基于领域本体的用户模型的研究 *

蒋秀林, 谢 强, 丁秋林
(南京航空航天大学 计算机科学与技术学院, 南京 210016)

摘 要: 目前大多数知识管理系统采用基于关键词或关键词向量空间模型表示用户的兴趣偏好。针对该方法不包含语义信息,很难准确表示用户感兴趣的信息,并且难于扩展,提出一种基于领域本体的用户模型。该模型利用用户访问量,采用改进的相似度算法,实现用户分类建立用户模型,体现用户个人偏好。最后将该模型应用于齐齐哈尔货车快速设计系统中,应用表明该模型能准确地反映用户兴趣,且提高了信息检索效率。

关键词: 用户模型; 本体; 领域本体; 相似度算法
中图分类号: TP311 **文献标志码**: A **文章编号**: 1001-3695(2012)02-0606-03
doi:10.3969/j.issn.1001-3695.2012.02.054

Research of user model based on domain ontology

JIANG Xiu-lin, XIE Qiang, DING Qiu-lin
(*College of Computer Science & Technology, Nanjing University of Aeronautics & Astronautics, Nanjing* 210016, *China*)

Abstract: At present the majority knowledge management systems use the handling of traffic system which still adopts the traditional retrieval mode which is the retrieval mode based keyword or based keyword rector to denote users' interest. The method does not contain semantic information, is difficult to accurately capture and extended users' interested information. In order to solve this problem, this paper proposed domain ontology based on user model, presented a domain ontology based on user model through the utilization of user access volume, used an improved similarity algorithm, realized of user classification based on user model and reflected the user's personal preferences. Finally it applied in Qiqiha'er freight vehicle fast design system, and had been proved improvement of the retrieval efficiency.

Key words: user modeling; ontology; domain ontology; semantic algorithm

0 引言

现在是一个信息大爆炸的时代,随着计算机技术的发展和进步,网络信息在不断地膨胀,如何快速准确地定位到需要的信息,如何提供个性化的服务,成为当前研究的热点和重点。虽然基于关键词或关键词向量空间模型的理论成熟,但这种方法缺乏语义信息,很难实现信息的共享和重用,因此这些模型不能准确地描述用户的个人兴趣。

为了解决上述缺陷,国内外许多系统引入本体概念,用来表示用户感兴趣的领域。但这些系统大多数只是一个概念层次结构,且用户模型是静态的,如 Smart Push[1]、OBWAN[2]等。由国内徐振宁等人提出的基于领域本体的动态用户模型,在学习更新过程中更多考虑的是浏览或检索文档中单个概念节点对用户兴趣度的影响,在一定程度上拆分了这些浏览或检索文档的本身语义信息。2005 年 Heckmann 等人[3]开发了通用用户本体 GUMO,用于智能语义丰富环境中的分布式用户模型的统一描述。但是在不同领域或不同应用中,对于相同的概念有不同的解释,要想实现 GUMO 在不同领域中的应用,就需要建立对不同领域的应用机制。为了增强语义信息、提高检索效率,本文引入本体,在领域本体的基础上构建用户模型,并将这一模型应用在齐齐哈尔货车快速设计系统中。

1 基于领域本体的用户模型的定义

1.1 领域本体

领域本体是专业性的本体,描述的是特定领域中概念与概念之间的关系,提供了某个专业学科领域中概念的词表及概念间的关系。它研究如何定义特定领域中的概念、概念之间的关系、发生的活动以及该领域的主要理论和基本原理。按照本体的表示形式,领域本体可以形式化地表示为一个五元组[4]:

$$\mathrm{Dom}O = (\mathrm{info}, \mathrm{Dom}C, \mathrm{Dom}R, \mathrm{rules}, \mathrm{functions})$$

其中:info 表示领域本体的基本信息,本文中主要存储用户登录名、所属部门、职务等基本信息;Dom$C$ 是该领域内相关概念的集合;Dom$R$ 是领域本体中概念存在的关系集;rules 表示在特定领域中既定的规则;functions 表示函数。

