计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
44(7):1151~1159, 2007

一种基于本体的个性化模式库建模方法

陈俊杰[1]　刘　炜[2]
(太原理工大学计算机与软件学院　太原　030024)
(chenjj@tyut.edu.cn)

A Modeling Method of User Profile Based on Ontology

Chen Junjie[1] and Liu Wei[2]
(*College of Computer and Software, Taiyuan University of Technology, Taiyuan* 030024)

Abstract The machine-made trait of search engines brings great trouble to people when they retrieve information. Therefore, the technology of user profile is introduced to solve the problem which makes search engine fit for people's demands of individuation and intelligence better. A modeling method of user profile based on ontology is advanced. And the innovation of this paper is to do modeling with a method of combining tree graphics and spatial graphics together, to set up ontology nodes in spatial graphics and to introduce the theory of interval valued fuzzy sets. In addition, the paper explains the method in theory, brings forward some correlative definitions and formulae, and designs an algorithm of founding and updating nodes based on ontology (LW-FUNO). The modeling method is helpful in overcoming the deficiencies of traditional modeling methods and is in favor of user profile's foundation, use and perfection. Finally, a well-ordered implement is used to prove correctness of the algorithm strictly, and then, time complexity of the algorithm, which is $T(n)=o(n^2)$, is analyzed. It is proved theoretically that the algorithm has the traits of correctness, validity and low time complexity. It is considered that the work in this paper is a useful attempt at the research of user profile.

Key words information retrieval; search engine; ontology; user profile; space figure

摘　要　搜索引擎的"千人一面"为人们信息检索时带来了很大的烦恼,个性化模式库技术的引入解决了这个问题,使得搜索引擎能够很好地满足人们的个性化、智能化需求. 提出一种基于本体的个性化模式库建模方法,通过树图和空间图相结合的方法来建模,在空间图中建立本体节点,并引入区间值模糊集理论,同时给出相关定义和公式,在对该方法进行理论分析的基础上,设计了一个实现算法. 这种建模方法对改进传统树形建模的不足有一定益处,更利于建立、使用和完善用户个性化模式. 理论分析证明,该算法具有正确性、有效性并且复杂性低的特点.

关键词　信息检索;搜索引擎;本体论;个性化模式库;空间图

中图法分类号　TP391

随着 Internet 及其相关技术的发展与成熟,人们已经进入信息量极大丰富的时代,Internet 已经成为人们获取所需资源和信息交流的主要场所. 怎样从这些浩如烟海的网页中快速、准确地找到自己需要的信息,就成了人们最关心的问题.

搜索引擎在帮助人们从网上获取所需信息中起到了重要的作用. 但是随着时间的推移,人们越来越发现已有的搜索引擎很难满足用户的个性化、智能化

收稿日期:2006-08-29;修回日期:2007-04-06
基金项目:国家自然科学基金项目(60472093);教育部科学技术研究重点项目(03020);山西省自然科学基金项目(20031038)

需求,这带来一系列的问题.比如不同领域的用户使用同样的关键词查询会得到同样的结果;有时用户会对欲查询内容没有明确概念,搜索引擎返回结果无的放矢,更加重用户的茫然.诸如此类的问题需要迫切解决,用户个性化模式库的研究应运而生成为当前热点.本文在以前工作的基础上参阅大量文献,提出一种用户个性化模式库的建立方法,在对该方法进行理论分析的基础上,设计一个算法来实现.本文所做工作对用户个性化模式库(以下简称模式库)的研究做了一些有益的尝试.

1 相关工作比较

文献[1-3]总结了当前流行的一些用户建模方法.如文献[1]介绍了多种用户建模方法,包括基于评价的、基于内容的和基于知识表示的方法,可以利用时间衰减函数、皮尔森相关算法、信息论方法、数据挖掘等来学习用户兴趣并建模.文献[2]介绍了个性化模式提取方法,包括回答问题法、日志交互法(可使用聚类、关联分析、神经网络等方法,分析日志、cookies、CGI、收藏夹等内容)、本体论引导的方法以及文档词频法(向量法).文献[3]介绍了用户模型的表示(基于向量空间模型的表示、基于神经网络的表示、基于用户-项目评价矩阵的表示、基于案例的表示、基于本体论的表示);用户模型的学习(TF-IDF 技术、贝叶斯分类器、决策树归纳、神经网络技术、聚类技术);用户模型的更新(信息增补技术、遗传算法、神经网络技术).

文献[4-8]采用文档向量法建模.如文献[4]采用向量空间模型表示 Web 页面和用户兴趣;利用模糊集合理论,建立一组模糊 IF-THEN 规则来建立用户模型;采用 Candidate/Rank 模式和 Widrow-Hoff 算法在线式学习对用户模型参数进行优化.文献[5]针对用户模型构造问题,结合手工定制建模与自动分析建模技术,采用空间向量模型表示法,提出了一种用户模型智能调整算法.包括:针对机器学习的新关键词处理;机器学习所得已包含在用户模型的关键词处理;对于用户模型中长期未被使用的关键词处理等内容.文献[6]建立、更新用户兴趣模型是在传统词频法基础上加入词条新鲜度,将用户兴趣用三元组表示(兴趣词条,兴趣权重,词条新鲜度).文献[7]确定学科领域类关键词分布和采用基于关键词提取的文档向量法获得用户个性特征.文献[8]采用 TF-IDF 文本向量表示法表示 HTML 文档,采用机器学习中的文本学习法进行用户模型的创建和更新.

