计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2017.20150822
54(2):405-414, 2017

GeoPMF:距离敏感的旅游推荐模型

张 伟 韩林玉 张佃磊 任鹏杰 马 军 陈竹敏
(山东大学计算机科学与技术学院 济南 250101)
(will_zhang2014@outlook.com)

GeoPMF: A Distance-Aware Tour Recommendation Model

Zhang Wei, Han Linyu, Zhang Dianlei, Ren Pengjie, Ma Jun, and Chen Zhumin
(*School of Computer Science and Technology, Shandong University, Jinan* 250101)

Abstract Although people can use Web search engines to explore scenic spots for traveling, they often find it very difficult to discover the sighting sites which match their personalized need well. Tour recommendation systems can be used to solve the issue. A good tour recommendation system should be able to provide personalized recommendation and take the time and cost factors into account. Furthermore, our investigation shows that often a user u will consider the distance between her/his habitual residence and the tour destination when she/he makes her/his travel plan. It is because that the travel distance reflects the effect of time and cost indirectly. Therefore, we propose a distance-aware tour recommendation model, named GeoPMF (geographical probabilistic matrix factorization), which is developed based on the Bayesian model and PMF (probabilistic matrix factorization). The main idea of GeoPMF is that for each user we try to get a most preferred travel distance span by mining her past tour records. Then we use it as a kind of weight factors added into the traditional PMF model. Experiments on travel data of Ctrip show that, our new method can decrease *RMSE* (root mean square error) nearly 10% compared with some baseline methods. And when compared with the traditional PMF model, the average decline on *RMSE* is nearly 3.5% in virtue of the distance factor.

Key words tour recommendation; recommender system; probabilistic matrix factorization (PMF) model; distance-aware; GeoPMF

摘 要 虽然目前旅游者可以利用Web搜索引擎来选择旅游景点,但往往难以获得较好符合自身需要的旅游规划.而旅游推荐系统是解决上述问题的有效方式.一个好的旅游推荐模型应具有个性化并能考虑用户时间和费用的限制.调研表明,用户在选择旅游景点时,目的地与用户常居地的距离常常是一个需要考虑的问题.因为旅行距离往往可以间接地反映了时间和费用的影响.于是,在贝叶斯模型和概率矩阵分解模型的基础上,提出一个旅行距离敏感的旅游推荐模型(geographical probabilistic matrix factorization, GeoPMF).主要思想是基于每个用户的旅游历史,推算出一个最偏好的旅游距离,并作为

收稿日期:2015-09-15;**修回日期**:2015-12-22
基金项目:国家自然科学基金项目(61272240,61672322);山东省自然科学基金项目(ZR2012FM037);微软国际合作基金项目(FY14-RES-THEME-25)
This work was supported by the National Natural Science Foundation of China (61272240, 61672322), the Natural Science Foundation of Shandong Province (ZR2012FM037), and the Microsoft International Cooperation Fund Project (FY14-RES-THEME-25).
通信作者:马军(majun@sdu.edu.cn)

一种权重,添加到传统的基于概率矩阵分解的推荐模型中.在携程网站的旅游数据集上的实验表明,与基准方法相比,GeoPMF 的 *RMSE*(root mean square error)可以降低近 10%;与传统概率矩阵分解模型(PMF)相比,通过考虑距离因子,*RMSE* 平均降幅近 3.5%.

关键词 旅游推荐;推荐系统;概率矩阵分解模型;距离敏感;GeoPMF 算法

中图法分类号 TP301

近年来,旅游已成为人们娱乐消遣的重要方式.据国家统计局网站发布的《2014 年国民经济和社会发展统计公报》① 显示,2014 年全年,我国出国游的人数达 1 亿人次,国内游达 36 亿人次.旅游已成为推荐系统[1]的重要应用领域之一.目前国内携程、途牛和去哪儿网等旅游网站收集了大量的用户反馈数据,为用户对景点的选择提供了依据.显然,若能通过旅游推荐系统,为用户提供更具个性化的推荐,将会极大地提高推荐系统的可用性.

关于旅游推荐已有不少工作.Ge 等人[2]认为旅行花费对景点选择有重要的影响,这里花费包括费用和时间.他们把旅行花费表示为一个〈时间,资金〉二元组.对于每个旅游者,都对应一个〈时间,资金〉二元组,用以表示用户的预期偏好;对于每个景点,也有一个〈时间,资金〉二元组,视为每个景点的固有属性.然后利用贝叶斯模型,将这 2 个二元组作为评分预测概率的先验条件进行建模,给出旅游推荐.在结合地理因素方面,Tobler[3]在对基于位置的社交网络(LBSN)的研究中,通过对用户移动设备 GPS 信息的记录,发现了一种签到地点的空间聚类现象[3],即个人游览地点趋向于聚在一起.在兴趣点(point-of-interest, POI)推荐的研究中,Ye 等人[4]提出了一种结合用户社交行为和地理因素的推荐模型,该模型是基于传统的协同过滤算法中对相似度的计算,首先找到与用户兴趣最近邻的 K 个用户,将这 K 个用户对该景点评分的加权平均作为评分的预测,只是在计算权值的时候结合了社交和地理信息.在考虑地理因素时,Ye 等人通过分析 Foursquare 和 Whrrl 数据集,也发现了空间聚类现象.进一步地,Ye 提出了一种指数模型来建模签到概率与距离的关系,并利用签到概率来计算新的权值.最终,该模型提高了兴趣点推荐的准确率.然而,这种模型不能很好地解决数据稀疏性问题,当有新数据加入时,还要重新计算权值.而且该模型需要计算每个用户去过的地点两两之间的距离,增大了计算量.Horozov 等人[5]提出一种基于权重的矩阵分解模型来解决这一问题.在用户特征向量和兴趣点特征向量的基础上,他们提出了用户活动区域矩阵和兴趣点影响力矩阵.指出兴趣点的影响力表现在用户到过某个景点再去周围景点的概率,是一种与距离有关的二维正态分布形式.Horozov 的模型是利用用户的签到信息,不包含用户的反馈打分,初始待分解矩阵中的元素是用户对每个景点的签到频次.

