基于情景和浏览内容的层次性用户兴趣建模①

孙海真, 谢颖华

(东华大学 信息科学与技术学院, 上海 201620)

摘 要: 用户兴趣建模是个性化服务的核心, 考虑到情景信息对用户偏好的影响, 对融和情景信息的用户行为日志数据进行深入研究, 提出了一种基于情景信息的用户兴趣建模方法. 该方法首先通过计算情景相似度来获得用户当前情景的近似情景集; 对“用户-兴趣项-情景”三维模型采用情景预过滤的方法降维处理. 然后根据用户浏览内容得到用户兴趣主题, 分析页面内容得到每种主题的兴趣关键词, 建立基于层次向量空间模型的用户兴趣模型. 实验结果表明, 本文提出的基于情景信息的用户兴趣模型对用户兴趣的预测误差控制在 9%以内, 是有效的.

关键词: 用户兴趣模型; 情景; 用户浏览内容; 文本聚类; TF-IDF

Hierarchical User Interest Modeling Based on Context and Browse Content

SUN Hai-Zhen, XIE Ying-Hua

(School of Information Science and Technology, Donghua University, Shanghai 201620, China)

Abstract: The user's interest model is the core component in a personalized services system. Considering the impact of context information on user interests, this paper deeply studies the user behavior log data based on context information, and proposes a user interest modeling method based on context information. First, we get the user's context set by calculating the context similarity, and reduce the dimension of the "user-interest item-context" 3D model through the method of context pre-filtering. Second, user browsing content forms interest topic, and web page content forms interest keyword. Then a hierarchical vector space model is set up based on the user profile. The experimental result shows that the prediction error of user interest degree is controlled within 9%, which is effective.

Key words: user interest model; context; user browsing content; text clustering; TF-IDF

个性化服务利用用户预先提供的数据或是利用数据挖掘等技术从用户的历史记录中收集用户偏好, 帮助用户获取感兴趣的信息, 避免了用户浏览大量无关资源而浪费时间. 用户兴趣模型的建立是个性化服务的核心, 资源推荐的准度和广度, 完全取决于用户建模表征用户兴趣的准确度和潜在用户兴趣的挖掘度.

用户兴趣建模一般包括两方面内容: 通过记录和分析用户浏览行为、浏览内容及用户反馈等收集用户信息并从中挖掘用户兴趣; 用合适的方法表示用户兴趣, 即建立用户兴趣模型, 并随用户兴趣变化动态更新用户兴趣模型[1].

传统的基于用户浏览行为的用户兴趣建模大部分只考虑用户和项目两个维度, 在一些融合诸如时间或位置情景的个性化服务中误差较大, 而基于情景信息的用户兴趣建模可有效的提高大数据时代个性化服务的精准度. 例如, Koren 提出一种融入用户时间情景信息的推荐模型 timeSVD++, 并将该算法在 Netflix 电影评分数据集上进行试验, 结果表明该模型的推荐精确度较未融入时间用户情境矩阵分解模型有了显著的提高[2]. Si 等人通过设定推荐系统服务中的用户情景信息为在线时刻、位置及心情三种类型, 并结合用户所感兴趣的主题关键词, 应用矢量模型构建了用户偏好模式, 来研究手机终端上关于图书的推荐服务问题[3]. Liu 等采用本体模型来表示用户情景信息, 并计算其

① 收稿时间:2016-04-06;收到修改稿时间:2016-05-05 [doi:10.15888/j.cnki.csa.005509]

用户情景信息之间的距离[4]. Shi 提出一种基于情绪特征的物品相似度的矩阵分解方法对情绪用户特征进行用户偏好建模[5]. 胡慕海对位置、时间、用户心情等多种用户情景信息, 提出了一个应用信息熵提取用户情景偏好特征的建模方式, 并通过超图模型将用户进行细分, 最后通过超图分割技术对用户偏好和用户偏好漂移进行识别与建模[6]. 王立才专门对情绪这类情景结合认知心理学的知识通过基于张量和高阶奇异值分解技术(Higher-order Singular Value Decomposition, HOSVD)进行用户偏好建模[7].

