软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(11):2556−2574 [doi: 10.13328/j.cnki.jos.004561]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

一种基于移动用户位置的网络服务推荐方法*

刘树栋[1,2], 孟祥武[1,2]

[1](智能通信软件与多媒体北京重点实验室(北京邮电大学),北京 100876)

[2](北京邮电大学 计算机学院,北京 100876)

通讯作者: 刘树栋, E-mail: bupt.mymeng@gmail.com, http://scs.bupt.edu.cn/cs_web

摘 要: 伴随着无线通信技术和智能移动终端的快速发展,基于位置的服务(location-based services,简称 LBS)以其移动性、实用性、随时性和个性化的特点,在军事、交通、物流等诸多领域得到了广泛的应用,成为最具发展潜力的移动增值业务之一.在一个基于位置的网络服务推荐框架的基础上,给出了一种基于位置的移动用户偏好相似度计算方法,同时证明了其满足近邻相似测度的一般性质;然后,提出一种符合社会学概念的信任值计算方法.把它们应用于基于移动用户位置的网络服务推荐过程中,从而形成了一种基于移动用户位置的网络服务推荐方法.该方法有效地提高了网络服务推荐的准确性和可靠性,同时缓解了推荐过程中可能存在的数据稀疏性以及冷启动问题.最后,通过公开的 MIT 数据集验证了该推荐方法的准确度和可行性.

关键词: 位置服务;个性化服务;相似度;推荐系统;协同过滤;信任关系

中图法分类号: TP18

中文引用格式: 刘树栋,孟祥武.一种基于移动用户位置的网络服务推荐方法.软件学报,2014,25(11):2556−2574. http://www.jos.org.cn/1000-9825/4561.htm

英文引用格式: Liu SD, Meng XW. Approach to network services recommendation based on mobile users' location. Ruan Jian Xue Bao/Journal of Software, 2014,25(11):2556−2574 (in Chinese). http://www.jos.org.cn/1000-9825/4561.htm

Approach to Network Services Recommendation Based on Mobile Users' Location

LIU Shu-Dong[1,2], MENG Xiang-Wu[1,2]

[1](Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia (Beijing University of Posts and Telecommunication), Beijing 100876, China)

[2](School of the Computer Science, Beijing University of Posts and Telecommunication, Beijing 100876, China)

Corresponding author: LIU Shu-Dong, E-mail: bupt.mymeng@gmail.com, http://scs.bupt.edu.cn/cs_web

Abstract: Along with the development of wireless communication technologies and smart mobile devices, location-based services (LBS), with its characteristics of mobility, practicality, momentary and personalization, has been widely applied in military, transportation, logistics etc, and it has become one of the most potential mobile value-added services. Based on a proposed framework of location-based network services recommendation, this paper first provides an approach to compute mobile users' preferences similarity from their geographic location, and proves that it satisfies the general properties of neighbor similarity measure. Then according with the concept of trust in sociology, a new method is presented for calculating trust value. By importing them into network services recommendation process, an approach of location-based network services recommendation is proposed, which effectively improves its accuracy and reliability, and mitigates data sparsity of users' similarity matrix and cold start users problem in recommendation process. Finally the proposed algorithm is proved to be more accurate and feasible in experiments by using the public dataset MIT.

Key words: location-based service; personalized service; similarity; recommender system; collaborative filtering; trust relationship

移动通信网的发展,为用户提供了一个更加丰富多彩的移动网络服务平台,实现了用户对网络信息资源随

* 基金项目: 国家自然科学基金(60872051); 北京市教育委员会共建项目
收稿时间: 2012-12-10; 修改时间: 2013-02-04, 2013-11-14; 定稿时间: 2014-01-10

时随地的获取与推送,使得为用户提供无处不在的移动网络服务成为可能.尤其是移动社会化网络的兴起,为用户在网络信息服务、共享、评论等方面提供了极大的帮助.与此同时,服务类型与服务内容的日新月异,有限的移动网络资源和硬件资源,为移动用户带来严重的移动信息过载问题.如何从浩瀚的移动网络环境中发现用户真正感兴趣的信息资源,丰富并满足移动用户对信息的个性化需求,逐渐成为移动通信网络中个性化服务领域亟待解决的技术难题.

近几年,推荐系统作为个性化信息服务的解决方案之一,在工业界和学术界都引起了广泛的关注.与传统的搜索引擎相比,推荐系统不仅注重搜索结果之间的关系和排序,而且还重点考虑用户的个性化偏好模型对搜索结果的影响.此外,普适计算理论的成功引入,使传统推荐系统不再仅仅关注“用户-项目”二元关系,而是将用户所处的上下文环境信息(如时间、位置、周围人员、情趣、活动状态、网络条件等等)一同考虑进来,形成“上下文-用户-项目”三元组系统,使得系统能够自动发现和利用各种上下文信息,满足用户随上下文信息变化而改变的个性化信息需求.例如,用户更乐意在上下班的公交车上看自己喜欢的小说/电影,而不是在办公室;与工作的办公室相比,用户更乐意在下班后的休闲娱乐广场了解周边促销广告.这一方面切实地满足用户体验,提高了用户满意率;另一方面增强了系统的适应性和推荐的精确度.因此,如何合理地提取基于用户上下文信息的个性化偏好,成为推荐系统的研究重点之一.

社会化网络用户之间的互动行为是人类社会行为的在线网络组织形式,间接地体现了网络用户之间的社会关系.人类社会关系信息对用户的行为习惯有非常重要的影响.比如,长辈对下辈的指导意见、同学之间的观点的相互参考与借鉴.目前,社会化网络的发展消除了亲密关系用户之间互动行为的区域限制,为用户之间的互动提供了极大的便利.位置服务与移动互联网的融合,特别是与移动社会化网络的融合,产生了与人们日常生活紧密联系、拥有具体场景化的位置服务,一方面,这些应用通过广大用户的参与及位置信息的分享,是后台的数据处理机制能够获取社会行为感知和分析的数据基础,在掌握人类群体行为规律、引导社会发展与进步等方面具有显著意义;另一方面增强了用户之间的互动频率,实现了用户之间的互动行为在时间上的单向性.例如,一个用户可以在某一时刻看到其朋友在上一时刻的留言信息及位置信息.

Pew 最新研究报告显示,58%的智能手机上网用户使用过位置服务,其中,用在导航和获得位置相关推荐的占 55%,分享自身位置信息的占 12%[1].所谓基于位置服务(location-based services,简称 LBS)[2]指通过移动终端和无线或卫星通信网络的配合,确定出移动用户的实际地理位置,从而提供用户需要的与位置相关的信息服务.基于位置服务主要包括手机导航、基于位置的社会化网络服务、智能交通、物流监控等等.

目前,以 facebook,twitter,Google+,MySpace 为代表的社交网络平台已有 20 多亿的用户,这些平台目前都已经具备了位置分享、位置签到、位置标识等位置服务的初级功能.位置服务与移动社会化网络的融合,形成移动互联网与传统互联网的无缝网络服务,通过分析用户行为的时间序列、行为轨迹和位置信息的标记组合,帮助用户与外部世界建立更加广泛而密切的联系,增强社交网络与地理位置的关联性,协助用户寻找朋友位置和关联信息,同时激励用户与位置相关的各种信息.这使得基于位置的社会化网络服务成为位置服务的核心内容,也为移动社会化网络的个性化信息服务提供了新的发展方向.ACM SIGSPATIAL GIS 已连续 3 年举办基于位置的社会化网络(location-based social networks,简称 LBSN)研讨会,二者结合也成为一种新的学术研究方向.这也恰恰符合移动互联网发展的 SOLOMO 的概念,即未来互联网发展将是社会化(social)、本地化(local)和移动化(mobile)的融合[3].

文献[4,5]将 GSNs 与位置追踪服务(比如 GPS)相结合,用户可以在不同的地点记录各种体验活动,特别是用户对地点或者活动的评分、评论.利用移动用户产生的大量的 GPS 服务数据,提出了一种在线原型系统,称为地理社会化推荐系统,将数据描述为三维张量,利用高阶奇异分解技术实现潜在语义分析和降维;同时,随着系统中数据的逐渐累积,用增量解决方案更新张量,实现了对移动用户的朋友、地点和活动的推荐.

在文献[6]中,为了实现与位置相关的活动推荐,鉴于有限的用户数据及用户-位置-活动评分数据矩阵的稀疏性,提出了 3 种基于协同过滤的移动推荐算法.第 1 种方法将所有用户的数据融合在一起,用集体矩阵分解模型提供普通的基于位置的移动推荐.第 2 种方法利用集体张量和矩阵分解模型实现了不同用户个性化推荐.第 3

种方法鉴于上述两种方法,为了获取尽可能高的准确率,首先对用户的位置/活动的表现直接进行优化排序,并依次做出推荐.对这 3 种方法,利用用户之间的相似度、位置特征、活动之间的相关性、用户-位置行为等信息,优化协同过滤算法,提供推荐的准确率.

文献[7]利用用户的位置及历史轨迹信息,将用户不同的行动轨迹与已定义的多种模式一一对应,建立用户行为与模式异构信息物理社会化网络(cyber-physical social network),利用随机游动理论为用户推荐近邻朋友.

Girardello 等人[8]利用用户当前的位置上下文,分析该位置附近手机应用程序的使用情况,将使用最频繁的应用程序推荐给用户,也只是考虑位置上下文和应用程序使用情况对应用程序选择的影响,并没有考虑移动用户个人偏好对应用程序选择的影响.

文献[9]通过统计分析移动用户访问过的商家网页,得到移动用户的偏好特征向量,使用余弦相似度来衡量用户描述文件与商家网页的相似度来衡量用户对商家的偏好,同时考虑移动用户与商家间的位置距离,距离远的商家,移动用户的偏好会降低.在将位置作为物理标识来使用时,主要利用位置间的距离衡量对移动用户偏好的影响,还可以进一步对位置进行抽象,如在家、在办公室、在户外等,通过对位置不同程度的抽象,更有利于推理和分类.