已有的研究大多是利用用户对地点的签到数据.利用签到的频次作为待分解的矩阵中的元素,或者将签到与否描述为一个布尔变量,利用形成的 0-1 矩阵计算用户相似度.这些方法利用的信息过少;在推荐上考虑用户的反馈不足;之前基于距离的推荐大多是景点之间的实地距离,而不是景点与用户之间的距离,个性化不强.针对上述问题,本文利用用户常居地到各个景点的距离这一地理信息,结合贝叶斯模型[6-7],提出一种针对旅游景点的推荐算法,即距离敏感的旅游推荐模型(geographical probabilistic matrix factorization, GeoPMF).其主要思想是基于每个用户的旅游历史,推算出一个最偏好的旅游距离,并作为一种权重添加到传统的基于概率矩阵分解的推荐模型中.我们模型中的目标函数是一个具有连续性的凸函数,能够利用随机梯度下降快速地训练模型.在携程网站的旅游数据集上的实验表明,与基准方法相比,GeoPMF 的 *RMSE*(root mean square error)可以降低近 10%;与传统概率矩阵分解模型(PMF)相比,通过考虑距离因子,*RMSE* 平均降幅近 3.5%.

1 基于距离因子的旅游推荐模型

1.1 GeoPMF 模型基本框架

较之于传统的推荐领域,如电影[8-10]、音乐[11-13]、在线商店[14],旅游推荐数据稀疏性问题更加严重.其主要原因在于用户旅游的频度较小.相对影视、音乐等活动,旅游的花费通常偏高,使得用户旅游的次数大大低于传统推荐领域的行为频次.我们将携程

① http://www.stats.gov.cn/tjsj/zxfb/201502/t20150226_685799.html

网站数据的统计结果与其他领域的数据集进行了对比分析,如表 1 所示.可以看出,对于前 4 个数据集,最稀疏的是 Ciao 数据集,其打分矩阵取值为空的元素占了 99.97%;相比而言,携程数据更加稀疏,仅是 Ciao 的 40%.

Table 1　The Sparsity Comparison Between Ctrip and Other Datasets

表 1　携程数据集与其他数据集稀疏性的比较

Dataset	User Number	Item Number	Rating Number	Sparsity /%
Epinions	40 163	139 738	664 824	0.051
FilmTrust	1 508	2 071	35 497	1.14
Flixster	53 213	18 197	409 803	0.04
Ciao	7 375	99 746	280 391	0.03
Ctrip	283 952	20 688	723 732	0.012

为了解决稀疏性问题,GeoPMF 采用矩阵分解的思路,并将距离因素考虑进来.在选择旅游景点时,用户会考虑景点与自身所在地之间距离的可接受范围.对于每一个用户,我们将景点划归为不同的距离区段,比如在 10 km 范围、10～20 km 范围等等,每一个距离区段用户选择的概率有差异;而且对每一个用户来说,都有一个最偏好的距离区段.GeoPMF 正是将这 2 个距离区段引入矩阵分解模型.图 1 给出本模型的实现方法.首先,我们经过数据预处理操作,从携程旅游数据中得到用户对景点的打分矩阵;然后,利用百度 LBS 开放平台根据景点地理信息获得其 GPS 信息,并计算每个用户-景点对之间的距离,得到距离区段矩阵;最后,将这 2 个矩阵作为 GeoPMF 模型的输入,通过随机梯度下降法训练出模型参数,最终输出用户预测评分矩阵.

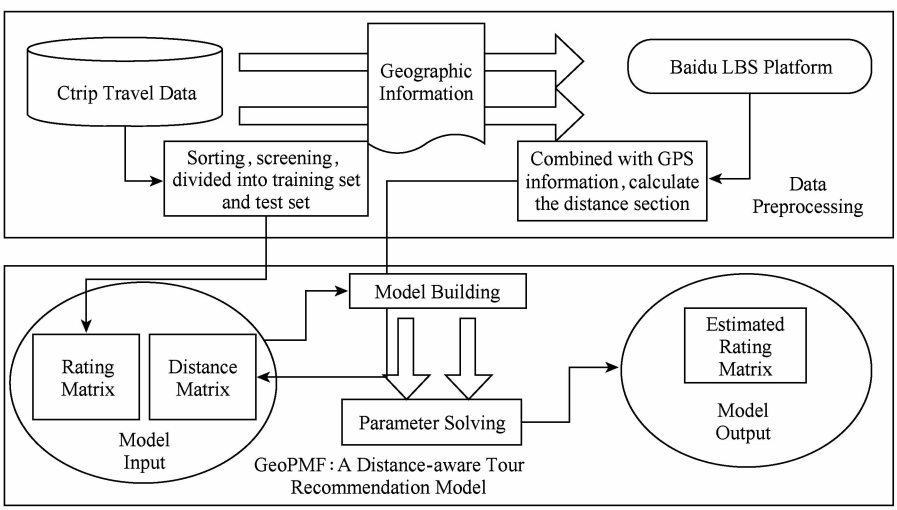


Fig. 1　The framework of GeoPMF model

图 1　GeoPMF 模型框架

1.2　距离对景点选择影响的研究

本文在携程网旅游数据中随机选取了部分用户,在地图上标注他们的旅游目的地,结果如图 2 所示.图 2 中用不同颜色的图标区分不同用户的旅游

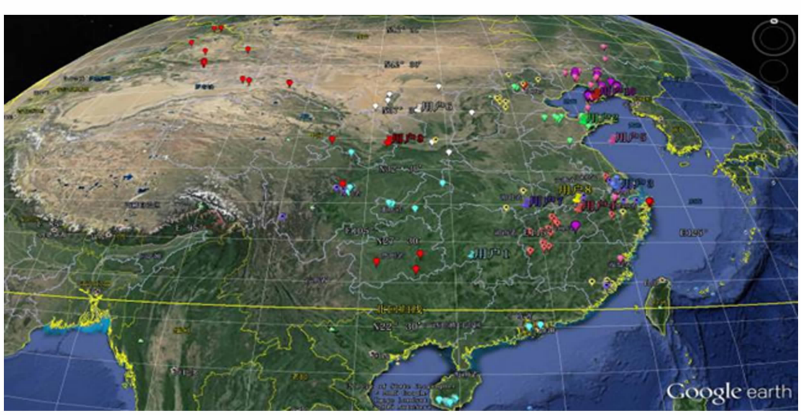


Fig. 2　Users' tourism destination spots distribution on Ctrip website

图 2　携程网站不同用户的旅游景点位置分布

历史,图钉用来标识用户的常居地.这些信息都是从携程网的旅游评论记录中获得.就旅游历史与用户常居地的相对距离来看,不同用户的行为差异较大.有些用户偏向仅去距离常居地较近的景点,如用户2、用户3和用户10.而像用户1、用户9,却偏向选择较远的景点.

