基于客户Web时空行为轨迹的兴趣点预测方法

陈冬林,夏琪,代四广

武汉理工大学电子商务与智能服务研究中心,武汉 430070

摘要 客户兴趣点预测是大数据环境下提高电子商务推荐精度的关键,针对现有客户兴趣预测未综合考虑客户多种行为和时序时间的影响问题。为研究一种基于客户Web时空行为轨迹的兴趣点预测方法,构建了包含客户、时间、行为和兴趣点四层子网的客户Web时空行为超网络模型,并引入行为影响因子,提出基于超边相似性的兴趣点预测算法,在建立连通矩阵的基础上,通过邻接矩阵计算、超三角形判定和超边相似度计算,得到相似度最高的超边,该超边对应的兴趣点即为预测结果。实验结果表明,该方法在时间误差允许范围内,兴趣点预测准确度随时间精度的减小而增加,与传统的标签预测方法相比,预测准确度由56.2%提高至74%。

关键词 兴趣点预测;Web时空行为;超网络;超边相似性

Web 3.0时代,大数据和物联网出现,网络成为客户需求的理解者和提供者,并以此为基础进行资源筛选和智能推荐。此背景下能够赢得客户青睐的一定是基于客户个性化行为、习惯信息聚合而构建的服务,因此电子商务平台的重点是运用个性化预测和推荐技术为客户提供基于其需求和偏好的个性化服务。前两代推荐系统研究集中在客户与客户(如协同过滤)、客户与商品项或标签(如内容推荐、标签推荐)之间,近年来集中在解决冷启动、数据稀疏性及提高精度的算法优化与改进方面[1-2]。王道平等[3]认为客户对冷门商品的行为更能说明兴趣度,可以引入热门物品权重系数来调整兴趣度;Krzywicki等[4]设计了跟踪客户对推荐商品列表的反馈行为进行排序调整的二次推荐机制;Scholz等[5]提出了客户购买意愿测度和基于效用的推荐系统。这些研究旨在提高推荐精度和客户满意度,但均以相对静态和较简单的客户行为信息为基础,未考虑大数据环境下精细化客户时空行为数据对兴趣变化的影响[6]。

目前,根据精细化的客户Web时空行为轨迹数据来进行精确的"预测式"推荐正受到广泛关注[7],如Amazon 2013年就推出"预测式"发货技术,它参考客户之前的订单、搜索记录、愿望清单、购物车,甚至鼠标悬停行为,在客户未购买之前就安排发货。尽管学术界争议"预测式"发货理论只是基于群体客户的预测而非针对个体行为的一种概念,但近年来人类现实时空行为的相关研究在交通导航、旅游地推荐等领域取得很大进展,这使得定量化研究Web时空行为轨迹成为可

收稿日期:2017-05-27;修回日期:2018-02-26

基金项目:国家自然科学基金项目(71172043);中央高校基本科研业务费专项(165215001);教育部留学回国人员科研启动基金项目(2013-693);湖北省教育厅科学技术研究项目(B2016403)

作者简介:陈冬林,教授,研究方向为个性化推荐及云计算,电子信箱:chendl@whut.edu.cn

引用格式:陈冬林, 夏琪, 代四广. 基于客户Web时空行为轨迹的兴趣点预测方法[J]. 科技导报, 2018, 36(7): 74-79; doi: 10.3981/j.issn.1000-7857.2018.07.011

