一种基于移动服务器端的树图建模方法

刘 炜 李舟军

(北京航空航天大学计算机学院 北京 100191)

摘 要 移动业务多媒体化和互联网化是移动服务发展的重要方向,但是高数据传输率的多媒体服务成本极高,在不增加硬件投入的前提下,为了降低成本,同时又不降低为用户提供服务的质量,可以将个性化服务应用于移动业务中。提出一种工作在服务器端的树图模型,它将移动终端用户的个性化需求与待推荐的互联网信息资源有机地联系在一起,以组织多播推送。引入区间值模糊集理论,给出建立树图模型的一系列相关定义及公式,在对该方法进行理论分析的基础上,设计了树图节点的建立与更新算法。理论分析证明,该算法具有确定性、能行性并且复杂性低的特点。

关键词 移动业务,个性化服务,区间值模糊集,树图模型,算法

中图法分类号 TP391 **文献标识码** A

Modeling Method of Tree Graph Based on Mobile Server-Side

LIU Wei LI Zhou-jun

(School of Computer Science and Engineering, Beijing University of Aeronautics & Astronautics, Beijing 100191, China)

Abstract Mobile multimedia and mobile Internet are the the important directions of mobile service development. However, it is the high cost with high data transmission rate of wireless multimedia communication service. Under the premise of no increasing the investment in hardware, the personalized service could be applied to mobile service not only reducing wireless multimedia communication cost but also without reducing quality of mobile service for users. It advanced a kind of model on tree graph which works on server-side. It integrated individual requirements of users in mobile terminal with recommended information resources on the Internet together in order to organize multicast push. The paper drew into the theory of interval valued fuzzy sets and explained the method of model on tree graph in theory, brought forward a series of correlative definitions and formulae on founding the model and designed an algorithm on founding and updating the node of the tree graph. Finally, it useed an implement of well-ordered to strictly prove the determinacy of the algorithm, and then, analyzed time complexity of the algorithm. On the theory, the algorithm is proved of the traits of determinacy, validity and low time complexity. It should be considered that the works is a useful attempt in the research fields.

Keywords Mobile service, Personalized service, Interval-valued fuzzy sets, Model on tree graph, Arithmetic

移动业务多媒体化和互联网化是移动服务发展的重要方向。随着移动通信终端处理能力、显示屏幕尺寸和存储能力的飞速增强,用户对通过无线方式直接访问互联网,以获取高服务质量的信息资源的需求日益强烈。然而,无线方式的数据传输速率的提高仍然是个瓶颈,目前,高数据传输率的多媒体服务需要占用大量稀缺的无线频谱资源,成本极高。

1 相关工作

个性化服务预测的质量决定了用户满意度的高低,是个性化服务应用于移动业务成功与否的关键。其中,用户模型的研究是个性化服务的基础与核心[1]。

国外研究者将用户模型引入移动通信中。G. P. Eleftheriadis 等人很早就看到对用户进行建模来提供个性化服务将是移动服务的一个发展方向[2]。德国汉诺威大学为移动用户建立偏好模型,通过移动终端 ID 号和无线信号增强器来捕获用户的运动信息[3]。Giuseppe Araniti 等人将用户个性化模型与 QoS 结合,研究无线网络多媒体的软服务质量机制[4]。加利福尼亚大学研究在 3G 网络中对用户群建立实时模型,为用户群预订资源以使其更好地获得服务质量(QoS)[5]。Spyros Panagiotakis 等人在移动环境下将用户位置信息等引入,以便更好判断用户所处环境[6]。G. Bartolomeo 等人研究为移动终端的用户建立模型,来安全地定制并获得服务[7]。多伦多大学收集手机用户偏好信息,为其提供有针对性的广告[8]。

北京邮电大学研究在移动运营中终端自动选择移动提供服务时的用户建模问题,该用户模型基于马尔可夫判决过程[9]。北京大学在智能手机上实现了一个智能接听系统,它学习用户对不同号码的接听习惯以建立用户个性化模型,从

到稿日期:2010-05-30 返修日期:2010-11-05 本文受国家自然科学基金项目(90718017)资助。

刘 炜(1977—),女,博士后,主要研究方向为智能信息处理,E-mail:mnb6000@buaa.edu.cn;李舟军(1963—),男,博士,教授,CCF 高级会员,主要研究方向为计算机科学理论、数据挖掘和生物信息学。