基于对旅游行为的观察,本文对该现象给出的解释是,用户选择景点之前,首先对要去的距离区段有一个基本的定位.前面提到的Ye等人[4]利用指数模型对景点实地距离与选择景点的概率进行了建模,但由于该模型本身具有计算概率值复杂、不能解决稀疏性等缺点,因此本文尝试通过新的方式对二者关系进行建模.首先,基于上述解释,我们认为景点所处的区段比实地距离更有考虑价值,鉴于此,在获取用户景点的经纬度信息后,我们计算出每个用户与去过的景点的距离,然后按照10 km为单位为这些景点进行区段划分.本文对不同区段内旅游数量统计处理,结果如图3所示.横坐标表示不同的距离区段,纵坐标是在每个区段内旅游频数.从图3中可以看出,用户在不同区段内旅游频次与距离区段有明显相关性.

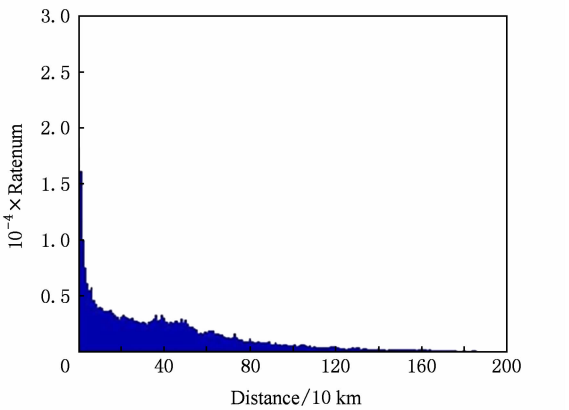


Fig. 3 The tourism frequency histogram in different distance sections

图3 不同距离区段内旅游频次直方图

然而,景点对用户的吸引力不仅在于旅游频次,还在于用户的评分,用户对景点的评分高低说明用户对该景点的喜欢程度.为了描述用户对不同区段景点的偏好,我们定义了一个概率函数,见式(1):

$$P(c)=\frac{\sum_{i,j\in\kappa}I_{ij}(c)\times r_{i,j}}{\sum_{i,j\in\kappa}r_{i,j}},\tag{1}$$

其中,c表示距离区段编号,以10 km为单位;κ表示用户i去过的景点集合;$I_{ij}(c)$为指示函数,当景点j位于用户i的第c个区段时为1,否则为0;r_{ij}是用户i对景点j的评分.我们用$P(c)$来估计用户对不同距离区段的喜好程度.统计结果如图4所示.横坐标为不同距离区段;纵坐标表示用户选择该区段的概率,即$P(c)$.从中看出,用户对不同区段内景点的喜好程度与距离区段也存在明显的相关性.而且,总体而言,用户更喜好距离较近的景点.

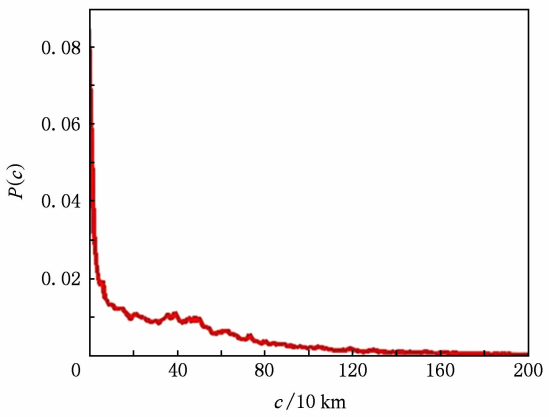


Fig. 4 The probability distribution of user preference with different tourist attractions

图4 用户对不同景点偏好的概率分布

经过上述统计分析,我们得出结论:景点所处的距离区段不仅对用户旅行目的地的选择有重要影响,也间接地影响了用户对去过景点的反馈评分.本文假设每个用户在旅游的时候心中有一个最偏爱的距离区段即d_i,它与景点对应的距离区段D_{ij}之间的偏差越小,用户选择的概率越大,给较高评分的概率也越大.因此,在1.3节中,我们将2个距离因子:用户最偏爱的距离区段c和表示景点属性的距离区段矩阵$\boldsymbol{D}$作为考虑因素,建立一个对旅行距离敏感的旅游推荐模型GeoPMF.

1.3 GeoPMF模型的形式化

GeoPMF将景点相对于每个用户所处的距离区段作为考虑因素.为此,本文引入距离区段矩阵$\boldsymbol{D}$,其中每一个元素D_{ij}表示相对于用户i的常居地来说,景点j所处的距离区段.用户i最偏爱的距离区段记为d_i.接着,我们将S_{ij}引入到矩阵分解模型中.S_{ij}表示用户i最偏爱区段d_i与景点j所处区段D_{ij}的相似度,取值范围是[0,1].区别于传统矩阵分解,我们对评分矩阵的分解见式(2):

$$\hat{R}_{ij}=S_{ij}\odot(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j).\tag{2}$$

图5是这种分解的图示.$\odot$表示Hadamard乘积算子,2个矩阵的Hadamard乘积含义如下:设$\boldsymbol{C}_{m\times n}=\boldsymbol{A}_{m\times n}\odot\boldsymbol{B}_{m\times n}$则$C(i,j)=A(i,j)\cdot B(i,j)$;

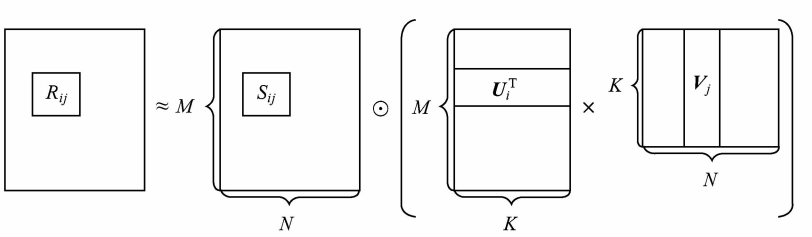


Fig. 5　Rating matrix decomposition of GeoPMF

图 5　GeoPMF 的评分矩阵分解

$\hat{\boldsymbol{R}}$ 是预测评分矩阵,维数为 $M\times N$;$\boldsymbol{S}$ 是距离区段相似度矩阵,维数为 $M\times N$;$\boldsymbol{U}$ 为用户特征矩阵,$\boldsymbol{V}$ 为景点特征矩阵,潜在因子数 K 决定了特征向量的维数.

