基于话题的信息空间模型与用户模型*)

The Information Space Model and User Model Based on Topic

何 军 周明天

(电子科技大学计算机科学与工程学院 成都610054)

Abstract Information filtering technology compares the user requirement with the incoming information stream. So it is necessary at first to better describe the features of information and user requirement. By excavating the concept of topic to be the basic logic unit and analyzing point of view, this paper analyzes information, the media of information and user requirement. Through adding the extension feature of user interested topic, we expand the traditional user requirement model that could reflect better the individuation and dynamic change of user requirement.

Keywords Information filtering, Feature extraction, User requirement model, Topic

信息网络中拥有大量并且不断增长的信息,而用户的信息需求又是各不相同、极具个性化的。为了帮助用户及时得到相关的信息,信息过滤技术依据用户不同的需求从信息流中选择用户需求的信息并及时送给用户,为用户提供个性化的服务,并能节省网络带宽[1,2]。在信息过滤环境中,需要有效地解决两个具有不确定性的问题:用户兴趣的改变和动态的信息流;同时,还要对用户兴趣及进入信息流进行匹配计算,以便将有用信息及时、准确地送到需要它们的用户处。为此就需要用一种模型化的方法有效地表示用户兴趣及信息空间,从而可对二者进行匹配计算。

用向量空间模型[3]来表示用户 profiles 和信息空间是一种基于关键词的被广泛应用和有效的方法(用户 profiles 文件表示用户的信息需求)。在向量空间模型中,信息被看作是一个多维空间中的矢量 D;同样的,用户 profiles 也被视为向量空间中的若干个分离的矢量——兴趣矢量 P。有了信息矢量和用户 profiles 矢量后,就可以通过计算两个矢量的 cos (D,P)值并确定相应的阈值来选择所需的文档;并可以通过计算 cos(D,D)之间以及 cos(P,P)之间的值将文档和用户 profiles 分类。

但按传统的向量空间模型来表示用户 profiles 和信息空间,只是注重了语义的内涵,缺乏对这些信息矢量和用户兴趣的内部结构及相互关系的描述,不利于对信息的分类组织,影响了过滤效率的提高,不利于系统按用户群进行过滤以便提高性能。本文从话题的角度,对信息、信息的载体和用户的兴趣需求进行了分析。通过发掘用户感兴趣话题的外延特征而扩展了传统的用户需求模型。

文章首先对信息及信息载体的基本概念和特征进行了分析,接着讨论了用户的需求和用户感兴趣话题的外延特征,在此基础上建立了一个扩展的用户需求模型。

1. 信息空间

本文讨论的信息空间是指由 Internet 上所有的电子化信息构成的信息集合。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。为了方便讨论,我们给出了一些基本定义。

定义1 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 Q。

媒体空间是信息的载体空间,是一种物理空间,其中的每条信息都具有大小、有效期、相对(相对于不同用户)价值等特性。目前已经出现了多种媒体形式来表达信息空间中的信息,它们有:文本、图形、图像、音频、视频或它们的联合形式等。这些媒体形式构成了媒体空间的物理组织。

定义2 信息空间的语义结构是指可将 Q 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。

本文中将话题作为划分信息空间的尺度,这样每一个子空间就表示了一个话题。

定义3 我们称话题的全集为信息的语义空间 Q',或称话题空间,表示为集合 $T=\{t_i \mid i=1,2,\cdots\}$,$t_i$ 表示话题 i。

显然,媒体空间 Q 可以按照集合 T 来进行语义划分。

语义空间 Q' 是逻辑空间(语义空间),物理空间 Q 是逻辑空间的载体。信息的语义空间与信息的媒体空间分别代表了信息空间的内在意义和外在表达形式。

性质1 Q' 中的每一个话题都有一个产生、发展、失效的过程。

随着人类社会的发展,信息空间中的信息也在高速地增长着。但信息空间中的不同话题的信息量会有不同的增长速度,即各子空间的增长速度是不均衡的,有的相对快些,有的相对慢些。在广播流中则对应为不同的话题有不同的流量、流密度(也即广播密度),并且每个话题的流量、流密度是随时间而变化的。