而针对每次来电自动选择响应方式[10]。作为国家科技支撑计划，浙江大学在普适计算环境下提出用户的智能影子模型，其中采用 BDP(Belief-Desire- Plan)模型对用户进行建模[11]。

国外研究多关注获取用户在移动网络中的环境信息及定制服务等方面内容，国内对此关注时间较短、研究较少，不过已将其作为未来的一个重要研究方向展开了研究。本文在以前工作的基础上[12]，讨论如何将待推荐内容与移动终端用户的个性化需求相匹配，以组织多播推送的问题。

2 树图模型

系统根据获取的互联网信息资源来为(移动)终端用户主动推送符合其需要的信息。终端与服务器端是系统工作的两端，考虑到终端的多样性及降低终端负担的需要，将系统主体设计在服务器端。服务器端包含为每个用户建立的个性化模式、树图、信息资源知识库等部件。移动终端定期向服务器端上报用户浏览信息，服务器端挖掘用户个性化偏好并建立小巧高效的个性化模式，同时网络信息资源获取并经处理后暂存在信息资源知识库中，树图是将用户的个性化需求与待推荐信息资源联系起来的桥梁，是系统的重要功能部件。

系统工作原理为：网络爬虫不断从互联网中挖掘各种信息资源，这些资源经一系列处理后分类暂存到相应的资源知识库中，再根据一定的策略从中分批抽取待推荐的信息资源。同时，系统建立的一个面向所有用户的大型树图依据改进的杜威十进分类法而来，具有规模大、分类全、可自动学习并更新等特征。树图具有两方面重要作用：一是这棵多叉树除根节点是抽象节点外，其余每个节点都代表不同的内容，待推荐的信息资源与之进行匹配查找以作推荐；二是树图中每个节点都包含了对该节点内容感兴趣的用户列表，这里每个用户都有一个唯一标识，一个用户可以出现在多个节点中，树图根据每个用户的个性化模式上报的兴趣点来构建用户列表。树图的学习与更新体现在两方面：一方面从各个用户个性化模式中学习用户兴趣点并更新维护用户列表；另一方面从互联网信息资源学习并扩展节点。这样树图就将互联网信息资源与移动终端用户兴趣点有机地联系在一起了。最后再将具有共同兴趣点的用户群建立用户群模型，依据地理位置等情况组织相应的信息资源进行推送。

树图初始值采用杜威十进分类法并对其作扩展而来。在用户未使用前，系统中会有一个初始化的树图，树图节点用四元组表示，各节点间由具有权重的有向边相连，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。此树图的初始值是依据杜威十进分类法(Dewey Decimal Classification，DDC)而来的，该分类法是世界上最广为采用的分类法，它来源于图书馆学分类，但用来组织网络上的资源分类具有很好的效果，并且还有一大长处，即能够扩充发展，因此本文将其作为树图的初始化来源。

本文对 DDC 做了适当改进。DDC 是用传统的学科来分类，总共以 10 个主要的学科(Main classes)来涵括所有的知识体系，每个大类下细分 10 类(Divisions)，然后又再分成 10 小类(Sections)。它的 10 个大类分别是：总类(Generalities)、心理学与哲学类(Psychology & Philosophy)、宗教类(Religion)、社会科学类(Social sciences)、语文类(Languages)、自然科学类(Pure sciences)、应用科学技术类(Applied science & Tech)、艺术、音乐与体育类(Arts，Music & Sports)、文学类(Literature)以及史地类(General geography & History)。另外，DDC 为每个学科给予一个以十进制特定数字表示的域来表示其范围，每个大类下可以以两层最多达 100 个小类来组成。本系统不使用这种方法，而是将每个都以一个十六位二进制数表示，则每一层的类都有 $2^{16}=65536$ 个，每一层上的类在初化时不需全都分完，这样就为每个用户在使用中因实际情况需要而做的扩充做了准备。二进制表示的另一个好处是使得本系统能够不局限于 DDC 的固定层数表示法，而可以随不同用户的不同使用情况做出灵活修改。

初始化后，随着每个用户的使用，根据模式库中建立的模式及对它们的分类总结，可以对树图进行修改。对用户兴趣域集中的类可以细化至多层，对其不感兴趣、很少访问的类则不做变动。

3 树图节点用户列表的建立与更新

3.1 节点的用户列表

从各个用户个性化模式中学习用户兴趣点并更新维护用户列表。每个节点有一张表，储存有对该节点对应信息感兴趣的用户，并根据用户兴趣程度、时间等因素的不同分别标识。另外，考虑到网络信息资源可能会被重复推送，需要设置标识，以免浪费传送资源。