综上所述, 情景化机制已经引起了国内外学者的广泛关注, 他们通过对用户位置、时间、业务需求的情景化挖掘, 基于不同的资源对象和情景来探讨用户兴趣的变化. 但这些研究主要集中在情景维度的某个方面, 缺少对用户兴趣表示及情景机制的完整描述. 本文的目的在于建立基于情景信息的用层次性户兴趣模型. 通过将用户情景进行系统的分类和识别, 将情景影响因子加入到用户兴趣建模过程中, 改进用户兴趣度的计算方法, 最后根据训练集和测试集的兴趣误差来验证模型的有效性.

1 融合情景的用户兴趣模型表示

1.1 用户兴趣三维建模

目前用户兴趣建模大多停留在二维上, 即用户维和项目维[8], 最终的用户兴趣度由用户和项目决定, 没有涉及到情景, 如时间、地点等. 本文加入情景维度来描述用户兴趣, 三维模型如图 1 所示.

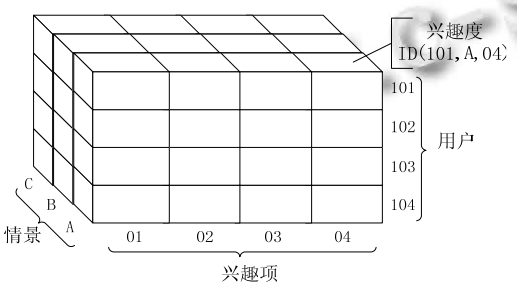

图 1 用户兴趣三维模型

“用户-兴趣项-情景”三维模型是一个三维的向量空间, 每个维度分别由各自的属性值组成的向量来表示, 图中 $ID(101, A, 04)$ 表示的就是在情景 A 下用户 101 对于兴趣项 04 的兴趣度. 可将用户兴趣模型形式化的表示为一个三元组:

$$UserModel = \{UserInfo, UerDoI, Context\}$$

其中 UserInfo 代表用户基本信息, 可以表示为 $UserInfo = \{UserID, Name, Sex, Age, Job\}$, 分别表示用户 ID、姓名、性别、年龄、职业; *UserDoI* 代表用户兴趣度(Degree of Interest); Context 表示情景维度.

1.2 基于层次的向量空间模型

本文基于情景的用户兴趣模型表示方法主要是对向量空间模型表示法(VSM)[9]进行改进, 由于传统的 VSM 表示方法是把所有种类的用户兴趣记录在同一个向量里, 并且很少考虑到用户所处的情景, 这样会导致不同情景、不同类别的兴趣特征项相互影响, 降低个性化服务的质量.

针对上述出现的问题, 本文基于情景信息提出层次性向量空间模型来表示用户兴趣. 其基本思想是: 1)将用户访问日志根据情景属性进行分类. 2)分别分析不同情景下的用户访问日志, 计算用户浏览网页的次数, 按照新闻、视频、调查、论坛、购物、社交、游戏给用户兴趣归类, 得到用户的兴趣主题. 3)通过页面 URL 获取页面内容信息, 提取文档中的关键词作为特征项用 VSM 来描述用户兴趣.

具体如图 2 所示, 底层是用户兴趣关键词, 第二层是划分的用户兴趣主题, 第三层是用户所处的情景, 最上一层是用户.

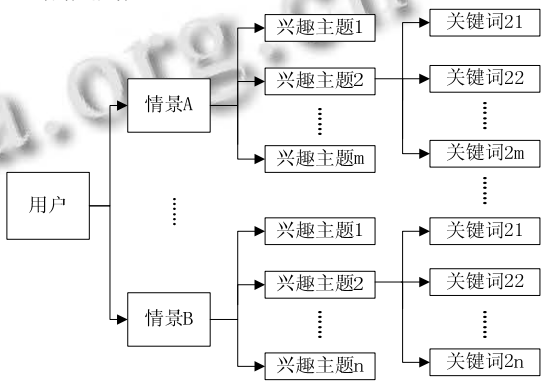

图 2 基于层次的用户兴趣模型表示结构图

如果用户在情景 A 下有 m 个不同的类别偏好, 即用户有 m 个兴趣主题, 那么情景 A 下用户兴趣模型可表示为如下结构的向量:

$$IModel = \{(T_1, W_1, n_1), (T_2, W_2, n_2), \cdots, (T_m, W_m, n_m)\} \quad (1)$$