文献[9-11]采用基于本体论的方法建模.如文献[9]使用本体论方法构造用户个性化模型,提出一套本体论标注方法并用于用户模型学习算法和兴趣度计算算法中;采用遗传算法对兴趣度变化量计算问题进行描述和表示.文献[10]采用基于 ontology 的用户描述文件表达方式和自动隐式学习算法来建立用户模型.文献[11]从抽象层次上论述抽取用户模式和日志来建立本体库的方法,用 XML 表示,使用 PART-A 和 IS-A 关系建立一系列公式并根据相关主题单位时间内出现次数来建立.

还有一些文献用另外的方法来建模.如文献[12]对建立多张表的方法进行用户建模做了理论研究.文献[13]采用 k-modes 聚类算法生成个性化模式.文献[14]对基于 Huffman 树形式的个性化模式表示方式做了研究.

2 构建模式库

笔者发现参阅的文献大多讲述的是模式库中模式的增、删、改,用到的方法最多的是向量特征法以及本体论法等,或两者的结合,还有一些采用其他的方法,如霍夫曼树等.另一个趋向是从抽象层次上论述,较多的用本体论的方法讨论模式库的建立.但是普遍来说,这些文献都是以树图的形式来建立模式库的.笔者认为树的形式简单明了,便于构建,对增、删、改有利,但是现实世界中的事物很少以这种方式表达,而且若以树图的方式存储,内容增加到一定程度后必然导致结构庞大,相关算法的时间和空间复杂度增大,这也正是为什么使用这些方法时,树图的大小必须限制在一定范围内,超过就必须剪枝的原因.根据以上分析,结合笔者的另一篇论文(《Research and Implementation of a User Interface Agent Module》),提出一种基于本体的树图和空间图两种层次表示的模式库建立方法.假设经过如下一些方法,已经获得了建立修改模式库需要的数据.这些方法包括从 HTML 或 XML 文档中根据权重的不同提取关键词向量,对网页超链接(Web 结构挖掘)、用户上网行为进行分析.这些行为包括使用本系统[15]或别的搜索引擎进行检索、浏览网页、收发邮件、使用论坛、浏览图片、在线观看、下载等(可通过使用 COM 技术,用活动模板库 ATL 实现与 IE 交互的 DLL 来实现[9]).

树图的节点是关键词,节点间用无向边连接,根节点是抽象节点,从上往下分叉展开,表示了现实世界中事物的从属关系. 模式库中的模式以空间图的方式表达,可分层次分主题建立若干张空间图. 图中节点采用本体形式表示,节点是通过对已获得的数据聚类而来. 每个节点是一个 N 元组,节点间由带权值的有向边相连. 允许节点间不相连,但不允许孤立节点的存在. 使用一些已存在的本体库(如 WordNet 等)作为依托,每个节点都有表示同义、反义、上位、下位等关系的多个属性. 这些节点是否在一个空间图中出现则视这个空间图表达的内容和完成的工作而定,这对下一步的智能推理工作很重要. 系统工作时,首先从树图中查找到某次查询或推理时需要的关键词,确定它所在的范围,然后转到相应的空间图. 这可使系统从整体上起到一个提纲挈领的作用,便于和真实世界保持一致. 针对树结构,笔者已在上一篇文章中详细论述,本文将专注于基于本体论的空间图的表示.

2.1 空间模式本体的建立

对上一节内容进一步深化,得出空间图的建模方法. 设想在用户还没使用前,系统根据树图中的较高层次的分类,分别建立若干领域的本体空间图,如体育、科技、艺术、建筑、医学等领域. 当然这些空间图初始时是比较简单的,它的作用一是体现了本系统是个通用型搜索引擎;二是在用户还没使用前,系统还没有用户个性化信息时,为用户刚开始使用的那段时间提供便利;三是这是系统的初始化,随着时间的推移,待系统获得足够多的使用信息时,可对这些空间图做进一步的更新,以适应用户的个性化需求.

在建模时,首先对已经获得的建立修改模式库需要的数据进行聚类,建立本体,然后在初始的或已有的某个空间图中找到相应的位置,进行修改,获得新的本体空间图. 随着用户的不断使用和系统对用户上网行为的不断分析,将会发现用户的若干兴趣剖面,这些剖面有些是长期的,有些是短期的. 将长期的涉及本体节点较多的那些节点分出,单独建立一个新的本体空间图. 短期的则穿插在已有的空间图中,一段时间不使用即可删除. 可对每个节点间的有向边设置时间记数,据此判断长期还是短期.