性质2 在一定时期内的总的话题数量是稳定的、有限的。

性质2说明,虽然信息空间和用户数都在不断高速增长,但在一定时期内的 $|T|$ 却有一定的稳定性、有限性。即无限信息媒体空间 Q 中的信息都可以映射到有限的话题空间中,如图1所示。因此当一个子网中的用户数越来越多、达到一定规模后,用户之间的兴趣会产生重叠、交叉,并使用户群的兴趣集合 ΣU_i 趋近于饱和,所以可以在子网入口处进行合成过滤,即对用户兴趣的并集进行过滤,从而提高过滤性能。

*)电子科学研究基金 DJ9.1.3资助。何 军 博士生,主要研究方向为分布信息系统、信息过滤技术、计算机网络及应用技术。周明天 教授,博士导师,主要研究方向为计算机网络、分布对象技术、并行分布处理和系统集成。

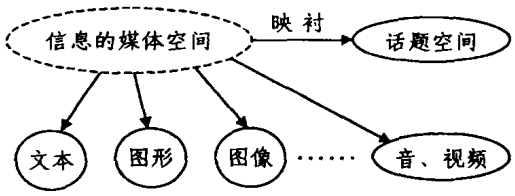

图1 信息空间结构图

通常人们首先将信息空间按内容划分成有限的若干类别，类是较为固定的，变化较慢的；然后在各个分类下再划分出众多的话题，话题的变化则较快。类的划分应遵循如下的原则：

1)类之间的交集应尽可能地小；

2)符合现实世界中人们的分类习惯。

每个话题就是一个话题对象。话题对象可以描述为：

```
Topic {
        Topic_Id,
        Topic_Name,
        Descripting,
 operation:
      Creat();
      Chang();
      Destroy();
 }
```

其中，creat()表示创建一个话题；chang()表示改变一个话题的属性；destroy()表示取消一个话题。

2. 文档特征及特征抽取

在 WWW 上大量的信息是以文档为单位来表示、记录并提供给用户的，这些文档包括文本、多媒体、图形、图像、音频、视频等多种形式。因此本文中对信息的处理也是以文档为单位的。

每一条信息都可以看作是对一个事件、一个状态的描述，描述的格式包括非结构化(如文本、超文本等)、半结构化(如电子邮件等)和结构化(如数据库等)的形式。计算机最善于处理的是结构化的信息。因此，为了使系统能提供高质量的信息服务，需要将非结构化和半结构化的信息转换成结构化的表示。这可以通过抽取出信息的特征组成结构化的特征描述数据来表示信息的内容，在此基础上进行信息过滤等处理。

定义4 文档中出现频率或重要程度大于等于某一给定阈值的特征项的集合称为文档的主特征。

定义5 文档中出现频率或重要程度小于某一给定阈值的特征项的集合称为文档的从特征。

定义6 主特征和从特征共同构成了文档的全特征。

主特征反映了文档的主要特性，但忽略了文档的非主要的特性，是一种不完全的文档特征信息，造成了文档部分信息的损失。为了完整地描述文档的全面的信息，还需要抽取文档的从特征信息，这样就可以完整地描述文档的信息了。

在已有的许多文档描述和搜索引擎中，通常只涉及到文档的主特征，例如关键词等。但有时一些用户对从特征也需要检索，以满足其个性化的需求。因此对文档采取全特征抽取，可以满足有各种 QoS 要求的查询、提高查全率。

文档的每一项特征都有权重以表示其重要程度，具体表示如下：

$$F(d)=\{(x_1,y_1),(x_2,y_2),\cdots\cdots,(x_n,y_n)\}$$

其中，F(d)表示文档 d 的特征集合，x_i 表示特征项 i，y_i 表示特征项 x_i 的权重。我们用 F_A 表示主特征集合，F_B 表示从特征集合，则文档的全息特征集合为：

$$F(d)=F_A(d)+F_B(d)$$

对于文本文档，F_A 和 F_B 为关键词和次关键词；对于图形、图像，F_A 和 F_B 为关键对象和次关键对象；对于视频，F_A 和 F_B 为关键帧和次关键帧。若从语义上来划分，F_A 和 F_B 分别为主要事件和次要事件。一个事件对象是一个复杂的数据结构，包含人物、物体等等。

