计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(6):236−239 [doi: 10.15888/j.cnki.csa.006380]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于动态用户画像的信息推荐研究①

刘　勇[1], 吴翔宇[1], 解本巨[2]

[1](青岛科技大学 信息科学技术学院, 青岛 266061)
[2](青岛惠泽创建信息新技术有限公司, 青岛 266000)

摘　要: 针对传统信息推荐方式精度偏低的问题, 引入用户画像作为推荐基础, 在深入研究文本分类和用户行为后, 提出一种基于动态用户画像的推荐方法. 该方法通过动态分析用户历史数据, 预测用户的兴趣变化趋势, 从而实现动态推荐. 离线实验证明, 该方法在预测用户偏好变化方面具有一定优势, 相较于传统的基于标签的信息推荐, 提高了推荐精度.

关键词: 信息推荐; 用户画像; 用户行为

引用格式: 刘勇,吴翔宇,解本巨.基于动态用户画像的信息推荐研究.计算机系统应用,2018,27(6):236–239. http://www.c-s-a.org.cn/1003-3254/6380.html

Research on Information Recommendation Based on Dynamic User Portrait

LIU Yong[1], WU Xiang-Yu[1], XIE Ben-Ju[2]

[1](Information Science and Technology Academy, Qingdao University of Science and Technology, Qingdao 266061, China)
[2](Qingdao HZCJ New Information and Technology Co. Ltd., Qingdao 266000, China)

Abstract: To solve the problem about low accuracy of traditional information recommendation method, this paper introduces the user portrait as the recommended basis, and proposes the recommendation method based on dynamic user portraits after further studying about the text classification and user behavior. By dynamically analyzing the user's historical data and predicting the user's interest trends, this method achieves dynamic recommendations. The off-line experiment improves that this method has some advantages in predicting user preference changes compared with the traditional label-based information recommendation, and it improves the recommendation accuracy.

Key words: information recommendation; user portrait; user behavior

“信息过载 (information overload)”[1]效应是互联网高速发展的副产物, 为了解决“信息过载”问题, 研究人员提供了两种方案, 一种是搜索引擎, 另一种为推荐系统. 前者通过检索关键字对海量信息进行筛选, 列出用户可能用到的信息列表供用户选择; 后者通过综合分析用户历史数据 (兴趣、行为)、用户所在场景、环境等, 把用户最为感兴趣的内容主动推送给用户, 相较于前者, 推荐系统的普适性、智能化程度以及精准度都存在一定的优势[2]. 现如今推荐技术已经成为学术研究的热点之一, 在社交网络、电子商务、广告投放等诸多领域独占鳌头[3].

目前主流的推荐方法可以分为: 基于内容的推荐 (content-based recommendation)、基于协同过滤的推荐 (collaborative filtering-based recommendation)、基于知识的推荐 (knowledge-based recommendation) 以及组合推荐 (hybrid recommendation)[4]. 针对应用场景和环境的不同可以选择不同的推荐方式.

在网络信息推荐领域, 用户和信息交互只存在浏览行为, 并没有对项目进行评分, 用户的兴趣偏好隐含在浏览历史当中, 所以必须通过分析用户行为来挖掘

① 收稿时间: 2017-09-23; 修改时间: 2017-10-28; 采用时间: 2017-11-02; csa 在线出版时间: 2018-05-28

用户的兴趣[5]. 传统的基于标签的信息推荐, 通过分析用户浏览记录给用户打上“兴趣-权重”标签然后进行推荐, 在一定程度上忽略了用户兴趣的变化趋势, 随着时间的推移, 推荐精度往往会降低, 影响用户体验.

所以, 要想提高信息推荐的精度, 需要一种可以随时间增长, 动态更新用户推荐候选项的方法, 基于这个设想, 本文提出一种基于动态用户画像的方法, 该方法建立在文本分类和用户行为动态分析的基础上, 后续相关实验证明了该方法的可行性.

1 文本处理和分类

在进行文本信息的推荐之前, 首先要对推荐集合中的文本进行处理. 本文使用支持向量机 (SVM) 文本分类法[6,7]对文本进行分类处理, 其核心思想是在 n 维向量空间内寻找一个最优分类的超平面, 表示为:

$$\omega^{\mathrm{T}}x+b=0$$

SVM 文本分类的主要步骤大致可以分为:文本特征提取、文本特征表示、文本分类.

