第 32 卷 第 14 期 计 算 机 工 程 2006 年 7 月
Vol.32 № 14 Computer Engineering July 2006
·人工智能及识别技术· 文章编号：1000—3428(2006)14—0169—03 文献标识码：A 中图分类号：TP311

基于本体的 Web 使用知识发现模型及应用

何 丽[1,2]，严冬梅[2]，韩文秀[1]
（1. 天津大学管理学院，天津 300072；2. 天津财经大学理工学院，天津 300222）

摘 要：本体在 Web 上的应用能够有效解决 Web 信息共享的语义问题。该文提出了基于 Web 本体和服务器日志文件的知识发现模型，主要讨论了用户访问行为的表示、语义用户分布的定义及发现算法。最后介绍了 Web 使用知识发现模型在 Web 个性化系统中的应用。
关键词：语义 Web；本体；Web 使用挖掘；语义用户分布

Web Usage Knowledge Discovery Model and Its Applications Based on Ontology

HE Li[1,2], YAN Dongmei[2], HAN Wenxiu[1]
(1. College of Management, Tianjin University, Tianjin 300072;
2. College of Science & Technology, Tianjin University of Finance & Economics, Tianjin 300222)

【Abstract】 The application of ontology in the world wide web provides an effective solution to share web information with awareness of semantics. This paper proposes a new model for knowledge discovery based on web log files and ontology, chiefly discusses the presentation of user access behaviors and the definition and discovery algorithm of semantic user profiles. Finally, it introduces an application framework of Web usage knowledge discovery model in Web personalization system.
【Key words】 Semantic Web; Ontology; Web usage mining; Semantic user profile

1 语义 Web 概述

本体在Web上的应用产生了语义Web[2]。与传统的Web相比，语义Web解决了Web上信息共享的语义问题，其目标是实现Web资源语义层次的共享和互操作。语义Web的创始人Tim Berners-Lee提出的语义Web体系结构从下到上分别为：Unicode, URL, XML, RDF(S), Ontology Vocabulary, Logic, Proof和Trust 7 个层次。其中，建立在Unicode、URL、XML、RDF(S)等语言标准上的本体层(Ontology Vocabulary)对实现语义Web的目标起着关键作用。它为领域概念模型的描述提供了丰富的原语，是进行知识推理和验证的基础。

标准的 Web 本体描述语言能够通过类、属性、公理和实例等元素来准确地描述 Web 上的元数据和 Web 对象，并提供基于分类层次的领域知识及其语义关联。因此，根据 Web 本体描述的类、属性、类之间的关联、属性之间的关联、类与属性之间的关联，以及类之间的外延关系和属性之间的外延关系(如，不相交、覆盖、等价) 等，可以准确地发现 Web 对象之间的关联和潜藏的语义知识。

Web本体描述语言是对Web本体进行定义和描述的一种语言。随着语义 Web研究的深入，Web本体描述语言已经从最初的RDF(S)、OIL、DAML 、DAML十OIL发展到OWL。OWL是W3C推荐的一种标准的Web本体描述语言，它在XML/RDF等已有标准的基础上，通过添加大量的基于描述逻辑的语义原语来描述和构造Web本体[3]。OWL本体抽象语法主要包含注释(annotations)、公理(axioms)和事实(facts)。OWL本体的主要内容在公理和事实中执行，OWL事实提供了个体(individuals)的结构化描述，每个个体由个体标识(individualID)和一组数据类型(type)和值(value)组成；OWL公理提供了类(class)和属性(properties)的描述。OWL类表示个体的集合，属性反映了个体与其他信息之间的语义关联。若用 C_1 和 C_2 来表示OWL定义的两个本体类，x_1 和 x_2 表示两个个体，P_1 和 P_2 为两个属性，对应的OWL主要公理构造符及描述逻辑语法如表 1 所示。