另外,设置一个节点阈值判断某节点是否需取出建立新的空间图,阈值可以是一个数值,也可以是一个百分数. 如果当前空间图某部分的节点在某段时间内增长较快较多,增长数超过阈值,或新增节点数与老图节点数相比超过某个百分比,则分出建立新的空间图. 是否分出和图中的总节点数有关,其遵循两原则:一是使节点数保持在一个范围,以使计算时间能保持在一个可接受的范围内;二是按兴趣剖面,特别是长期兴趣剖面划分成独立空间图,便于查找、推理,以及与用户的交互修改. 节点间边是有向的,是可变化的,边的方向代表节点间的逻辑指向或用户浏览方向. 边上设置时间记数,逐渐衰减,有刺激会增强,减到阈值下该边将被取消. 长期兴趣剖面的阈值应小于短期兴趣剖面的阈值,以使长期兴趣边的保留时间长于短期兴趣边. 节点用本体表示,含有的属性中也许有在另一个空间图的节点,这样就建立了不同空间图之间的联系. 需要时可将多个空间图合并成一个.

2.2 节点的本体表示

定义 1. 节点的本体论表示为一个五元组:

$$ON(M, Do, R, AT, IS),$$

其中,M:节点的元信息描述.

Do:节点的域,可包括多个不同的子域,$Do = \{d_i \mid \exists i, j, 1 \leqslant i \leqslant n, 0 \leqslant j \leqslant n, d_i \cap d_j = \varnothing, n$ 为域中子域的个数$\}$.

R:关系,是一个二元组$\langle R_r, R_w \rangle$,R_r 是本体的关系,包括上位、下位、同义、反义、多义等;R_w 是依托的本体库中的和这些关系相对应的词,这些词具有可扩充性.

AT:节点的属性.

IS:节点的实例.

2.3 定义与公式

首先给出域、子域、空间、子空间和从子域到子空间映射的定义.

定义 2. 设 D 为域,$D = \{d \mid d \in C\}$,C 为概念.

定义 3. 若 $D' \subseteq D$,则称 D' 是 D 的子域.

定义 4. 设 G 为拓扑空间,即设 G 是不空的集合,如果 G 中的某些子集被定义为开集,且满足下列条件:

1) G 与空集是开集;

2) 任意多个开集的并是开集;

3) 有限个开集的交是开集,

则这些开集称为 G 上的拓扑结构,G 为拓扑空间.

定义 5. 若 $G' \subseteq G$,则称 G' 是 G 的拓扑子空间.

定义 6. 设存在 ξ,$\xi: D \rightarrow G$,则称 ξ 是从 D 到 G 的一个映射.

公式 1. 边的时间衰减公式(LW-FTDE):

边的时间衰减公式考虑和两个因素有关,一个是从建立到当前过去的自然时间,另一个是这段时间中的对该边的激励,即每一次的使用起到一个增强的作用.

设 t 表示从建立到当前过去的自然时间,以 min 为单位. 这是因为考虑到用户的兴趣会有长期中期与短期之分,长期兴趣可以以天为单位,中期兴趣可以以 h 为单位,而短期兴趣特别是只查询一两次的兴趣则应该以 min 为单位,结合起来考虑,以 min 为单位比较合适. 但是这样对中长期兴趣会使得数值很大,故采用对数的形式表达.$\ln(t+1)$是一个离散函数,$t=0,1,2,\cdots,\ln(t+1)\in[0,+\infty)$. 随着 t 的增加,w 的值会逐渐减少.

设 s 表示激励,$s\in\mathbb{N}$ ($\mathbb{N}$ 为自然数),每激励一次,w 的值都会增加.

设 w 是边的时间衰减权值,是关于 t 和 s 的函数,则有 $w\propto 1/t$,$w\propto s$.

$$W=\begin{cases}e^{-\frac{\ln(t+1)}{k\times s}}, & t\neq 0,\\ 1, & t=0.\end{cases}\tag{1}$$

$w\in(0,1]$,k 是系数,$k>1$,w 采用以 1/e 为底的自然对数形式表示是要使结果落在(0,1)上.

公式 2. 聚类中心与节点的区间值模糊集映射公式.

给出公式 2 之前,先简单介绍一下区间值模糊集理论. 人们已经发现,很多信息中都存在着不确定性,模糊方法是目前处理信息不确定性的主要方法之一,处理模糊信息时常常需要使用区间来表示. 本文引入区间值及区间值模糊集的概念:

定义 7. 用 I 表示单位闭区间[0,1],称包含于闭区间[0,1]的闭区间 $\bar{a}=[a^-,a^+]$ 为区间值,I 上的区间值全体记为 $[I]$,即 $[I]=\{[a^-,a^+]\mid a^-\leqslant a^+,a^-,a^+\in I\}$.

定义 8. 设 X 是一非空普通集合,称映射 $A:X\to[I]$,$x\to[A^-(x),A^+(x)]$ 为 X 上的区间值模糊集,X 上所有的区间值模糊集记为 $IF(X)$.

考虑本文所提问题,假设获得用户行为的有用数据并已形成聚类,因为聚类问题很少能获得一个精确的值,这也是不符合现实世界的,所以这个聚类结果中心可以用一个区间值模糊集来表示:$A(x)=[A^-(x),A^+(x)]$,$x\in G$. 同理,本文设节点也用一个区间值模糊集来表示:$B(x_i)=[B^-(x_i),B^+(x_i)]$,$x_i\in G$,$i$ 代表该空间图中的某个节点.