总之，各类文档的特征均可以划分为两种类型的特征：语义特征和物理特征。语义特征表示文档的含义和所从属的话题等，主要以表示文档含义的关键词等来表征；物理特征包括文档的长度、作者、产生时间等。

3. 用户兴趣模型

3.1 基本概念

对信息用户而言，他们具有大量、分散、动态等特点。每个用户都因其自身的个性特征而有不同的信息需求，体现在对内容的类型要求不同、信息获得的实时性高低要求不同和信息表现方式的要求不同等方面。用户的信息需求也是不断发展变化的，即所需信息的类型等不断变化，用户随时都可能对新的内容产生兴趣，也可能对曾经感兴趣的内容不再需求。由于一条信息通常都是从属于一个或多个话题的，同时用户的兴趣也是可以按话题来划分的，所以我们依据话题来建立用户兴趣的模型，并将用户兴趣表示为话题的集合。

用户模型是指描述用户行为、特性、习惯等各项特征的模型。用户兴趣模型是指描述用户感兴趣的信息范围的模型。可以将用户兴趣模型按三个维度来划分[1]：

1)*短期/长期维*，反映用户信息需求的时间长短特性；

2)*明显/隐含维*，反映用户兴趣模型是由用户直接说明还是从用户的行为中抽取出来；

3)*个人/群组维*，反映用户兴趣模型是用于单个用户的还是一个模型用于一组用户。

用户模型可以由一些变量来表示，如用户最近的行为、用户对进入文档的响应、用户的位置等，从依据这些变量所收集的统计信息中可以得到用户兴趣模型。本文主要研究用户的较稳定的、变化较少的长期兴趣。

3.2 用户兴趣程度度量

已有的许多信息服务系统，如 Yahoo、Google、Researchindex 等，多采用基于分类和关键词的较简单的数据模型。在这种方法中，每一条信息与一个或多个类相关，一组关键词与一条或多条信息相关。类和关键词也用于说明用户查询或用户 profiles。为了更精确地反映用户的兴趣需求，本文做了进一步的探讨。

我们设：用户 u 对某一个话题 t 感兴趣的程度为 $p(t/u)$，话题集合为 T{t}。则用户 u 的话题兴趣向量为：$U=\langle p(t_1/u),p(t_2/u),\cdots,p(t_i/u)\rangle, i\in N$，由该向量构成用户的 profiles 文件。

一个文档 d 对某一话题 t 的从属程度为 $p(d/t)$，该值的取得可以是先抽取出文档的特征，然后与某话题特征进行隶属度计算，从而得到 $p(d/t)$值。

用户 u 对某一进入文档感兴趣的程度为 $i_u(d)=p(t/u)p(d/t)$，$i_u(d)$就是进行信息过滤时用来与用户阈值进行比较的数值。

用户阈值是指在过滤文档时，对文档进行匹配计算所得的值的最低要求。用户阈值既可以由用户代理通过观察、记

求、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，也可以由用户自己手动调整。用户阈值是动态变化的，由于随着时间的推移，用户对某一个感兴趣的话题研究也越来越深入，其阅读速度也必然越来越快，信息的需要量相应增加，此时就应适当地调整阈值。

当用户对某一个话题感兴趣的程度 $p(t/u)$ 越高时，即使一个文档与该话题的相关度较低，也会引起用户的注意。所以，此时的用户阈值应设置得低些。

定理1 对任何一个 $p(d/t)\geqslant 0$，当 $p(t/u_i)\leqslant p(t/u_j)$ 时，（其中 $0\leqslant i, 0\leqslant j$），必有 $i_{u_i}(d)\leqslant i_{u_j}(d)$ 成立。