设评分的估计值与真实值之间存在误差为 ε,并假设 ε 服从高斯分布,则

$$R_{ij}\sim\mathcal{N}(S_{ij}\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j,\sigma^2),\tag{3}$$

其中 $\mathcal{N}(R_{ij}\mid\mu,\sigma^2)$ 是满足均值为 μ、方差为 σ^2 的高斯分布.

S_{ij} 的定义基于以下思想:对于用户去过的景点,所处的距离区段 D_{ij} 与 d_i 的差值会影响用户的反馈评分,二者偏差越小,用户给高分的可能性越大;对于用户没有去过的景点,D_{ij} 与 d_i 偏差越小,用户选择该景点作为旅游目的地的可能性也越大.因此,可采用欧氏距离来计算相似度,见式(4).对于每一个 S_{ij},表示用户最偏爱距离区段 d_i 与景点所处距离区段 D_{ij} 的近似程度,值越大,二者越近似,用户选择该景点的概率越高.

$$S_{ij}=S(d_i,D_{ij})=1-\|d_i-D_{ij}\|^2.\tag{4}$$

根据极大似然估计的思想,假设 R_{ij} 之间是独立同分布的,我们得到用户评分矩阵的似然函数为式(5):

$$p(\boldsymbol{R}\mid\boldsymbol{U},\boldsymbol{V},d,\boldsymbol{D},\sigma_{\boldsymbol{R}}^2)=\prod_{i=1}^{M}\prod_{j=1}^{N}[\mathcal{N}(R_{ij}\mid S(d_i,D_{ij})\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j,\sigma_{\boldsymbol{R}}^2)]^{I_{ij}},\tag{5}$$

其中 $\mathcal{N}(R_{ij}\mid S(d_i,D_{ij})\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j,\sigma_{\boldsymbol{R}}^2)$ 是 R_{ij} 的概率分布,符合均值为 $S(d_i,D_{ij})\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j$、方差为 $\sigma_{\boldsymbol{R}}^2$ 的高斯分布.I_{ij} 是指示函数,当用户 i 对景点有评分时,即 R_{ij} 存在时,其值为 1,否则为 0.为了防止过拟合,假设 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 服从高斯先验分布.基于最大化贝叶斯后验的思想,得到 $\boldsymbol{U},\boldsymbol{V},d$ 的后验形式,见式(6):

$$p(\boldsymbol{U},\boldsymbol{V},d\mid\boldsymbol{R},\boldsymbol{D},\sigma_{\boldsymbol{R}}^2,\sigma_{\boldsymbol{U}}^2,\sigma_{\boldsymbol{V}}^2)\propto p(\boldsymbol{R}\mid\boldsymbol{U},\boldsymbol{V},d,\boldsymbol{D},\sigma_{\boldsymbol{R}}^2)p(\boldsymbol{U}\mid\sigma_{\boldsymbol{U}}^2)p(\boldsymbol{V}\mid\sigma_{\boldsymbol{V}}^2),\tag{6}$$

其中 $p(R\mid\boldsymbol{U},\boldsymbol{V},d,\boldsymbol{D},\sigma_{\boldsymbol{R}}^2)$ 为似然函数,见式(5);$p(\boldsymbol{U}\mid\sigma_{\boldsymbol{U}}^2)$,$p(\boldsymbol{V}\mid\sigma_{\boldsymbol{V}}^2)$ 是 $\boldsymbol{U},\boldsymbol{V}$ 的先验概率.最终,通过最大化对数后验求得 $\boldsymbol{U},\boldsymbol{V}$ 和 d,见式(7):

$$\begin{aligned}(\boldsymbol{U}^*,\boldsymbol{V}^*,d^*)&=\max_{\boldsymbol{U},\boldsymbol{V},d}\ln p(\boldsymbol{U},\boldsymbol{V},d\mid\boldsymbol{R},\boldsymbol{D},\sigma_{\boldsymbol{R}}^2,\sigma_{\boldsymbol{U}}^2,\sigma_{\boldsymbol{V}}^2)=\\&-\frac{1}{2\sigma_{\boldsymbol{R}}^2}\sum_{i=1}^{M}\sum_{j=1}^{N}I_{ij}(R_{ij}-f(\boldsymbol{U}_i,\boldsymbol{V}_j,d_i,D_{ij}))^2-\\&\frac{1}{2\sigma_{\boldsymbol{U}}^2}\sum_{i=1}^{M}\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{U}_i-\frac{1}{2\sigma_{\boldsymbol{V}}^2}\sum_{j=1}^{N}\boldsymbol{V}_j^{\mathrm{T}}\boldsymbol{V}_j-\\&\frac{1}{2}\Big[\Big(\sum_{i=1}^{M}\sum_{j=1}^{N}I_{ij}\Big)\ln\sigma_{\boldsymbol{R}}^2+M\ln\sigma_{\boldsymbol{U}}^2+N\ln\sigma_{\boldsymbol{V}}^2\Big]+C,\end{aligned}\tag{7}$$

其中,C 是一个与参数无关的常量.

使上述目标函数最大化,等价于最小化公式:

$$\begin{aligned}E=&\frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{N}I_{ij}(R_{ij}-S(d_i,D_{ij})\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j)^2+\\&\frac{\lambda_{\boldsymbol{U}}}{2}\sum_{i=1}^{M}\|\boldsymbol{U}_i\|_{\mathrm{F}}^2+\frac{\lambda_{\boldsymbol{V}}}{2}\sum_{j=1}^{N}\|\boldsymbol{V}_j\|_{\mathrm{F}}^2,\end{aligned}\tag{8}$$

其中 $\lambda_{\boldsymbol{U}}=\dfrac{\sigma_{\boldsymbol{R}}^2}{\sigma_{\boldsymbol{U}}^2}$,$\lambda_{\boldsymbol{V}}=\dfrac{\sigma_{\boldsymbol{R}}^2}{\sigma_{\boldsymbol{V}}^2}$.