文献[16]指出,对于两个区间值模糊集所表示的模糊区域,其间相应的拓扑关系程度也为区间值. 所以,聚类结果中心与节点的比较问题可以用两个模糊区域的相交程度来解决,给出下面公式:

$$P_{c_i}=[\bigvee_{x\in G}\{A^-(x)\wedge B_i^-(x)\},\ \bigvee_{x\in G}\{A^+(x)\wedge B_i^+(x)\}],\tag{2}$$

$P_{c_i}\in[I]$,如果 $P_{c_i}=[1,1]$,则表示两个模糊区域一定相交 如果 $P_{c_i}=[0,0]$,则表示两个模糊区域完全不相交.

公式 3. 设置新边判定公式(LW-FNES):

根据式(2)所得的值,来设计式(3). 设 $P_{c_i}=[p_i^-,p_i^+]$,因为式(3)是确定是否在新老节点间设置新边的公式,它需要得出一个数值,与预先给定的阈值 β 比较来决定是否设边,所以考虑 S_i 和两个因素有关:一个是 $0<\frac{p_i^++p_i^-}{2}<1$,另一个是 $0<p_i^+-p_i^-<1$,当 $p_i^+\neq p_i^-$ 时,有 S_i 与 $\frac{p_i^++p_i^-}{2}$ 同方向变化,S_i 与 $(p_i^+-p_i^-)$ 反方向变化. 考虑到 $\frac{p_i^++p_i^-}{2(p_i^+-p_i^-)}\in\left(\frac{1}{2},+\infty\right)$,为将此区间映射到一个有限集合上,取它的反正切函数 $\arctan\frac{p_i^++p_i^-}{2(p_i^+-p_i^-)}$,再将其映射到 $\left(\frac{2}{\pi}\times\arctan\frac{1}{2},1\right)$ 上. 另外再考虑当 $p_i^+=p_i^-$ 时的情况,则可以得到下列公式:

$$S_i=\begin{cases}\frac{2}{\pi}\times\arctan\frac{p_i^++p_i^-}{2(p_i^+-p_i^-)}, & p_i^+\neq p_i^-,\\ p_i^+=p_i^-, & p_i^+=p_i^-\end{cases}\tag{3}$$

其中 $S_i\in I$.

公式 4. 新边权值公式(LW-FWNE):

据式(3)求出的 S_i 判定是否设置新边,并对设置的新边赋予权值. f_i 考虑与两个因素有关:S_i 与 W_i. 并且 f_i 与 S_i,W_i 同方向变化. 故设计下列公式:

$$f_i=\mu\times S_i-l/W_i,\ f_i\in(0,1),\tag{4}$$

其中,l 是一个系数,$0<l<W_i<1$,$l\in\mathbb{R}$,l 的选取使得 l/W_i 保持在(0,1)间. μ 也是一个系数,$\mu>0$,$\mu\in\mathbb{R}$. μ 和 l 的选取使得 f_i 保持在(0,1)间.

3 本体表示的节点的建立与更新算法

根据前文所做的工作,可以得出本体表示的节点的建立与更新算法. 假设已获得用户行为的有用

数据,再按照某个聚类算法,形成聚类. 这个聚类结果中心和树图中的关键词比较,确定属于哪个子域,然后找到相应的空间图,分别与每个节点相比较,并按照本体论的区间值模糊集映射到闭区间[0,1]上(式(2)). 按照所得值来看,若超过某个阈值则可认为它属于这个节点,将其加入到该节点(比如所得值为0.95,超过阈值,则可加入到相对应的节点,同时修改老节点的值);若对不止一个节点的阈值都超过,则都加入并修改它们;若都没有超过阈值,则将其作为一个新的节点加入该空间图中,同时根据获得的数据和定义的节点生成该节点. 计算所得的数值则作为生成新节点和各个老节点新边的依据和权值(式(3))(超过某个阈值就设置边,根据两个节点的上下位关系设置边的方向),再在新边上设置权值(式(4)).

修改过的老节点也需要修改已有的边,也根据前边计算所得的数值经过一些计算,修改老边的权值. 如果有两个及以上的老节点被修改,则检查这些老节点间是否有边相连,若有则修改这些老边的权值,若没有则建立新边,这里按照修改后的老节点的值来重新计算;若只有一个老节点被修改,则只对该节点已有的那些边进行修改. 超过某个阈值就建新边,小于某个阈值就删除该边,同时修改边的时间计数. 这样就完成节点的表示与更新算法. 这里要注意的是,老节点结合了聚类结果,老节点被修改了边也会发生变化,但变化必然不大(否则就建立新节点了),所以不必按新节点重新来计算,那样代价将会变大.

3.1 算 法

算法1. 本体表示的节点的建立与更新算法(LW-FUNO).