能[8-9]。因此,根据客户复杂的Web时空行为轨迹数据预测兴趣点,对于提高客户推荐精度的研究具有理论价值和现实意义。国内外基于时序和兴趣演化规律的研究已初有成果。孙光福等[10]运用SequentialMF推荐算法,分析了客户时序购买行为;Bao等[11]通过建立概率矩阵分解模型,研究了时序社交网络行为,如发表微博或与他人建立好友关系,用于挖掘客户在微博中的潜在兴趣,更准确地预测客户兴趣;Koren等[12]提出基于客户-产品-时序的TimeSVD++算法,将时序行为加入客户、产品的特征向量中,有效预测了兴趣漂移;Li等[6]同时考虑客户的长期与短期兴趣,基于时间敏感度权重调整客户的阅读兴趣度;Feng等[13]对现有的SPCR、TACF、TimeSVD++时序环境兴趣演化方法进行改进,提出结合社交行为的时间重叠的社区兴趣演化模型,将时间因素融入了兴趣预测中。综合分析客户Web时空行为及相关文献,现有研究存在的主要问题是[14-16]:未考虑客户兴趣点在时序时间内的变化;对行为的考虑仅集中在评分、购买和社交行为上,未对客户搜索、浏览、转发、收藏、加购物车等复杂行为进行有效整合;传统方法不适用于复杂Web时空行为轨迹预测。鉴于上述问题,本文运用复杂网络中的超网络理论,研究基于客户Web时空行为轨迹的兴趣点预测方法,以提高推荐精度。

1　客户Web时空行为超网络模型的构建

借鉴超网络舆论引导方法[17-18],根据客户在不同Web空间上的历史行为轨迹数据,考虑客户兴趣受其他客户、时间、行为、兴趣点等因素的影响,构建由客户、时间、行为和兴趣点四层子网组成的客户Web时空行为超网络模型。以客户U、时间T、行为B、兴趣点I四元组$\{U,T,B,I\}$表示客户Web行为轨迹,模型的四层子网中,客户子网以客户为节点、客户之间的影响关系为边,如社交亲密度;时间子网以时间点为节点、时间点之间的先后关系为边;行为子网以客户的行为关键词为节点、各个行为关键词的包含关系为边,如购买包含浏览,但不一定包含收藏;兴趣点子网以各空间中各类商品的叶子结点为节点、商品之间的转化关系为边,如显示屏与主机。以客户、时间点、商品的叶子节点和客户的行为关键词为节点,各子网之间的关系为超边(如客户u_1在时间t_1浏览(b_1)了商品i_1),建立的客户Web时空行为超网络模型如图1所示。

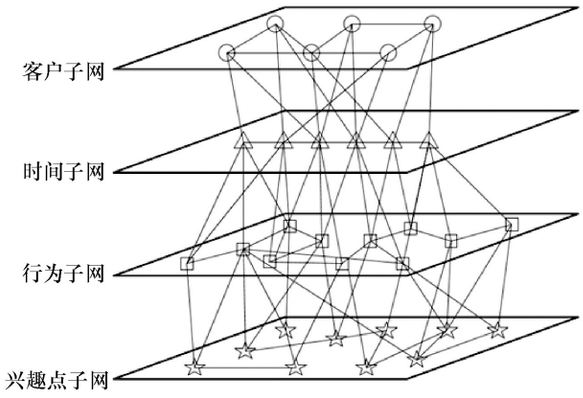

图1　客户Web时空行为超网络模型

Fig. 1　Super-network model of Web time-space customers behavior

客户Web时空行为超网络模型中,超边(SE)是各层子网中1个或多个节点的具有相互唯一性的网络链接,可描述不同客户在不同时间点的行为轨迹如图2所示。四层子网通过超边相互关联,超边SE分别经过各层子网的1个或多个节点,随着最小时间单位和客户组合的增加而不断增加。

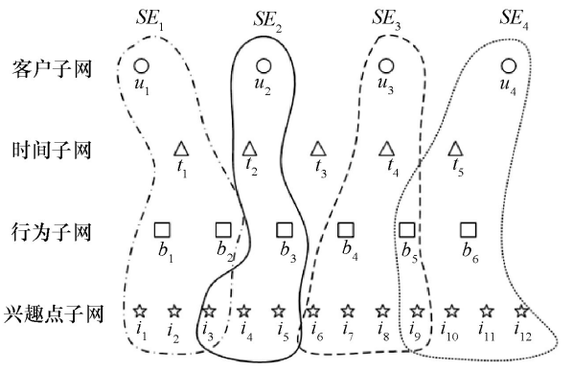

图2　客户Web时空行为轨迹

Fig. 2　Trajectory of Web space-time customer behavior

在客户Web时空行为超网络模型的基础上,提出基于超边相似性的兴趣点预测算法。超链路预测(superlink prediction)是指在超网络结构上,对未知超链接(现在或过去已经存在但未被描述或发现的超边)的预测和未来超链接(未来会出现的超边)的预测。在超网络中,由3条超边构成的三角形称为超三角形(super triangle),构成超三角形3个顶点的3个子网节点V_i、V_j、V_k满足条件$V_i \in SE_m$、$V_j \in SE_n$、$V_k \in SE_p$,如图3所示。超边相似性(superedge similarity)是指2条超边的相似程度,2