定义 1(树图节点表示)　设节点由四元组 $TN(M,U_c,U_d,S)$ 表示，其中

M：节点的元信息描述，是一个三元组$\langle M_i,M_n,M_f\rangle$。$M_i$ 是节点的唯一标识(节点号)，M_n 是节点名，M_f 是该节点的父节点号；

U_c：对节点具有稳定的长期兴趣的用户群，是一个二元组$\langle U_{ci},U_{cw}\rangle$。$U_{ci}$ 是用户唯一标识，U_{cw} 是用户对此节点的兴趣权重；

U_d：对节点具有非稳定、非长期兴趣的用户群，如短期兴趣和新兴趣等，是一个二元组$\langle U_{di},U_{dw}\rangle$。$U_{di}$ 是用户唯一标识，U_{dw} 是用户对此节点的兴趣权重；

S：推荐队列，是一个二维 N 元数组，$S=N[S_i][j]$，$(i\in[1..n],j=0,1)$，S_i 是资源标识，j 由 0、1 中的一个表示，标识一定时期内某资源是否被发送过。若是新资源，加在数组尾，如数组已满，则删除数组头元素，余者累进一位，新资源加在数组尾部；若数组中已有该资源，则将 j 改变。

树图的节点的数据结构设计如下：

```
Node-Structure
{
  M{Id: //标识符(节点号)
    Node-name://节点名
    Father-node:} //父节点号
  Uc{〈Uid1://用户标识符
     LInterest-Weight1:〉,//长期兴趣权重
     〈Uid2:
     LInterest-Weight2:〉,…}
  Ud{〈Uid1://用户标识符
     SInterest-Weight1:〉,//短期等兴趣权重
     〈Uid2:
     SInterest-Weight2:〉,…}
  S{〈Rid1: //资源标识符
```

Send-label1:〉,

//资源发送标识,Send-label=0,1

〈Rid2:

Send-label2:〉,…,

〈RidN:

Send-labelN:〉}}

3.2 用户列表的更新

用户列表的更新主要体现在用户对节点表示信息的兴趣权重的更新,分别从长期兴趣和非长期兴趣两方面加以讨论。

1)长期兴趣点权重的更新

对长期兴趣点来说:一是一个用户可能有多个长期兴趣点,且这些兴趣点对该用户的重要程度可能不同,表现在权重上的不同;二是一个用户的所有兴趣点在一个更新周期内不一定都能得到用户响应,由此会造成权重的变化;三是长期兴趣点是相对来说比较稳定的兴趣点,因此可能会对该兴趣点的周边兴趣点产生一定的影响。

用户列表中的兴趣权重的更新要考虑三方面因素:一是各个用户模式上报的权重;二是相邻节点是否也具有该用户的兴趣点;三是树图中该节点及其相邻的节点间有向边的权重。

第一个因素是指在每个周期内,各用户个性化模式将向上层——树图上报该用户在本周期内的长期、非长期兴趣点权重的变化情况,树图由此更新各节点的用户列表中的兴趣权重,以作为系统进行个性化推荐的依据。

第二个因素考虑与该节点相邻的周边节点是否有同一个用户的兴趣点,由于在一个周期内的更新中各节点用户兴趣权重的变化可能会造成互相影响,因此指定这种更新是从根节点开始、从上至下逐层修改,以避免混乱。同时根据相邻节点中用户是否具有相应的兴趣点区别对待。

第三个因素是与相邻节点间有向边的权重,二、三两个因素结合起来作为该节点与相邻节点间关系对兴趣点的影响。对原来不包含某用户的相邻节点,若计算出的权重超过阈值,则可将该用户加入到相邻节点的用户列表中。这是由于相邻节点关系往往比较密切,对某一节点感兴趣的用户也可能会对它的周边节点产生兴趣,因此可作为用户兴趣点信息推荐的扩展。

本文设计长期兴趣点权重更新公式如下:

$$F_{POI}=\frac{2}{\pi}\cdot\arctan\left[\frac{\lambda\cdot\sum_{i=0}^{n}(U+N_i\cdot E_i)+(1-\lambda)\cdot m\cdot U}{(n+m+1)\cdot U}\right] \tag{1}$$

