

移动端个性化旅游推荐系统模型设计

陈传敬, 陈琳

(山东科技职业学院, 山东 潍坊 261053)

摘要: 本文在基于现有的个性化推荐研究基础, 提出了用户情境和文本内容组合推荐的方法, 并构建了用户兴趣矩阵模型, 设计了一套个性化推荐系统。本系统可以依据移动用户的偏好为用户提供推送服务。该系统由移动端及云平台服务器构成, 移动端对用户信息进行采集, 并展示推荐数据; 服务器端收集来自移动端的数据, 展开数据分析, 构建用户兴趣模型, 结合文本处理技术获取推荐内容, 推送给移动用户。

关键词: 文本处理; 个性化推荐; 情景; 兴趣矩阵

中图分类号: TP311 **文献标识码:** A

Model Design of the Personalized Tourism Recommendation System Based on Mobile Terminal

CHEN Chuanjing, CHEN Lin

(Shandong Vocational College of Science and Technology, Weifang 261053, China)

Abstract: Based on the previous studies of personalized recommendation, this paper proposes a combined recommendation method of user context and text content, constructs a user interest matrix model, and designs a personalized recommendation system. This system can provide push service for users according to their preferences. The system is composed of mobile terminal and cloud platform server, in which the mobile terminal collects user information and displays recommendation data. The server collects data from mobile terminal, launches data analysis, builds user interest model, obtains recommendation content with text processing technology, and pushes it to mobile users.

Keywords: text processing; personalized recommendation; scenario; interest matrix

1 引言(Introduction)

随着移动用户需求的不断提升, 现有的个性化推荐逐渐不能满足用户。因此, 将个性化推荐与移动网络技术的结合成为优化的选择, 研究个性化推荐, 提取用户偏好及个人信息、推算服务矩阵模型、文本相似度匹配、优化相似度算法有助于提高用户认可度。

与桌面计算机相比, 在移动环境中工作的移动电话, PAD等设备有着高实时性和方便性的特点, 但局限于屏幕尺寸、网络传输及信息处理能力。移动用户的位置和周边的环境都有可能随时间变化。因此需要提取移动用户情景信息数据学习, 深入探索移动用户兴趣点, 并过滤不符合用户需求的信息, 提高服务移动用户的质量。

Kbaier^[1]等人采用混合推荐方法; LeiShi^[2]等使用决策树创建用户组, 向用户提供推荐; Chiliguano^[3]等提出卷积神经

网络提取事物特征; Qiudan Li等人提出了基于协同过滤的移动商务服务, 采用了多维矩阵模型; 基于Android平台开发的AppBrain通过Android API接口来检测用户近期安装的应用, 并查找同类应用向用户推荐; 基于iPhone平台开发的一款应用AppsFire, 采用了基于社会化信息流的方法进行推荐。

本文研究提出了移动环境下个性化推荐矩阵模型系统(Personalized Recommendation Matrix Model System), 并对该系统的结构、框架、核心技术等进行研究。系统获取位置等情景信息, 混合内容推荐个性化推荐技术分析客户行为, 构建用户偏好矩阵模型, 提高智能化推荐的性能。

2 相关工作(Related research)

个性化推荐系统是智能化的用于挖掘有效数据的平台。推荐系统可分为三个部分^[4], 用户模型构建、推荐矩阵模型构建、推荐运用。用户模型构建中, 采集信息方式有隐式浏览

信息、显式浏览信息、关键字和行为属性、用户评分、用户书面评价、用户历史关注。推荐矩阵模型构建可采用多种推荐算法，常用的推荐算法有基于内容的推荐、基于协同过滤的推荐及多种推荐技术结合的推荐等技术。

基于内容的推荐^[5]，根据浏览历史构建用户兴趣模型，对项目内容特征提取构建项目特征模型，分析得出用户兴趣模型和项目特征模型之间的匹配度，根据匹配程度做出推荐。其特征集是文本关键词的集合，进行文本词频率及其逆文档频率的数据统计。其优点是效率高，不存在数据稀疏性或冷启动。缺点是不利于特征集提取，无法处理复杂的文本特征，对音频、视频信息无法特征推荐。