条超边重合的部分越多越相似。在超网络中2条超边的相似性由这2条超边所共同构成的超三角形的个数决定,超三角形个数越多越相似。根据超边相似性原理,计算超三角形的个数衡量超边相似度,通过相似度最高的超边预测未来可能出现的新超边及其兴趣点。

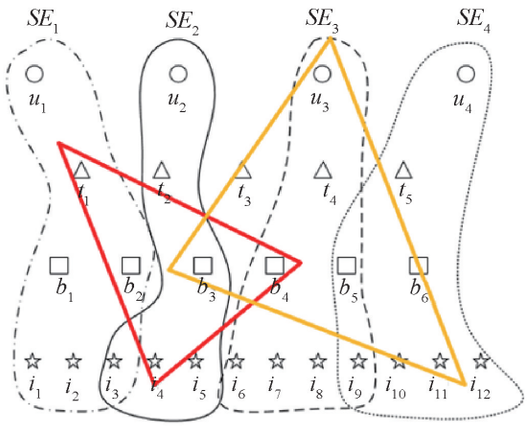

图3 超三角形示意

Fig. 3 Super triangle diagram

2 基于超边相似性的兴趣点预测算法

2.1 建立超网络的连通矩阵

Web兴趣超网络的连通矩阵$\boldsymbol{H}$是一个m行n列的矩阵,每一行代表1条超边SE_m,每一列代表四层子网中的一个节点V_n,m为超网络中所有超边的总个数,n为四层子网所有节点的总个数。若超边SE_m中包含节点V_n且节点V_n为非行为子网节点,则所对应的h_{mn}的值为1;若超边SE_m中包含节点V_n但节点V_n为行为子网节点,则所对应的h_{mn}的值为相应行为影响因子值;若超边SE_m中不包含节点V_n,则所对应的h_{mn}的值为0。建立的连通矩阵为

$$H=\begin{array}{c} \\ SE_1 \\ SE_2 \\ \vdots \\ SE_m \end{array}\begin{array}{c} \begin{array}{cccc} V_1 & V_2 & \cdots & V_n \end{array} \\ \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ h_{m1} & h_{m2} & \cdots & h_{mn} \end{bmatrix} \end{array} \tag{1}$$

考虑7种客户行为,根据不同行为对客户兴趣的影响程度不同(从浏览到搜索购买,再到评论转发,表示客户兴趣逐渐增强),设定各行为类型的影响因子,如表1所示。

2.2 计算超网络的邻接矩阵

四层子网中节点之间超边连接的情况,用节点的邻接矩阵$\boldsymbol{A}_{SE}$表示,即

表1 行为影响因子对照表

Table 1 Influencing factors of behaviors

行为类型	浏览	搜索	收藏	加购物车	购买	评论	转发
影响因子	0.2	0.4	0.6	0.7	0.8	0.85	0.95

$$\boldsymbol{A}_{SE}=\boldsymbol{H}^{\mathrm{T}}\cdot\boldsymbol{H}-\boldsymbol{D} \tag{2}$$

邻接矩阵$\boldsymbol{A}_{SE}$中的元素$(A_{SE})_{ij}$表示包含节点V_i与V_j之间连接超边的情况,用于判断超三角形中的超边存在与否,即若$(A_{SE})_{ij}$不为0,则节点V_i与V_j之间存在超边。在式(2)中,$\boldsymbol{D}$是一个对角矩阵,对角元素表示包含节点V_i的超边个数,非对角元素均为0。在式(2)中,减去对角矩阵$\boldsymbol{D}$是为了使邻接矩阵$\boldsymbol{A}_{SE}$的对角元素为0,也就是去掉节点V_i与自身的连边,相当于直接将对角元素归0。