式中,设 F_{POI} 表示某节点的用户列表中某用户的长期兴趣权重的更新值,$F_{POI}\in(0,1]$;λ 是系数,$\lambda\in(0,1]$;设 U 是本周期内该用户个性化模式上报的对应于该节点的兴趣点的权重更新值,$U\in(0,1]$;设 N_i 表示该节点的相邻节点的用户列表存在同一个用户时的长期兴趣权重,$N_i\in(0,1]$;设 E_i 表示该节点与对应的相邻节点间有向边的权重,$E_i\in(0,1]$;设 n 表示该节点的相邻节点的用户列表中存在同一个用户时的相邻节点数,$n\in N$;设 m 表示该节点的相邻节点的用户列表中不包括同一个用户时的相邻节点数,$m\in N$。

式(1)将用户模式上报的兴趣点权重及相邻节点与其相互影响都考虑在内,同时也考虑到这三个因素的重要程度的差异,因此不仅能得到较好的权重,也便于用户兴趣点的扩展推荐。

2)非长期兴趣点权重的更新

非长期兴趣点又可细分为短期兴趣点、新兴趣点等。对长期、非长期兴趣点的识别及划分是用户个性化模式的任务(另文讨论),本文对非长期兴趣点权重的更新主要考虑短期兴趣点的删除满足条件及新兴趣点的变化情况。

短期兴趣点的删除考虑与两个因素有关:一个是用户个性化模式定期上报的权重,另一个是时间。在一定时间后,若用户个性化模式上报的该兴趣点这两个因素值较低,或因用户在本周期内对该兴趣点并未关注从而并未上报,则当最终计算得出的权值低于一定阈值后,该用户将在此兴趣点的用户列表中被删除。

新兴趣点开始会被设定为权重的最高值 1,一定时间后根据用户对其感兴趣的程度,判定它应升级为该用户的长期兴趣点还是短期兴趣点,或者被删除。

设计非长期兴趣点权重更新公式如下:

$$W=\begin{cases}e^{-\frac{\ln(T+1)}{\alpha\cdot S}}, & T\neq 0\\ 1, & T=0\end{cases} \tag{2}$$

式中,设 T 表示兴趣点从建立到当前过去的自然时间,采用对数形式表示以避免数值过大,$\ln(T+1)$ 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 表示用户个性化模式上报的权重,$S\in(0,1]$;设 W 表示该兴趣节点的用户列表中某用户的非长期兴趣权重的更新值,是关于 T 和 S 的函数,则有 $W\propto 1/T$,$W\propto S$,$W\in(0,1]$;α 是系数,$\alpha>1$;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。

4 树图节点的建立与更新算法

系统从获得的互联网信息资源中学习并扩展已有树图中的节点。获得互联网信息资源并经一系列清洗处理后存储在信息资源知识库中(另文讨论),知识库中存储的信息资源和树图均采用相同的分类法并使用同一套标识。树图由于经过不断的更新与细化,使得节点的分类与标识相对知识库中信息资源的分类与标识更扩展、更细致;为便于和树图节点比较,信息资源采用和它具有包容关系的、最相近的、尽可能低层的分类层级来标识,例如"猩猩"这一资源在分类法层级中有"生物"→"动物"→"哺乳动物"→"灵长类哺乳动物"这几层,根据最相近及低层原则,选择"灵长类哺乳动物"作为其分层标识。这样就为信息资源与树图进行比较、为用户推荐及对树图进行更新提供了可能。

获得知识库中依一定的策略取出的信息资源,然后根据信息资源的标识号,找到基于相同分类法的标识号与之对应的树图节点,再在该节点作为根节点构成的子树中,将信息资源与各节点进行比较:首先进行精确查找,搜索是否有与该信息资源完全匹配的节点,若有则将资源加入到该节点的推荐队列中,同时停止执行算法;若没有则进行模糊匹配,从子树的根节点开始从上往下逐层比较,同层依次比较后取比较值最大的节点作为父节点,继续与此父节点的下层子节点进行比较,同层比较值较小的其余节点的子节点不再参与下一轮的比较。比较值若超过一定的阈值(该阈值是取值在 0 到 1 之间的小数,其值比较接近 1),则信息资源属于此节点,加入

节点的推荐队列中，同时停止执行算法；比较值若小于此阈值，则继续执行算法直到子树最底层；如果这些比较值出现一个或多个明显较大，同时其它的值较小的情况，则将该信息资源代表的信息主题提取出，作为比较值最大的节点的子节点加入到树图中，同时将该资源加入到新节点的推荐队列中；若这些比较值中较大的值差别不是很大，则在子树中选择这些较大比较值个数最多的直系兄弟节点层，将该资源代表的信息主题提取出，作为新的兄弟节点加入，同时将该资源加入到新节点的推荐队列中。由此完成树图节点的建立与更新。该过程如图 1 所示。