本文采用 Protégé -4.1-beta 工具中的 OWL 语言来描述该领域中知识本体。为了维护这些知识本体,将它们存储在知识本体库中。

1.2 基于领域本体的用户模型

用户模型是实现个性化服务的基础和核心,是一种面向算

收稿日期: 2011-08-21; **修回日期**: 2011-09-30 **基金项目**: 国家自然科学基金资助项目(70871061,71171117)

作者简介: 蒋秀林(1987-),女,安徽安庆人,硕士研究生,主要研究方向为知识工程、信息安全(solinlee@163.com);谢强(1972-),男,副教授,博士,主要研究方向为知识工程、信息系统与信息安全、人机交互;丁秋林(1935-),男,所长,教授,博士,主要研究方向为知识工程、信息系统与信息安全、人机交互等.

法、具有特定结构、形式化的用户描述。本文讨论的用户模型采用如下形式的定义:

$$\mathrm{DOBUM} = (A_{\mathrm{user}},\ \mathrm{Dom}O,\ AS,\ R_{\mathrm{user}})$$

其中:A_{user}是用户属性集合,包括基本信息集和知识背景集,而基本信息集包括用户名、职称、部门、电话、住址等,知识背景集则记录用户的个人偏好即常访问的关键词;Dom$O$ 是前面定义的领域本体(本文针对用户感兴趣的领域本体);$AS = ((c_1, s_1),(c_2, s_2),\cdots,(c_n, s_n))$,$c_1, c_2, \cdots, c_n$ 是领域本体中概念的集合,n 是领域本体中相关概念的个数,$s_1, s_2, \cdots, s_n$ 是用户对领域本体中每个概念的访问量;R_{user}表示用户间关系的集合,反映用户间的层次结构。

从用户模型的形式化定义可以看出,用户模型主要涉及的是用户感兴趣的概念、概念的属性以及属性之间的关系,因此采用本体概念层次树结构来存储。通过下面三个步骤建立基于领域本体的用户模型:

a)用户登录后,分析日志获取用户的基本信息和访问本体概念层次树中叶子节点的访问量。

b)通过本体的推理和扩展技术以及 a)中获取的访问量计算非叶子节点的访问量。

c)合并叶子节点的访问量向量和非叶子节点的访问量向量,用 V 来表示,其中 $V = (v_1, v_2, \cdots, v_n)$,$v_i(1 \leqslant i \leqslant n)$ 表示用户对第 i 个概念节点感兴趣的程度,n 表示本体概念层次树中节点的个数。

提取本体概念层次树中的叶子节点,它们的访问量向量用 V' 表示,$V' = (v_1', v_2', \cdots, v_t')$,其中 v_i' $(1 \leqslant i \leqslant t)$ 表示用户对第 i 个概念叶子节点感兴趣的程度,t 表示本体概念层次树中叶子节点的个数。变量 v_i'的计算公式为

$$V_i' = \log_2(\sum_{R \in l_i} F_R) / \sum_{j=1}^{t} \log_2(\sum_{R \in l_i} F_R) \tag{1}$$

其中:F_R 表示用户访问叶子节点 l_i 的 R 资源的次数。通过式(1)可以计算出叶子节点访问量向量。

本体概念层次树中父子节点之间存在语义关系,基于这一点,可以利用本体的推理技术,然后根据叶子节点的访问量计算出树中非叶子节点的访问量。下面给出非叶子节点访问量的计算公式(式(2))。假设树中有 t 个叶子节点,$P_1, P_2, \cdots, P_t$ 表示从根节点到所有叶子节点的最短路径,$(n_{i0}, n_{i1}, \cdots, n_{iv})$ 表示路径 P_i 从根节点到叶子节点所经过的节点,用 $S(n_{iu})$ 表示路径 P_i 中节点 $n_{iu}(0 \leqslant u \leqslant v)$ 的访问量,计算公式为

$$S(n_{iu}) = \begin{cases} \alpha * S(n_{i(u+1)}) / N(n_{i(u+1)}) + 1 & 0 \leqslant u \leqslant v \\ V_{iu}' & u = v \end{cases} \tag{2}$$