基于协同过滤的推荐^[6]，有相同兴趣的用户或物品构成集合，根据相似程度进行推荐。如果一个物品是和当前用户有着很大兴趣相似度的用户所喜欢的物品，那么当前用户也往往会喜欢该物品；如果用户想购买某件物品，而另一件物品和这一物品有很大相似性，那用户也会购买另一物品。

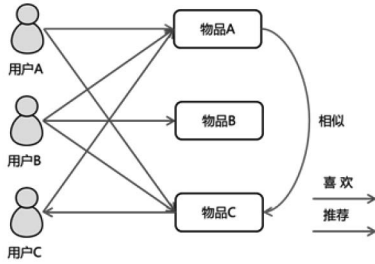


图1 协同过滤原理图

Fig.1 Principle diagram of collaborative filtering

协同过滤算法的优点是，可以用于文字内容推荐，也可以用于视频、音乐、相片等非文本内容，用途广。缺点是存在数据稀疏性、冷启动及扩展性问题。

针对上述问题本文提出了基于移动用户情境和文本内容组合推荐的方式，弥补上述缺点，提高推荐系统精确性、稳定性。

3 信息的收集处理(Information collection and processing)

移动信息收集处理包括用户信息、项目信息、情景信息。用户信息可以是用户的喜好或特征；项目信息是系统中待推荐项目附加的各种属性；情景信息综合了用户的周围环境，如当前位置、天气、时间。通过提取用户的个人信息、项目信息和周围情景，并交服务器模型处理，对用户潜在喜好推测，达到推荐目的。

移动用户信息包括基本信息和行为信息。每个用户基本信息可以用一维矩阵表示， $P_i = \{Id_i, Name_i, Sex_i, Age_i, Edu_i, Married_i, DESC_i, IMEI_i\}$ ， P_i 代表第*i*个用户的信息矩阵，矩阵中存储用户编码、姓名、性别、年龄、教育程度、婚否、描述、移动终端码。行为信息是指用户对项目的浏览次数、点击趋势、地理坐标。用户行为信息一维矩阵表示为 $A_i = \{Id_i, Click_i, Favorites_i, Share_i, Time_i, Latitude_i, Longitude_i\}$ ，

A_i 代表第*i*个用户的行为矩阵，矩阵中存储用户编码、点击次数、是否收藏、是否分享、查阅时间、经度、维度，行为信息能体现用户兴趣。

项目信息是将被推荐的项目的特征属性的信息，覆盖面广，可以是文本、歌曲、视频或者是一件商品。项目属性会因项目而异，如餐厅的推荐，项目信息是指餐厅的种类、菜系、价格等因素；如旅游地点，项目信息是门票价格、景点地点、优惠时间、开放时间等；如音乐推荐，项目信息是作曲家名称、歌手名称、歌曲风格等。本文以旅游项目为例建立项目特征矩阵模型。

$$\text{Project} = \begin{Bmatrix} p_1^1 & p_1^2 & p_1^3 & \dots & p_1^j \\ p_2^1 & p_2^2 & p_2^3 & \dots & p_2^j \\ p_3^1 & p_3^2 & p_3^3 & \dots & p_3^j \\ p_4^1 & p_4^2 & p_4^3 & \dots & p_4^j \\ p_5^1 & p_5^2 & p_5^3 & \dots & p_5^j \end{Bmatrix} = \{p_1^1, p_1^2, p_1^3, \dots, p_1^j\} = \begin{Bmatrix} p_1^j \\ p_2^j \\ p_3^j \\ \dots \\ p_5^j \end{Bmatrix}$$

项目信息模型矩阵为用户二维空间矩阵，横轴*i*表不同的兴趣领域，如吃、住、娱、游、购；数轴*j*代表*i*领域下不同的实体项目，实体项目又包含众多特征属性。

$$P_i^j = \{Id_i^j, Name_i^j, Picture_i^j, Price_i^j, Traffic_i^j, OpenTime_i^j, Tele_i^j, Desc_i^j\}$$