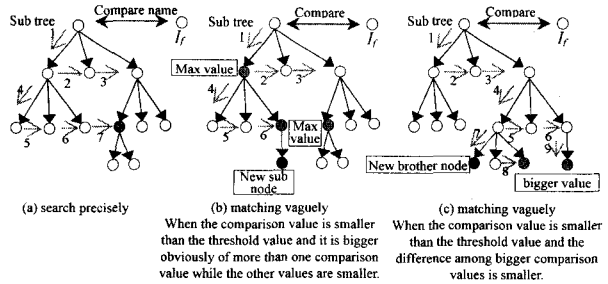

图 1　算法执行流程示意图

在给出模糊匹配公式前，先引入区间值及区间值模糊集的概念。

定义 2(区间值定义)　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]=$\{[a^-,a^+]|a^-\leqslant a^+,a^-,a^+\in I\}$。

定义 3(区间值模糊集定义)　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，由于要将信息资源与子树节点进行模糊匹配，而从知识库中取出的信息资源来源于网络，是从庞杂的网络信息资源聚类而来，而聚类问题很少能获得一个精确的值，因此信息资源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，设节点也用一个区间值模糊集来表示：$B(x_i)=\lfloor B^-(x_i),B^+(x_i)\rfloor,x_i\in G,i$ 代表该树图子树中的某个节点。

文献[13]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，信息资源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：

$$P_a=\lfloor \bigvee_{x\in G}\{A^-(x)\wedge B_i^-(x)\},\bigvee_{x\in G}\{A^+(x)\wedge B_i^+(x)\}\rfloor \quad (3)$$

比较值 $P_a\in[I]$，[I]表示单位闭区间。如果 $P_a=[1,1]$，则表示两个模糊区域一定相交；如果 $P_a=[0,0]$，则表示两个模糊区域完全不相交。

4.1　算法(LW-TGNO)

输入：从知识库中抽取的信息资源 I_f，树图 T_r；

输出：更新后的树图 T_r'；

1：获得 I_f

2：在 T_r 的节点中查找分类标识号与 I_f 相同的节点 T_n

3：以 T_n 为根节点，构造子树 TR_t

4：for all $i:i\in[1..n]$ do // TR_t 有 n 个节点，I_f 与 TR_t 中的节点依次进行比较

5：　比较 I_f-Name 与 TR_{ti}-Name

　　//I_f-Name 表示信息资源名，TR_{ti}-Name 表示第 i 个节点名

6：　if I_f-Name=TR_{ti}-Name then

7：　　$RQ_{j+1}^{(i)}=I_f,j=j+1,j\in N$ //将 I_f 加入到节点 TR_{ti} 的推荐队列 $RQ^{(i)}$ 的末尾

8：　　Goto 35://结束算法

9：　end if

10：end for

11：*VagueMatch* //进行模糊匹配

12：for all $k:k\in[1..m]$ do // TR_t 有 m 层节点

13：　for all $l:l\in N$ do

　　// TR_t 中某节点具有 l 个子节点

14：　　使用式(3)，比较 I_f 与 $TR_{ts(l)}^{(k)},s=1,2,\cdots$，比较值记为 $W_{s(l)}^{(k)}$

　　//s 表示第 $k-1$ 层节点与 I_f 比较后具有最大比较值的节点，$TR_{ts(l)}^{(k)}$ 表示该节点的下一层(第 k 层)子节点

15：　　比较 $W_{s(l)}^{(k)}$ 与预先给定的阈值 α

16：　　if $W_{s(l)}^{(k)}\geqslant\alpha$ then

　　//比较值大于阈值 α 时

17：　　$RQ_{j+1}^{TR_{ts(l)}^{(k)}}=I_f,j=j+1,j\in N$

　　//将 I_f 加入到节点 $TR_{ts(l)}^{(k)}$ 的推荐队列 $RQ^{TR_{ts(l)}^{(k)}}$ 的末尾

18：　　Goto 35://结束算法

19：　end if

20：　end for

21：　找出 Max$TR_{ts}^{(k)}$ //找出本层和 I_f 比较后具有最大比较值的第 s 个节点

22：end for

　//下面是当比较值小于阈值 α 时

23：if $W_{s(l)}^{(k)}<\alpha$ then

24：　if $W_q=\{W_{s(l)}^{(k)}\}$ 中的比较值一个或多个明显较大，同时其它比较值较小时 then

　　// $W_q=\{W_{s(l)}^{(k)}\}$ 记为比较值集合

25：　　找出 W_q 中最大值 Max$W_{s(l)}^{(k)}$，并找出其对应的节点 $TR_{ts(l)}^{(k)}$

26：　　将 I_f 代表的信息作为节点 $TR_{ts(l)}^{(k)}$ 的新建子节点 $TR_{t(TR_{ts(l)}^{(k)})(l+1)}^{(k+1)}$ 加入到子树 TR_t 中