目前,基于位置的网络服务推荐研究领域大都围绕移动用户位置变化的轨迹及位置特征挖掘,以轨迹匹配的方式寻找相似用户,并以此给出服务推荐策略.很少将位置信息作为一种特殊的上下文信息与传统推荐系统相结合,实现基于位置上下文的推荐[10].鉴于此,本文将位置信息引入到移动通信网用户偏好提取及相似度的计算过程中,考虑移动用户的位置对用户对网络服务信息使用的特征的影响,提出了一种基于位置的用户网络服务特征相似度计算方法;为了缓解移动用户基于位置相似用户矩阵的稀疏性及用户的冷启动问题,利用移动用户之间的通信记录信息,提出了一种适合社会心理学概念的信任值计算方法,构建移动用户的通信信任网络,提高推荐的成功率和准确率.

基于位置的移动网络服务推荐在具备传统推荐系统的一般特性的同时,还具有移动性强、易受上下文因素的影响及实时性和互动性强的特征.这里将移动用户位置信息、用户偏好及用户之间的信任关系相结合,为移动用户提供个性化网络服务推荐,一方面符合移动位置服务中的随时、随地为任何人和任何事提供个性化信息服务的宗旨;另一方面,充分考虑了个性化服务推荐中,用户实际需求不仅受其个性化偏好的影响,而且受其所处的社会互动网络关系及上下文环境影响的特点.

1 背景知识

1.1 推荐系统的基本知识

目前,主流的推荐算法包括以下几种:基于内容的推荐、协同过滤推荐、基于知识推荐和混合推荐[11].其中,基于“集体智慧”思想的协同过滤是目前应用最多的一种推荐方法,该方法首先从用户以往历史数据中提取用户对项目的偏好信息,根据具有相似项目偏好的用户之间喜欢的项目的差异,将相似用户喜欢的项目推荐给彼此.因此,任何提取用户偏好及用户之间的相似度计算就成为协同过滤算法中的关键技术.

传统协同过滤算法主要包括以下几步:(1) 建立用户数据模型;(2) 用户偏好相似测度的计算;(3) 近邻用户的选择;(4) 产生预测.

1.1.1 用户数据模型的建立

用户对各种服务项目使用的历史数据是建立用户数据模型的基础,挖掘与分析每个用户使用的所有服务项目的评分,提取出用户-服务 $n\times m$ 阶实数矩阵 R,其中,n 代表所有用户的数量,m 代表所有用户的服务项目的数量.矩阵 R 中的元素 $r_{i,j}$ 表示用户 i 对服务项目 j 的评分,例如在 MovieLens 数据集中,所有用户对电影的评分用 0~5 之间的整数表示,0 表示用户没有评分,1~5 表示用户的喜好程度的高低.

1.1.2 用户偏好相似测度的计算

目前,除了基于优化决策的非负矩阵分解方法之外,基于服务项目评分协同过滤推荐系统中,都要用到用户偏好相似测度的计算,其中主流的相似度计算方法主要有 3 种:

第 1 种为余弦相似度[12],定义如下:

$$sim(x,y)=\cos(x,y)=\begin{cases}\dfrac{\sum_{s\in S_{xy}} r_{x,s}\cdot r_{y,s}}{\sqrt{\sum_{s\in S_{xy}} r_{x,s}^2\cdot\sum_{s\in S_{xy}} r_{y,s}^2}}, & |S_{xy}|\geqslant 2\\ 0, & |S_{xy}|<2\end{cases},$$

其中,$sim(x,y)$表示用户 x 和用户 y 之间的相似度,$r_{x,s}$ 表示用户 x 对项目 s 的评分,$S_{x,y}$ 表示用户 x 和用户 y 共同评分的项目集合.

第 2 种为 Pearson 相关系数法[13],详细定义如下:

$$sim(x,y)=\begin{cases}\dfrac{\sum_{s\in S_{xy}}(r_{x,s}-\overline{r}_x)\cdot(r_{y,s}-\overline{r}_y)}{\sqrt{\sum_{s\in S_{xy}}(r_{x,s}-\overline{r}_x)^2\cdot\sum_{s\in S_{xy}}(r_{y,s}-\overline{r}_y)^2}}, & |S_{xy}|\geqslant 2且H\neq 0\\ 0, & |S_{xy}|<2或H=0\end{cases},$$

其中, $H=\sqrt{\sum_{s\in S_{xy}}(r_{x,s}-\overline{r}_x)^2\cdot\sum_{s\in S_{xy}}(r_{y,s}-\overline{r}_y)^2}$,$\overline{r}_x$ 和 $\overline{r}_y$ 表示用户 x 和用户 y 对所有项目评分的均值.

第 3 种是修正的 Pearson 相关系数法[13],详细定义如下:

$$sim(x,y)=\begin{cases}\dfrac{\sum_{s\in S_{xy}}(r_{x,s}-r_{mid})\cdot(r_{y,s}-r_{mid})}{\sqrt{\sum_{s\in S_{xy}}(r_{x,s}-r_{mid})^2\cdot\sum_{s\in S_{xy}}(r_{y,s}-r_{mid})^2}}, & |S_{xy}|\geqslant 2且H_{mid}\neq 0\\ 0, & |S_{xy}|<2或H_{mid}=0\end{cases},$$

其中,r_{mid} 表示修正中值, $H_{mid}=\sqrt{\sum_{s\in S_{xy}}(r_{x,s}-r_{mid})^2\cdot\sum_{s\in S_{xy}}(r_{y,s}-r_{mid})^2}$.

1.1.3 近邻用户的选择

一般来说,对近邻用户的选择标准有两个:一是控制为每个用户所选择的近邻用户相似度的大小,即,选择相似度大于指定阈值的近邻用户;二是控制为每一个用户所选择的近邻用户数量,最常用的是 Top-N 方法,即,为所有用户选择相同数目的近邻用户.

1.1.4 产生预测

在推荐产生的过程中,就是计算用户未来对某一服务项目的预测评分,一般采用加权平均值法:

$$P_{u,i}=\overline{r}_u+\frac{\sum_{x\in NB}(r_{x,i}-\overline{r}_x)\cdot sim(u,x)}{\sum_{x\in NB}|sim(u,x)|},$$

其中,u 是目标用户,NB 是目标用户 u 的近邻用户集.

1.2 上下文信息与传统推荐算法的融合

引入上下文信息的上下文感知推荐系统是目前该领域研究的热点之一,其中,将上下文信息恰当地融合到传统推荐算法中,一方面符合上下文信息的现实意义;另一方面要对推荐结果有实质性的帮助,提供推荐的准确率,或者满足用户特定需求体验等等.在文献[14]中,阐述了上下文信息与传统推荐算法的两种融合方案(如图 1 所示),并指出:目前的已有研究重点考虑上下文信息在推荐生成过程中的作用(如图 1(a)[14]所示),而忽略了上下文信息对用户偏好的影响(如图 1(b)[14]所示).

本文将用户位置信息视为一种特殊的上下信息,融合到用户的偏好提取与推荐生成的过程中,考察位置信息对用户偏好及推荐结果的影响;同时,为了解决数据稀疏性及冷启动用户的问题,将移动用户的好友关系信息引入到推荐生成过程中.

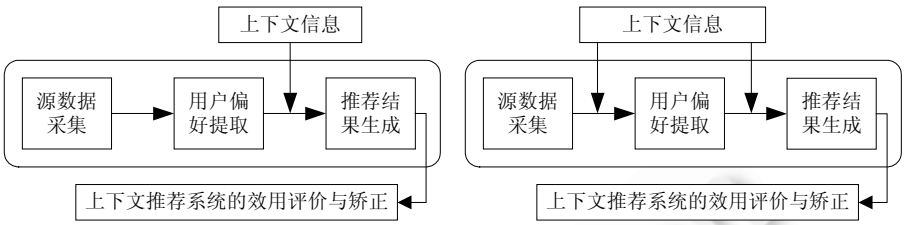

Fig.1 Schemes of fusion between contextual information and traditional CF

图 1 上下文信息与传统协同过滤的融合方案

2 移动通信网中基于位置的用户-网络服务特征模型

2.1 基于位置的用户-网络服务基本模型

与传统互联网用户相比,移动通信网中,用户的最大特征是用户位置随时间的随机性变动,正是这种位置的变动性,才使得基于移动用户不同位置的服务推荐成为可能.但是,如何认知用户随位置变化对信息的个性化需求的变化规律,准确地提取移动用户随位置变化的个性化信息需求偏好模型,将成为基于位置的移动通信网信息推荐服务的关键.在本文中,我们将根据用户位置随时间的周期性变化,学习用户对信息的个性化需求随位置的变化规律,提取用户对个性化信息需求偏好模型.

2.1.1 移动通信网中的基本数据模型

移动用户在一定时间周期范围(一天、一周或者一个月)内,其所处的地理位置是在不断地变化,同时,在不同的地理位置所需求的信息服务也是不同的.也就是说,移动用户的个性化信息需求随着用户所处地理位置变动而改变.但是在多个时间周期(几天)内,移动用户的地理位置的变化存在一定的规律性,这是我们提取用户对信息需求的个性化偏好模型及挖掘不同移动用户之间相似的行为规律的前提条件.比如,一个智能手机用户在每天早上出门之前,上网查看当天的天气情况;在上班路上会进同一家餐厅吃早餐,坐地铁的时候会听音乐;而下班后会进超市购物,而晚饭都会进健身房锻炼身体等等.