1.1 文本特征提取

首先对文本进行分词, 然后对文本中的停用词和单字词进行过滤. 计算给定文本中词语的词频:

$$tf_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

其中, $n_{i,j}$为该词在文本d_j中出现的次数, 分母为总词语量. 选取特征之前, 假设每个特征具有独立性, 然后选取若干词频较高的词语作为该文本的特征词集合, 将文本表示为 n 维特征向量:

$$X=(x_1,x_2,x_3,\cdots,x_n)$$

其中, n为选取词语的个数.

1.2 文本特征表示

特征词的权重使用 TF-IDF 公式进行计算,公式如下所示:

$$w_{ik}=\frac{tf_{ik}*\log(N/n_k+0.01)}{\sqrt{\sum_{i=k}[tf_{ik}*\log(N/n_k)+0.01]^2}} \tag{2}$$

其中, tf_{ik}表示词t_k在文本中的词频, N为所有的文本, n_k为包含特征词t_k的文本数.

1.2.1 归一化处理

因为文本长度偏差会影响特征词的权重计算, 所以要进行归一化处理, 将选定的特征词权重规范到一定区间内, 其公式为:

$$tf^*=\frac{tf-\min}{\max-\min} \tag{3}$$

其中, tf^*为标准化后的词频, min为该特征词在所有文本中的最小词频, max为最大词频.

1.3 文本分类

通过上述过程计算就可以得到文本的 n 维特征向量:

$$V=(t_1,w_1(d);t_2,w_2(d);t_3,w_3(d);\cdots;t_n,w_n(d)) \tag{4}$$

其中, t_i表示特征词, $w_i(d)$为该词在文本中的权重, n为特征词的个数.

由上述步骤就可以将文本量化为可进行计算的数据结构, 然后通过相似度计算就可以确定目标文本的所属分类, 经文本处理和分类后的文档集可以表示为如下的向量集:

$$documentCollection\{document_i(a_1,a_2,a_3,\cdots,a_n)|i,n\in N^*\}$$

其中, $documentCollection$为向量余弦相似度较高的一类文本集合, N^*为不含 0 的自然数集.

2 动态用户画像

2.1 用户画像

用户画像是通过综合分析用户数据, 抽象出的一个可代表用户各项维度的标签模型, 其中维度一般包括: 人口统计学维度、兴趣维度和商业维度, 形式如图 1 所示.

图 1 用户画像

在本文的研究场景中, 用户浏览过程一般是匿名的, 所以可得到的用户人口统计学维度信息较少, 而用户的浏览历史数据相对容易获取, 这使得建立用户兴趣模型成为可能.

2.2 引入动态用户画像

随着时间的推移, 用户浏览历史的文本集合不断

扩大, 用户的兴趣标签权重会发生变化, 如果不能更新兴趣标签的权重, 会导致推荐精度下降, 所以必须通过一定的方式更新用户的兴趣标签集.

为了方便表示用户兴趣维度, 本文采用一个 n 维元组[8]表示用户兴趣标签集, 其形式表示为:

$$UserPortrait =< label_1, label_2, label_3, \cdots, label_n > \quad (5)$$

其中$label_i$代表用户兴趣标签.

用户兴趣维度的 n 维特征向量可表示为:

$$U = (p_1, p_2, p_3, \cdots, p_n) \quad (6)$$

设用户兴趣标签集为查询向量:

$$U = queryVector = \{u_1, u_2, \cdots, u_n\} | n = 1, 2, \cdots \quad (7)$$

通过计算查询向量和推荐集中的文档向量的相似度, 来决定是否将文本推荐给目标用户. 本文选用余弦相似度来度量两向量的相似程度:

$$\cos(\vec{w}_d, \vec{w}_q) = \sum_{i=1}^{k} w_{i,d}, w_{i,q} / (\sqrt{\sum_{i=1}^{k} w^2{}_{i,d}} \sqrt{\sum_{i=1}^{k} w^2{}_{i,q}}) \quad (8)$$

随着用户的浏览集合不断扩大, 能代表用户兴趣的标签的权重会随时间推移而发生变化, 通过动态的分析用户的浏览行为, 在一定程度上可以预测用户的兴趣变化.