证明：∵ $p(d/t)\geqslant 0$

又∵ $0\leqslant p(t/u_i)\leqslant p(t/u_j)$

∴ $0\leqslant p(t/u_i)p(d/t)\leqslant p(t/u_j)p(d/t)$

即：$i_{u_i}(d)\leqslant i_{u_j}(d)$

证毕

定理1说明，对属于某个话题的一个文档，对该话题感兴趣程度高的用户对该文档的感兴趣程度也相对较高。

3.3 话题的内涵与外延

每个概念、话题都包含有内涵和外延两个属性。由于用户的个人特征、个性等各不相同，因此不同的用户对同一话题的外延的要求也必然各不相同。许多信息服务系统的用户兴趣模型只是考虑了用户感兴趣话题的内在含义[4~7]，而没有注意用户感兴趣话题的外延范围的大小，因而根据这种用户兴趣模型提供给用户的信息会有一部分是用户不需要的，造成系统资源和用户精力的浪费。本文在结合话题内涵的基础上，通过加入用户感兴趣话题的外延来建立用户兴趣模型，可以更精确地反映用户的个性化需求，从而可以进一步提高服务质量。

内涵表示有关话题的内容、内在含义、解释。外延表示话题涉及范围的大小，它由时间、作者、位置等特征因素来确定。针对每个外延特征，都有相应的用于确定范围的限定算符和用于比较操作的比较算符。示例如下：

用户感兴趣的话题：

{内涵：
　　关键词：
　　…
　外延：时间：$t_1\leqslant t\leqslant t_2$；
　　作者：$\langle a_1, a_2, a_3\rangle$；
　　…
}

对于文本类型的文档，其内涵可以由关键词表示；对于超媒体类型的文档，其内涵可以根据媒体类型的不同（如音频、视频等）采用其它的表示方法（如颜色直方图等），MPEG-7标准已经为此制定了许多规则。

由此可以得到有关用户兴趣范围的判定规则的表达式，如：

{Keywords $=\langle w_1, w_2, w_3\rangle$
$t_1\leqslant t\leqslant t_2$；
作者 $=\langle a_1, a_2, a_3\rangle$；
}，其中 t 可以表示文档的产生时间或其他时间。

因为用户的兴趣需求总是在不断变化，所以用户可以自定义新的话题特征及相应的比较操作（也可由系统自动完成）。每个用户可以根据话题类生成各自的感兴趣的话题对象来描述用户的个性化的兴趣需求。话题对象的生成、维护可以是人工的，也可以由系统根据用户的行为自动生成。用户的一个或多个感兴趣话题对象共同构成用户的兴趣 profile 文件。这种方式既考虑了用户感兴趣话题的内涵，又考虑了不同用户对话题外延的不同需求，因而能更好、更精确地反映用户的需求。而传统的基于关键词匹配的方法只是一种基于话题内涵的比较方法，缺乏对外延的限制，使信息系统所提供服务的精确度不够高。

用户对外延的需求也是在不断动态变化的。用户当前需求的外延始终是整个话题外延的子集。而由于信息量的不断迅速增长的特性，整个话题外延也是在不断增长变化的。随着时间的推移、用户阅读过程的进行、用户兴趣的变化，使得用户对外延的需求也在变化。

由于话题对象的内涵和外延总是在不断地变化，同时为了提高系统服务器端的性能和灵活性，需要对话题对象进行操作，以便使系统中的话题数量保持在合适的范围之内。这些操作包括：新建、更新、拆分、合并、继承、撤消等。其中，对已有的话题对象是否进行拆分、合并、继承等操作主要是依据各话题对象所过滤得到的文档集合来进行判断。

例如拆分操作，设有话题对象 A 按属性 m（m 既可以是内涵属性也可以是外延属性）可以分为两个子话题对象 B1、B2，即 $\rightarrow$m 和 m 话题，则可将 $\rightarrow$m 和 m 加入到新产生话题对象的内涵属性或外延属性中。用“-”表示拆分运算符，则有：