2.3 判定超网络中的超三角形

超边SE_1、SE_2是否能共同构成超三角形,可从这2条超边包含的所有节点中任意选3个节点V_i、V_j、V_k,由式(3)、式(4)计算$(A_{SE_1})_{ij}$和$(A_{SE_2})_{jk}$进行判定,即

$$\delta_{xy}=\begin{cases}1 & x=y\\ 0 & x\neq y\end{cases} \tag{3}$$

$$(A_{SE_1})_{xy}=H_{zx}\cdot H_{zy}-\delta_{xy}\cdot H_{zx} \tag{4}$$

若同时满足3个条件则超边SE_1、SE_2可以共同构成一个超三角形,否则不能构成。条件为:$(A_{SE_1})_{ij}$值不为0,即表示节点V_i和V_j通过超边SE_1;$(A_{SE_2})_{jk}$值不为0,即表示节点V_j和V_k通过超边SE_2;邻接矩阵$\boldsymbol{A}_{SE}$中节点V_i和V_k所对应的元素$(A_{SE})_{ik}$不为0。

2.4 计算超网络中超边的相似度

超边的相似度(Sim)指超网络中任意2条超边的相似程度。超网络中任意2条超边SE_1和SE_2的相似度,由式(5)、式(6)计算,即

$$\mathrm{sign}(x)=\begin{cases}1 & x>0\\ 0 & x=0\\ -1 & x<0\end{cases} \tag{5}$$

$$Sim_{SE_1SE_2}=\frac{\frac{1}{6}\sum_{V_i,V_j,V_k\in SE_1\cup SE_2}\mathrm{sign}\left[\sum_{\alpha,\beta,\gamma\in V_i,V_j,V_k}(A_{SE_1})_{\alpha\beta}(A_{SE_2})_{\beta\gamma}(A_{SE})_{\alpha\gamma}\right]}{C^3_{|SE_1\cup SE_2|}} \tag{6}$$

式中,分子表示超边SE_1与超边SE_2所能共同构成的超三角形个数;分母$C^3_{|SE_1\cup SE_2|}$为组合数,表示从超边SE_1与SE_2所包含的所有节点中任取3个节点的组合数,即2条超边能共同构成的超三角形的最大个数。

将相似度值归一化处理得到超边的相似性矩阵$\boldsymbol{A}_{sim}$,其中相似性矩阵$\boldsymbol{A}_{sim}$中的元素$(A_{sim})_{mn}$表示超边SE_m与SE_n的相似度,$(A_{sim})_{mn}$越大,SE_m与SE_n相似度越高,SE_m与SE_n之间存在的共同兴趣点越多。

2.5　算法步骤

输入所有客户截止当前时刻在Web购物网站上的四元组$\{U,T,B,I\}$行为数据集,按下列步骤进行。

Step 1:将四元组$\{U,T,B,I\}$行为数据集转换成Web兴趣超网络的连通矩阵$\boldsymbol{H}$。

Step 2:获取连通矩阵$\boldsymbol{H}$,运用式(2),计算超网络的邻接矩阵$\boldsymbol{A}_{SE}$。

Step 3:运用式(4),计算$(A_{SE_\alpha})_{ij}$和$(A_{SE_\beta})_{jk}$($\alpha,\beta\in[1,m]$),其中α、β为m个超边中的任意2个超边的编号。

Step 4:获取$(A_{SE_\alpha})_{ij}$、$(A_{SE_\beta})_{jk}$、$(A_{SE})_{ik}$,并判定$(A_{SE_\alpha})_{ij}$、$(A_{SE_\beta})_{jk}$、$(A_{SE})_{ik}$的值是否均不为0,若均不为0,则运用式(6)计算任意2个超边的相似度$Sim_{SE_\alpha SE_\beta}$。

Step 5:获取α、β的所有可能取值对应的$Sim_{SE_\alpha SE_\beta}$,构成相似性矩阵$\boldsymbol{A}_{sim}$,其中的元素$(A_{sim})_{\alpha\beta}$等于$Sim_{SE_\alpha SE_\beta}$。

