of Science and Technology of China,2005,6. (in Chinese)

[15] 文丽. 基于 UCL 的大规模并播系统接收端设计[D]. 四川绵阳:西南科技大学,2004,6.
Wen Li. Receiving end experiment research of large scale parallel broadcasting system based on UCL[D]. Southwest University of Science and Technology,2004,6. (in Chinese)

作者简介:

邢　玲　女,1978 年出生于四川省攀枝花市,现为北京理工大学电子工程系博士研究生,2002 年毕业于西南科技大学电子工程专业,获学士学位;2005 年于中国科学技术大学通信与信息系统专业获硕士学位,主要研究方向为智能信息处理与主动服务技术. E-mail:xingling@bit.edu.cn

马建国　男,1957 年出生于四川省梓潼县,博士,西南科技大学信息工程学院副院长,教授,博士生导师,研究方向为信息系统技术,主持国家 863 计划项目与国家自然科学基金多项研究课题,已出版著作四本,在电子学报等学术期刊发表学术论文 70 余篇.

李幼平　男,1935 年出生于福建省厦门市,中国工程院院士,中国工程物理研究院研究员,西南科技大学信息工程学院院长,1957 年南京工学院无线电专业毕业,1957 至 1959 在清华大学无线电系研修多路通信与遥测,此后在成都电讯工程学院担任教师,1964 年 10 月,调往中国工程物理研究院,开始了核武器研究生涯,曾获得多种奖励,其中包括国家科技进步一等奖、国家发明二等奖、国防科技重大成果一、二、三等奖,1999 年获香港何梁何利基金技术科学奖,近年来在信息共享技术方面开展了研究,首先提出统一内容定位和大规模并播技术,在此基础上提出了营造双结构互联网的理念.