-A(m)＝{B_1(m)，B_2($\rightarrow$m)}，其中 m 是内涵属性并且

B_1：内涵＝A：内涵＋m，

B_1：外延＝A：外延，

B_2：内涵＝A：内涵＋$\rightarrow$m。

B_2：外延＝A：外延。

若 m 是外延属性，则有：

B_1：内涵＝A：内涵，

B_1：外延＝A：外延＋m，

B_2：内涵＝A：内涵。

B_2：外延＝A：外延＋$\rightarrow$m。

上式中，-A(m)表示对话题对象 A 按属性 m 进行拆分操作。

用“＋”表示合并运算符，则有：

A＋B＝C，并且

C：内涵＝A：内涵 ∪ B：内涵，

C：外延＝ A：外延 ∪ B：外延。

用 * 表示继承运算符，则有：

C＝ * A，并且

C：内涵＝{a_i，c_j｜ $a_i\in$ A：内涵，$i\in$ Integer；c_j为新内涵元素，$j\in$ Integer }，

C：外延＝{a_i，c_j｜ $a_i\in$ A：外延，$i\in$ Integer；c_j 为新外延元素，$j\in$ Integer }。

3.4 用户需求模型

用户 profile 文件表达了用户对信息的需求和兴趣，是信息过滤系统的关键组件。用户向系统提供了一次 profile 文件后，就可以及时、连续地收到与他相关的数据，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。另外，基于用户的兴趣是随时间而变化的事实，需要及时修改 profile 文件以反映用户信息需求的变化。

用户兴趣由其感兴趣的话题类组成（如足球比赛是一个话题类）。用户感兴趣的当前话题对象是类的实现，如球赛是一个类，则某年的球赛就是该类的一个实现。

在上文的基础上，我们提出了一个基于话题的用户需求模型，其 profile 格式为：

user { $topic_1$, threshold, limited;
$topic_2$, threshold, limited;
…
$topic_n$, threshold, limited}

简记为$\langle p_1, p_2, \cdots, p_n\rangle$，其中 limited 表示用户感兴趣话题 $topic_i$ 的外延限制，threshold 表示用户阈值。阈值 threshold 用于判别用户对一篇文档是否感兴趣，属于对内涵的要求。

相应的有下面的判定规则：

规则1

对文档 D 与用户 u，若 D 与话题 $T_i(i\in N)$的相关值 $R>=user_u:threshold_{T_i}$，并且 D 符合用户 u 对话题 T_i 的外延限制，则称 D 是用户 u 所需要的文档。

文档与话题都有一个相关值，这个值是统一的，与用户无关的，在0和1之间。Threshold 也取0和1之间，但每个用户依据自己的需要其 threshold 值各不相同，形成了个性化需求。

3.5 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：

$$Q=U+M$$

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'。用户兴趣向量$\langle p_1, p_2, \cdots, p_n\rangle$中元素选取的优劣、数量多少直接影响到 U' 与 U 的接近程度和信息过滤算法的效率。也是评价信息过滤效果、进而调整用户 profile 文件的重要依据。向量$\langle p_1, p_2, \cdots, p_n\rangle$中元素的选取应该既是较少的以便使计算量较少，同时又是较优的以便能准确描述用户的需求。

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)的信息的特征从向量$\langle p_1, p_2, \cdots, p_n\rangle$中去除或是修改其特征值；将新发现的用户感兴趣的信息的特征加入到向量$\langle p_1, p_2, \cdots, p_n\rangle$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动态变化的兴趣需求，使 U' 能逐渐逼近 U。

小结 基于关键词的用户需求模型只注重了用户需求的个性，而未注重用户需求的共性，不便于进行协同过滤和对用户群的过滤，虽可减轻网络流量，但却使系统要处理的用户模型的数量随用户规模的扩大而成正比扩大，不适合于大规模信息过滤系统。此外，尽管用户的话题需求是较为长期、稳定的，但由于不断地进行阅读活动，用户对话题的外延的要求却是随时间变化的。因此本文从话题的角度并加进了外延限制而建立的用户兴趣模型，既能反映用户需求的共性，又能反映用户的个性化特征，使系统可以较为灵活地处理用户需求，在为单个用户过滤信息的同时，还可以通过改变外延限制而使系统只需以较少的模型数量来统一地表示用户群的信息需求，高效地为用户群过滤信息，增强了系统的可伸缩性，适合于大规模信息过滤系统。