其中:参数 a 表示推理因子,是通过实践得出,本文中 $\alpha = 1.8$;$n_{i(u+1)}$ 节点表示路径 P_i 中 n_{iu} 节点的儿子节点;$N(n_{i(u+1)})$ 表示树中 $n_{i(u+1)}$ 节点的兄弟节点的个数。

通过式(2)可以计算出所有路径中节点的访问量,因此树中的节点 $n_u(1 \leqslant u \leqslant t)$ 的访问量计算公式为

$$S(n_u) = \sum_{u=1}^{t} S(n_u) \tag{3}$$

综上分析,可以获得非叶子节点的访问量向量 $V'' = (v_1'', v_2'', \cdots, v_r'')$。其中:$v_i''$ $(1 \leqslant i \leqslant r)$ 表示用户对第 i 个概念非叶子节点感兴趣的程度,r 表示本体概念层次树中非叶子节点的个数。

2　基于领域本体的用户模型的构建

由于登录系统的用户多且身份关系复杂,但是他们之间会存在一定的相似性,例如某些用户属于同一领域。为了提高检索效率和个性化服务质量,本文利用改进的相似度算法对用户进行分类,使同一领域的用户兴趣和需求尽可能相同,这样就可以对同一领域的用户建立一种用户模型,为这一领域的用户提供相似的服务。当某一用户的兴趣需求发生变化时,系统只需对该用户所在用户模型进行更新。

2.1　DOBUM 的构建过程

基于领域本体的用户模型的建立过程包括用户的基本信息的获取、领域本体的构建、相似用户分类、用户模型构建这几个步骤。首先通过用户登录和用户浏览所得的日志文件搜集用户的基本信息,并对这些信息进行处理;其次利用这些信息和领域本体库使用本体编辑工具建立领域本体;再次通过改进的相似度算法对用户进行分类;最后对每类用户建立用户模型。由于这是个动态的过程,构建过程中产生的数据量比较大,因此需要利用更新算法对每类用户模型进行更新,及时淘汰不感兴趣的信息,及时推荐用户感兴趣的信息。整个过程如图 1 所示。

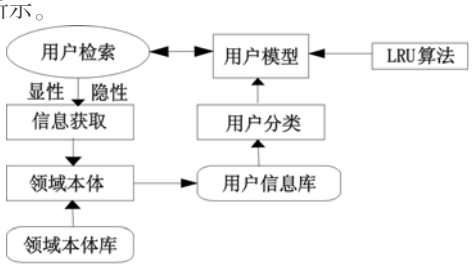


图1　DOBUM的构建过程

2.2　用户分类过程

通常情况下,用户多且复杂,但是他们之中有些属于同一领域,因此可以将这些用户进行分类,因为属于同一领域的用户在检索时往往具有一定的相似性,这样可以利用这一特征来推荐用户感兴趣的信息。在 2.1 节中介绍了用户模型的定义,本文通过本体概念层次树中的用户访问量来计算用户相似性。一般情况下,本体概念层次树中有几百个节点,但是用户感兴趣的只是一小部分,因此就需要解决数据稀疏性问题。数据越稀疏也就难获取相似的用户以及相似的资源,虽然目前已经提出很多不同的计算相似度的算法,常用的就包括基于余弦的相似度公式[5]、基于相关系数的相似度公式[6]和调整的余弦相似度公式[7]三种,但是它们都还存在许多缺陷,针对这些问题,本文提出一种改进的相似度算法。