本节涉及的基本数据集有:

(1) 移动用户集:即移动通信网中所有用户的集合,用 U 来表示.

(2) 移动用户的地理位置集合:即所有移动用户可能处于地理位置信息,用 Z 来表示.这里的地理位置的定义是宽泛的,比如在家、在路上、在办公室等等.

(3) 移动网络服务集合:即提供给用户的所有移动网络服务集合,用 S 来表示.

(4) 在所有时间周期内,每个移动用户位置变化序列矩阵:根据用户活动次数,将每一个时间周期分隔成 N 段,则移动用户在一个时间周期内所处地理位置的变化序列为 $z_i,i=1,2,\ldots,N$,在所有的 M 个时间周期内,每个移动用户的位置变化序列矩阵为

$$P_t=(z_{ij})_{M\times N},\ i=1,2,\ldots,M,\ j=1,2,\ldots,N,\ t\in U.$$

(5) 在每个时间周期内,每个移动用户在所有地理位置上应用移动网络服务矩阵为一个 $N\times L$ 维矩阵:

$$Q^{t,k}=(s_{nl}^{tk})_{N\times L}=\begin{bmatrix} s_{11}^{tk} & s_{12}^{tk} & \cdots & s_{1l}^{tk} \\ s_{21}^{tk} & s_{22}^{tk} & \cdots & s_{2l}^{tk} \\ \vdots & \vdots & \vdots & \vdots \\ s_{n1}^{tk} & s_{n2}^{tk} & \cdots & s_{nl}^{tk} \end{bmatrix},\ t\in U,\ k=1,2,...,M.$$

2.1.2 基于位置的移动用户偏好特征

在本节中,首先给出基于位置的移动用户偏好模型的定义,结合余弦相似度的计算方法,提出一种基于位置的用户偏好相似测度计算方法,并验证其有效性.

定义 1. 基于位置的移动用户偏好模型是一个二元组 $p=(L,S)$,S 是一个关于移动用户使用的网络服务项目的多维向量,表示用户在某一位置 L 上使用的网络服务 $S=(s_1,s_2,\ldots,s_L)$,s_i 表示移动用户对第 i 个网络服务项目的使用情况(比如是对该网络服务体验评分,或者使用次数等等).

定义 2. 在一个时间周期内,移动用户随位置变化的网络服务偏好模型 $P=(p_1,p_2,\ldots,p_M)$.具体三维模型如图 2 所示.

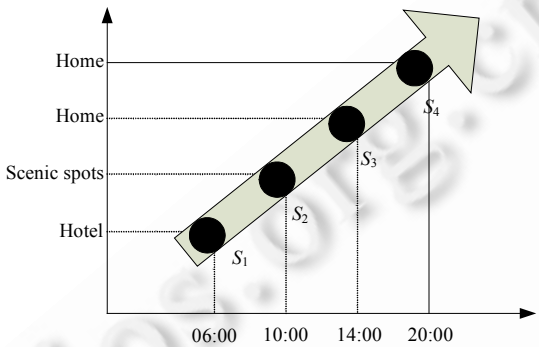

Fig.2 Location-Based preference of a mobile user

图 2 移动用户基于位置的偏好特征

在一个时间周期(比如 1 天)内,随着时间的变迁,用户位置在不停地变化.例如,将一天时间划分为 4 个时间段(午夜:1,上午:2,下午:3,晚上:4),该用户 u_1 的位置变化情况为(在家:1,办公室:2,超市:3,在家:1),而另一个移动用户 u_2 的位置变化情况为(宾馆:4,风景区:5,在家:1,在家:1),这一天上午的这段时间内,两个用户使用的服务信息分别为 $(s_1^1,s_2^1,...,s_L^1)$ 和 $(s_1^2,s_2^2,...,s_L^2)$,那么两个移动用户在这一天上午这一段时间段内基于位置的特征模型分别为 $m_{11}^1=(1,1,(s_1^1,s_2^1,...,s_L^1))$ 和 $m_{14}^2=(1,4,(s_1^2,s_2^2,...,s_L^2))$,其中,$s_m^n$ 表示第 n 个移动用户在对第 m 个网络服务项目的评分或者使用次数.依次可以提取这两个移动用户在这一天时间内其他时段的使用的网络服务特征.

定义 3. 设两个移动用户 u_x 和 u_y 分别在位置 L_x 和 L_y 的对所有网络服务项目的应用特征为 $p_x=(L_x,S_x)$和 $p_y=(L_y,S_y)$,S_x 和 S_y 分别是这两个移动用户在位置 L_x 和 L_y 使用的所有网络服务多维特征向量,经归一化处理,使它们具有相同的长度.其中,$dis(L_x,L_y)$是这两个用户所在的位置之间的距离,可以根据不同实际应用条件,选择不同的距离计算公式(比如欧式距离、汉明距离等),则基于位置的移动用户网络服务偏好相似度定义为

$$sim(u_x,u_y)=\frac{1}{e^{dis(L_x,L_y)}}\cos(s_x,s_y).$$

显然,一方面,当两个移动用户在相同位置上时,他们之间距离为 0,即 $dis(L_1,L_2)=0$,此时,$\frac{1}{e^{dis(L_x,L_y)}}=1$;对于移动用户的任意两个不相同的位置,由于 $dis(L_x,L_y)>0$,因此 $0<\frac{1}{e^{dis(L_x,L_y)}}<1$. 对任意的 $u_x\in U,u_y\in U$,当且仅当 $x=y$ 时,$sim(x,y)=1$.所以,任意的两个移动用户之间的相似度 $sim(u_1,u_2)\in(0,1)$.

另一方面,对任意的 $U_x,U_y,U_z\in U$,上述定义满足以下不等式:

$$sim(u_x,u_y)sim(u_y,u_z)\leqslant[sim(u_x,u_y)+sim(u_y,u_z)]sim(u_x,u_z).$$

证明:

因为 $\cos(x,y)\cos(y,z)\leqslant[\cos(x,y)+\cos(y,z)]\cos(x,z)$且$\forall x\in R^*$, $0<\frac{1}{e^x}\leqslant 1$,

所以,$\frac{sim(x,y)\cdot sim(y,z)}{[sim(x,y)+sim(y,z)]\cdot sim(x,z)}=\frac{\cos(x,y)\cdot\cos(y,z)}{[\cos(x,y)e^{dis(y,z)}+\cos(y,z)e^{dis(x,y)}]sim(x,z)}$

$$=\frac{\cos(x,y)\cdot\cos(y,z)e^{dis(x,z)}}{[\cos(x,y)e^{dis(y,z)}+\cos(y,z)e^{dis(x,y)}]\cos(x,z)}$$

$$\leqslant\frac{\cos(x,y)\cdot\cos(y,z)}{[\cos(x,y)+\cos(y,z)]\cos(x,z)}\cdot\frac{1}{e^{\min(dis(x,y),dis(y,z))}}$$

$$\leqslant 1\cdot\frac{1}{e^{\min(dis(x,y),dis(y,z))}}$$

$$\leqslant 1.$$

因此,上述不等式成立.

所以,定义 3 满足近邻测度中的相似性测度的基本性质[15],是一种有效的近邻相似测度. □

2.2 移动通信网中用户之间的信任值

这里将利用移动用户之间通信记录(离线数据)的次数及时长建立移动用户之间好友关系,并把好友关系视为一种特殊的邻居关系.这里的好友关系分为直接好友关系和间接好友关系两种.

2.2.1 直接好友关系

在移动通信网中,不同用户之间的互动关系主要有语音通信和短信通信两种方式,两个用户之间通信次数及时长是衡量用户好友关系的重要性的关键指标.

定义 4. 假设在一定时间周期内,用户 $u_x\in U$ 与所有用户的通信过程中,总的主叫通话次数为 T,总的主叫通话时间为 P,发出短信息的总数为 Q.在这一段时间内,用户 $u_x\in U$ 与用户 $u_y\in U$ 的通信过程中,总的主叫通话次数为 t_{xy},总的主叫通话时间为 P_{xy},发送给用户 y 的总的短信息数量为 q_{xy},则用户 x 对用户 y 的信任值定义如下:

$$trust_{x\to y}=\frac{t_{xy}}{T}+\frac{p_{xy}}{P}+\frac{q_{xy}}{Q}.$$

同理,可以计算用户 u_y 对用户 u_x 的信任值 $trust_{y\to x}$.按照这样的方法,可以依次计算所有用户相互之间的信任值,并可以形成一个关于所有用户之间的相互直接信任矩阵 Tr_1.

2.2.2 间接好友关系

社会学中认为[16]:信任既是情感的又是理性的;既是关系的根据又是关系的结果;既是一种外在的理性制度,又是一种内在的情感道德.社会人之间天然存在的各种持久的相互依赖关系及密切的内部联系,导致的由熟知而自然而然产生的直接信任关系,具有强烈的感情色彩.两个社会陌生人通过中间媒介而产生的间接信任关系,在没有制度性约束的条件下,一方面是一种受社会道德影响的社会理性行为,又是一种夹杂着个人感情的依托关系.这种间接性社会信任关系是衡量一个社会整体道德水平的重要依据.

由此可以看出:用户之间的间接信任关系一方面受用户所处的社会环境的影响,一方面受其个人情感因素的影响.因此,这里用一个用户对他的所有好友平均信任度,表示该用户对他所处的社会环境情感认知程度;而将其所有好友用户对他的平均信任度,视为该用户社会声誉度,表示所有用户对该用户社会理性信任程度.这里将两者加权平均,计算用户之间的间接信任关系.这种间接信任值的确定与计算,充分考虑了两个用户之间共同好友对信任值的影响,同时还兼顾了用户的社会关系背景对信任值的反馈影响.