参 考 文 献

1 Allen R B. User models: Theory, Method, and practice, Int. J. Man-Machine Studies, 1990, 32: 511～543
2 Stadnyk I, Kass R. Modeling User's Interest in Information Filters. Communication of the ACM, 1992, 35(12): 49～50
3 Lee D L, Chuang H, Seamons K, Document Ranking and the Vector Space Model. IEEE Software, 1997, 14(2)
4 Foltz P W, Dumais S T. Personalized Information Delivery: An Analysis of Information Filtering Methods. Communications of ACM, 1992, 35: 51～60
5 tauritz D. Adaptive Information Filtering as a Means to overcome Information Overload, M. Sc. thesis, Department of Computer Science Leiden University, the Netherland, sep. 1996
6 Resnick P, Varian H R. Recommender Systems. COMMUNICATIONS of the ACM, 1997, 40(3): 56～58
7 Rodriguez-Mula G, et al. Collaborative value filtering on the Web. Computer Networks and ISDN Systems, 1998, 30: 736～738

(上接第122页)

最后由站点的分析员定义阈值，对推荐结果进行排序、筛选、合并，确定给用户的推荐商品集为：USER(u)←{j|Rec(j)>α, j∈IS}。阈值的确定可以根据需要灵活使用不同的方法，例如可以确定 $\alpha=\frac{1}{|J|}\sum_{j\in J}\mathrm{Re}\,c(j)$。这部分产生的推荐主要可以使站点提高交叉销售收益。

此外，对当前用户 u 还有另一方面的推荐集。该用户对站点访问的 session 对象中的最后一个页面 p 可以认为代表用户的最新兴趣，用户之所以访问这个页面 p，就是因为该页面所涉及的物品内容。所以运用信息抽取方法[3]，找出该页面涉及的物品有 $k_1, \cdots\cdots, k_z$，$1\leqslant z\leqslant|IS|$，而且有 $w(k_1, p), \cdots, w(k_z, p)$。则此时衡量物品 m 是否可以推荐给用户 u 的关于物品 m 推荐值的计算公式为：$\mathrm{Re}\,c(m)=\sum_{i=1}^{z}w(k_i, p)\times sim(k_i, m)$。由此公式可以得到当前对于用户 u 的另一个推荐集为：USER(u)←{m|Rec(m)>β, m∈IS}。

小结 在电子商务的新形势下，为了提高客户的忠诚度我们需要深入分析、了解客户，建立关于客户的兴趣偏好等知识模式，然后以这些建立的知识模式为依据向客户提供个性化的服务。本文则在当前已有的研究基础上采用 item-to-item correlation 和 user-to-item correlation 相结合的方法，确定针对当前用户的推荐策略，并根据需要建立了物品间相似模式、用户-物品模式等。

参 考 文 献

1 Cooley R, Mobasher B, Srivastava J. Grouping Web Page References into Transactions for Mining World Wide Web Browsing Patterns. In: Proc. of ICTAI 1997
2 Han Jiawei, Kamber M. data mining
3 Craig A. Knoblock. Accurately and Reliably Extracting Data from the Web: A Machine Learning Approach. IEEE Data Engineering Bulletin, Volume 23
4 Customer Data Quality. http://WWW.firstlogic.com
5 Balabanovic M, Shoham Y, Yun Y. An Adaptive Agent for Automated Web Browsing. http://WWW.citeseer.com
6 [美]Berson A, Smith S, Thearling K 著，贺奇等译. 构建面向CRM的数据挖掘应用
7 Schafer J B, et al. E-Commerce Recommendation Applications. Data Mining and Knowledge Discovery, 2001, 5(1/2): 115～153
8 Nahm U Y, Mooney R J. Text Mining with Information Extraction. In the proceeding IJCAI-2001
9 Buchner A G, et al. An Internet-enabled Knowledge Discovery Process. http://www.citeseer.com
10 Karypis G. Evaluation of Item-Based Top-N Recommendation Algorithms: [Technical Report CS-TR-00-46]. Computer Science Dept., University of Minnesota. 2001