输入:已获得的用户行为的有用数据 D_t,树图 T_r,空间图 G_r;

输出:修改过的空间图 G_r';

① 将 D_t 按照模糊聚类算法,形成模糊聚类,生成模糊聚类中心 D_c;
② 将 D_c 与 T_r 中的叶节点比较,确定所属子域;
③ 找到相应空间图 G_r;
④ for all i :$i\in[1..n]$ do /* G_r 有 n 个节点,D_c 与 G_r 中的每个节点进行比较 */
⑤ 比较 D_c 与 G_{r_i};/* G_{r_i} 代表某个节点 */
⑥ 使用式(2),将比较结果按照区间值模糊集理论映射到[I]上;
/* 下面将式(2)所得结果 P_{c_i} 与阈值 α 相比较判定是生成新节点还是修改老节点 */
⑦ if $P_{c_i}\geqslant\alpha$ then /* 超过阈值 α,修改老节点 */
⑧ if 只有一个节点使得 $P_{c_i}\geqslant\alpha$ then
⑨ 将 D_c 加入 G_{r_i} 并修改 G_{r_i} 的值 /* 超过阈值 α 可认为 D_c 属于 G_{r_i},将其加入到该节点 */
⑩ *OldSideOne*
⑪ end if
⑫ if 不只有一个节点使得 $P_{c_i}\geqslant\alpha$ then
⑬ 设 $\exists i_1,i_2,\cdots$,且 $i_1\neq i_2\neq\cdots$,对应节点为 G_{ri_1}, G_{ri_2},…
⑭ 将 D_c 加入 G_{ri_1},G_{ri_2},…,并修改 G_{ri_1},G_{ri_2},…的值;
⑮ *OldSideNotOne*
⑯ end if
⑰ else /* 没有超过阈值 α,生成新节点 */
⑱ 将 D_c 作为一个新的节点加入该空间图 G_r 中;
⑲ 根据获得的数据 D_t 和节点的定义生成该节点;
⑳ *NewSide*;
㉑ end if
㉒ end for
/* 下面根据式(3)在新节点和老节点间设置新边 */
㉓ *NewSide*
㉔ 使用式(3)与预先给定的阈值 β 比较;
㉕ if $S_i\geqslant\beta$ then
㉖ 在新节点与该老节点间设置新边;
㉗ 使用式(4)得出新边权值并赋值给新边;
㉘ 根据两个节点的上下位关系设置边的方向;
㉙ end if
/* 下面对老边进行修改与更新 */
㉚ *OldSideNotOne*
㉛ if G_r 中有两个及以上的老节点 G_{ri_1}, G_{ri_2},…被修改 then
㉜ 检查这些节点间有无边:
㉝ if 有边 then
㉞ 修改老边的权值,使用式(1),$s:=s+1$;
㉟ 再使用式(4)得出修改后的边的权值;
㊱ else
㊲ 建立新边;

㊳ 使用式(2)按照修改后的老节点重新计算 P_{c_i};

㊴ 再使用式(1)(3)(4)计算新边的权值;

㊵ end if

㊶ end if

㊷ *OldSideOne*

㊸ if G_r 中只有一个老节点 G_{r_i} 被修改 then

㊹ 对 G_{r_i} 已有的那些边进行修改;

㊺ 使用式(1),$s := s+1$,再使用式(4)得出修改后的边的权值;

㊻ end if

㊼ Return G'_r.

3.2 算法分析

算法具有5条重要的特性:

1) 输入数据. 每个算法都应当有0个或多个输入.

2) 输出数据. 每个算法都应当有1个或多个输出(即算法必须得到结果).

3) 确定性. 指算法中的每一个步骤都应当是确定的.

4) 有穷性. 一个算法必须在有限的步骤内结束.

5) 有效性. 算法的每个步骤都应当能有效执行,并能得到确定的结果.

满足了上述5条特性的算法就是一个好的、正确的算法. 本文提出的本体表示的节点的建立与更新算法(LW-FUNO)具有输入数据,即已获得的用户行为的有用数据 D_t、树图 T_r 和空间图 G_r;也具有输出数据,即修改过的空间图 G'_r. 下面对该算法从正确性、有穷性及有效性上分别加以分析.

3.2.1 算法正确性分析

算法的正确性是指算法的每一步必须是有确定意义的. 周培德教授指出,若一个算法满足良序原则,则该算法是正确的(详见文献[17-18]).

定义9. 良序定义. 设$\prec$是集合 S 上的一个关系,并且满足以下性质:

1) 给定 S 中的 X,Y,Z,如果 $X \prec Y$(称 X 先于 Y),$Y \prec Z$,则有 $X \prec Z$;

2) 给定 S 中的 X,Y,以下3种可能性中有且只有一种为真:

$$X \prec Y;X = Y;Y \prec X;$$

3) 如果 A 是 S 中任意一个非空子集,则 A 中必有一个元素 X,使得对于 A 中的所有 Y,都有 $X \prec Y$ 成立,

则称$\prec$是集合 S 上的一个良序.

这3条性质为下文叙述的方便,可以分别称其为良序的传递性、惟一性及非空性.

定理1. 若一个良序的子句集 G 能够推导出 $X_1 \prec X_n$,也即 $X_1 \rightarrow X_n$,则这种推导过程可以表示为 $G \cup \{X_1, \sim X_n\}$ 是一个不可满足的子句集.

定理2. 设 P 是算法的开始语句,Q 是算法的结束语句,若一个算法是正确的,则其子句集 G 能够推导出 $P \rightarrow Q$.

推论1. 若一个算法是正确的,则其子句集 $G \cup \{P, \sim Q\}$ 是不可满足的子句集.