Step 6:获取相似性矩阵$\boldsymbol{A}_{sim}$,选取相似度值最大的超边,其涉及的兴趣点即为最终推荐的集合R。

输出最终推荐结果R。

3　实验验证与分析

3.1　实验数据与评估准则

实验数据采集于数据堂网站(www.datdatang.com)的开放数据,该数据集共记录了780名客户2015年6月至7月在团购网站大众点评上的30631条行为数据,包括429种商品(兴趣点),以及浏览、购买、收藏和加购物车4种行为。

评估标准的选择是验证预测算法可行性的关键部分。本实验采用时间精度P_t与兴趣点预测准确度P_i的综合分析结果。其中,时间精度P_t参考心理学遗忘曲线规律和武器系统中的射击精度定义,如图4所示,点O表示无时间误差的点,即此时实际兴趣点与预测兴趣点相吻合;点F表示时间误差为±1的点,即实际兴趣点与预测兴趣点误差一个单位时间;点G表示时间误差为±2的点,点L表示时间误差为±3的点;点W表示时间误差太大或未预测出或预测错误的点;预测兴趣点越接近圆心O,时间误差越小,时间精度P_t越高。兴趣点预测准确度P_i衡量算法的预测准确度及有效性,用公式(7)表示(算法预测出的兴趣点中与实际相一致的兴趣点个数占预测总个数的比重),即

$$P_i=\frac{PI\cap AI}{PI} \tag{7}$$

式中,PI为预测兴趣点;AI为实际兴趣点。

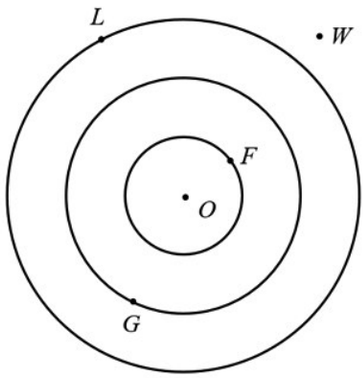

图4　时间精度示意

Fig. 4　Schematic diagram of time accuracy

3.2　结果与分析

为证明本文预测方法的的可行性和有效性,采用上述实验数据,先使用Excel将原始数据集处理成包含客户编码(Idenity, ID)、时间(按天提取)、行为和兴趣点(商品ID)的行为轨迹四元组$\{U、T、B、I\}$集,再使用MATLAB软件进行实验得出预测结果,如图5所示。

由图5最活跃客户兴趣点预测对比可见,对最活跃客户来说,预测缺失的兴趣点占比不到18%,而74%的兴趣点能够在时间误差范围内被预测。

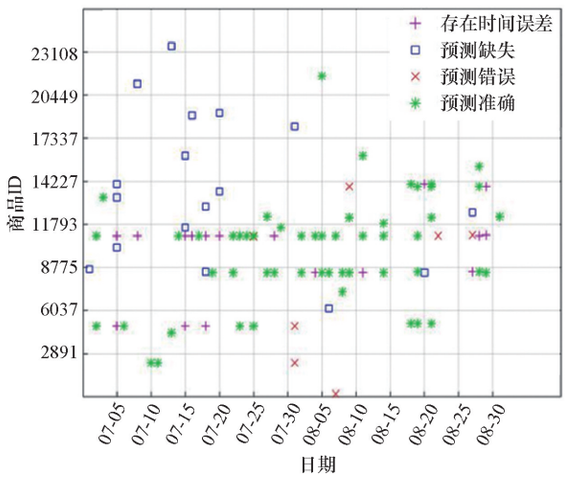

图5　最活跃客户兴趣点预测对比

Fig. 5　Comparison of point prediction of the most active customer interest

兴趣点预测准确度随时间误差变化的曲线如图6所示。当时间精度最高,即时间误差为0时,P_i=54.2%;当时间误差为±1时,P_i=58.3%;当时间误差为±2时,P_i=66.7%;当时间误差为±3时,P_i=70.1%。随着时间误差的增大(在允许范围内),相应兴趣点预测准确度趋于稳定,最终稳定值为P_i=74%。