其中, T_i 为第 i 个主题特征向量, W_i 为主题权重, n_i 为第 i 个主题包含文档实例数量($1 \le i \le m$), W_i 初始化如

下:

$$W_i = I(p_1) * I(p_2) \cdots * I(p_{q_i}) \quad (2)$$

其中 $I(p_k), k = 1,2,\mathrm{L}\ ,q_i$, 表示用户对网页 p_k 的兴趣度.

若 T_i 类包含 m 个兴趣关键词条, 则 T_i 可表示为:

$$T_i = \{(k_{i1}, w_{i1}), (k_{i2}, w_{i2}), \cdots, (k_{im}, w_{im})\} \quad (3)$$

其中, (k_{ij}, w_{ij}) 是 T_i 类的第 j 个关键词条, k_{ij} 为关键词, w_{ij} 为关键词的权重.

2 用户兴趣建模方法

2.1 情景建模

2.1.1 情景模型定义及分类

情景维度模型是表示情景综合信息的模型, 用户偏好会随所处情景(如时间、地点、环境、用户状态等)的不同而发生变化, 因此建立用户兴趣模型时需要考虑到用户情景.

研究分析顾君忠[10]对情景信息的分类方法, 本文将情景信息划分为 3 个情景维度, 表示为 *Context* = {*User Context*, *Time Context*, *Spatial Context*}.

(1) 用户情景(User Context)指用户的概要信息、社会地位等. 从用户的信息表中我们可以获得用户的年龄、性别、职业等信息. 用户维情景可以表示为 *User Context* ={*Age*, *Sex*, *Job*}.

(2) 时间情景(Time Context)指用户与系统发生交互的时间, 可根据具体需要按照不同的分层粒度对时间情景进行组织. 时间维情景可以表示为 *Time Context* = {*Date*, *DayOfWeek*, *TimeOfWeek*, *Month*, *Quarter*, *Year*}. 其中 *DayOfWeek* ={"*Mon*", "*Tue*", "*Wed*", "*Thu*", "*Fri*", "*Sat*", "*Sun*"}; *TimeOfWeek* = {"*Weekday*", "*Weekend*"}.

(3) 空间情景(Spatial Context)指用户与系统发生交互时所处地点的相关信息. 可以表示为 *Spatial Context*={*Province*, *City*}.

不同的情景信息对用户兴趣的影响程度不尽相同, 因此在建立模型之前需要分析与用户兴趣密切相关的有效情景, 在情境建模时可以视具体情况适当减少或增加一些维度.

2.1.2 情景预过滤

情景预过滤(contextual pre-filtering)是利用当前情景信息过滤掉与当前情景无关的用户数据, 从而构建与当前情景相关的数据集合[11]. 简单来说, 如果分析一个只在星期六上网的人的兴趣, 则只需过滤出每周六的所有用户的评测数据来构造兴趣模型即可, 然而这种方法存在缺点, 太精确的情景信息可能不够实用. 比如, 对于星期六或者星期日去看电影的用户来说, 情景信息其实差别不大; 但与星期三(工作日)相比, 那就不同. 所以在过滤情景信息时, 不应该把周日的数据也给过滤掉. 此外, 精确过滤后的数据量相对来说有所减少, 导致数据稀疏问题. 因此在实际兴趣建模过程中会使用情景泛化处理来解决上述问题.

本文在建立用户兴趣模型前首先通过时间情景对用户浏览行为的日志数据进行预过滤, 考虑到过度细化的缺陷, 时间维情景划分方式为: *TimeOfWeek* = {"*Weekday*", "*Weekend*"}.

2.1.3 情景后过滤

情景后过滤(contextual post-filtering)不会在输入数据和建模时考虑情景信息, 而是在生成用户兴趣项列表时根据情景信息进行如下处理: 1)过滤掉不相关的兴趣项. 2)调整列表中兴趣项的排序.