表 1 OWL 提供的主要公理构造符

公理构造符	描述逻辑语法	公理构造符	描述逻辑语法
SubClassOf	$C_1 \subseteq C_2$	SubPropertyOf	$P_1 \subseteq P_2$
EquivalentClass	$C_1 \equiv C_2$	EquivalentProperty	$P_1 \equiv P_2$
DisjointWith	$C_1 \cap C_2 = \{\}$	InverseOf	$P_1 \equiv P_2^-$
SameindIividualAs	$\{x_1\} \equiv \{x_2\}$	SymmetricProperty	$P \equiv P^-$
DifferentFrom	$\{x_1\} \neq \{x_2\}$	TransitiveProperty	$P^+ \subseteq P$

利用 Web 本体提供的类、属性，及其公理所提供的推理规则，用户可以准确地使用数据中隐藏的知识，并且可以用本体来表示发现的知识，以实现知识在语义层的共享。

2 用户访问行为表示

Web 服务器日志文件记录了 Web 用户的浏览行为，一般包括访问日志、引用日志和代理日志。根据站点结构和 Web 本体，可以把一个 Web 站点表示为一组 Web 对象的集合。如果某站点上有 s 个可访问的 Web 对象，则该站点的对象集合可描述为：$OL=\{obj_1, obj_2, ..., obj_s\}$。根据 Web 页面和 Web 对象之间的映射关系，对服务器中的日志记录进行预处理，可

基金项目：天津市自然科学基金资助项目（033611011）
作者简介：何 丽(1969—)，女，讲师、博士生，主研方向：Web 数据挖掘，Web 决策支持系统；严冬梅，讲师、博士；韩文秀，教授、博导
收稿日期：2005-10-19 **E-mail：**renke21@vip.sina.com

以将服务器日志记录表示为 $L=<uid,\{obj,time\}^n>$ 形式。其中，uid 代表一个访问用户，obj 表示该用户访问的 Web 对象，$time$ 表示 uid 在 obj 上的浏览时间。

定义 1(用户访问行为) 用户 A 的访问行为可描述为 $A=<uid_A,\{(l_A.obj,hits,time)\}^n>$。其中，$l_A \in L$，$n \geq 1$，$hits$ 表示用户对 $l_A.obj$ 的访问次数，$time$ 表示用户对 $l_A.obj$ 的访问时间。

在 Web 这个开放的环境中，用户的浏览行为隐藏着用户的访问偏好，具有相似访问偏好的用户通常具有相似的浏览特征。在数据预处理阶段，可以使用聚类和关联规则发现等方法对历史用户进行聚类，使具有相似访问行为的用户聚集形成若干个相似用户群。关于用户聚类方法的研究成果有很多，在此不再赘述。

定义 2(用户分布 pr_i) 用户分布 pr_i 描述了用户群 i 的访问偏好，可表示为

$$pr_i=\{<obj_1,h_{i1}>,<obj_2,h_{i2}>,...<obj_s,h_{is}>\}$$

其中，h_{ij} 表示该用户群 i 对对象 obj_j 的访问偏好。一般情况下，用户在某对象上的访问偏好与用户在该对象上的访问时间和访问次数成正比，可以将 h_{ij} 定义为

$$h_{ij}=\frac{hits_{ij}\cdot time_{ij}}{\sum_{1\leq k\leq s} hits_{ik}\cdot times_{ik}} \tag{1}$$

其中，$hits_{ij}$ 和 $time_{ij}$ 分别表示该用户群 i 在 obj_j 上总的访问次数和访问时间。没有被访问的 Web 对象在 pr_i 中的偏好权重为 0。

假设某站点中所有历史用户经过聚类后产生了 t 个不同的用户群，根据式(1)和 Web 日志，可以建立所有用户群和所有 Web 对象之间的访问偏好关联矩阵 $M_{t,s}$：

$$M_{t,s}=\begin{bmatrix} h_{11} & h_{12} & ... & h_{1s} \\ h_{21} & h_{22} & ... & h_{2s} \\ ... & ... & .. & .. \\ h_{t1} & h_{t2} & ... & h_{ts} \end{bmatrix}$$

