第42卷 第6A期
2015年6月

计 算 机 科 学
Computer Science

Vol. 42 No. 6A
June 2015

移动终端泛在情境适应的网络音乐推送研究

张秀玉

(福建信息职业技术学院传媒艺术系 福州 350003) (中央财经大学信息学院 北京 100081)

摘 要 当前信息推送模型主流结合用户兴趣或移动终端模式,缺乏对终端情境模式与用户兴趣适应整体支持的理论解决方案。为改善音乐推送现状,实现面向移动终端的网络音乐推送,提出了一种面向终端情境的音乐推送模型。首先给出了终端情境的概念定义;其次,提出了一种基于贝叶斯网络和协同过滤的音乐推送模型;然后,定义了音乐推送模型的构建过程;最后,通过实例情境验证了面向终端情境的音乐推送效果,表明了该模型适用、有效。

关键词 终端情境,音乐推送,贝叶斯网络,协同过滤,推送模型

中图法分类号 TP311.5 文献标识码 A

Research on Web Music Push Model of Mobile Terminal Ubiquitous Context Adaptation

ZHANG Xiu-yu

(Media Art Department, Fujian Polytechnic of Information Technology, Fuzhou 350003, China)

(School of Information, Central University of Finance and Economics, Beijing 100081, China)

Abstract The main current information push model is based on user interests or terminal situation, lacking theoretical support to the two elements together. To improve the approach of pushing music, realizing mobile-termination-oriented pattern, this paper presented a music push model that focused on terminal situation. Firstly, the paper defined the concept of terminal situation. Secondly, the paper put forward the music push model based on Bayesian network and collaborative filtering algorithm. Next, the paper defined this model's establishment process. Finally, the paper validated the effects of terminal-situation-oriented model through living examples, indicating that this model is more applicable and effective.

Keywords Terminal situation, Pushing music, Bayesian network, Collaborative filtering algorithm, Push model

随着移动互联网技术的迅速发展,以及移动智能终端功能的日益强大,人们对移动终端的广泛使用性逐渐体现出来,甚至可以称为对移动终端的一种依赖。移动音乐市场庞杂的音乐资源,已远远超过一个用户的音乐需求和自主选择能力。在一般的应用情境下,用户通常会在某个极其有需求的时候,在音乐平台中根据关键字搜索的办法搜寻网络音乐,通过衡量音乐播放次数、用户评价、排名或口碑等方式,选择最终收听的网络音乐。但虾米音乐等终端音乐播放器搜索网络音乐时难以充分结合网络音乐的好评度和访问量。围绕这个问题,目前已经出现的解决方案通常根据用户此前的下载历史等个性化信息,推荐其可能感兴趣的网络音乐,提高网络音乐的"露脸"概率,这是精准推送的方法之一。然而这种解决方案存在的问题之一,就是用户访问音乐平台和访问音乐的不确定性。

目前的用户兴趣建模通常是基于用户检索过程中的检索结果来定制推荐,基于该用户过往的搜索内容和浏览结果,进行算法设计。其用来衡量用户兴趣的数据源通常为用户浏览页面行为和反馈信息、用户浏览日志信息、用户浏览页面之间的超链接信息等。用来计算兴趣度的重要变量通常为用户的浏览行为——页面浏览时间、拉动滚动条的次数、保存页面的概率、是否重复访问同一页面等。这些基于PC端的用户兴趣模型研究内容,也部分适用于本文所考虑的移动端情况。

研究内容都是基于用户信息检索过程的行为分析,通过检索行为、点击行为和浏览行为的记录和学习,构建用户在当前情境下的选择和偏好模型,进行记录,以便在相似情境出现的情况下提供符合用户兴趣的网络资源推荐。移动互联网的用户兴趣模型搭建内容则较为空白,其所选择的变量和核心算法与PC端差别较大。在移动互联网信息推送方面,更多的研究成果已经体现在应用层面[1],如社交型的新鲜事、关注人信息推送;多媒体型的电视节目提醒;应用型的背单词、天气信息推送等等。如上文所言,这些推送都建立在用户主动关联和设置的前提下,产品拥有较强的应用属性[2]。

通过实际案例观察表明,用户使用移动端搜索和浏览页面的行为习惯必然与PC端不同[3]。因此,要更加准确地挖掘用户兴趣并进行信息推送,必然要对以上变量进行修正,添加基于移动端页面浏览特点的新变量,并构建新的模型。移动互联网用户行为习惯的分析,更新了传统意义上数据源的范围,特别针对"碎片化"移动互联网的使用特点,提出基于碎片时间段的精准推送策略[4]。因此,本文的研究目的在于,根据移动终端用户的行为特征和使用习惯,进行终端情境模型搭建,构建网络音乐的推送模型,并进行算法设计和验证。