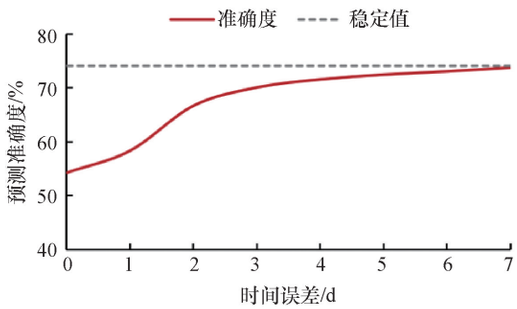

图6 预测准确度随时间误差变化曲线

Fig. 6 Prediction accuracy changes with time errors

为证明本文方法的优越性,分别采用基于客户Web时空行为轨迹的兴趣点预测方法(本文方法)和传统的标签预测方法[19](传统方法)进行预测计算,比较两种方法在时间误差允许范围内的预测准确度,其中传统方法选取其预测准确度的极值。分析结果如图7所示,可以看出,本文方法明显优于传统方法。

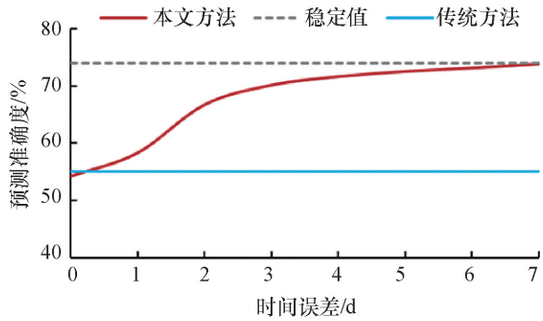

图7 两种方法预测准确度对比

Fig. 7 Comparison of two methods in forecasting accuracy

4 结论

随着大数据和物联网的快速发展,客户兴趣的"预测式"分析成为当前研究的热点之一。现有研究对行为的考虑仅集中在评分和购买上,未对客户搜索、浏览、转发、收藏、加购物车等行为进行有效整合,缺乏客户兴趣在时序时间内变化的研究。本文提出基于客户Web时空行为轨迹的兴趣点预测方法,以各空间(购物网站)上的海量数据为基础,融合时序条件下多空间精细化的客户Web行为,建立客户Web时空行为超网络模型,运用基于超边相似性的超链路预测算法进行客户兴趣点预测,并在评估准则中加入时间精度这一新标准。本文方法在时间误差允许范围内,兴趣点预测准确度随时间精度的减小而增加,最终趋于稳定值74%。与传统的标签预测方法相比,本文方法明显提高了预测准确度,为客户个性化推荐的建模和预测研究提出了可供借鉴的新思路。但是鉴于目前网站之间的信息孤岛问题,基于全网数据预测客户在Web时空行为轨迹的兴趣点是未来的研究方向。

参考文献(References)

[1] Bobadilla J, Ortega F, Hernando A, et al. Recommender systems survey[J]. Knowledge-Based Systems, 2013, 46(1): 109-132.

[2] Lika B, Kolomvatsos K, Hadjiefthymiades S. Facing the cold start problem in recommender systems[J]. Expert Systems with Applications, 2014, 41(4): 2065-2073.

[3] 王道平, 李志隆, 杨岑. 基于热门物品惩罚和用户兴趣变化的知识推送算法[J]. 系统工程, 2014(1): 118-123.
Wang Daoping, Li Zhilong, Yang Cen. Knowledge push algorithm based on hot item punishment and user interest change[J]. Systems Engineering, 2014(1): 118-123.

[4] Krzywicki A, Wobcke W, Kim Y S, et al. Collaborative filtering for people-to-people recommendation in online dating: Data analysis and user trial[J]. International Journal of Human-Computer Studies, 2015, 76: 50-66.

[5] Scholz M, Dorner V, Franz M, et al. Measuring consumers´ willingness to pay with utility-based recommendation systems[J]. Decision Support Systems, 2015, 72: 60-71.

[6] Li L, Zheng L, Yang F, et al. Modeling and broadening temporal user interest in personalized news recommendation[J]. Expert Systems with Applications, 2014, 41(7): 3168-3177.