27：　　$RQ_1^{TR_{t(TR_{ts(l)}^{(k)})(l+1)}^{(k+1)}}=I_f$

　　//将 I_f 加入到新节点 $TR_{t(TR_{ts(l)}^{(k)})(l+1)}^{(k+1)}$ 的推荐队列 $RQ_1^{TR_{t(TR_{ts(l)}^{(k)})(l+1)}^{(k+1)}}$ 的队首

28：　else

29：　　$W_q=\{W_{s(l)}^{(k)}\}$ 中较大比较值差别不多

30：　　在子树中选择具有个数最多的这些较大比较值的直系兄弟节点层，这些直系兄弟子节点对应的上层父节点记为 $TR_{ts(l)}^{(k)}$

　　//给这些同层直系兄弟节点新增一个兄弟节点

31：　　将 I_f 代表的信息作为节点 $TR_{ts(l)}^{(k)}$ 的新建子节点 $TR_{t(TR_{ts(l)}^{(k)})(l+1)}^{(k+1)}$ 加入到子树 TR_t 中

32：　　$RQ_1^{TR_{t(TR_{ts(l)}^{(k)})(l+1)}^{(k+1)}}=I_f$

33：　end if

34：end if

35：Return T_r'

4.2　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)能行性：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可行的算法。本文提出的树图节点的建立与更新算法具有输入数据，即从知识库中抽取的信息资源 I_f 和树图 T_r；也具有输出数据，即更新后的树图 T_r'。因此，下面将对该算法从确定性、有穷性(主要是复杂性)及能行性上分别加以分析。

4.2.1 算法确定性分析

算法的确定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[14,15])。

定义 4(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。

为以下叙述的方便，这三条性质可以分别称为良序的传递性、唯一性及非空性。

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。

证明：参见文献[15]中定理 3.35 的证明。

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 $P\rightarrow Q$。

证明：参见文献[16]中定理 4.2.20 的证明。

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。

下面对本文提出的树图节点的建立与更新算法构建子句集 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 2,3 是顺序执行关系，分别记为 A_1,A_2；

3)语句 4－10 是一个循环结构，其中语句 5 与后面的语句是顺序结构，记为 A_3，语句 6－9 是一个 IF 语句，其中语句 7,8(语句 8 是一个 Goto 语句)是顺序结构，分别记为 A_4,A_5；

4)语句 11－34 是模糊匹配 VagueMatch：其中语句 11 是一个入口函数，记为 A_6，语句 12－22 是一个嵌套循环结构，语句 13－20 是内层循环结构，其中语句 14,15 是顺序执行关系，分别记为 A_7,A_8，语句 16－19 是一个 IF 语句，其中语句 17,18(语句 18 是一个 Goto 语句)是顺序结构，分别记为 A_9,A_{10}，语句 21 和前面的内层循环结构是顺序执行关系，记为 A_{11}；语句 23－34 是一个嵌套分支选择结构，嵌套了 IF 比较值一个或多个明显较大且其它比较值较小，记为 A_{12}(语句 25－27)，*ELSE* 较大的比较值差别不大，记为 A_{13}(语句 29－32)；

5)语句 35 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

$$\begin{aligned}G=&\{(P\rightarrow A_1)\wedge(A_1\rightarrow A_2),A_2\rightarrow A_3,\vee_{i=4,6}(A_3\rightarrow A_i),\\&A_4\rightarrow A_5,A_6\rightarrow A_7,A_7\rightarrow A_8,\vee_{i=9,11}(A_8\rightarrow A_i),A_9\rightarrow\\&A_{10},\vee_{i=6,12}^{13}(A_{11}\rightarrow A_i),\vee_{i=5,10,12}^{13}(A_i\rightarrow Q)\}\\=&\{\sim P\vee A_1,\sim A_1\vee A_2,\sim A_2\vee A_3,\sim A_3\vee A_4,\sim A_3\\&\vee A_6,\sim A_4\vee A_5,\sim A_6\vee A_7,\sim A_7\vee A_8,\sim A_8\vee A_9,\\&\sim A_8\vee A_{11},\sim A_9\vee A_{10},\sim A_{11}\vee A_6,\sim A_{11}\vee A_{12},\\&\sim A_{11}\vee A_{13},\sim A_5\vee Q,\sim A_{10}\vee Q,\sim A_{12}\vee Q,\sim A_{13}\\&\vee Q\}\end{aligned}$$