下面对本文提出的LW-FUNO算法构建子句集 G,并对该算法语句进行分析:

1) 语句①为算法的开始,记为 P;

2) 语句②③是顺序执行关系,分别记为 A_1,A_2;

3) 语句④~㉒是一个循环结构,其中语句⑤⑥是顺序结构,分别记为 A_3,A_4,该循环结构内部含有嵌套分支选择结构,即 if(超过阈值 α,修改老节点(语句⑧~⑯),嵌套了 if 只有一个节点使得 $P_{c_i} \geqslant \alpha$,记为 A_5(语句⑧~⑪),并包含一个对 *OldSideOne* 的调用;及 if 不只有一个节点使得 $P_{c_i} \geqslant \alpha$,记为 A_6(语句⑫~⑯),并包含一个对 *OldSideNotOne* 的调用),else(没有超过阈值 α,生成新节点(语句⑱~⑳),记为 A_7,并包含一个对 *NewSide* 的调用);

4) 语句㉓~㉙是被调用的 *NewSide*,其中语句㉔与后面的语句是顺序结构,记为 A_8,语句㉕~㉙是一个 if 语句,记为 A_9;

5) 语句㉚~㊶是被调用的 *OldSideNotOne*,这又是一个嵌套分支选择结构,其中语句㉜与以后的语句是顺序结构关系,记为 A_{10},然后 if 有边,记为 A_{11}(语句㉞㉟),else 建立新边,记为 A_{12}(语句㊲~㊴);

6) 语句㊷~㊻是被调用的 *OldSideOne*,这是一个 if 结构,记为 A_{13};

7) 语句㊼是结束语句,记为 Q.

由以上对本算法语句的分析,可以证明该算法是正确的.

证明. 本算法子句集

$$G = \{(P \rightarrow A_1) \wedge (A_1 \rightarrow A_2), A_2 \rightarrow A_3, A_3 \rightarrow A_4, \bigvee_{i=5}^{7}(A_4 \rightarrow A_i), A_5 \rightarrow A_{13}, A_6 \rightarrow A_{10},$$

$$\bigvee_{i=11}^{12}(A_{10}\rightarrow A_i),(A_7\rightarrow A_8)\wedge(A_8\rightarrow A_9),$$

$$\bigvee_{i=9,11}^{13}(A_i\rightarrow Q)\}=$$

$$\{\sim P\vee A_1,\sim A_1\vee A_2,\sim A_2\vee A_3,$$

$$\sim A_3\vee A_4,\sim A_4\vee A_5,\sim A_4\vee A_6,$$

$$\sim A_4\vee A_7,\sim A_5\vee A_{13},\sim A_6\vee A_{10},$$

$$\sim A_{10}\vee A_{11},\sim A_{10}\vee A_{12},\sim A_7\vee A_8,$$

$$\sim A_8\vee A_9,\sim A_9\vee Q,\sim A_{11}\vee Q,$$

$$\sim A_{12}\vee Q,\sim A_{13}\vee Q\},$$

则由良序的传递性(定义9)及定理1,可知子句集 G 中有 $(P\rightarrow Q)$,即这种推导过程可以表示为 $G\cup\{P,\sim Q\}$ 是一个不可满足的子句集,则再由定理2及推论1可知本算法是正确的. 证毕.

3.2.2 算法复杂性分析

算法的"有穷性"指"在合理的范围之内"的有限步骤. 如果让计算机执行一个历时千年才结束的算法,算法尽管有穷,但超过了合理的限度,该算法也是无用的. 因此,算法"有穷性"的体现主要就是算法复杂度,包括时间复杂度和空间复杂度. 由于硬件技术的发展,现代算法的空间问题对计算机的要求不高,一般较少讨论,故算法复杂度主要体现在它的时间复杂度上,本文将对本体表示的节点的建立与更新算法的时间复杂度进行分析.

设 $L_0=\max(|D_j|),j=1,2,\cdots,r,D_j$ 表示某次收集到的个性化信息模糊聚类生成 D_c 并与树图 T_r 比较确定所属子域,并调出相应空间图 G_r 这个过程的一次执行时间;设 L_1 为修改一个老节点所需的时间;设 L_2 为建立一个新节点所需的时间;设 L_3 表示为新节点设置一个新边所需的时间(包括建边、赋权值及设置方向);设 L_4 表示对被修改的老节点的已有老边的修改所需的时间;设 L_5 表示在被修改的两个老节点间建立新边所需的时间.

对本算法进行分析,发现它的执行可以分为两步:第1步是对空间图前处理,即 D_j;第2步是对空间图的处理,包括对节点及边的处理,即 L_1,L_2,L_3,L_4,L_5,这5种处理对一次 D_j 的本算法执行不会同时出现,而可能会出现2或3种. 因此如果对这5种情况同时出现计算得出的算法复杂度必然大于算法的实际复杂度. 实际上,算法的执行时间除上述这些外,还有比较、判断等时间,由于这些时间与 $L_i(i=1,2,3,4,5)$ 相比很小,这里为方便计算略去不计.

1) G_r 中只有一个老节点被修改时的情况

设 G_r 中有 n 个节点,这种情况包含了空间图前处理、从 G_r 中挑选一个老节点来修改及该老节点与 G_r 中所有其他老节点间所有可能老边的修改时间的最大值,即 $L_0+L_1+(n-1)L_4$.

2) G_r 中不只有一个老节点被修改时的情况

这种情况包含了空间图前处理,从 G_r 中挑选多于一个的老节点来修改,每个老节点与 G_r 中所有其他老节点间所有老边的可能修改及新边的可能设置的最大值(假设与其他所有老节点都修改边及建新边各一次,实际上这两种情况是互质的,因此可以取两者平均值),即 $L_0+n[L_1+\frac{n-1}{2}(L_4+L_5)]$.