张秀玉(1964—),女,硕士,副教授,主要研究方向为信息服务与网络科学,E-mail:37042633@qq.com。

1 相关定义

在移动情境内涵的问题上，国内外有很多不同的结论[5]。Dey 将必要情境信息分为位置、基础设施或资源、用户、环境、实体、时间几个部分[3]。Lieberman 等人认为必要的情境信息有用户、环境和应用 3 个方面[2]，其中，用户相关情境包括活动、位置、标志、描述；与环境的相关情境包括功能、维护、能源等。Hakkila 等人提出，情境信息分为物理环境、链接设备、用户行为、偏好和社会背景 5 个方面[4]。Scmidt 等人从人因学的角度将情境信息模型表示为与人相关的情境信息和与物理相关的情境信息。在国内的研究中多数将情境信息大致分为自然环境、设备环境和用户环境 3 项[1]。而在更多的研究中，一般根据所讨论问题的应用性来对情境信息进行界定和分类。

在上文对移动终端特征的分析中提到，对于向用户推荐一首歌曲而言，不能再单纯地以用户评分、过往的日志信息等显性信息来衡量，而应该衡量包括歌曲的实际收听时间、单位时间内的重复播放次数、个性化服务使用情况等在内的因素以及由用户所处在的特定情境而产生的外部信息，以此来形成模型化的反馈评价体系，对协同过滤算法进行修正。所以，在前人研究的基础上，本文将传统的基于用户_项目的两维模型拓展为用户_项目_情境的三维模型，其中情境信息包括用户情境、任务情境、物理情境。用户情境包括用户年龄、性别、社会地位等基本信息，以及历史搜索和下载情况等；任务情境包括用户收听一首歌曲的一切影响因素，包括带宽、网络连接性、可用资源以及设备情况等；物理环境包括时间、位置、天气、日期等。这些所有情境的外化表现在该目标用户在过往的音乐收听中表现出的个性化移动终端情境，包括收听时间、次数、频率、偏好的时间点等。而这些外化的情境，可以通过隐形或显性的测量手段得到，通过算法确定为该用户的音乐类别偏好。为方便算法展开，为各情境定义如下符号，并注明相应的移动终端特征，如表 1 所列。

表 1 移动端收听音乐情境的分类

情境类别	情境信息	符号	对应移动特点
用户情境	用户$_i$的生理条件（年龄、性别、社会地位）、用户$_i$的过往搜索/下载历史	U	移动网络普及程度高，年龄、性别分布零散，社会地位差别较大。过往下载/搜索历史衡量用户兴趣
任务情境	网络连接性、通信开销、通信带宽、使用设备	C	移动端下载对于流量、带宽、移动设备的要求不同
物理情境	时间、地点、天气、日期	T	设备移动性、使用时间碎片化、在某些特定的地点与时间对于特定的音乐有更加强烈的偏好
用户评价情境	过往用户评分	I	移动端评分机制便利
音乐收听情境	音乐的实际收听时间、是否重复收听、对于同一类别的音乐收听所占的比重、是否搜索相似的音乐	V	音乐的实际收听时间和是否重复收听比单纯的音乐评分更能体现是否符合用户偏好

2 终端情境的特征度量

为了在算法设计和实验验证环节更清晰地定义变量和获取实验数据，有必要对上文提到的用户移动端情境特征进行更详细的描述、分类和度量方法辨析。根据表 1 对基于移动端收听音乐情境分类，除任务情境所需要的数据信息来自于网络音乐推送使用条件，其他用户情境、物理情境、用户评价情境、音乐收听情境均来自于用户行为数据。移动端用户行为数据最普遍的存在形式是用户日志。用户在网站上的行为将会在网站运行过程中产生原始日志(raw log)，并按照用户行为情况分类汇总成会话日志(session log)，每个会话表示一次用户行为和对应的服务。比如，目标用户在音乐播放器中搜索某首歌曲的名称进行查询，播放器根据自己的推荐方式反馈查询结果，就称为服务对查询结果生成了日志。如果用户点击了某首反馈结果中的歌曲，这个行为将会被记录在点击日志中，称为该用户的历史点击信息。日志可以记录用户的各种行为，为电子商务网站常用的日志形式有网页浏览日志、购买日志、点击日志、评分日志等。在移动端，同样借用日志的形式来解决本文所需要的用户行为数据获取问题。

用户反馈行为的形式有很多分类，通常使用显性反馈行为和隐性反馈行为两种。显性反馈行为一般是从相关性评估者处获取的，即目标用户对搜索结果的直接评价，是较明确的对结果的喜好程度。其在网页端视频网站页面中极为常见，通常为用户对该视频的评分，这种直接表达的显性反馈可直接影响其他相关评估者对搜索结果的选择。在一些歌曲榜单推荐平台，也能见到这种显性反馈机制，如用户评分、用户评论等。在移动音乐推送上，显性反馈行为数据有两个明显的不足之处。第一，与网页端内容不同，移动端音乐的收听者的精力重点放在收听歌曲上，而鲜有用户在下载收听歌曲后返回下载平台进行歌曲评分及评价。第二，返回评价的用户通常为极端使用用户：要么他对此首歌极其满意，要么他极其不满意。不管何种心理，这样得到的反馈数据都是不精准的。这就造成了真实数据大量缺失。