定义 5. 假设在一定的时间周期内,用户 $u_x\in U$ 与用户 $u_y\in U$ 之间没有任何方式的通信联系,即,不存在直接的好友关系,设用户 u_x 的好友集合为 S_x,用户 u_y 的好友集合为 S_y,则 $S_x\cap S_y=\varnothing$.若存在一系列的用户集 $S_k,k\in Z^*$,使得 $S_x\cap S_1\cap\ldots\cap S_y\neq\varnothing$,即,用户 u_x 和用户 u_y 之间存在信任连通路,则用户 u_x 和用户 u_y 之间可能存在间接好友关系.设 $\overline{trust_{x\to i}}$ 表示用户 u_x 对其所有信任好友的平均信任度:

$$\overline{trust_{x\to i}}=\frac{\sum_{i\in S_x}trust_{x\to i}}{|S_x|}.$$

用户 u_y 的所有好友用户对他的平均信任度为

$$\overline{trust_{j\to y}}=\frac{\sum_{j\in S_y} trust_{j\to y}}{|S_y|},$$

则用户 u_x 与用户 u_y 间接信任度定义为

$$\widetilde{trust}_{x\to y}=\alpha\cdot\overline{trust_{x\to i}}+\beta\cdot\overline{trust_{j\to y}},0<\alpha,\beta<1,\alpha+\beta=1.$$

同理,可以计算所有其他用户之间可能存在的间接好友信任度,从而形成一个关于所有用户之间的间接好友信任度矩阵 Tr_2,则最终的所有用户之间的好友信任度矩阵为

$$Trust=\begin{bmatrix} Tr_1 & 0\\ 0 & Tr_2\end{bmatrix}.$$

3 移动通信网中基于用户位置的网络服务推荐

3.1 基于移动用户位置的网络服务推荐基本框架

基于位置的移动通信网络服务推荐框架如图 3 所示.

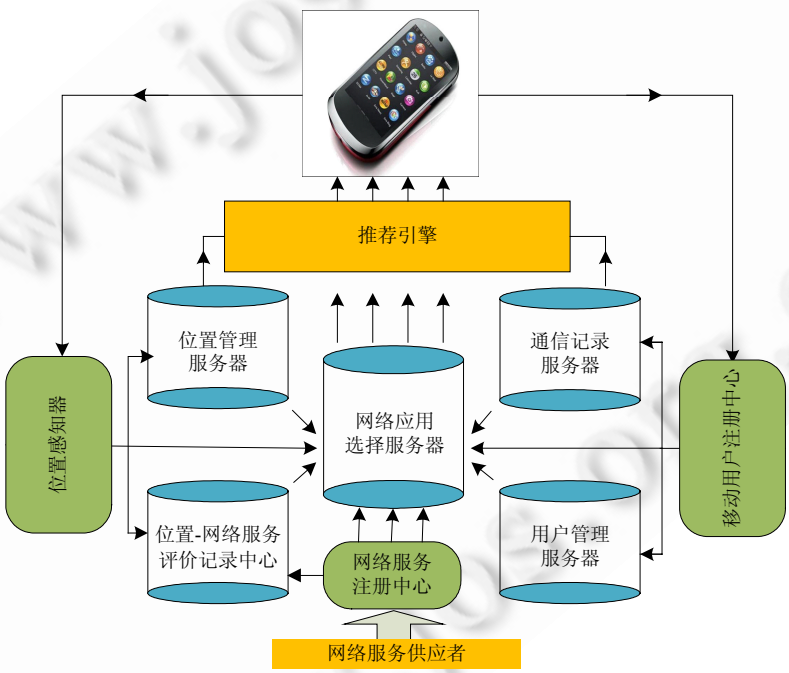

Fig.3 Framework of network services recommender system based on mobile users' location

图 3 基于移动用户位置的网络服务推荐系统框架

(1) 移动用户注册:这是整个推荐系统的基础部件,只有注册用户才能对网络服务信息进行使用与评价,是这个推荐系统的用户数据源.

(2) 网络服务注册:为网络服务的供应者提供网络服务注册与上传功能,是整个推荐系统的网络服务数据源.

(3) 移动用户位置感知器:实时感知移动用户地理位置信息随时间的变化情况,是基于位置的网络服务系统关键部件之一.

(4) 通信记录:记录所有移动用户之间相互通信信息,是好友关系挖掘的数据源.

(5) 位置管理:记录移动用户地理位置随时间变化的情况,为网络服务推荐引擎提供位置信息,并在推送

过程中提供位置指引信息.

(6) 位置-网络服务评价记录:记录用户在所有地理位置上对网络服务的使用情况与评价情况,是基于位置的用户偏好数据源,是整个推荐系统的核心组件之一.

3.2 基于移动用户位置的网络服务推荐

基于移动用户位置的网络服务推荐算法的具体步骤如下:

1. 建立移动用户基于位置的偏好模型

在移动通信网中,移动用户位置信息随着时间的变化而改变,这种位置的变迁诱导了移动用户偏好的改变.因此,移动用户基于位置的偏好模型的建立就是要以时间-位置为主线,在训练数据中提取用户随时间-位置变动而导致的不同网络服务集.因此,移动用户基于位置的偏好模型建立步骤如下:

Step 1. 对于一个移动用户 u_x,根据定义 1,在某一地理位置 L_x 上提取该用户使用的网络服务集合 S_x,得到该用户在这个位置上使用的网络服务特征 $p_{L_1}=(L_x,S_x)$.

Step 2. 在这个时间周期内的其他不同位置上,执行步骤 Step 1,提取用户 u_x 在这些位置上使用的网络服务特征 $p_{L_2},...,p_{L_M}$.

Step 3. 在所有的时间周期内,执行步骤 Step 1 和 Step 2,若在不同时间周期内的相同位置上,用户 u_x 使用了相同的网络服务项目,则计算该用户对该项网络服务的平均评价值为该用户在所有时间周期内对该网络服务的整体评价;否则,将一个时间周期上使用的网络服务评价作为该用户在所有时间周期内对该网络服务的整体评价.得到该用户在整个训练集上的基于位置的全局偏好特征 $(\overline{p_{L_1}},\overline{p_{L_2}},...,\overline{p_{L_M}})$.

Step 4. 对所有的移动用户,重复执行 Step 3,提取他们在整个训练集上的全局偏好矩阵 P.

2. 基于位置的相似度计算

在基于位置的所有用户全局偏好矩阵 P 中,首先利用定义 3 计算任意两个移动用户之间的基于位置的相似度,并将所有位置上的任意两个移动用户之间的平均相似度视为这两个用户之间的全局相似度,从而可以计算出整个训练集上的所有用户之间的全局相似度矩阵 Sim.

3. 移动用户之间信任值的计算

移动用户之间的信任关系主要包括直接信任关系和间接信任关系,其中,直接信任关系由用户之间直接的通信关系形成,间接信任关系由用户之间信任传播而形成,直接信任关系和间接信任关系共同组成了移动用户之间的通信信任网络.这里,用信任值作为评价用户之间信任关系紧密程度的标准.

(1) 直接信任值的计算

Step 1. 删除训练集中对所有网络服务评价记录为空的用户,抽取训练集中对所有网络服务信息评价记录非空的所有用户,作为组成整个通信信任网络的用户群 U.这一步是为了删除无用的用户数据信息.

Step 2. 根据定义 4,以用户 u_x 为例,计算与该用户存在直接通信关系的信任值,并写入直接信任矩阵 Tr_1 中.在这一步中,如果用户 u_x 的被叫用户只有 1 个,则将用户 u_x 视为孤立用户,并删除这种信任关系.

Step 3. 对 U 中每一个用户,执行 Step 2,得到完整的直接信任矩阵 Tr_1.

(2) 间接信任值的计算

间接信任关系是从社会学的角度,从好友的好友集为用户寻找可能的信任关系.因此,好友的好友集的查找与组建是计算间接信任值的关键.在所有用户集 U 中,以 u_x 为例,间接信任值的计算步骤如下:

Step 1. 从用户的直接信任矩阵 Tr_1 中查找出用户 u_x 的所有直接好友用户集 $Frienfs_x$.

Step 2. 计算用户 u_x 对其好友用户集 $Frienfs_x$ 中的每一个用户的信任值及用户 u_x 对其所有好友用户的平均信任值 $\overline{trust}_{x\to i}$.

Step 3. 设用户 u_x 的任意一个好友 $u_y\in Frienfs_x$,$u_z\in Frienfs_y$ 是用户 u_y 的一个直接信任好友,从直接信任矩阵 Tr_1 中,计算出用户 u_z 的所有直接信任好友对其的平均信任值 $\overline{trust}_{j\to z}$.

Step 4. 根据定义 5 计算出用户 u_x 对用户 u_z 的间接信任关系 $trust_{x\to z}$,并将其写入间接信任矩阵 UTr_1,称为

第 1 层间接信任矩阵.

Step 5. 对于所有用户集 U 中除 u_x 外的其他用户,依次执行 Step 1,Step 2 和 Step 3,计算出所有用户之间的第 1 层间接信任值,并写入间接信任矩阵 UTr_1.

Step 6. 对于第 1 层间接信任关系中的所有用户,依次执行 Step 1,Step 2,Step 3 和 Step 4,那么就可以计算出所有用户之间的第 2 次间接信任矩阵 UTr_2.同理可以计算出第 3 层、第 4 层的间接信任矩阵 UTr_3,UTr_4.

Step 7. 将上述步骤中计算得到的各个层次的间接信任矩阵 UTr_1,UTr_2,UTr_3 和 UTr_4 依次写入矩阵 Tr_2 中,从而形成最终的信任矩阵 Tr.