首先给出几个定义:

$$R_{ij} = \mathrm{num}_{bij} / \mathrm{num}_{ij} \tag{4}$$

其中:num_{bij}表示被两个相似用户(所在部门、职务、权限等相同的用户)共同访问的资源的数量,num_{ij}表示不被这两个用户共同访问的资源的数量。

$$\mathrm{Dis}(i,j) = \sqrt{\sum_{u \in U_{ij}} (R_{iu} - R_{ju})^2} \tag{5}$$

其中:U_{ij}表示两个相似用户共同访问的资源;R_{iu}表示 i 用户访问 u 资源的访问量;R_{ju}表示 j 用户访问 u 资源的访问量;Dis(i, j)表示一般资源之间的访问量相似度。

$$S(i,j) = R_i^{\mathrm{T}} R_j / (R_i^{\mathrm{T}} R_i + R_j^{\mathrm{T}} R_j - R_i^{\mathrm{T}} R_j) \tag{6}$$

式(6)[8]是 Tanimoto coefficient 提出的,计算两个向量之间的角度,其中R_i、R_j表示两个用户 i、j 访问 n 维资源空间的访问量向量。

根据以上定义,下面提出改进的相似度算法,计算公式为

$$\mathrm{sim}(i,j) = w \times R_{ij}/\mathrm{e}^{\mathrm{Dis}(i,j)} + (1-w) \times S(i,j) \quad 0 < w < 1 \qquad (7)$$

在本文中 w 取值为 0.4。下面举两个例子来说明此算法的优越性,取向量 $x=(0,4,0,0)$,$y=(0,1,0,0)$,$z=(0,2,0,0)$,通过文献[5]中余弦的相似度公式可以看出这三个平行向量之间没有区别,但是利用本文提出的式(7)计算出 x、y 之间的相似度为 $\mathrm{sim}(x,y)=w/3\times\mathrm{e}^3+(1-w)\times 4/13$,$x$、$z$ 之间的相似度为 $\mathrm{sim}(x,y)=w/3\times\mathrm{e}^2+(1-w)\times 8/12$,明显可以看出 $\mathrm{sim}(x,z>\mathrm{sim}(x,y)$。

又例如:$x=(0,7,7,0)$,$y=(0,7,2,0)$,$z=(0,7,4,1)$,由于 x 向量的特殊性,$\mathrm{sim}(x,y)$、$\mathrm{sim}(x,z)$不能利用文献[6,7]中的基于相关系数的相似度公式和调整的余弦相似度公式计算,但可以用本文的相似度公式计算,而且本文的相似度计算公式还可以通过调整 w 值来适应不同的情况。因此根据改进的相似度公式可以获得相似的用户,从而实现用户分类。

2.3 DOBUM 更新过程

在用户不断访问系统的过程中,用户模型在不断的变化,由于服务器的容量有限,需要采取一定的算法对用户模型进行维护,对那些不经常访问的资源进行遗忘,对那些用户感兴趣的但不存在此模型中的资源进行扩展。本文采用文献[9]中的 LRU 算法,更新过程如图 2 所示。

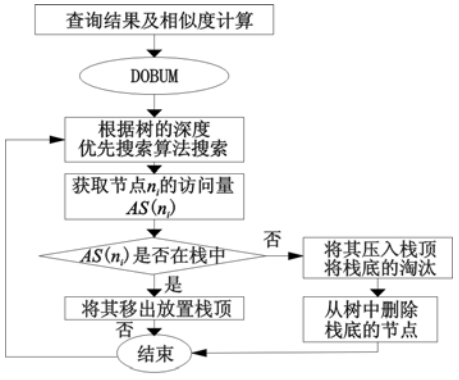
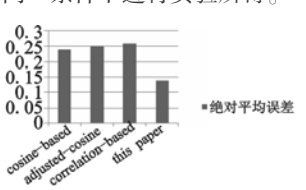