为了更好的预测用户的兴趣的变化趋势, 提高推荐的精度, 本文使用贝叶斯动态线性模型对用户兴趣维度进行预测, 下面给出模型定义:

观测方程: $y_t = F_t' \theta_t + v_t, v_t \sim N[0, V]$

状态方程: $\theta_t = G_t \theta_{t-1} + w_t, w_t \sim N[0, W_t]$

初始先验: $(\theta_0 | D_0) \sim N[m_0, C_0]$

其中, y_t为t时刻的观测值;θ_t为未知的状态向量;F_t为已知的n维向量, 用来描述观测数据和状态之间的关系; v_t为观测误差值;w_t为状态误差值, 且v_t和w_t相互独立.

对用户浏览集合U进行采样,

时间间隔为: $t = t_i - t_{i-1}$

观测序列为: $y_{t_0}, y_{t_1}, y_{t_2}, \cdots, y_{t_n}$

由模型定义给出其一步预测和后验分布:

(1) $t-1$时刻的后验分布: 对于均值m_{t-1}和方差矩阵C_{t-1}有$(\theta_{t-1} | D_{t-1}) \sim N[m_{t-1}, C_{t-1}]$

(2) t时刻的先验分布为: $(\theta_t | D_{t-1}) \sim N[a_t, R_t]$, 其中 $a_t = G_t m_{t-1}, R_t = G_t C_{t-1} G_t' + W_t$

(3) 一步预测分布: $(y_t | D_{t-1}) \sim N[f_t, Q_t]$, 其中 $f_t = F_t' a_t, Q_t = F_t' R_t F_t + V_t$

(4) t时刻的后验分布: $(\theta_{t-1} | D_{t-1}) \sim N[m_t, C_t]$, $m_t = a_t + A_t e_t$, $C_t = R_t - A_t A_t' Q_t$, 其中$A_t = R_t F_t Q_t^{-1}$, $e_t = y_t - f_t$

通过后验信息不断修正先验信息, 求得预测值, 根据预测值更新用户画像兴趣维度的标签权重, 从而更新用户兴趣集U,由公式 (5) 计算更新后的兴趣集与文档向量的余弦相似度, 当相似度大于 0.6 时将该文本信息推荐给用户.

3 实验分析

3.1 实验数据

本文通过网络爬虫抓取某信息分享平台 200 用户的交互信息, 其中包括个人主页信息、关注人 URL、收藏、个人动态、关注领域等信息. 为了高效的提取网页文本信息, 使用行块分布函数对网页文本内容进行抽取.

本文分别对传统的基于标签的信息推荐 (Based-on Label Recommendation, BLR) 和基于动态用户画像的推荐 (Based-on Dynamic User Portrait Recommendation, BDUPR) 进行实验, 前者直接通过分析全部的用户浏览集合进行用户的偏好计算, 后者通过对用户浏览集进行分时段采样, 动态计算用户的兴趣偏好.

3.2 评价标准

预测准确度有 3 类[9]: 评分预测准确度评测、使用预测准确度评测、物品排名预测准确度评测. 在本文的应用场景中, 推荐并不预测用户对项目的偏好 (评分), 而是用户是否点击 (收藏、关注等) 被推荐的信息, 所以选用“使用预测准确度评价”作为本文方法的评价标准.

为用户推荐的内容可能有下列几种情况:

表 1 结果分类

	被推荐	未被推荐
浏览	真阳性数 (tp)	假阴性数 (fn)
没有浏览	假阳性数 (fp)	真阴性数 (tn)

通过统计上表数值, 计算如下比率:

查准率: $precision = \frac{\#tp}{\#tp + fp}$

查全率: $recall = \frac{\#tp}{\#tp + \#fn}$

在实际推荐中, 用户的浏览量十分有限且推荐集中的文本数量较多, 所以这里选用查准率作为验证标准.