4. 相似矩阵与信任矩阵的融合

相似矩阵与信任矩阵融合的最终目的是为推荐目标用户甄选邻居用户,这些邻居用户使用的所有网络服务是推荐给该用户的候选集.因此,相似矩阵与信任矩阵的融合方式对最终的推荐结果有决定性的影响.在传统的协同过滤推荐算法中,Top-N 方法是最常用近邻选择方法.将信任关系信息引入到协同过滤推荐算法中,设用户 u_x 的相似用户集合为 Sim_x,信任好友用户集合为 Tr_x,则用户的邻居用户集 NB_x=Sim_x∪Tr_x,其中,近邻相似权重的处理方法如下:

$$\omega(x,y)=\begin{cases}\dfrac{2\cdot sim(x,y)\cdot trust(x,y)}{sim(x,y)+trusy(x,y)}, & y\in NS_x\cap NT_x\\ sim(x,y), & y\in NS_x, y\notin NT_x\\ trust(x,y), & y\notin NS_x, y\in NT_x\end{cases}.$$

5. 产生预测

在这个过程中,预测用户对任意网络服务项目的兴趣度.

假设用 $P_{x,i}$ 表示用户 u_x 对网络服务项目 i 的预测评分:

$$P_{x,i}=\begin{cases}\overline{r}_x+\lambda\sum_{y\in NB_x}(r_{y,i}-\overline{r}_y)\cdot\omega(x,y), & \overline{r}_x\neq 0\\ \lambda\sum_{y\in NB_x}r_{y,i}\cdot\omega(x,y), & r_x=0\end{cases},$$

其中, $\lambda=\dfrac{1}{\sum_{y\in NB_x}\omega(x,y)}$.

3.3 数据稀疏性及冷启动问题

数据稀疏性及冷启动问题是目前推荐系统中亟待解决的两个难题之一,众多学者在此方面做了大量的研究工作.目前为止,针对这两个问题的主要解决方法有两种.一种是在用户偏好相似协同过滤的基础上,考虑社会化网络关系信息对用户行为偏好的影响,实现社会化协同推荐[17,18],缓解了用户偏好相似协同过滤中用户相似矩阵可能存在的数据稀疏性及冷启动问题.另一种是由 Chen 等人[19,20]提出的基于矩阵分解的特征推荐方法.该方法是在传统矩阵分解获取用户、项目潜在因子(latent factor)的基础上,融合用户、项目及全局特征,从事物本源特征上学习用户喜好,从根本上遏止问题发生的可能性.

针对上述类似问题,借鉴第 1 种解决方法,分析用户的移动通信信息,将用户好友关系信息作为一种特殊的近邻关系融入到相似矩阵中,形成最终的推荐候选近邻用户集,一方面扩大了推荐候选近邻集的数目,缓解了单一相似近邻用户集作为推荐候选近邻集的稀疏性;另一方面,相对于相似近邻的冷启动用户,可以从可能的通信好友集中为其挑选推荐候选,从而在一定程度上降低冷启动用户存在的可能性.

3.4 性能分析

本算法的主要步骤是基于移动用户位置的相似度的计算、用户之间间接信任值的预测以及对结果的排序算法.一般情况下,余弦相似度的计算复杂度为 $O(N^2)$,根据定义 3,将移动用户的位置上下文信息加入到相似度计算中的计算复杂度为 $O(kN^2)$,其中,$0<k\leqslant 1$,远远小于 N.根据定义 4,移动用户之间的间接信任关系的计算复杂度为 $O(\max(\alpha,\beta)\cdot N)$.一般排序算法的计算复杂度为 $O(N^2)$.因此,算法的整体复杂度为 $O(N^2)$.

4 实验与结果分析

本节主要描述实验数据、设计方案、评价准则及结果分析.实验使用的硬件环境和系统软件见表 1.

Table 1 Hardware and system software

表 1 硬件环境和系统软件

操作系统	内存	CPU	开发语言和工具	数据库
Windows XP	2GB	2.8GHz	JDK1.5.0_04, Eclipse3.2, Matlab7.0	MySQL5.0

4.1 实验数据

本文实验首先采用麻省理工学院多媒体实验室 MIT 收集的数据集[21]进行实验.该数据集包括 106 个移动用户在 2004 年 7 月~2005 年 6 月共 12 个月使用各项网络服务、通信记录的信息以及各种上下文信息.虽然从这些数据集上不能直接得到移动用户基于位置的偏好信息、好友关系信息以及明确的地理位置信息,但是这些数据集中蕴含移动用户基于位置的偏好信息以及好友关系信息.因此,首先必须对这 12 个月内的所有移动用户行为信息统计学习,明确移动用户随时间而改变的位置信息,挖掘用户之间由于相互通信而产生的好友关系信息,提取移动用户随时间-位置变化的偏好模型.

4.2 实验评价标准

本文采用绝对平均偏差 MAE 和 *P@R* 两个评价标准来衡量网络服务推荐算法的准确度,其中,

- 绝对平均误差 MAE 定义为

$$MAE=\frac{\sum_{i=1}^{N}|p_i-q_i|}{N},$$

其中,N 表示推荐的网络服务个数,p_i 表示对网络服务用户体验预测,q_i 表示用户的实际评价.MAE 体现算法预测值与用户实际体验值之间的差异,所以 MAE 的数值越小,表明网络服务推荐算法的准确度越高.

- *P@R*[22]定义为

$$P@R=\frac{\#\text{relevant services in the Top-}R\text{ services}}{R}=\frac{|\{s_i\in S\mid s_i\in rec\_set\wedge p_i\in \text{Top-}R\}\cap\{s_j\in S\mid s_j\in used\_set\wedge q_j\in \text{Top-}R\}|}{R},$$

即,预测的用户可能常用的 Top-R 项网络服务占用户实际常用的 Top-R 项网络服务的比值,表示对用户常用 Top-R 项网络服务的预测正确率.

4.3 实验步骤及实验结果与分析

4.3.1 实验预处理

(1) 实验数据的预处理

基于原始 MIT 移动数据集在用户使用网络服务及位置信息随时间的变化特征,为了满足本实验的要求,首先需要对原始 MIT 数据集进行预处理,主要包括两个方面的内容:一是在时间维度上对用户使用网络服务信息的规约处理;二是对移动用户位置信息随时间变化的规约处理以及在语义上一一对应.

在时间维度上对用户使用网络服务信息的规约处理,就是在更大的时间跨度上对移动用户使用网络服务进行重新统计分析:原始 MIT 数据集上,移动用户使用网络服务随时间推延的变动频率大,而且在时间上的跨度非常不均匀.例如,第 10 个用户 $user_{10}$ 在 2004 年 9 月 10 日在 10:18:36,10:18:43,10:18:58 的 3 个时间点上依次使用的网络服务为(phone,mrc,context_logs),在随后的 1 个多小时内,没有该用户使用网络服务的任何记录,直到 12:35:47 才有该用户使用网络服务的下一条记录.这样的统计记录可能比较符合移动用户实际使用网络服务的变化特点,但是按照这样的时间粒度划分的统计结果,并不适用于一般推荐系统对输入数据的基本要求.任何推荐系统都不可能在每过 1 秒钟或者每隔几秒钟就为用户提供一次推荐服务,一方面,这不符合人们在实际生活

中对推荐服务的使用需求;另一方面,从本质上讲,也不符合推荐系统产生及发展的原始动力.因此,需要在更大的时间跨度上对用户使用的网络服务信息进行统计分析.这里,首先以小时为单位,对每个用户在每 1 个小时内使用的网络服务进行统计,然后将一天内的 24 小时,以 6 个小时为一个划分单元分成 4 个时间段(0~6,6~12,12~18,18~24),并重新统计在这个 4 个时间段内用户使用网络服务信息.在连续数天的时间内,可以提取出每个用户在一天的 4 个时间段内使用网络服务特征.

对移动用户位置信息随时间变化的规约处理以及在语义上一一对应:在原始 MIT 数据集上,移动用户的地理位置信息是模糊的、不易理解的 IDS 信息,这种 IDS 信息与用户生活中语义上的地理位置信息存在着单一的对应关系,而不是一一对应的.例如,第 10 个用户 $user_{10}$ 中的 IDS 值为 5119.6029 代表该用户的家,但是对于其他用户而言,该信息可能代表其他语义信息.此外,与移动用户使用的网络服务随时间变化的特点一样,在原始 MIT 数据集上,代表用户位置信息的 IDS 变动频率同样很大,同时,波动次数很高.例如,在第 10 个用户 $user_{10}$ 位置信息记录表中,在 2004 年 9 月 10 日 17 点~18 点的 1 个小时内,用户的 IDS 信息在 5119.6029 和 5119.4034 两个值之间交替变动了 45 次,平均 1.33 分钟就要变动一次.显然,这不符合人们实际生活常识.导致这种现象的原因可能是多方面的,这里不再赘述.因此,为了适应实验研究,需要对这些数据进行数据清理以及在时间维度上的规约处理.首先,在每 1 个小时内的移动用户所处的次数最多的 IDS 代表该用户在这个时间段内的位置信息.用同样的方法,可以统计出在一天的 24 小时内的 4 个不同时段(0~6,6~12,12~18,18~24),移动用户的位置信息.

经过上述两个过程,以时间为纽带,提取出了每个用户在每一天内的 4 个不同时段在语义上的位置信息以及在每一个位置上使用网络服务的信息.由于该数据集中用户对其所使用的网络服务项目没有显示的评价,只有所有用户使用网络服务使用次数,在一个地理位置上,某个时段内,所有用户对所有网络服务的使用情况中,使用次数大于 20 只占 3.37%,因此,这里将次数大于 20 的所有网络服务项目的使用次数压缩映射到 20,一方面可以避免奇异项目点的存在,另一方面可以避免过大的数值产生较大波动的 MAE 值.这样,所有用户在一个地理位置上的某一个时段内,对所有网络服务项目所有偏好,在数值上的表示范围为[1,20].