图2 LRU更新算法

3 实验结果与分析

本文以齐齐哈尔货车知识管理系统为实践背景,以提高查询非结构化文档的检索效率。在实验中,本文选择了 20 位用户测试了 1 个月,选择了其中 2 051 条记录进行计算和分析。本文选择绝对平均误差和覆盖范围作为评估的标准。首先图 3 显示了本文提出的改进的相似度算法比常用的余弦相似度算法、基于相关系数的相似度算法以及调整的余弦相似度算法三种算法在过滤结果中的优势,它明显在绝对平均误差上相对其他算法有改进。

图 4 是四种算法在查准率方面的比较结果。查准率采用公式:准确度 = 正确过滤的资源数/应有过滤的资源数,在上面同一条件下进行实验所得。



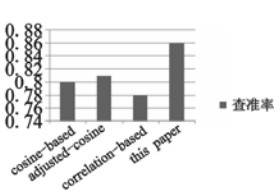


图3 四种算法的绝对平均误差比较 图4 四种算法的查准率的比较

从实验结果可以看出,本文提出的改进的相似度算法具有明显的较高的查准率,而且在过滤中提供给用户的信息误差比其他的三种算法都低。

4 结束语

本文提出的基于领域本体的用户模型相对于传统的用户模型考虑了语义信息,使得在个性化服务系统中能够更好地向用户推荐个性化的信息,提高了用户查询的准确率。本文的用户模型采用本体概念层次数的结构进行存储,并提出改进的相似度算法。通过实验结果证明该算法有明显的改进,在查询准确率上有明显提高,在查询绝对平均误差方面明显比其他三种常用算法降低了。在本文的基础上,下一步力争在查询召回率方面进行研究,提高个性化服务系统的个性化推荐效率。

参考文献:

[1] KURKI T, SULONEN R. Agents in delivering personalized content based on semantic metadata[C]//Proc of 1999 AAA I Spring Symposium Workshop on Intelligent agents in Cyberspace. Stanford: AAA I Press,1999: 84-93.

[2] PRETSCHNER A, GAUCH S. Ontology based personalized search [C]//Proc of the 11th IEEE International Conference on Tools with Artificial Intelligence. Washington: DC IEEE Computer Society, 1999: 391-398.

[3] HECHMANN D, BRANDHERM B, SCHMITZ M, *et al*. GUMO: The general user model ontology[C]//Proc of International Conference on User Modeling. Berlin: Springer-Verlay,2005: 428-432.

[4] 张付志, 李伟静, 朱彩云. 基于领域本体的跨系统个性化服务用户模型[J]. 计算机工程,2009, 35(13):31-33.

[5] SARWAR B, KARYPISARYPIS G, KONSTAN J, *et al*. Item-based collaborative filtering recommendation algorithms [C]//Proc of the 10th International World Wide Web Conference. New York: ACM Press,2001:285-295.

[6] DESHPANDE M, KARUPIS G. Item-based Top-N recommendation algorithms[C]//Proc of ACM Trans on Information Systems. New York: ACM Press,2004:143-177.

[7] KARYPIS G. Evaluation of item-based top-n recommendation algorithms[C]//Proc of the 10th International Conference on Information and Knowledge Management. New York: ACM Press, 2001: 247-254.

[8] LI Jin-zhong. Patter recognition introduction[D]. Beijing: Higher Education Press, 1994.

[9] 陈媛,苟光磊. 个性化服务用户模型研究[J]. 计算机工程与设计, 2008,29(9):2413-2416.

[10] LI Xiao-jian, CHEN Shi-hong. Research on personalized user model based on semantic mining from educational[C]//Proc of International Joint Conference on Artificial Intelligence. Washington DC: IEEE Computer Society,2009:589-594.

[11] 尹红利,王新军. 基于本体的个性化信息检索系统模型研究[D]. 济南:山东大学,2006.

[12] 于强,周良,丁秋林. 基于用户浏览行为的用户模型调整算法研究[J]. 计算机与数字工程,2010,38(11):122-126.

[13] ZHANG Xin, LIU Wei-long. Improving recommender system by using ontology[C]//Proc of the 5th Wuhan International Conference on E-Business. 2006:1266-1271.