例如采用传统的用户兴趣建模方法得到用户兴趣列表, 假设用户对新闻类的网站感兴趣, 考虑到用户所处的空间情景(如城市), 可以直接过滤掉与当前情景关联概率小的项目, 得到情景优化后的兴趣列表.

2.2 基于 PV 提取用户兴趣主题

网页浏览次数 PV(Page View)[12]是统计互联网用户浏览网页的次数, 通过分析 url 的类别, 归类得到用户兴趣主题.

url 访问频率 uf(url visit frequency): 表示 url 被用户访问的频繁程度, 计算公式如下:

$$uf = \frac{u_i}{\sum_{i \in T} u_i} \quad (4)$$

其中 u_i 表示第 i 条 url 的 PV 值, T 为用户访问的所有 url 集合. uf 大的 url 说明用户访问频繁, 对这类网站的兴趣度越高.

2.3 基于网页内容提取用户兴趣关键词

2.3.1 文本特征项的提取

在对文档进行特征提取之前, 需要先进行文本信息的预处理——特征词条的选择. 从自然语言理解的角度来看, 名词及名词短语、动词及动词短语是一个文本的核心, 它们的简单组合可以作为整个文档的简单表示. 本文采用中国科学院计算机研究所研制的汉语分词系统 NLPIR 进行分词[13].

对页面文档进行处理并提取特征词的步骤如下:

(1) 通过页面 URL 获取页面内容信息, 清除页面中网页标签信息, 进行页面清洗, 将其转化为文本文档;

(2) 调用字典模块对文档进行分词, 将文档转化为词序列;

(3) 根据停用词表去除词序列中的停用词;

(4) 计算每一个词的权重;

(5) 根据权值大小对词进行降序排列, 选取前 n 个词作文档的特征词集合.

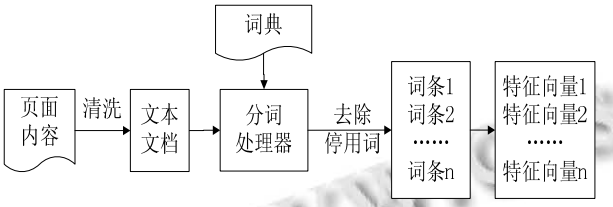

图 3 文本特征向量抽取

2.3.2 TF-IDF 算法计算关键词权重

一个文档集 T 中的某个文档 D, 对于 D 中的关键词 W 来说, W 在 T 中除了 D 之外的其它文档中出现的次数越少, W 对于 D 的区分度就越高. 因此, 如果 D 中有两个关键词 W_1 和 W_2, 它们在 D 中出现在频率一样, 而 W_1 在文档集合 T 中的其它文档中出现地次数比 W_2 少, 则对于文档 D 来说, W_1 的权重应该大于 W_2 的权重. 即: 一个关键词的权重与它在一个文档中出现的频率 *tf*(Term Frenquency)成正比, 与它在文档集中其它文档中出现的频率 *idf*(Inverse document frequency)成反比[14], 该计算方式表示为:

$$w(t,d)=\frac{tf(t,d)\times\log(N/n_t+0.01)}{\sqrt{\sum_{w\in d}[tf(t,d)\times\log(N/n_t+0.01)]^2}} \quad (5)$$

其中, $w(t,d)$为词 t 在文本 d 中的权重, $tf(t,d)$为词 t 在文本 d 中的词频, N 为训练文本的总数, n_t 为训练文本集中出现词语 t 的文本数, 分母为归一化因子.

2.3.3 改进的文本特征聚类算法

在基于浏览内容进行数据挖掘分析中, 由于用于挖掘的数据源是文本文档集合, 而挖掘的目的是发现用户的兴趣类型及兴趣浓度. 采用 K-means 方法进行聚类分析时, 由于预先不知道用户的兴趣种类, 即不知道进行 K-means 聚类的 K 值, 因此无法直接采用. 而且 K-means 方法中初始聚类中心的选取直接影响到最后的聚类结果, 并且很容易陷入局部最优解. 层次凝聚法能够生成层次化的嵌套簇, 准确度较高. 但在每次合并时, 需要全局地比较所有簇之间的相似度, 并选出最佳的 2 个簇, 因此执行速度较慢, 不适合大量文件的集合.