矩阵 $M_{t,s}$ 中的每一行向量代表一个用户分布，表示一个用户群在该站点所有对象上的访问偏好情况。

3 Web 使用知识发现模型

根据 Web 本体描述的信息和资源之间的关系，在已建立的用户访问偏好关联矩阵 $M_{t,s}$ 的基础上，通过建立基于 Web 本体的使用知识发现模型，可以获得不同用户群在语义层的访问偏好。

3.1 Web 对象与本体类之间的映射模型

定义 3(类实例) 假定

$$A_c=\{a_1^C,a_2^C,...,a_n^C\}$$

表示本体中定义的类 C 的属性集合，若对象 obj 具有属性

$$A_o=\{a_1^{obj},a_2^{obj},...,a_n^{obj}\}$$

或存在属性 $b^{obj}\in A_o$，$b^{obj}\notin A_C$，但 b^{obj} 是 $a_i^C,(1\leq i\leq n)$ 的一个等价属性，那么对象 obj 是类 C 的一个实例。

在语义 Web 环境下，每个用户群感兴趣的 Web 对象可能分布在不同本体类对应的领域中。根据 Web 本体描述的类与属性之间的关联以及本体中对个体的描述，可以直接建立 Web 对象到本体类之间的显式映射模型。

若 Web 本体描述的类集合为 Ω，则 Web 对象列表 OL 与 Ω 之间的映射模型定义为 $\Gamma:OL\rightarrow\Omega$。

$$\Gamma(obj_i)=\begin{cases} X,\text{若}obj_i\text{与类}X\text{相关} \\ \Omega_{root},\text{否则} \end{cases} \tag{2}$$

其中，$X\in\Omega$ 表示一个本体类，Ω_{root} 代表本体中定义的根类，代表特定领域范围的最高层概念。若 obj_i 是类 X 的一个类实例或是本体中描述的 X 的一个个体，则 obj_i 与 X 相关。

根据 $\Gamma(obj_i)$ 定义的映射模型，可以实现对 Web 对象的分类。在实际应用中，为了显式地获得用户在不同语义层的访问情况，还需要建立 Web 对象到本体类的反向映射模型 $\sigma:\Omega\rightarrow 2^{OL}$。

$$\sigma(X)=\{obj\mid obj\in OL,\Gamma(obj)=X\text{或}\Gamma(obj)=Y,Y\text{是}X\text{的等价类}\} \tag{3}$$

Web 本体不仅提供了 Web 对象与类之间的语义关联，还提供了类之间的外延关联。可以借助于 Web 本体提供的类与其超类、子类之间的关系，实现用户访问对象在语义层的扩充。

$$\begin{aligned} \sigma^{-1}(X)&=\sigma(X)\cup\bigcup_{Y\text{是}X\text{的子类}}\sigma(Y) \\ \sigma^{+1}(X)&=\sigma^{-1}(X)\cup\bigcup_{Y\text{是}X\text{的超类或}X\text{与 Y 具有共同的基类}}\sigma(Y) \end{aligned} \tag{4}$$

$\sigma^{-1}(X)$ 和 $\sigma^{+1}(X)$ 分别表示 σ 的向下和向上扩展，它可以在语义层对用户的访问对象进行扩充，扩展 Web 检索系统的查询范围。图 1 是与某 Web 本体对应的一个概念模型。假如某用户的访问偏好为结点 development and programming 对应的 Web 对象，则通过 $\sigma^{-1}(X)$ 和 $\sigma^{+1}(X)$ 进行扩展，可以发现该用户在 language 、database 和 soft engineer 等概念上的访问兴趣，并从语义上扩充和增强对用户服务。