实体特征矩阵包括，实体编码、名称、图片信息、价位、交通状态、开放时间、联系方式、描述，实体项目矩阵将项目信息模型延伸为三维矩阵模型，作为项目特征集。

情境是用户附近环境的相关信息的集合体，可以是时间、天气、用户位置、服务的网络、移动终端，且能对用户决策造成影响。本文采用时间、位置情景信息构建情境模型。

$$\text{Context}_i = \{Id_i, UserId_i, Dtime_i, Latitude_i, Longitude_i\}$$

情境模型为一维矩阵模型，包括编码信息、用户编码、时间、经度、维度。可记录用户浏览信息时间、位置等信息。本文采用移动蜂窝网络基站定位来确定用户的位置，时间信息和终端状况可直接通过移动用户终端获取，将这些情景信息通过移动网络传递给服务器端。

4 用户模型的构建(Construction of user model)

本文的用户模型综合考虑了用户、文本和情景三个要素，采用矩阵、用户兴趣度及向量空间模型方式结合表示。该模型能够分析用户的历史情景，得到个体行为模型，对历史浏览项目学习，将其文本信息作为训练语料，提取项目文本特征词并建立特征词矩阵，用以标识特征词及其对应的概率关系，通过建立的特征词矩阵计算用户与待推荐项目文本的相似概率，进行相似度计算得到用户对待推荐项目的兴趣度，从待推荐项目中找到合适的推荐项目。综合分析情景和用户的兴趣度，二者结合构成本模型。

4.1 情境计算

情景可以认为由*n*个情景属性构成，定义成：

$$\text{Context}(C) = (C_{1c}, C_{2c}, \dots, C_{ic}, \dots, C_{nc})$$

上式中 $C_{1c}, C_{2c}, \dots, C_{ic}, \dots, C_{nc}$ 分别对应着情景 $C_1, C_2, \dots, C_i, \dots, C_n$

的属性值。

用户当前状态的情景和历史情景的相似值可定义成:

$$\text{Sim}(\text{Context}(C), \text{Context}(H)) = \frac{\text{Count}(C_{ic}) + \dots + \text{Count}(C_{ic}) + \dots + \text{Count}(C_{nc})}{n * N}$$

$\text{Count}(C_{ic})$ 代表在用户以往的情景信息 $\text{Context}(H)$ 里, 情景属性值的数目。

4.2 文本特征提取

每个实体项目都有对应的文本内容, 可对其特征提取获取特征词和特征词权, 并重构成的向量空间的模型。

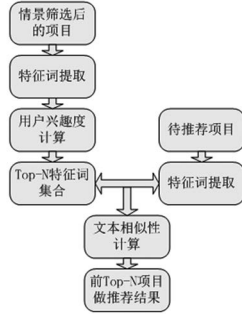


图2 基于文本处理

Fig.2 The text-based treatment

一篇文档 D_i 可以表示成:

$$D_i = \{(t_{i1}, w_{i1}), (t_{i2}, w_{i2}), \dots, (t_{ij}, w_{ij}), \dots, (t_{im_i}, w_{im_i})\}$$

文档 D_i 是第 i 个项目的文本内容, t_{ij} 是第 i 个项目文本中的第 j 个特征词, w_{ij} 是第 i 个项目中的第 j 个特征词在当前文本 D_i 中所占权重值, m_i 是文档 D_i 中的所含有的特征词的数目。

采用TF-IDF方法统计特征词权重的式子定义为:

$$w(t, \bar{d}) = \frac{tf(t, \bar{d}) * \log(N / n_t + 0.01)}{\sqrt{\sum_{w \in \bar{d}} [tf(w, \bar{d}) * \log(N / n_w + 0.01)]^2}}$$

其中, $\bar{d}$ 是文本, t 是一个词语, $w(t, \bar{d})$ 为 t 在 $\bar{d}$ 中的所占的权重, $tf(t, \bar{d})$ 为 t 在 $\bar{d}$ 里的词频, N 是所有应用到的训练文本的总的数目, n_t 为所有文本中有 t 出现的文本数目, 也可以说成是 t 的文档频度。