综合考虑这两种聚类方法的优缺点, 提出一种改进的文本聚类方法, 具体过程如下:

对于给定的文件集合 $D=\{d_1,d_2,\cdots,d_i,\cdots d_n\}$:

(1) 将 D 中的每个文件 d_i 看作是一个具有单个成员的簇 $c_i=\{d_i\}$;

(2) 任选其中一单个成员簇 c_i 作为聚类的起点;

(3) 在其余未聚类的样本中, 找到与 c_i 距离满足条件的 d_j (可以是与 c_i 距离最近的点, 即相似度 $sim(c_i,d_j)$ 最大的 d_j; 也可以是与 c_i 距离不超过阈值 d 的点, 及相似度 $sim(c_i,d_j)\geq d$ 的任意 d_j). 将 d_j 归入 c_i 形成一个新的簇 $c_k=c_i\cup d_j$;

(4) 重复步骤(3), 直至与 c_i 距离最近的 d_k 与 c_i 之间的距离超过阈值 d, 则认为聚完了一类;

(5) 选择一个未聚类的单个成员簇, 重复步骤(3)和(4), 开始新一轮的聚类, 直到所有的单个成员簇 c_i 都参与了聚类, 最终聚成 K 类;

(6) 采用 K-means 算法进行修正, 得到修正后的文本聚类结果.

算法流程图如下:

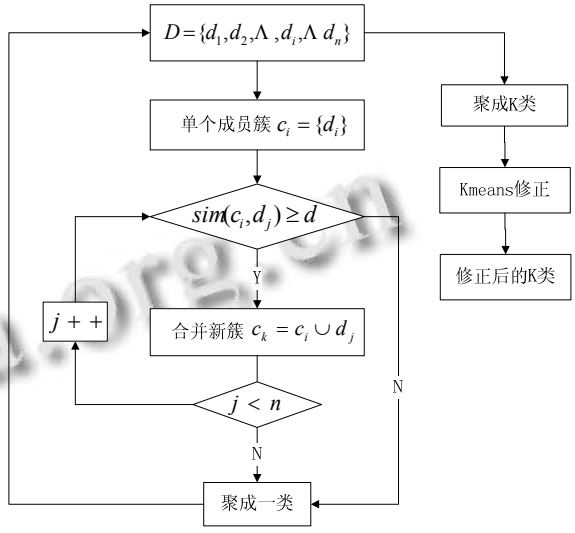

图 4 改进的文本聚类算法流程图

3 实验步骤及结果

3.1 实验步骤

本文采用的数据集是 CNNIC(http://cnnicdata.datatang.com/)数据堂提供的数据集. 该数据集包含用户连续 4 周内访问电脑软件及浏览网页的行为日志. 实验原始数据总时长为 28 天: 取前 21 天数据作为训练数据(Training Data), 用以预测后 7 天(测试数据, Test Data)用户的兴趣分布.

Step1. 情景信息预过滤: 将用户浏览网页的行为

日志按时间情景分为工作日和周末两类.

Step2. 数据预处理: 从用户行为日志中过滤出用户 ID、浏览的网址及用户浏览次数.

Step3. 提取用户兴趣主题: 统计数据集合中域名级别的网站, 分析网站的类别, 过滤掉搜索网站、导航网站等不能明确反映用户兴趣爱好的网站, 得到新闻、视频、调查、论坛、购物、社交、游戏七大兴趣主题.

Step4. 提取兴趣关键词: 将每种主题的url转化为文本文档, 采用NLPIR进行分词, 通过TF-IDF算法计算关键词的权重, 将文本文档用向量表示, 并进行聚类分析, 得到每个主题下关键词的权重, 进而得到每种主题的权重, 建立基于层次的用户兴趣模型.

3.2 实验结果

实验中抽取三个用户的浏览行为特征, 采用第二章的方法分析训练数据集, 得到用户对每种主题的兴趣度, 如表 1 所示.