式(8)就是 GeoPMF 最终的目标函数.我们利用随机梯度下降法(stochastic gradient descent,SGD)学习得到参数 $\boldsymbol{U},\boldsymbol{V},d$.

GeoPMF 的概率模型图如图 6(b).较之于模型 PMF(图 6(a)),本文在预测评分时,引入距离因子 d_i 和距离区段矩阵 $\boldsymbol{D}$.

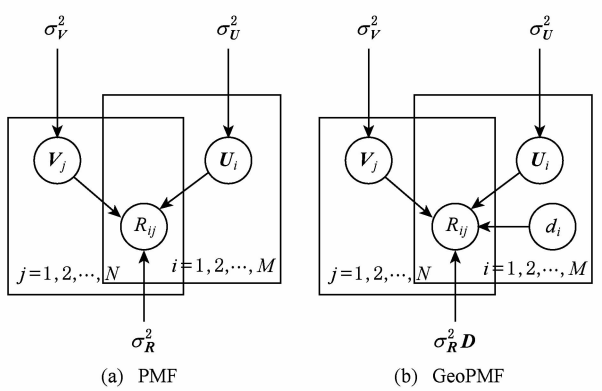


Fig. 6　Graphical models

图 6　图模型

2 实 验

2.1 数据集

1) 携程网旅游数据. 本文实验数据集采用携程网旅游攻略的用户评论信息. 数据集包含用户节点283 952 个、景点节点 20 688 个、用户打分 723 732 个,见表 1 所示.

2) 获取地理信息. 根据景点节点的名称信息,使用百度地图提供的开放 API,生成景点以及用户常居地的经纬度坐标. 距离选取 10 km 为步长,每 10 km 表示一个区段. 我们计算了每个用户常居地到他去过的景点之间的距离,确定景点所属的距离区段用以形成距离区段矩阵 $\boldsymbol{D}$.

3) 生成训练集测试集. 本文采用按时间分割的方式划分测试集训练集,见图 7 所示. 首先,去掉评论次数少于 3 条的用户的所有评分数据;然后,按照每个用户评论时间的顺序对评分数据排序;最后,按照 2:1的比例将每个用户前 2/3 的评分作为训练集,剩余的作为测试集,并且对于训练集中的每个用户,保证在测试集中至少有一个评分数据.

经过数据处理,我们最终得到 3 个数据文件:训练集文件(xctour_train. txt)、测试集文件(xtour_test. txt)和距离区段文件(distance_section. txt). 训练集和测试集所包含用户数、景点数以及评分数等统计信息,见表 2. 距离区段文件保存了每个用户去过的所有景点所属的距离区段信息,共包含 300 677 个距离区段数据.

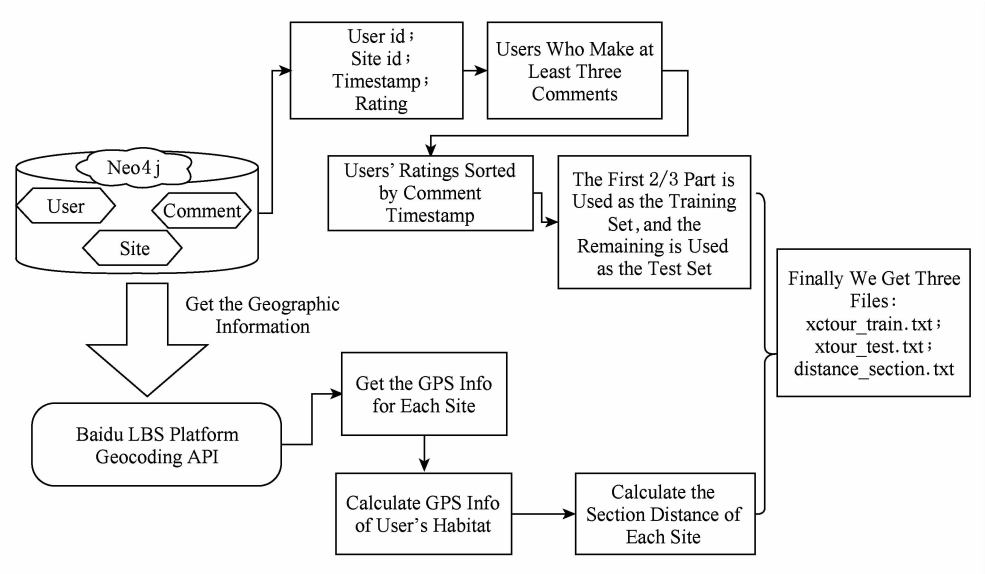


Fig. 7 Preprocessing on Ctrip dataset

图 7 携程数据集预处理

Table 2 Statistics of Ctrip Dataset

表 2 携程数据集统计信息

Dataset	File Size/MB	User Number	Site Number	Rating Number or Section Number	Min Score or Min Distance	Max Score or Max Distance	Set Ratio /%
Training Set	2.46	31 408	17 177	193 541	1	5	64.37
Test Set	1.37	31 408	20 451	107 136	1	5	35.63
Section File	4.07	31 408	20 588	300 677	0	458	

2.2 基准方法

1) GlobalAverage. 用户评分矩阵所有真实值的平均值作为评分预测值.

2) ItemAverage. 对某一景点的评分等于该景点收到的所有评分的平均值.

3) SVD. 传统 SVD 算法[15]利用了数学中奇异值分解的思想. 其评分估计值等于用户特征向量与景点特征向量的乘积,即 $\hat{R}_{ij}=\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j$,损失函数是根据最小化 $RMSE$ 的目标得到,最终优化的损失函数为式(9):

$$E=\frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{N}I_{ij}(R_{ij}-\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j)^2, \tag{9}$$

SVD 是一种最基本的矩阵分解模型.