用户在多次使用移动旅游客户端后, 其浏览过的项目文本的集合可以定义为:

$$D_i = \{(t_{i1}, w_{i1}), (t_{i2}, w_{i2}), (t_{i3}, w_{i3}), \dots, (t_{ij}, w_{ij}), \dots, (t_{im_i}, w_{im_i})\}$$

用户浏览或点击过的第 i 个项目文本为记为 D_i , 其中 t_{ij} 用来表示第 i 个项目文本中出现的第 j 个特征关键词, w_{ij} 表示第 j 个特征关键词在项目文本 D_i 中所占有的权重, 还有 m_i 代表在项目文本 D_i 的特征关键词的总数。

项目文本集合的 D 的特征词权重矩阵可以表示成:

$$D = \begin{bmatrix} w_{11} & \dots & w_{1m_1} \\ \dots & w_{ij} & \dots \\ w_{n1} & \dots & w_{nm_n} \end{bmatrix}$$

上述矩阵中, w_{ij} 代表第 i 个文本中的第 j 个特征词的权重值的大小。

4.3 兴趣度计算

用户的兴趣度可定义为:

$BI = \alpha * s(P) + \beta * c(P) + \chi + \xi$, 其中 ξ 服从正太分布 $N(0, \sigma^2)$ 。

BI 为与 $s(P)$ 、 $c(P)$ 相关的随机变量, $s(P)$ 和 $c(P)$ 分别代表用户对项目的浏览次数和评分的多少, α 、 β 、 χ 是和 $s(P)$ 、 $c(P)$ 不相干的参数, ξ 代表着随机产生的误差, 该方程为一个多线的正态线性回归方程。

将用户兴趣度 BI 运用到文本权重矩阵:

$$w_{ij}' = w_{ij} + \frac{w_{ij}}{\sum_{j=1}^{m_i} w_{ij}} * BI_i$$

上式中, BI_i 为移动手机用户对第 i 个文本的兴趣度。将用户兴趣度和特征词的权重相关联, 可计算用户对各项目兴趣度, 得到最佳推荐。

5 结论(Conclusion)

本文采用情景构建移动用户模型, 作为为个性化推荐依据。计算用户当前情景和历史情景的相似值, 和当前情景相似值最高的历史情景下的项目信息优先推荐给用户。运用自然语言处理知识, 对项目文本内容作分词处理, 提取特征关键词并计算各个特征词在文本中的权重值, 建立特征词矩阵。根据用户行为信息, 分析用户浏览、分享、收藏项目的情况, 结合特征词矩阵构建用户的兴趣特征集合。将用户兴趣集合中的文本和待推荐项目作相似性的计算, 根据得出相似度值做顺序排序, 向用户推荐。

参考文献(References)

- [1] Kbaier M E B H, Masri H, Krichen S. A personalized hybrid tourism recommender system[C]. International Conference on Computer Systems and Applications. Hammamet: IEEE, 2018: 244-250.
- [2] Shi L, Zhao WX, Shen YD. Local representative-based matrix factorization for cold-start recommendation[J]. ACM Transactions on Information Systems, 2017, 36(2): 1-28.
- [3] CHILIGUANO P, FAZEKAS G. Hybrid music recommender using content-based and social information[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016: 2618-2622.
- [4] Bobadilla J, Ortega F, Hernando A, et al. Recommender systems survey. Knowledge-Based Systems, 2013(46): 109-132.
- [5] SHI Y, LARSON M, HANJALIC A. Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges[J]. ACM Computing Surveys (CSUR), 2014, 47(1): 1-45.
- [6] Breese J S, Heckerman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering[J]. Uncertainty in Artificial Intelligence, 2013, 98(7): 43-52.

作者简介:

陈传敬(1989-), 男, 硕士, 助教. 研究领域: 计算机应用。

陈琳(1992-), 男, 硕士, 助教. 研究领域: 图形图像处理。