表 1 用户对兴趣主题的兴趣度

用户	101		102		103	
主题	workday	weekend	workday	weekend	workday	weekend
新闻	0.4294	0.3332	0.2146	0.1172	0.4308	0.4624
购物	0.2902	0.2282	0.4679	0.3309	0	0
论坛	0.1929	0.2069	0	0	0	0
社交	0.0757	0.2084	0.0239	0.0818	0.3418	0.2816
调查	0	0	0.2472	0.3529	0.2274	0.1063
游戏	0.0118	0.0233	0	0	0	0
视频	0	0	0.0465	0.1172	0	0.1494

从表 1 中可以看出用户 101 在工作日对于新闻和购物类比较感兴趣, 周末增加了社交类的兴趣; 用户 102 在工作日对购物类有着浓厚的兴趣, 周末转向了调查类的网站; 用户 103 在工作日对新闻和社交比较感兴趣, 周末新增了视频类的兴趣. 因此在个性化服务中考虑到情景信息可以发现用户不同情景下的兴趣倾向, 从而改善用户的体验.

对测试数据集进行分析得到每种主题的误差如表 2 所示.

表 2 兴趣主题的绝对误差

兴趣主题	新闻	购物	社交	调查	游戏	视频
训练集	1.0000	0.6627	0.5098	0.4698	0.0177	0.1575
测试集	1.0000	0.6981	0.4919	0.4296	0.0000	0.0764
绝对误差	0.0000	0.0354	0.0178	0.0402	0.0177	0.0811

从表 2 中可以看出, 根据前面提到的兴趣度计算方法计算得到的用户兴趣度与测试集中用户兴趣度绝对误差控制在 9%以内, 由此可以验证本文提出的基于情景信息的用户兴趣模型是合理及有效的.

4 结语

本文将情景信息融和到用户兴趣建模过程中, 结合情景预过滤的思想, 将用户兴趣三维模型降维处理, 建立基于层次的向量空间模型, 并改进文本特征聚类算法. 分析训练集和测试集的用户兴趣, 得到用户兴趣预测误差, 实验结果表明误差控制在 9%以内, 表明该算法的可行性和有效性. 目前只考虑到单维度的静态情景信息, 下一步的工作将研究多维度情景和动态情景对用户兴趣的影响.

参考文献

1 南智敏.基于网页兴趣度的用户兴趣模型体系研究[硕士学位论文].上海:复旦大学,2012.

2 Koren Y. Collaborative filtering with temporal dynamics. Communications of the ACM, 2010, 53(4): 89–97.

3 Si H, Kawahara Y, Kurasawa H, et al. A context-aware collaborative filtering algorithm for real world oriented content delivery service. Proc. of ubiPCMM, 2005.

4 Liu D, Meng XW, Chen JL. A framework for context-aware service recommendation. 10th International Conference on Advanced Communication Technology (ICACT 2008). IEEE. 2008, 3. 2131–2134.

5 Shi Y, Larson M, Hanjalic A. Mining mood-specific movie similarity with matrix factorization for context-aware recommendation. Proc. of the Workshop on Context-Aware Movie Recommendation. ACM. 2010. 34–40.

6 胡慕海.面向动态情境的信息推荐方法及系统研究[博士学位论文].武汉:华中科技大学,2011.

7 王立才.上下文感知推荐系统若干关键技术研究[博士学位论文].北京:北京邮电大学,2012.

8 邢晓兵.面向用户兴趣的用户浏览行为分析方法及应用[硕士学位论文].沈阳:东北大学,2013.

9 郝水龙,吴共庆,胡学钢.基于层次向量空间模型的用户兴趣表示及更新.南京大学学报,2012,2:190–197.

10 顾君忠.情景感知计算.华东师范大学学报(自然科学版),2009,5:1–20,145.

11 刘海鸥.云环境用户情景兴趣的移动商务推荐模型及应用研究[博士学位论文].秦皇岛:燕山大学,2013.

12 赵钦森.基于用户行为的动态推荐系统算法研究及实现[硕士学位论文].成都:电子科技大学,2013.

13 ICTCLAS 中文分词系统官方网站.http://ictclas.org/.

14 蒋晨.基于用户情景感知的动态兴趣模型及其应用[硕士学位论文].武汉:华中师范大学,2014.