4) PMF. 由 Salakhutdinov 等人[16]首先提出,其概率模型图见图 6(a). 他假设预测评分与真实评分之间存在高斯噪声,并假设 $\boldsymbol{U},\boldsymbol{V}$ 满足均值为 0 的高斯分布. 最终得到的损失函数为式(10):

$$E=\frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{N}I_{ij}(R_{ij}-\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j)^2+\frac{\lambda_U}{2}\sum_{i=1}^{M}\|\boldsymbol{U}_i\|_{\mathrm{F}}^2+\frac{\lambda_V}{2}\sum_{j=1}^{N}\|\boldsymbol{V}_j\|_{\mathrm{F}}^2. \tag{10}$$

5) SocialMF. 由 Jamali 和 Ester[17]提出,将社交网络中的信任关系结合到矩阵分解中,其目标函数形式为式(11):

$$E=\frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{N}I_{ij}(R_{ij}-\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j)^2+\frac{\lambda_U}{2}\sum_{i=1}^{M}\|\boldsymbol{U}_i\|_{\mathrm{F}}^2+\frac{\lambda_V}{2}\sum_{j=1}^{N}\|\boldsymbol{V}_j\|_{\mathrm{F}}^2+\frac{\lambda_T}{2}\sum_{i=1}^{M}\Big(\Big\|\boldsymbol{U}_i-\sum_{v\in N_i}T_{i,v}\boldsymbol{U}_v\Big\|_{\mathrm{F}}^2\Big), \tag{11}$$

其中,$\boldsymbol{T}$ 表示信任关系矩阵,当用户 v 关注用户 i 时,$T_{i,v}=1$;N_i 表示用户 i 所关注的其他用户的集合. 通过加入信任关系这一特征,Jamali 和 Ester 通过实验证明该方法能显著降低 *RMSE*. 在携程旅游数据中也能够取得用户之间的关注信息,而且 GeoPMF 和 SocialMF 都是以矩阵分解为基础,区别在于选取的上下文信息以及建模形式不同,因此我们将 SocialMF 也作为比较对象进行实验.

上述所有的推荐算法都在我们处理过的携程训练集 xctour_train.txt 上进行实验.

2.3　评价指标

在推荐领域,评价一个推荐算法预测评分的好坏,常用的评价指标是 *RMSE*,用来表示估计评分的误差,定义为式(12):

$$RMSE=\sqrt{\frac{\sum_{ij}(R_{ij}-\hat{R}_{ij})^2}{|Test|}}, \tag{12}$$

其中,R_{ij} 是用户 i 对景点 j 的评分,$\hat{R}_{ij}$ 是用户 i 对景点 j 的预测评分,$|Test|$ 是测试集中评分的数量.

2.4　参数设置

PMF,SVD 正规项 $\lambda_U=\lambda_V=0.001$,GeoPMF 正规项设置为 $\lambda_U=\lambda_V=0.01$. d 的每一项利用景点距离区段均值进行初始化,即 d_i 初值为 $\boldsymbol{D}$ 对应行向量元素的均值. 矩阵 $\boldsymbol{U},\boldsymbol{V}$ 中元素取值服从均值为 0、标准差为 0.1 高斯分布.

2.5　结果比较

1) GeoPMF 与基准方法及传统矩阵分解的比较. 考虑特征向量 $\boldsymbol{U}_i$ 和 $\boldsymbol{V}_j$ 的维数 K,即潜在因子数会对结果造成影响,我们设置了不同的特征向量维数进行实验,得到图 8 中的结果. 最下面的一条线是 GeoPMF 的结果. 总体来看,矩阵分解方法要比基准方法效果好. 基准方法 GlobalAverage 和 ItemAverage 是直接利用均值进行预测,所以 *RMSE* 并不发生变化,在图 8 中表现为直线. 而 PMF 和 SVD 区别仅在于正规项的加入,所以 2 条曲线几乎一致. 在每个维度下,GeoPMF 的结果都要优于其他方法. 横向来看,对于 GeoPMF,SocialMF,SVD 来说,随着特征向量维数的增加,*RMSE* 先减少后增加,均在维数为 5 达到最优. 随着特征向量维数的增加,GeoPMF 的结果与 PMF 和 SVD 之间差距逐渐增大. 当特征向量维数为 5 时,*RMSE* 降低幅度近 1%,在达到稳定状态时,*RMSE* 降低幅度达到 5%. SocialMF 的 *RMSE* 在特征向量维数为 5 时达到最优,但最优值也要稍差于 GeoPMF,且维数继续增加时,*RMSE* 剧烈升高,SocialMF 实验结果恶化. 最终实验结果显示,较之于基准方法,GeoPMF 的 *RMSE* 平均降幅为 9%,最优值降幅为 10%;较之于矩阵分解方法 PMF 和 SVD,*RMSE* 平均降幅为 3.5%,最优值降幅为 1%.

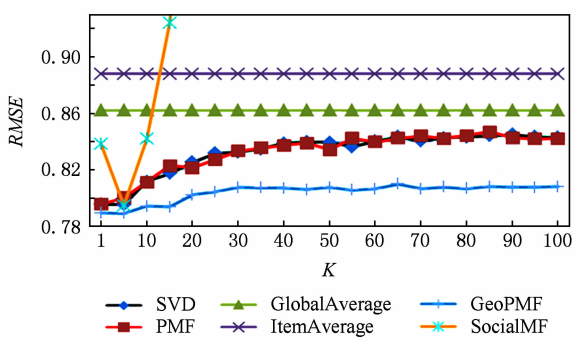


Fig. 8　Impact of dimensionality K on *RMSE*

图 8　特征向量维数 K 对 *RMSE* 的影响

虽然从上述实验结果我们看到 GeoPMF 模型的优越性,但是为了验证 GeoPMF 实验结果是真正优于基准方法,还是因为优化过程的随机初始化等导致的性能提高,本文对图 8 中实验结果进行了显著性检验[18]. 我们对 PMF 和 GeoPMF 的实验数据进行显著性分析,表 3 是对 2 组数据进行独立 T 检验的结果. 从结果中看出,显著性为 0.005,说明二者方差存在显著性差异,在方差不等的情况下,双尾显著性为 0.000;而当显著性小于 0.05 时,认为配对样本之间存在显著差异,即后测与前测之间存在显著差异,说明 GeoPMF 对于 *RMSE* 的降低效果显著.