则由良序的传递性(定义 4)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，得证。

4.2.2 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，本文将对树图节点的建立与更新算法的时间复杂度进行分析。

设 $L_0=\max(|T_j|),j=1,2,\cdots,r$，$T_j$ 表示某次获得 I_f 并在树图 T_r 的节点中查找分类标识号与 I_f 相同的节点 T_n，再以 T_n 为根节点构造子树 TR_t 这个过程的一次执行时间(r 表示树图 T_r 有 r 个节点)；设 $L_1=\max(|N_i|),i=1,2,\cdots,n$，$N_i$ 表示 I_f-*Name* 与 TR_{ti}-*Name* 进行比较，直到找出比较值一致的节点并将 I_f 加入到该节点的推荐队列的时间(n 表示该子树 TR_t 有 n 个节点)；设 L_2 表示 I_f 与子树节点进行一次模糊比较的时间；设 L_3 表示 I_f 加入一个节点的推荐队列的时间；设 L_4 表示将 I_f 代表的信息主题作为一个新节点加入到子树中相应处所需的时间。

对本算法进行分析，发现它的执行可以分为两步：第一步是对子树前处理，即 T_j；第二步是对子树的处理，是指精确查找，即 N_i，或模糊匹配。由于时间复杂度是按所需时间的最大值来考虑的(即最坏情况下)，因此这里的 T_j 和 N_i 分别取其上界值计算，即 $L_i(i=0,1)$。实际上，算法的执行时间除上述这些外，还包括比较、判断等时间，由于这些时间与 $L_i(i=0,1,2,3,4)$相比很小，这里为方便计算略去不计。以下对本算法的几种不同执行情况按照最坏情况时间复杂性进行分析：

① 精确查找

这种情况包含了子树前处理、找出 I_f-*Name* 与 TR_{ti}-*Name* 相同的节点并将 I_f 加入到该节点的推荐队列的时间的最大值，即 L_0+L_1；

② I_f 与节点模糊匹配，当比较值大于阈值 α 时

这种情况包含了子树前处理、精确查找最大时间值、I_f 与 TR_t 节点进行模糊匹配的时间及其加入节点推荐队列的

时间，设 TR_t 有 n 个节点，算法最多将 TR_t 中每一层都执行完，每一层(第二层除外)都只选择一个节点继续下层模糊匹配，所以按照所有节点都执行计算得出的算法复杂度必然大于算法的实际复杂度，即最坏情况下时间复杂度小于 $L_0+L_1+nL_2+L_3$；

③ I_f 与节点模糊匹配，当比较值小于阈值 α，且比较值一个或多个明显较大，同时其它比较值较小时

这种情况包含了子树前处理、精确查找最大时间值、I_f 与 TR_t 每一层的节点都进行模糊匹配的时间、作为一个新节点加入到子树中及加入节点推荐队列的时间，仍按照所有节点都执行模糊匹配来计算，即最坏情况下时间复杂度小于 $L_0+L_1+nL_2+L_3+L_4$；

④ I_f 与节点模糊匹配，当比较值小于阈值 α，且较大比较值差别不大时

这种情况同情况③类似，包含了子树前处理、精确查找最大时间值、I_f 与 TR_t 每一层的节点都进行模糊匹配的时间、作为一个新节点加入到子树中及加入节点推荐队列的时间，仍按照所有节点都执行模糊匹配来计算，即最坏情况下时间复杂度小于 $L_0+L_1+nL_2+L_3+L_4$；