(2) 实验训练集与测试集的选择

为了验证本文所提出的方法的有效性,在 MIT 数据集上的实验中,选择在 2004 年 7 月~2005 年 3 月这 8 个月内,106 个用户位置移动的网络服务使用数据作为训练集(其中有 4 个用户在网络服务记录表或者位置记录表为空,也就是说,实际上只有 102 个用户),将 2005 年 4 月~2005 年 6 月这 3 个月内的数据作为测试集.

(3) 基于移动用户位置信息相似度计算

在原始的 MIT 数据集中,没有显著的有关移动用户位置的坐标信息.在上述(1)的数据预处理阶段,只是将标识用户位置的 IDS 信息从语义上对应到了移动用户实际生活中的位置标识,从物理意义上不能将这些位置信息一一对应到便于数学计算的距离坐标信息,从而不能利用传统的坐标距离计算方法(例如欧式距离或者汉明距离等)和定义 3 基于用户位置的相似度计算方法.因此,充分考虑人们实际生活中的语义位置信息在服务推荐系统中的重要意义,合理地利用该数据结构特点,这里给出一种在语义意义上的距离计算方法:

$$dis(Location_1, Location_2) = \begin{cases} 0, & \text{if } (Location_1 = Location_2) \\ \infty, & \text{otherwise} \end{cases}.$$

(4) 信任值的计算

根据定义 4,可以直接计算移动用户之间由于相互通信而形成的直接好友关系.利用定义 5 计算移动用户之间间接的好友关系时,根据著名的六度分割理论(又称六度空间理论或者小世界理论等),一个人和世界上的任何其他人之间所间隔的人不会超过 6 个,即,世界上任意的两个陌生人之间,通过最多不超过 6 个人就能彼此认识,因此在寻找用户之间是否存在共同的好友用户时,只需寻找相隔 6 个用户之间是否存在共同好友.因此,定义 5 中只需 $k \in [1,3], k \in Z^*$ 即可.

4.3.2 实验及结果分析

1. 位置上下文信息对移动用户偏好的影响

根据文献[23]中提出的上下文移动用户行为波动率的理论,验证位置信息对移动用户偏好的影响.实验中,

在位置信息作为单一上下文的约束条件下,移动用户的波动率,即,移动用户对某项网络服务使用量的变化幅度可以表示为

$$vol_{t,s_j}=\frac{\frac{1}{n_t}\sum_{c_{tk}\in Z_t}\left|Volume(u_x,s_j,z_{tk})-\overline{Volume(u_x,s_j,Z_t)}\right|}{\overline{Volume(u_x,s_j,Z_t)}},$$

其中,

- Z_t表示移动用户 u_x达到所有位置的集合;
- z_{tk}表示集合 Z_t中的一个实例;
- n_t表示集合 Z_t中包含的实例个数;
- $Volume(u_x,s_j,z_{tk})$表示移动用户 u_x在位置 z_{tk}约束下对移动网络服务 s_j的使用量;
- $\overline{Volume(u_x,s_j,Z_t)}=\frac{1}{n_t}\sum_{c_{tk}\in Z_t}Volume(u_x,s_j,z_{tk})$表示移动用户 u_x在 Z_t中所有位置上下文约束下对网络服务项 s_j的平均使用量.

波动率越大,说明移动用户行为受位置上下文的影响就越大.图 4 为 MIT 数据集中,移动用户 10 在位置上下文约束下该用户的波动率变化示意图.从图中可以看出:移动用户使用的不同网络服务项目受位置上下文影响的程度有较大的差别.在不同月份,Menu 的波动率的变化幅度较大;相对而言,Phone 的波动率的变化幅度较小,但是在所有的时间段内,Phone 项的波动率居高不下,整体平均值要大于 0.5.这说明移动用户对 Phone 项的使用情况受位置上下文的影响较大.从整体上来说,对于所有网络服务项目,在不同的数据段内,该用户行为受位置上下文的影响的波动率的平均值要大于 0.4,说明位置上下文对该用户行为偏好有较大的影响;同时,这也从一个侧面说明了本文中提出的基于位置的用户偏好提取及相似度计算是切合现实情况,具有实际意义.

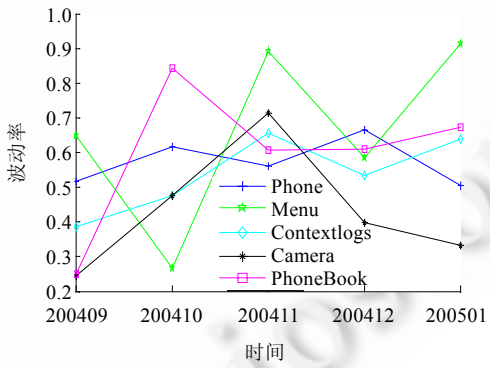

Fig.4 Volatility-Time history of mobile user 10 behaviors under location context

图 4 移动用户 10 在位置上下文约束下行为的波动率随时间的变化

2. 数据稀疏性及用户冷启动问题的实验分析

(1) 位置上下文信息对用户相似度矩阵稀疏性的影响

在传统的协同过滤算法中,计算用户的相似度矩阵时只考虑用户-项目二维向量,而没有考虑各种上下文信息对用户偏好及用户之间相似度的影响.在本文中,将对用户偏好有重要影响作用的位置上下文信息引入到用户相似度的计算过程中,这势必对用户相似度矩阵的稀疏性等方面有一定的影响.本实验的主要目的正在于此.

从表 2 可以看出:将位置上下文信息引入到用户相似度的计算,使得用户的相似用户数量有不同程度的下降.其中最为明显的是 $User_3$:在没有考虑位置上下文信息(传统的协同过滤)的相似矩阵中,其相似用户数量为 88;而在本文提出的基于位置上下文信息的相似矩阵中,其相似用户数量为 28,下降的幅度达到了 68.18%.这势必极大地增强了用户相似向量的稀疏性,也就是为相似用户的寻找增加了难度.对于整个用户集,通过类似的对

比实验计算,使用定义 3 计算得到的相似度矩阵的稀疏性,要比传统的协同过滤算法得到的相似度矩阵的稀疏性高 1.6%.整体上来说,将位置上下文信息引入到相似度的计算过程中,相似矩阵稀疏性的增强幅度并不是非常的大,对于个别用户而言(例如 $User_{10}$),这样的增强幅度甚至完全可以忽略.事实上,这表明位置上下文对所有用户个性化偏好的影响是一致的.这也从一个侧面说明定义 3 是合理的.

Table 2 Contrast on the number of similar users

表 2 相似用户数对比

	Similar users (no location)	Similar users (with location)	Decline (%)
$User_3$	88	28	68.18
$User_5$	89	88	1.12
$User_{10}$	89	89	0
$User_{70}$	89	85	4.49
$User_{100}$	19	17	10.53

(2) 用户信任关系信息对近邻矩阵稀疏性的影响

用户对网络服务项目的使用,可能存在较少历史数据及个体用户之间应用偏好的差异性,很难根据用户之间的历史行为数据找到偏好相似的用户,从而导致用户相似度矩阵存在稀疏性较高的问题.同时,从表 2 可以看出:将用户的位置上下文信息引入到相似度的计算过程中,使得这个问题进一步恶化.在该实验中,将讨论用户信任关系信息的引入对缓解这个问题的影响程度.

从表 3 可以看出:与单一的相似用户集作为近邻用户集相比,加入信任好友用户形成最终的近邻用户集,在数量上都有一定程度的增加,在降低近邻相似矩阵的稀疏性的同时,能够缓解用户的冷启动问题.例如表 3 中,$User_{31}$,$User_{95}$ 和 $User_{104}$ 在加入信任好友用户集前后,其近邻用户集在数量上分别增加了 7.5%,4.29%和 35.5%.这也相当于,这些用户的近邻相似矩阵的稀疏性分别降低了 7.5%,4.29%和 35.5%.此外,$User_{66}$ 的相似用户集个数为 0,也就是说,在相似近邻关系下,此用户是一个完全的冷启动用户.在加入信任用户集后,其近邻用户集个数增到 28,因此在这种条件下,这个用户不再是冷启动的了.把位置信息引入到相似度的计算过程中,由于需要进行位置匹配,从而增加用户近邻相似矩阵的稀疏性(见表 2);同时,把信任用户集与相似用户融合在一起,形成最终的近邻用户集,能够在一定程度上降低近邻相似矩阵的稀疏性,同时帮助缓解可能存在的冷启动用户问题.

Table 3 Contrast on the number of neighbor users

表 3 近邻用户数对比

	Similar users	Neighbor users	Increase (%)
$User_3$	28	28	0
$User_{31}$	80	86	7.5
$User_{66}$	0	28	280
$User_{95}$	70	73	4.29
$User_{104}$	45	61	35.5

3. 对参数α和β的评估

为了得到更加准确的移动用户之间的间接信任预测值,该实验将对定义 5 中的参数α和β进行评价,考察α和β(α+β=1)之间不同取值比例对移动用户之间的间接信任预测值的影响.采用的评价指标是预测信任值与真实信任值之间的绝对平均误差 MAE 和 NDCG(normalized discounted cumulative gain)[24,25],其中,NDCG 是用来衡量信任值排序的准确度.NDCG 数值在 0 和 1 之间,越接近于 1,算法的排序结果越好.因此,把α和β之间的比值分别设置为 1:4,2:3,1:1,3:2 和 4:1,在训练集中计算预测信任值和真实信任值之间的绝对误差 MAE 和 NDCG,结果如图 5(a)和图 5(b)所示.由此可知,当α和β之间的比值为 2:3 时,对移动用户间接信任预测值达到最优.此外,为了得到α和β之间更加精确的比例分配,还可以把对所有用户间预测信任值与真实信任值之间的绝对平均误差 $\sum_u MAE_u$ 和 $\sum_u NDCG_u$ 作为最优目标值,在训练集中分别把当 $\sum_u MAE_u$ 和 $\sum_u NDCG_u$ 达到最小值和最大值时的α和β值作为最优比例分配值.