3) G_r 中生成新节点时的情况

这种情况包含了空间图前处理、建立一个新节点及该新节点与 G_r 中所有老节点间建立可能新边的时间的最大值,即 $L_0+L_2+nL_3$.

4) 完成算法的一遍执行(既一次建立或修改个性化模式库)的基本操作所需时间小于

$$L_0+L_1+(n-1)L_4+L_0+n[L_1+\frac{n-1}{2}(L_4+L_5)]+$$

$$L_0+L_2+nL_3=3L_0+(n+1)L_1+L_2+nL_3+$$

$$\frac{1}{2}(n^2+n-2)L_4+\frac{1}{2}(n^2-n)L_5.$$

这里的 $L_i(i=1,2,3,4,5)$ 是完成某种基本操作对应的时间,可看做是常数,$L_0=\max(|D_j|),j=1,2,\cdots,r$,为空间图前处理取的上界值(实际 D_j 只有在极小概率下才会取到 L_0),故也可看做是一个常数,所以上式与 n^2 是同阶的,如果记上式为 $T(n)$,则有 $T(n)=O(n^2)$. 又因为 $T(n)$ 是对 $L_i(i=1,2,3,4,5)$ 同时出现时所得出的,实际情况必然小于此,所以本算法的时间复杂度可以表示为 $T(n)=o(n^2)$.

3.2.3 算法有效性分析

一个算法是否有效,从理论上还没有一种方法能够证明[18]. 从它的含义上看,算法有效性是指算法的每个步骤都应当能有效执行,也即算法中描述的操作都是可以通过已经实现的基本运算执行有限次来实现的. 本文对该算法的正确性及复杂度的分析正是基于对其每条语句及每个基本操作而来的,因此对本算法的正确性及有穷性的证明也就间接证明了它的有效性.

上文对本文提出的本体表示的节点的建立与更新算法从一个算法应该具有的5个特性方面分别加以分析证明,由此得出本算法是一个正确的、好的算法的结论.

4 结束语

搜索引擎的发展为人们从网上获取所需信息带来了很大的便利,但是它们的“千人一面”又对人们的检索带来了很多烦恼.随着时代的发展,人们越来越需要一种具有个性化、智能化的搜索引擎,个性化模式库的研究是解决这个问题的一种方法.

本体论自从被引入计算机领域以来,已被探索性地应用到了很多方面.因为本体将词汇间的关系做了深入的表达,同时也由于它的可扩充性和可重用性,在搜索引擎中应用本体具有光明的前景.本文在基于本体的基础上,提出了一套用户个性化模式库的建模方法,独创一系列公式以支持,并设计了一个算法来实现.本文的创新之处在于以树图和空间图相结合的方法对个性化模式库建模,在空间图中建立本体节点,并引入了区间值模糊集理论.相信本文所做工作对用户个性化模式库的研究做了一些有益的尝试.

参 考 文 献

[1] Yang fengqin. Method of building hierarchical user interest model:[Master dissertation][D]. Changchun: Northeast Normal University, 2004 (in Chinese)
(杨凤芹. 建立层次结构用户兴趣模型的方法:[硕士学位论文][D]. 长春:东北师范大学, 2004)

[2] Liu Yanqing, Tian Xuan, Su Guilian. Personalized information retrieval research survey on Internet [J]. Computer Engineering and Design, 2004, 25(5):772-775 (in Chinese)
(刘艳青, 田萱, 苏桂莲. 基于 Internet 的个性化信息检索技术的研究[J]. 计算机工程与设计, 2004, 25(5):772-775)

[3] Wu Lihua, Liu Lu. User profiling for personalized recommending system—A review [J]. Journal of the China Society for Scientific and Technical Information, 2006, 25(1): 55-62 (in Chinese)
(吴丽花, 刘鲁. 个性化推荐系统用户建模技术综述[J]. 情报学报, 2006, 25(1):55-62)

[4] Dai Xuewu. The research on neural-networks-based user modeling and Web information filtering:[Master dissertation][D]. Chongqing:Southwest University, 2004 (in Chinese)
(代学武. 基于神经网络的用户建模和 Web 信息过滤研究:[硕士学位论文][D]. 重庆:西南师范大学, 2004)

[5] Zhao Zhongmeng, Yuan Wei, He Shili, *et al*. Research on the intelligent adjustive algorithm for user profile in personalized search engine [J]. Computer Engineering and Applications, 2005, (24):184-187 (in Chinese)
(赵仲孟, 袁薇, 何世丽, 等. 个性化搜索引擎中用户模型智能调整算法的研究[J]. 计算机工程与应用, 2005, (24):184-187)

[6] Chen Liang, Li Xuemei, Chen Shifu. Design and implementation of a personalizing meta search engine AIPMSE [J]. Application Research of Computers, 2003, (12):136-138, 145 (in Chinese)
(陈亮, 李雪梅, 陈世福. 个性化元搜索引擎 AIPMSE 的设计与实现[J]. 计算机应用研究, 2003, (12):136-138, 145)