[7] Felfernig A, Jeran M, Ninaus G, et al. Toward the next generation of recommender systems: Applications and research challenges[M]. Berlin: Springer International Publishing, 2013: 81-98.

[8] Wang P, Zhou T, Han X P, et al. Modeling correlated human dynamics with temporal preference[J]. Physica A: Statistical Mechanics & Its Applications, 2014, 398(4): 145-151.

[9] Barabási A. The origin of bursts and heavy tails in human dynamics[J]. Nature, 2005, 435(7039): 207-211.

[10] 孙光福, 吴乐, 刘淇, 等. 基于时序行为的协同过滤推荐算法[J]. 软件学报, 2013(11): 2721-2733.

Sun Guangfu, Wu Le, Liu Qi, et al. Recommendations based on collaborative filtering by exploiting sequential behaviors[J]. Journal of Software, 2013(11): 2721-2733.

[11] Bao H, Li Q, Liao S S, et al. A new temporal and social PMF-based method to predict users' interests in micro-blogging[J]. Decision Support Systems, 2013, 55(3): 698-709.

[12] Koren Y. Collaborative filtering with temporal dynamics[J]. Communications of the ACM, 2010, 53(4): 89-97.

[13] Feng H, Tian J, Wang H J, et al. Personalized recommendations based on time-weighted overlapping community detection[J]. Information & Management, 2015, 52(7): 789-800.

[14] Bobadilla J, Ortega F, Hernando A, et al. Recommender systems survey[J]. Knowledge-Based Systems, 2013, 46(1): 109-132.

[15] 周涛, 韩筱璞, 闫小勇, 等. 人类行为时空特性的统计力学[J]. 电子科技大学学报, 2013, 42(4): 481-540.
Zhou Tao, Han Xiaopu, Yan Xiaoyong, et al. Statistical mechanics on temporal and spatial activities of human[J]. Journal of University of Electronic Science and Technology of China, 2013, 42(4): 481-540.

[16] Jalali M, Mustapha N, Sulaiman M N, et al. WebPUM: A web-based recommendation system to predict user future movements[J]. Expert Systems with Applications, 2010, 37(9): 6201-6212.

[17] 田儒雅, 刘怡君, 牛文元. 舆论超网络的领袖引导模型[J]. 中国管理科学, 2014, 22(10): 136-141.
Tian Ruya, Liu Yijun, Niu Wenyuan. Leader-guiding model of online opinion supernetwork[J]. Chinese Journal of Management Science, 2014, 22(10): 136-141.

[18] Liu Y, Li Q, Tang X, et al. Superedge prediction: What opinions will be mined based on an opinion supernetwork model [J]. Decision Support Systems, 2014, 64(3): 118-129.

[19] 李兴华, 陈冬林, 杨爱民, 等. 基于用户兴趣-标签的混合推荐方法研究[J]. 情报学报, 2015(5): 466-470.
Li Xinghua, Chen Donglin, Yang Aimin, et al. A study of mixed recommendation method based on user interest-tag[J]. Journal of the China Society for Scientific and Technical Information, 2015(5): 466-470.

Method of interest points prediction based on customer web temporal behavior trajectory

CHEN Donglin, XIA Qi, DAI Siguang

Research Center for E-Business and Intelligent Services, Wuhan University of Technology, Wuhan 430070, China

Abstract Interest point is the key to improving the accuracy of e-commerce recommendation under big data environment. However, the existing predictive research ignores the comprehensive impact of various customers' behaviors and time series on interest points. In order to make up this gap, the article sets up a customer Web space and time super network model which involves four subnets: customer, time, behavior and interest point, and establishes the influence factors of behavior. Then, based on similarity of superlink prediction method and the establishment of connectivity matrix, the adjacency matrix is calculated and super triangle judgement is made, so that the most similar super edge and the best prediction results of interest points are obtained. Finally, experiment shows that the precision of interest prediction gets better with the decrease of time accuracy within the allowable range of time error. Compared with the traditional method of label prediction, the prediction accuracy is improved from 56.2% to 74%.

Keywords interest points prediction; Web time-space behavior; super network; superedge similarity

（责任编辑　韩星明）