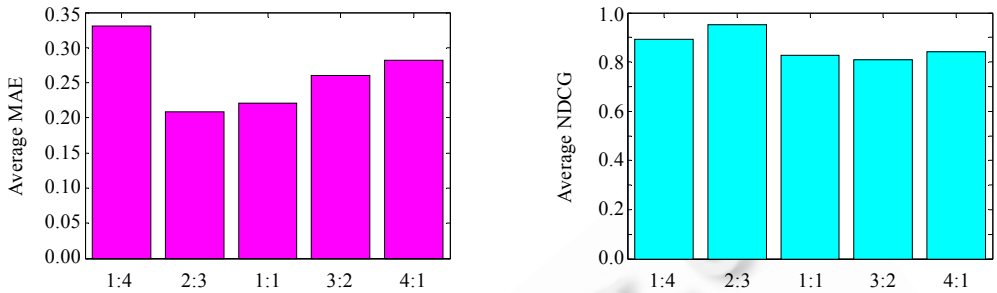

Fig.5 Average MAE and NDCG of trust with different ratio between α and β

图 5 α和β取不同的比值时,关于信任度的 MAE 和 NDCG 平均值

4. 信任用户中可能存在的噪声问题

在通信信任用户集中可能存在与用户兴趣偏好不相同的噪声现象,例如,一位教授和一位快递员之间由于需要配送快件而产生突发的通信关系,虽然这两个体之间存在通信信任关系,但是他们之间的兴趣偏好可能存在较大的差异.如果把此类用户也加入到推荐候选的近邻用户集中,那么可能会降低推荐的准确率,影响推荐算法性能.由定义 4 可知,在通信信任用户集中,此类用户的信任值往往都偏小.我们分别选择最大信任值、中间信任值和最小信任值的 5 个信任好友用户作为推荐近邻用户集,为近邻相似关系下的冷启动用户产生推荐结果(见表 4).从表 4 中可以看出:当推荐个数分别为 3 和 4 时,在这 3 种情况下推荐准确率还没有太大的区别;但是随着推荐个数的增加,在信任值较小的情况下,推荐准确率出现较大幅度的下滑.也就是说,用户间的偏好相似性出现较大的偏差,那么这些信任值较低的用户在通信信任集中极有可能是噪声.为了避免这种噪声问题,一方面可以在直接信任值的计算过程中删除通信次数较少的用户间的信任关系,此外,在间接信任值的计算过程中并没有无休止的间接信任关联查找,而是根据六度分割理论,只判断相隔最多 6 个用户之间仍然存在共同好友才视为彼此为间接好友关系;另一方面,在一般 KNN 协同过滤推荐中,对近邻用户的选择是按照相似值的大小依次取前 N 个,因此,在这个过程中很有可能选不到那些信任值较小的近邻用户.

Table 4 *P@R* of three different sets of trusted users

表 4 3 种不同信任用户集的 *P@R*

	Top 5	Middle 5	Last 5
*P@*3	0.795 2	0.789 3	0.789 3
*P@*4	0.765 0	0.765 0	0.743 8
*P@*5	0.678 6	0.666 7	0.632 0
*P@*6	0.623 1	0.610 3	0.569 9

5. 性能及效率分析

实验 1. 该实验主要对移动用户位置信息对用户偏好及推荐结果的有效性影响进行实验验证,基准对比方法为传统的协同过滤(traditional collaborative filtering,简称 TCF).实验以训练集为输入,分别以传统的协同过滤方法和提出的基于位置的协同过滤方法(based location collaborative filtering,简称 BLCF)提取用户偏好,计算用户之间的相似度.将最近邻用户数 N 的值依次取 10,15,20,25,30,35,40,分别为目标用户产生推荐项目集,并按照预测值的大小将推荐网络服务项目集进行排序,对比测试集中用户实际应用的网络服务项目,分别计算出 MAE 和 *P@R* 的平均值.

实验 2. 该实验以上述实验 1 为基准对比,将移动用户之间的信任关系信息引入到基于位置的协同过滤算法(based location collaborative filtering,简称 BLCF)中,考察将信任关系信息融合到基于位置的协同过滤算法(based location collaborative filtering with trust,简称 BLCFT)的实验对比效果、效率以及数据稀疏性与用户冷启动问题对最终实验结果的反馈影响.实验首先以训练集为输入,以基于位置的协同过滤方法提取用户偏好,并计

算用户之间的通信信任值;同时,参考对α和β的实验评估结果,将参数α和β分别取值 0.6 和 0.4,把用户相似用户集与信任好友集融合在一起,得到近邻用户集.然后,将最近邻用户数 N 值依次取 10,15,20,25,30,35,40,分别为目标用户产生推荐项目集,并按照预测值的大小将推荐网络服务项目集进行排序,对比测试集中用户实际应用的网络服务项目,分别计算出 MAE 和 $P@R$ 的平均值.同时,以近邻用户数 N=10 为基数,以 5 为其稳定增加的步长,为每位目标用户提供推荐服务,考察前后两种条件下平均耗时增长率的变化情况.

实验 3. 在实验 1 和实验 2 的基础上,如同文献[19,26],把实验 2 中得到的近邻信息引入到矩阵分解过程(matrix factorization,简称 MF)中,分别计算出 MAE,$P@R$ 的平均值和平均耗时的增长率(实验结果如图 6~图 8 所示),并与本文方法进行实验对比.

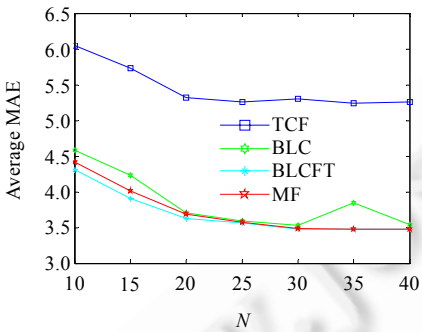

Fig.6 MAE of TCF, BLCF, BLCFT and MF

图 6 TCF,BLCF,BLCFT 和 MF 的 MAE 值

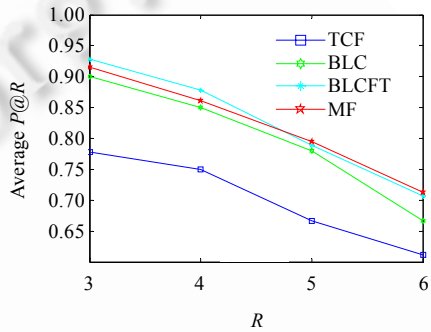

Fig.7 Average $P@R$ of TCF, BLCF, BLCFT and MF

图 7 TCF,BLCF,BLCFT 和 MF 的 $P@R$ 平均值

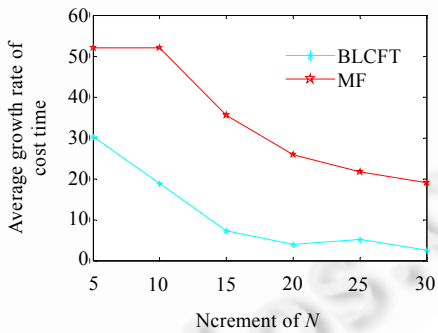

Fig.8 Growth rate of cost time

图 8 耗时的平均增长速率

实验结果分析:

(1) 图 6 和图 7 显示:无论是以 MAE 为评价标准还是以 $P@R$ 为评价标准,基于位置的协同过滤推荐算法都要优于传统的协同过滤推荐算法.这说明本文提出的基于移动用户位置的网络服务推荐策略是可行的、有意义的.

(2) 从图 6 可以看出:在 Top-N 值较小(即数据较稀疏)时,TCF 中有较大的 MAE 值;而且在 Top-N 值较大(数据较稠密)时,TCF 中的 MAE 值同样不够理想.其主要原因在于:传统的协同过滤推荐算法没有考虑环境中上下文的相似度;而 BLCF 在 Top-N 值较小(即数据较稀疏)时,同样有较大的 MAE 值,其主要原因在于将位置因素加入到相似度的计算过程中,进一步稀疏了用户相似度矩阵,增加了算法在启动阶段寻找相似邻居的难度,从而造成了较大的误差值 MAE.相比之下,BLCFT 无论是在数据稀疏还是稠密条件下,在准确度方面都有较好的性能,而且具有较小的波动性.由此可见,该算法确实能够

提高在数据稀疏性条件下的推荐准确度.

(3) 整体而言,图6中的所有MAE值都要大于3,这说明预测值与真实值之间的平均误差大于3.单纯从数值层面上讲,这样的实验结果是糟糕的.然而从图7显示的结果来讲,3种方法中*P@R*平均值最小的也要大于0.6,显然,这是一种相对理想的实验结果.此外,以BLCFT方法为例,随着*R*值的不断变大,*P@R*值的不断变小而不会上升.其主要原因在于:对所有移动用户来说,最常使用的网络服务项目相对比较集中,使得当*R*的数值为3和4时有较高的*P@R*值.但是对于同一个用户,在不同的时间周期内,最常用的网络服务项目使用情况有较大的浮动,从而造成了较大的MAE预测误差.这也充分说明了:受移动环境下输出能力等不确定性因素的影响,移动网络服务具有不稳定性的特点.