3.3 实验结果

推荐列表长度对多用户平均查准率具有较大的影

响, 如果取值太则小无法说明推荐方法的可行性, 取值过大会造成结果难以预估[2](查准率可能偏大也可能偏小). 图 2 为推荐列表长度与查准率的关系:

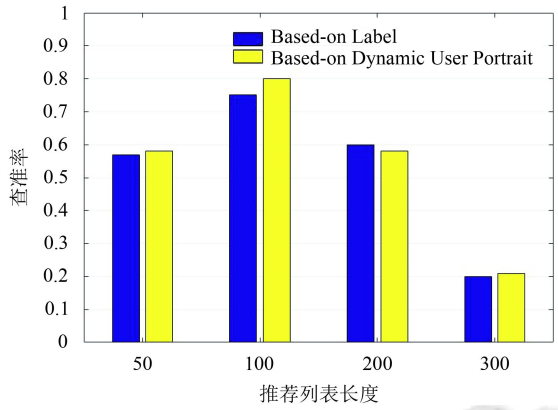

图 2 查准率

测试列表长度的实验条件是优化过的, 所以查准率可能较高. 经过多次试验最后设定列表长度为 100, 那么两种推荐方式的查准率和时间序列的关系如图 3 所示.

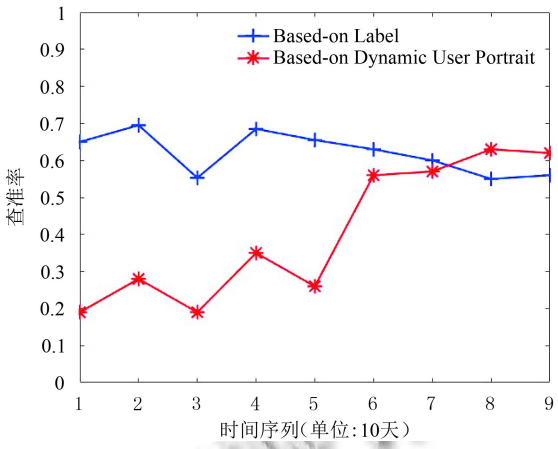

图 3 查准率对比

从图中可以看出,在时间序列的一开始, BDUPR 的查准率并不理想, 这是因为训练集数量较小, 所以查准率偏低, 但随着时间序列的增长, 训练集不断增加, 成上升趋势, 开始趋于平稳. BLR 在推荐的一开始具有较高的查准率, 这是因为 BLR 在推荐进行之前就将全部的用户历史数据量化, 用于构建用户的兴趣模型, 其文档训练集数据要多于 BDUPR 的分时段采样训练集合, 所以在靠前的推荐周期当中, BLR 的推荐准确率要高于 BDUPR, 但随着时间推移, 用户兴偏好会发生一定的变化, 致使 BLR 的推荐准确率降低.

两种方法在进行实验过程中都存在一定的数据波动, 可能有以下几个原因: 用户群体中部分用户的兴趣点差距很大, 导致平均查准率偏低; 由于网络信息更新速度较快, 用户兴趣波动较大.

4 结束语

本文研究的基于动态用画像的推荐在捕捉用户兴趣变化方面较静态推荐有一定优势, 但在新异推荐方面还存在不足: 建立用户兴趣模型的数据全部来源于用户和服务器的历史交互数据, 对于那些用户之前从未接触过的信息, 存在冷启动问题.

参考文献

1 王国霞, 刘贺平. 个性化推荐系统综述. 计算机工程与应用, 2012, 48(7): 66–76.
2 吴丽花, 刘鲁. 个性化推荐系统用户建模技术综述. 情报学报, 2006, 25(1): 55–62.
3 孟祥武, 刘树栋, 张玉洁, 等. 社会化推荐系统研究. 软件学报, 2015, 26(6): 1356–1372. [doi: 10.13328/j.cnki.jos.004831]
4 许海玲, 吴潇, 李晓东, 等. 互联网推荐系统比较研究. 软件学报, 2009, 20(2): 350–362.
5 史艳翠, 戴浩男, 石和平, 等. 一种基于时间戳的新闻推荐模型. 计算机应用与软件, 2016, (6): 40–43. [doi: 10.3969/j.issn.1000-386x.2016.06.010]
6 王正鹏, 谢志鹏, 邱培超. 语义关系相似度计算中的数据标准化方法比较. 计算机工程, 2012, 38(10): 38–40. [doi: 10.3969/j.issn.1007-130X.2012.10.008]
7 张征杰, 王自强. 文本分类及算法综述. 电脑知识与技术, 2012, 8(4): 825–828, 841.
8 王智囊. 基于用户画像的医疗信息精准推荐的研究[硕士学位论文]. 成都: 电子科技大学, 2016.
9 Ricci F, Rokach L, Shapira B, *et al*. Recommender Systems Handbook. 2nd ed. 北京: 机械工业出版社, 2015