Table 3 T-test Result in SPSS

表 3 SPSS T-检验结果

GeoPMF	Variance Equality Test		T-test Mean Value Equality						
	F	Significant	t	Degrees of Freedom	2-tailed Significant	Mean Difference	Standard Error Difference	95% Confidence Interval of Difference	
								Lower Limit	Upper Limit
Homogeneity of Variance	8.486	0.005	−14.408	78	0.000	−0.031 418	0.002 180	−0.035 759	−0.027 077
Non Homogeneity of Variance			−14.408	55.911	0.000	−0.031 418	0.002 180	−0.035 787	−0.027 050

接着,我们比较不同算法 *RMSE* 随迭代次数的变化. 根据上述实验结果,我们将特征向量维数固定于 5. 实验结果如图 9 所示. 从图 9 可看出,GeoPMF 效果也要优于其他推荐算法,当算法收敛时,*RMSE* 达到 0.79,较之于基准方法和 PMF 分别有 10% 和 1% 的提高,并且也稍优于 SocialMF 方法. 总体来看,随着迭代次数的增加,GeoPMF 的 *RMSE* 不断降低,收敛后较之于 PMF 和 SVD,更加稳定. 另外,可以看出,而 SVD 由于没有引入正规项,当迭代次数达到 30 时,*RMSE* 出现上升趋势,说明存在过拟合现象.

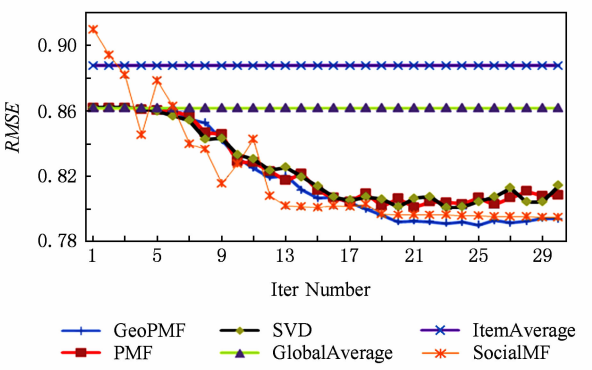


Fig. 9 Impact of iter number on *RMSE* ($K=5$)

图 9 迭代次数对 *RMSE* 的影响($K=5$)

2) 距离区段可视化. $\boldsymbol{d}$ 是在模型假设中定义的区段向量,其中的每一个元素 d_i 代表用户最偏好距离。我们通过随机梯度下降学习矩阵 $\boldsymbol{U}$,$\boldsymbol{V}$ 的同时,也学习得到 $\boldsymbol{d}$. 为了直观地展示距离区段这一距离因子,我们对 $\boldsymbol{d}$ 的学习结果和用户已经去过的景点区段进行了可视化分析,如图 10 所示. 横坐标表示随机选取的 13 位用户. 每一位用户对应纵轴的一列散点集合,我们用 D_u 表示与用户对应的一列点集. 其中,每一列的每一个星型符号表示用户去过的景点所属距离区段即 D_{ij},菱形表示 GeoPMF 模型学习得到的用户最偏好区段 d_i. 注意,在训练开始前,$\boldsymbol{d}$ 中元素是用 $\boldsymbol{D}$ 中对应的每一行距离区段均值进行初始化的. 从图 10 中看出,在训练结束后,菱形落在星型符号集中分布的区域周围,即 $\boldsymbol{d}$ 更加靠近用户最常去的距离区段,这与人们的经验一致.

3) 模型效率. 表 4 是对矩阵分解算法运行时间的统计结果. 从表 4 可看出,GeoPMF 运行时间较之于 PMF 和 SVD 有所增加. 由于算法引入距离区段矩阵,并且在学习过程中要同时学习距离区段向量 $\boldsymbol{d}$,使得性能相对 PMF 和 SVD 来说有所降低. 但这种运行时间的增加相对于 *RMSE* 的降低来说是在可接受范围之内的. 而 SocialMF 的运行时间较之于 GeoPMF 增加了近 3 倍,且从前面的实验结果看,GeoPMF 的实验结果也要稍优于 SocialMF,这也更加体现了 GeoPMF 的优越性.

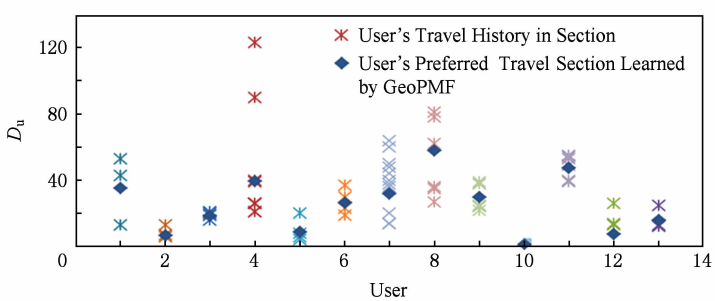


Fig. 10 An illustration of distance section

图 10 景点距离区段的图示

Table 4　The Runtime of Recommendation Algorithms

表 4　推荐算法运行时间

Recommender Algorithm	Runtime/ms	Speedup to PMF
PMF	402	1
SVD	591	1.47
GeoPMF	1031	2.56
SocialMF	3857	9.59

3　总结及未来工作

本文中,我们对携程网旅游数据进行统计分析,证明景点所处的距离区段在旅游目的地选择中是一个重要的考虑因素.据此,我们提出了一种基于距离因子的旅游推荐模型 GeoPMF,从矩阵分解的角度研究了旅游推荐算法,目的是降低评分估计误差.我们结合 PMF,将用户最偏爱距离区段和景点实际所处的距离区段作为考虑条件,纳入概率分解模型.这样做的好处是,我们就既考虑用户对景点本身的偏好,同时考虑了用户对距离区段的偏好.在最终的实验结果中,*RMSE* 降低到 0.79.通过与基准方法的比较,证明了 GeoPMF 对降低 *RMSE* 有显著效果.同时,GeoPMF 对用户旅游景点的选择上也有一定指导意义.