(4) 在近邻用户数稳步增长的条件下,在为每一位用户提供网络服务推荐时增加了推荐候选的搜索范围,肯定是要消耗更多的时间.正如图8所示,近邻用户数每增加5个,耗时的平均增长率都要大于2.5%.但当近邻用户数*N*的增加量达到20以上时,耗时的平均增长率逐渐趋于稳定.也就是说,在近邻用户数增加到30以后,在此基础上,近邻用户数的增加量对平均耗时增长率的影响趋于稳定.这也表明,此时近邻用户数量的增加不再是导致耗时增长的重要原因.此外,用户之间相似度及信任值的计算完全可以在线下完成的情况下,从平均耗时效率上考虑,说明该算法对增加用户数量的敏感程度不是很高.

(5) 如图6和图7所示,在此实验数据集上,虽然随着*R*值的增大,基于矩阵分解的方法在*P@R*上略优于本文提出的方法,但是整体而言,本文提出的方法在MAE和*P@R*两项性能指标上稍有优势.然而在耗时的增长率上,基于矩阵分解的方法明显要高于本文提出的方法,这在实时性要求较高的移动网络服务环境下是不可忽视的优点.

5 总 结

以位置服务为代表的地理空间信息及应用服务产业已经成为当前信息服务行业的重要组成部分,特别是在移动通信终端与互联网逐渐融合的当下,为用户增加位置维度,提供各种与人们日常生活紧密联系、拥有具体场景化的位置服务应用,具有重要的研究意义和极高的实用价值.本文在基于位置的移动用户偏好信息结构建模的基础上,提出一种基于位置的移动用户相似度计算方法,从相似测度的概念及基本性质上,证明了这种方法是行之有效的.然后,提出一种符合社会学概念的信任值计算方法.将基于位置的相似度引入到网络服务推荐选择的过程中,并与信任度相结合,构成基于移动用户位置的网络服务推荐方法.该方法有效提高了网络服务的推荐的准确性和可靠性,同时缓解了推荐过程中可能存在的数据稀疏性以及冷启动问题.在真实的MIT数据集上的实验结果表明,该方法能够切实有效地提高个性化移动网络服务推荐的精确度.下一步主要工作包括:挖掘移动用户位置历史轨迹、检测移动用户位置的变化规律、研究移动用户位置预测方法以及基于预测位置的网络服务推荐方法.

References:

[1] Pew Research. 28% of U.S. adults use mobile and social location-based services. 2011. http://pew research.org/pubs/2096/mobile-social-location-based-services-geosocial-social-media-location-tagging

[2] Jensen CS, Christensen AF, Pedersen TB. Location-Based services: A database perspective. In: Proc. of the Scandinavian Research Conf. on Geographical Information Science. New York: ACM Press, 2001. 59−68.

[3] Murphy M, Meeker M. Top 10 mobile Internet trends. 2011. http://www.kpcb.com/team/index.php?Matt Murphy

[4] Papadimitriou A, Symeonidis P, Manolopoulos Y. Geo-Social recommendations. In: Proc. of the ACM PeMA 2011. New York: ACM Press, 2011.

[5] Symeonidis P, Papadimitriou A, Manolopoulos Y. Geo-Social recommendations based on incremental tensor reduction and local path traversal. In: Proc. of the 3rd ACM SIGSPATIAL Int'l Workshop on Location-Based Social Networks (LBSN 2011). New York: ACM Press, 2011. 89−96. [doi: 10.1145/2063212.2063228]

[6] Zheng VW, Cao B, Zheng Y, Xie X, Yang Q. Collaborative filtering meets mobile recommendation: A user-centered approach. In: Proc. of the 24th AAAI Conf. on Artificial Intelligence (AAAI 2010). Melo Park California: American Association for Artificial Intelligence, 2010. 236−241.

[7] Yu X, Pan A, Tang LA. Geo-Friends recommendation in GPS-based cyber-physical social network. In: Proc. of the 2011 Int'l Conf. on Advances in Social Networks Analysis and Mining. Washington: IEEE Computer Society, 2011. 361−368. [doi: 10.1109/ASONAM.2011.118]

[8] Girardello A, Michahelles F. AppAware: Which mobile applications are hot? In: Proc. of the Human Computer Interaction with Mobile Devices and Services (MobileHCI 2010). New York: ACM Press, 2010. 431−434. [doi: 10.1145/1851600.1851698]

[9] Yang WS, Cheng HC, Dia JB. A location-aware recommender system for mobile shopping environments. Expert Systems with Applications, 2008,34(1):437−455. [doi: 10.1016/j.eswa.2006.09.033]

[10] Meng XW, Hu X, Wang LC, Zhang YJ. Mobile recommender systems and their application. Ruan Jian Xue Bao/Journal of Software, 2013,24(1):91−108 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4292.htm [doi: 10.3724/SP.J.1001.2013.04292]

[11] Xu HL, Wu X, Li XD, Yan BP. Comparison study of Internet recommendation system. Ruan Jian Xue Bao/Journal of Software, 2009,20(2):350−362 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3388.htm [doi: 10.3724/SP.J.1001.2009.03388]

[12] Gediminas A, Alexander T. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Trans. on Knowledge and Data Engineering, 2005,17(6):152−162. [doi: 10.1109/TKDE.2005.99]

[13] George T, Merugu S. A scalable collaborative filtering framework based on co-clustering. In: Proc. of the 5th IEEE Int'l Conf. on Data Mining. Washington: IEEE Computer Society, 2005. 625−628. [doi: 10.1109/ICDM.2005.14]

[14] Wang LC, Meng XW, Zhang YJ. Context-Aware recommender systems: A survey of the state-of-the-art and possible extension. Ruan Jian Xue Bao/Journal of Software, 2012,23(1):1−20 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4100.htm [doi: 10.3724/SP.J.1001.2012.04100]

[15] Sergios T, Konstantinos K. Pattern Recognition. 4th ed., San Diego: Academic Press, 2009.

[16] You H. A sociological research on the relationship between emotion and trust—A construction and testification of structural equation model in trust relationship [Ph.D. Thesis]. Wuhan: Wuhan University, 2009 (in Chinese with English abstract).

[17] Ma H, King I, Lyu MR. Learning to recommend with social trust ensemble. In: Proc. of the SIGIR 2009. New York: ACM Press, 2009. 203−210. [doi: 10.1145/1571941.1571978]

[18] Liu F, Lee HJ. Use of social network information to enhance collaborative filtering performance. Expert Systems with Applications, 2010,37(7):4772−4778. [doi: 10.1016/j.eswa.2009.12.061]

[19] Lu Q, Yang D, Chen T, Zhang W. Informative household recommendation with feature-based matrix factorization. In: Proc. of the CAMRa 2011. New York : ACM Press, 2011. 15−22. [doi: 10.1145/2096112.2096116]

[20] Chen T, Tang L, Liu Q, Yang D. Combining factorization model and additive forest for collaborative followee recommendation. In: Proc. of the KDD-Cup Workshop 2012. New York: ACM Press, 2012. http://www.kddcup2012.org/workshop

[21] Eagle N, Pentland A, Lazer D. Inferring social network structure using mobile phone data. Proc. of the National Academy of Sciences, 2009,106(36):15274−15278. [doi: 10.1007/978-0-387-77672-9_10]

[22] Wang LC, Meng XW, Zhang YJ, Shi YC. New approaches to mood-based hybrid collaborative filtering. In: Proc. of the ACM CAMRa 2010. New York: ACM Press, 2010. 28−33. [doi: 10.1145/1869652.1869657]

[23] Wang LC, Meng XW, Zhang YJ. A cognitive psychology-based approach to user preferences elicitation for mobile network service. Acta Electronica Sinica, 2011,39(11):2547−2553 (in Chinese with English abstract). [doi: 10.3724/SP.J.2011.02403]

[24] Jarvelin K, Kekalainen J. Cumulated gain-based evaluation of IR techniques. ACM Trans. on Information Systems, 2002,20(4):422−446. [doi: 10.1145/582415.582418]

[25] Qiao XQ, Yang C, Li XF, Chen JL. A trust calculating algorithm based on social networking service users' context. Chinese Journal of Computers, 2011,12(34):2403−2414 (in Chinese with English abstract). [doi: 10.3724/SP.J.2011.02403]

[26] Koren Y. Factorization meets the neighborhood: A multifaceted collaborative filtering model. In: Proc. of the 14th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2008. 426−434. [doi: 10.1145/1401890.1401944]

附中文参考文献:

[10] 孟祥武,胡勋,王立才,张玉洁.移动推荐系统及其应用研究.软件学报,2013,24(1):91−108. http://www.jos.org.cn/1000-9825/4292.htm [doi: 10.3724/SP.J.1001.2013.04292]

[11] 许海玲,吴潇,李晓东,阎保平.互联网推荐系统比较研究.软件学报,2009,20(2):350−362. http://www.jos.org.cn/1000-9825/3388.htm [doi: 10.3724/SP.J.1001.2009.03388]

[14] 王立才,孟祥武,张玉洁.上下文感知推荐系统研究进展.软件学报,2012,23(1):1−20. http://www.jos.org.cn/1000-9825/4100.htm [doi: 10.3724/SP.J.1001.2012.04100]

[16] 游泓.情感与信任关系的社会学研究——一个信任关系结构方程模型的建构和验证[博士学位论文].武汉:武汉大学,2009.

[23] 王立才,孟祥武,张玉洁.移动网络服务中基于认知心理学的用户偏好提取方法.电子学报,2011,39(11):2547−2553. [doi: 10.3724/SP.J.2011.02403]

[25] 乔秀全,杨春,李晓峰,陈俊亮.社交网络服务中一种基于用户上下文的信任度计算方法.计算机学报,2011,12(34):2403−2414. [doi: 10.3724/SP.J.2011.02403]

刘树栋(1984－),男,山东沂南人,博士生,主要研究领域为推荐系统,网络服务.
E-mail: bupt.mymeng@gmail.com

孟祥武(1966－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为网络服务,通信软件,推荐系统.
E-mail: mengxw@bupt.edu.cn