⑤ 由于以上 4 种情况是互斥的，故完成算法的一遍执行的时间应取以上各情况的最大值，即最坏情况下时间复杂度小于 $L_0+L_1+nL_2+L_3+L_4$。

这里的 $L_i(i=2,3,4)$ 是完成某种基本操作对应的时间，可看作是常数；$L_0=\max(|T_j|), j=1,2,\cdots,r$，为子树前处理取的上界值(实际 T_j 只有在极小概率下才会取到 L_0)，故也可看作是一个常数；但是考虑到 L_1 取值的复杂性(完成的非基本操作)，则 L_1 不能看作是常数。如果记上式为 $T(n)$，则本算法最坏情况下的时间复杂度可以表示为 $T(n)=O(L_1+nL_2)$，是一个多项式。

4.2.3 *算法能行性分析*

一个算法是否有效，从理论上还没有一种方法能够证明[15]。从它的含义上看，算法能行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。本文对该算法的确定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，因此对本算法的确定性及有穷性的证明也就间接证明了它的能行性。

以上对本文提出的树图节点的建立与更新算法从理论的角度进行了分析证明，由此可以看出本算法具有确定性、能行性并且复杂度低的特点。

结束语 移动业务多媒体化和互联网化是移动服务发展的重要方向，但是高数据传输率的多媒体服务成本极高。在不增加硬件投入的前提下，为了降低成本，同时又不降低为用户提供服务的质量，可以将个性化服务应用于移动业务中。

本文关注如何将待推荐内容与移动终端用户的个性化需求相匹配，以组织多播推送的问题。提出了一种工作在服务器端的树图模型，引入区间值模糊集理论，给出相关定义和公式，在对该方法进行理论分析的基础上，设计了一个算法来加以实现。理论分析证明该算法具有确定性、能行性并且复杂度低的特点。今后将进一步在基于本体的解释能力与适应能力良好的用户个性化模型、用户模型与待推荐内容的相似性度量及计算方法、动态用户群模型建模及用户群模型预测精度度量等方面进行深入研究。

参 考 文 献

[1] Cufoglu A, Lohi M, Madani K. A Comparative Study of Selected Classification Accuracy in User Profiling[C]// San Diego, CA. Seventh International Conference on Machine Learning and Applications. 2008: 787-791

[2] Eleftheriadis G P, Theologou M E. User Profile Identification in Future Mobile Telecommunications Systems[J]. IEEE Network, 1994: 33-39

[3] Erbas F, Kyamakya K, Steuer J, et al. On the User Profiles and the Prediction of User Movements in Wireless Networks[C]// The 13th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications. 2002: 2282- 2286

[4] Araniti G, Meo P D, Iera A, et al. Adaptively Controlling the QoS of Multimedia Wireless Applications Through "User Profiling" Techniques[J]. IEEE Journal on Selected Areas in Communications, 2003, 21(10): 1546-1556

[5] Pandey V, Ghosal D, Mukherjee B. Exploiting User Profiles to Support Differentiated Services in Next-Generation Wireless Networks[J]. IEEE Network, 2004, 18(5): 40-48

[6] Panagiotakis S, Koutsopoulou M, Alonistioti A, et al. Context Sensitive User Profiling for Customised Service Provision in Mobile Environments[C]// IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications. 2005: 2014-2018

[7] Bartolomeo G, Berger F, Eikerling H J, et al. Handling User Profiles for the Secure and Convenient Configuration and Management of Mobile Terminals and Services[C]// Proceedings of the 16th International Workshop on Database and Expert Systems Applications(DEXA'05). 2005

[8] Bila N, Cao Jin, Dinoff R, et al. Mobile User Profile Acquisition through Network Observables and Explicit User Queries[C]// The Ninth International Conference on Mobile Data Management. Beijing, China, 2008: 98-107

[9] 黄海清，王永峰. 基于代理的用户偏好建模研究[J]. 哈尔滨工业大学学报，2007，39(7)：1163-1165

[10] 罗匡，谭继志，王衡，等. IICAS：一个基于用户偏好的智能接听系统[C]// 第四届和谐人机环境联合学术会议. 武汉，中国，2008：588-597

[11] 张犁，潘纲，李石坚，等. 智能影子(SmartShadow)：一个新的普适计算模型[C]// 第四届和谐人机环境联合学术会议. 武汉，中国，2008：175-182

[12] 陈俊杰，刘炜. 一种基于本体的个性化模式库建模方法[J]. 计算机研究与发展，2007，44(7)：1151-1159

[13] 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

[14] 周培德. 算法设计与分析[M]. 北京：机械工业出版社，1985

[15] 杜亚军. 搜索引擎智能行为的研究及实现[D]. 成都：西南交通大学，2005

[16] 邱小平. 关于办公信息系统智能化的研究[D]. 成都：西南交通大学，2003