在未来的工作中,我们会将 GeoPMF 应用于其他旅游网站的数据以及其他包含地理信息的数据集,用来验证该模型的适应性.另外,我们的 GeoPMF 也有一定局限性,首先,我们模型选择用户的常居地是一个定值,在现实生活中,用户的地理位置往往伴随着迁徙行为,比如一个用户常居地从一个省份到另一个省份;其次,当用户到达一个景点进行旅游时,常常会对所在目的地的周边景点也产生兴趣.另外,除了考虑物理距离,还应考虑交通的便利性.对于以上情况,我们会以 GeoPMF 为基础,结合景点选择中的各种影响因素,提出一种更具泛化能力的模型,为旅游者的行程做出更好的规划.

参　考　文　献

[1] Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions [J]. IEEE Trans on Knowledge and Data Engineering, 2005, 17(6): 734-749

[2] Ge Y, Liu Q, Xiong H, et al. Cost-aware travel tour recommendation [C] //Proc of the 17th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2011: 983-991

[3] Tobler W. A computer movie simulating urban growth in the detroit region [J]. Economic Geography, 1970, 46: 234-240

[4] Ye M, Yin P, Lee W C, et al. Exploiting geographical influence for collaborative point-of-interest recommendation [C] //Proc of the 34th ACM SIGIR Int Conf on Research and Development in Information Retrieval. New York: ACM, 2011: 325-334

[5] Horozov T, Narasimhan N, Vasudevan V. Using location for personalized poi recommendations in mobile environments [C] //Proc of the Int Symp on Applications Internet. Los Alamitos, CA: IEEE Computer Society, 2006: 625-636

[6] Ji Junzhong, Liu Chunnian, Sha Zhiqiang. Bayesian belief network model learning, inference and applications [J]. Computer Engineering and Applications 2003, 39(5): 24-27 (in Chinese)
(冀俊忠, 刘椿年, 沙志强. 贝叶斯网模型的学习、推理和应用[J]. 计算机工程与应用, 2003, 39(5): 24-27)

[7] Cheng Lanlan, He Pilian, Sun Yueheng. Study on Chinese keyword extraction algorithm based on naive Bayes model [J]. Journal of Computer Applications, 2005, 25(12): 2780-2782 (in Chinese)
(程岚岚, 何丕廉, 孙越恒. 基于朴素贝叶斯模型的中文关键词提取算法研究[J]. 计算机应用, 2005, 25(12): 2780-2782)

[8] Lekakos G, Caravelas P. A hybrid approach for movie recommendation [J]. Multimedia Tools & Applications, 2008, 36(1/2): 55-70

[9] Biancalana C, Gasparetti F, Micarelli A, et al. Context-aware movie recommendation based on signal srocessing and machine learning [C] //Proc of the 2nd Challenge on Context-Aware Movie Recommendation. New York: ACM, 2011: 5-10

[10] Mirza B J, Keller B J, Ramakrishnan N. Studying recommendation algorithms by graph analysis [J]. Journal of Intelligent Information Systems, 2003, 20(2): 131-160

[11] Cano P, Koppenberger M, Wack N. Content-based music audio recommendation [C] //Proc of the 13th Annual ACM Int Conf on Multimedia. New York: ACM, 2005: 211-212

[12] Chen H, Chen A L P. A music recommendation system based on music data grouping and user interests [C] //Proc of the 10th Int Conf on Information and knowledge Management. New York: ACM, 2001: 231-238

[13] Li Ruimin, Lin Hongfei, Yan Jun. Mining latent semantic on user-tag-item for personalized music recommendation [J]. Journal of Computer Research and Development, 2014, 51(10): 2270-2276 (in Chinese)
(李瑞敏, 林鸿飞, 闫俊. 基于用户-标签-项目语义挖掘的个性化音乐推荐[J]. 计算机研究与发展, 2014, 51(10): 2270-2276)

[14] Lee K C, Kwon S. Online shopping recommendation mechanism and its influence on consumer decisions and behaviors: A causal map approach [J]. Expert Systems with Applications, 2008, 35(4): 1567-1574

[15] Koren Y. Factorization meets the neighborhood: A multifaceted collaborative filtering model [C] //Proc of the 14th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2008: 426-434

[16] Salakhutdinov R, Mnih A. Probabilistic matrix factorization [C/OL] //Proc of the Advances in Neural Information Processing Systems. 2007: 1257-1264 [2015-11-16]. http://papers.nips.cc/paper/3208-probabilistic-matrix-factorization.pdf

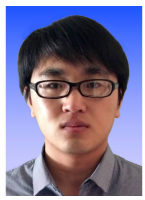
[17] Jamali M, Ester M. A matrix factorization technique with trust propagation for recommendation in social networks [C] //Proc of the 4th ACM Conf on Recommender Systems. New York: ACM, 2010: 135-142

[18] Zhou Yuzhu, Jiang Fenghua. The regression analysis of the experimental DATAS and the remarkable examination [J]. Physical Experiment of College, 2001, 14(4): 43-46 (in Chinese)
(周玉珠, 姜奉华. 实验数据的一元线性回归分析及其显著性检验[J]. 大学物理实验, 2001, 14(4): 43-46)

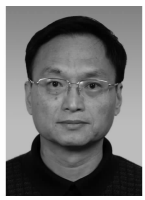
Zhang Wei, born in 1993. PhD candidate in Shandong University. Student member of CCF. His main research interests include information retrieval, tweet summarization and recommender system.

Han Linyu, born in 1992. Master candidate in Shandong University. Student member of CCF. Her main research interests include information retrieval, Web data mining and recommender systems (zhangdianlei11@gmail.com).

Zhang Dianlei, born in 1993. Master candidate in Shandong University. Student member of CCF. His main research interests include information retrieval, data mining and recommender systems (zhangdianlei11@gmail.com).



Ren Pengjie, born in 1990. PhD candidate in Shandong University. Student member of CCF. His main research interests include information retrieval, data mining.



Ma Jun, born in 1956. Professor and PhD supervisor in Shandong University. Senior member of CCF. His main research interests include information retrieval, data mining, parallel computing, natural language processing.

Chen Zhumin, born in 1977. Associate professor and master supervisor in Shandong University. Senior member of CCF. His main research interests include Web information retrieval, data mining, and social computing (chenzhumin@sdu.edu.cn).