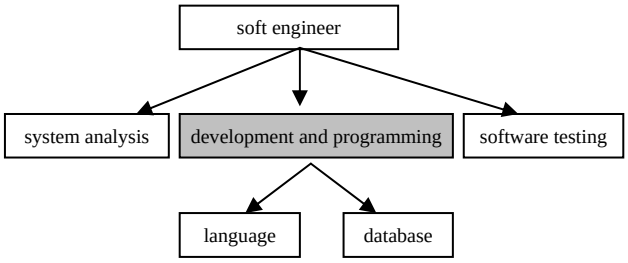

图 1 Web 本体中的一个概念模型

3.2 语义用户分布及发现算法

根据建立的类与 Web 对象之间的双向映射模型，可以获得每个用户群在语义层的访问偏好。

定义 4(语义用户分布 spr_i) 若某站点本体描述的类的集合为 $\Omega=\{X_1,X_2,....,X_k\}$，则语义用户分布 spr_i 可表示为

$$spr_i=\{<X_1,w_{i1})>,<X_2,w_{i2}>,...<X_k,w_{ik}>\}$$

w_{ij} 表示类 X_j 在语义用户分布 spr_i 中的权重，$w_{ij}\in[0,1]$。

语义用户分布 spr_i 反映了用户 i 在语义层的访问偏好。根据访问偏好关联矩阵 $M_{t,s}$，和本体类与 Web 对象之间的映射模型，可以建立对应的群分布矩阵 $M'_{t,k}$：

$$M'_{t,k}=\begin{bmatrix} w_{11} & w_{12} & ... & w_{1k} \\ w_{21} & w_{22} & ... & w_{2k} \\ ... & ... & .. & .. \\ w_{t1} & w_{t2} & ... & w_{tk} \end{bmatrix}$$

$M'_{t,k}$ 的每一个行向量对应一个语义用户分布。根据定义 4，一个语义用户分布是基于本体类的一个权重向量，权重 w_{ij} 为 $\sigma(X_j)$ 映射的所有对象在 pr_i 中的权值之和。$M'_{t,k}$ 的生

成算法描述如下：

输入：用户分布矩阵 $M_{t,s}$ 和站点本体；

输出：语义用户分布矩阵 $M'_{t,k}$；

步骤：

(1)初始化 $M'_{t,k}$ 各元素的值为 0；

(2)建立 Web 对象与本体类之间的映射

FOR $i = 1$ to k DO /*初始化各类映射的对象集合*/

$\sigma(X_i) = \phi$; EndFor

FOR $j = 1$ to s DO

$X_r = \Gamma(obj_j)$; /* $X_r \in \Omega$ */

$\sigma(X_r) = \sigma(X_r) \cup \{obj_j\}$;

/*将 web 对象归入不同本体类 */

(3)建立语义用户分布矩阵

FOR each $pr_i \in M_{t,s}$ DO

FOR each $obj_l \in OL$ DO

$M'_{t,k}(ij) = M'_{t,k}(ij) + \sum_{obj_l \in X_j} M_{t,s}(il)$;EndFor

EndFor

语义用户分布矩阵 $M'_{t,k}$ 反映了所有用户在语义层的访问偏好。为了实现基于 Web 的知识共享，可以反过来用本体描述语言来表示通过知识模型获得的知识。假设某用户只访问了图 1 中的概念节点 development and programming 及其子概念的相关对象，且相应的语义用户分布的权重向量为{<language,0.52>,<database,0.48>}，则可用 XML 语法来描述该语义用户分布如下：

```
<concept>
<name>development and programming </name>
<concepttermlist>
    <term>
    <weight>1</weight>
      <name> development and programming</name>
    </term>
    <term>
    <weight>0.52</weight>
      <name> langue</name>
    </term>
    <term>
    <weight>0.48</weight>
      <name> database</name>
    </term>
    </concepttermlist>
</concept>
```

4 Web 使用知识发现模型在 Web 个性化系统中的应用框架

要实现基于 Web 的个性化服务，首先需要解决两个关键问题：(1)如何有效地描述用户的服务请求；(2)如何实时、准确地获得和反馈反映用户服务请求的 Web 信息。