[7] Wang Haoming, Zhang Yuexian, Wu Zhijun, *et al*. Research of the Chinese meta-search engine model based on intelligent agent [J]. Computer Engineering and Applications, 2005, (31):154-156, 204 (in Chinese)
(王浩鸣, 张曰贤, 吴志军, 等. 基于智能 Agent 的中文元搜索引擎模型研究[J]. 计算机工程与应用, 2005, (31):154-156, 204)

[8] Ma Yan, Zou Xianchun, Bao Junjie, *et al*. A design of the model of intelligent cell searching engine on Internet [J]. Journal of Chongqing Normal University (Natural Science Edition), 2004, 21(3):15-18 (in Chinese)
(马燕, 邹显春, 包骏杰, 等. 一种互联网智能元搜索引擎模型的设计[J]. 重庆师范大学学报(自然科学版), 2004, 21(3):15-18)

[9] Li Yong. Ontology-based personal user modeling technology and application in the intelligent information retrieval:[Master dissertation][D]. Changsha: National University of Defense Technology, 2002 (in Chinese)
(李勇. 智能检索中基于本体的个性化用户建模技术及应用:[硕士学位论文][D]. 长沙:国防科学技术大学, 2002)

[10] Huang Guojing, Cui Zhiming. Ontology-based personalized meta search engine [J]. Microelectronics & Computer, 2004, 21(12):100-103 (in Chinese)
(黄国景, 崔志明. 基于 Ontology 的个性化元搜索引擎研究[J]. 微电子学与计算机, 2004, 21(12):100-103)

[11] Yuefeng Li, Ning Zhong. Web mining model and its applications for information gathering [J]. Knowledge-Based Systems, 2004, (17):207-217

[12] Li Guangjian. The research and design of personalized Internet information retrieval system:[Ph D dissertation][D]. Beijing: The Graduate University of the Chinese Academy of Sciences, 2002 (in Chinese)
(李广建. 个性化网络信息检索系统的研究与实现:[博士学位论文][D]. 北京:中国科学院研究生院, 2002)

[13] Tian Xuan, Liu Xiyu, Meng Qiang. Realization of personalized intelligent Web retrieval [J]. Computer Engineering and Applications, 2003, (1):195-197 (in Chinese)
(田萱, 刘希玉, 孟强. 实现 Web 页面的智能个性化检索[J]. 计算机工程与应用, 2003, (1):195-197)

[14] Tian Xuan, Meng Xiangguang, Liu Xiyu. Research on representation of user profile in intelligent information retrieval [J]. Journal of the China Society for Scientific and Technical Information, 2004, 23(1):21-26 (in Chinese)

(田萱,孟祥光,刘希玉.智能信息检索中个性化模式的表示形式研究[J].情报学报,2004,23(1):21-26)

[15] Chen Junjie, Liu Wei. A framework for intelligent meta-search engine based on agent [C]. The 3rd Int'l Conf on Information Technology and Applications(ICITA'05), Sydney, 2005

[16] Yu Qiangyuan, Liu Dayou, Ouyang Jihong. Topological relations model of fuzzy regions based on interval valued fuzzy sets [J]. Acta Electronica Sinica, 2005, 33(1):187-189 (in Chinese)
(虞强源,刘大有,欧阳继红.基于区间值模糊集的模糊区域拓扑关系模型[J].电子学报,2005,33(1):187-189)

[17] Zhou Peide. Design and Analysis of Computer Algorithm [M]. Beijing: Machine Industry Publishing House, 1985 (in Chinese)
(周培德.算法设计与分析[M].北京:机械工业出版社,1985)

[18] Du Yajun. Study and implementation on intelligent action of search engine:[Ph D dissertation] [D]. Chengdu: Southwest Jiaotong University, 2005 (in Chinese)
(杜亚军.搜索引擎智能行为的研究及实现:[博士学位论文][D].成都:西南交通大学,2005)

Chen Junjie, born in 1956. He is currently professor and Ph. D. supervisor at the College of Computer and Software Engineering, Taiyuan University of Technology, Taiyuan, Shanxi. He received his Ph. D. degree in computer science from Beijing Institute of Technology in 2003. His main research interests include data mining, search engine and ontology.
陈俊杰,1956 年生,博士,教授,博士生导师,主要研究方向为数据挖掘、搜索引擎和本体.

Liu Wei, born in 1977. Received her bachelor's degree in computer science from China University of Mining and Technology in 2001, Xuzhou, Jiangsu. And now she is a Ph. D. candidate at the College of Computer and Software, Taiyuan University of Technology, Taiyuan, Shanxi. Her current research interests include ontology, meta-search engine and grid.
刘 炜,1977 年生,博士研究生,主要研究方向为本体、元搜索引擎与网格(liuwei_xx@sina.com).

Research Background

The technology of user profile is introduced to solve the problem that the machine-made trait of search engine brings great trouble to people when they retrieve information. This paper advances a modeling method of user profile based on ontology. And it designs a modeling in a method of combining tree graphics and spatial graphics together, sets up ontology nodes in spatial graphics and introduces the theory of interval valued fuzzy sets. Our project is supported by the National Natural Science Foundation of China (60472093), the Grand Science and Technology Research Program of Ministry of Education (03020) and the Natural Science Foundation of Shanxi Province (20031038).