传统的 Web 个性化方法(如合作过滤方法(Collaborative Filtering, CF))虽然能够在一定程度上解决这两个问题，但它们大都是基于事务数据库而实现的，不能为用户提供基于 Web 语义的、更加灵活的个性化服务。在语义 Web 环境下，可以使用语义用户分布来获得用户在不同领域的访问偏好，从而为用户提供准确而灵活的智能化服务。语义用户分布在 Web 个性化系统中的应用框架如图 2 所示。

该个性化系统可分成 3 个主要处理阶段：数据预处理，语义知识发现和在线推荐。其中，数据预处理和语义知识发现是两个离线部件。

数据预处理阶段根据现有的 Web 日志记录和站点本体，使用传统的 Web 使用挖掘(聚类、关联规则等)技术将用户的访问记录转换为定义 1 描述的用户分布，并建立访问偏好关联矩阵 $M_{t,s}$；语义知识发现阶段在 $M_{t,s}$ 的基础上生成语义用户分布矩阵 $M'_{t,k}$。

在线推荐阶段的主要任务是根据语义用户分布矩阵 $M'_{t,k}$ 描述的用户访问偏好，准确及时地反馈在线用户感兴趣的 Web 信息或服务。该阶段主要由 3 个部分组成：

(1)获得当前用户分布。在线推荐引擎首先收集活动用户访问的 Web 对象，并根据已建立的知识发现模型和 Web 本体将当前用户的活动会话转换成基于本体类集合 Ω 的当前用户分布。

(2)模式匹配。通过预设一个最小信息过滤阈值 Δ，将当前用户分布与语义用户分布矩阵 $M'_{t,k}$ 中的每个语义用户分布进行匹配，并将匹配结果大于 Δ 的语义用户分布归入推荐集合。

(3)将推荐集合中的语义用户分布实例化成真实的 Web 对象并以交互的方式推荐给活动用户，用户可以根据系统推荐的结果选择自己需要的信息或服务，或进一步修正自己的查询。语义用户分布是基于本体类集合 Ω 的群分布，对符合匹配结果的语义用户分布可以根据式(3)转换成真实的 Web 页面集合，必要时还可以根据式(4)对推荐的 Web 对象集合进行扩充。

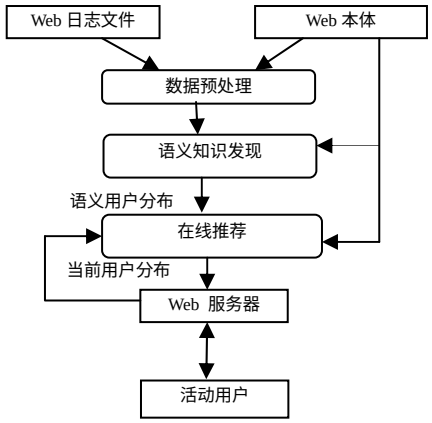

图 2 知识模型在 Web 个性化服务系统中的应用框架

基于Web本体的知识发现模型在Web个性化系统中的应用，一方面可大大降低传统的个性化系统在模式匹配过程的计算负担；另一方面，可利用本体类和对象之间的双向映射模型解决传统Web个性化系统中的“新项目”问题[8]，并可在其基础上根据本体描述语言提供的类与属性间的关系，实现基于本体属性的更细粒度、交互式的Web个性化服务。

5 结束语

随着电子商务的不断发展和语义 Web 研究的深入，基于本体的 Web 使用知识发现和智能化服务将成为未来一段时间内 Web 领域的一个新的研究热点。本文提出的基于本体的 Web 使用知识发现模型，能根据 Web 对象的内在属性及 Web 本体描述的类与属性之间的关联，为用户活动提供更加准确的解释和推理能力，在基于语义 Web 的智能信息检索和智能电子商务的个性化服务中具有一定的实际意义和参考价值。

（下转第 201 页）