相对于显性反馈行为的不足，隐性反馈行为是通过对用户行为的直接捕捉和推断所得出的数据。常见的隐性反馈包括用户停留在页面的时间、连接的点击次数等等。在不告知用户其选择将被应用于评价的反馈基础上，这种数据将更加真实可信。然而其缺点在于，对于分析的问题，隐性反馈行为给出的答案将不那么明确。当某用户停留在某首歌曲时间较长时，不知道他是对此首歌曲的兴趣度更高，还是基于其他的客观原因。

表 2 从几个方面比较了显性反馈行为和隐性反馈行为。

表 2 显性反馈行为和隐性反馈行为的比较

特征	显性反馈行为	隐性反馈行为
反馈结果	明确	不明确
数据数量	较少	大量
数据准确程度	准确性较差	较准确
实时读取	实时	有延迟
正负反馈	都有	只有正反馈

根据显性反馈行为和隐性反馈行为的特点，对上一节中提到的与本文相关的用户使用情境进行进一步分类及反馈内容的辨析，如表 3 所列。其中，任务情境来自于网络音乐推送使用条件，不在用户行为数据层面抓取信息。

根据两种反馈行为的优劣特点及所需要的各种情境下数据的特点，选择相应的反馈行为，如表 4 所列。

表 3　各种情境下显性反馈行为和隐性反馈行为的内容

情境类别	情境信息	显性反馈行为	隐性反馈行为
用户情境	用户 $_i$ 的生理条件（年龄、性别、社会地位）	用户对自身情况的描述	用户的搜索、下载日志
	用户 $_i$ 的过往搜索/下载历史	用户对下载情况的描述	用户的搜索、下载日志
物理情境	时间	用户对收听时间的描述	用户对某首歌曲的收听日志
	天气	用户对天气类型偏好的描述	用户在不同天气情况下歌曲的收听情况
	地点	用户对使用地点的描述	用户对某首歌曲的收听日志
	日期	用户对使用日期的描述	用户对某首歌曲的收听日志
用户评价情境	过往用户评分	用户对某歌曲的评分及评价	用户对某首歌曲的收听日志
音乐收听情境	歌曲的收听率（对于某首歌实际收听的时间占歌曲总时长的比例）	用户对收听率的描述	用户对某首歌曲的收听日志
	周内重复收听的频率	用户对周内使用频率的描述	用户对某首歌曲的收听日志
	歌曲的偏好度（周内对于某首歌曲的收听时间占总音乐收听时间的比例）	用户对歌曲偏好度的描述	用户对某首歌曲的收听日志

表 4　各种情境下显性反馈行为和隐性反馈行为的选择情况

情境类别	情境信息	显性反馈行为	隐性反馈行为
用户情境	用户 $_i$ 的生理条件（年龄、性别、社会地位）	√	
	用户 $_i$ 的过往搜索/下载历史		√
物理情境	时间	√	
	天气	√	
	地点		√
	日期	√	
用户评价情境	过往用户评分	√	
音乐收听情境	歌曲的收听率		√
	周内收听的频率		√
	歌曲的偏好度		√

3　算法设计

3.1　基于移动特征的算法模型搭建

3.1.1　基于用户_项目评价的用户兴趣矩阵表示

在已经研究成型的用户兴趣模型中，基于用户_评价矩阵（U-I 矩阵）的表示方式是其中运用比较广泛的一种。该方法用矩阵 R_{mn} 表示用户模型，其中通过 $User$ 和 $Item$ 来确定特定用户 $_u$ 对特定项目 $_i$ 的偏好程度 $_p$。其模型表示如图 1 所示。

$$UI=\begin{bmatrix} u_1p_1 & u_1p_2 & u_1p_3 & \cdots & u_1p_{n-1} & u_1p_n \\ u_2p_1 & u_2p_2 & u_2p_3 & \cdots & u_2p_{n-1} & u_2p_n \\ u_3p_1 & u_3p_2 & u_3p_3 & \cdots & u_3p_{n-1} & u_3p_n \\ u_4p_1 & u_4p_2 & u_4p_3 & \cdots & u_4p_{n-1} & u_4p_n \\ u_5p_1 & u_5p_2 & u_5p_3 & \cdots & u_5p_{n-1} & u_5p_n \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ u_mp_1 & u_mp_2 & u_mp_3 & \cdots & u_mp_{n-1} & u_mp_n \end{bmatrix}$$

图 1　用 U-I 模型表示的用户兴趣模型

根据上文对用户情境、物理情境的分析，此处可将 $User\ dimension$ 定义为：$User=\langle U_{age}, U_{sex}, U_{status}, U_{history}\rangle$。$U_{age}$，$U_{sex}$，$U_{status}$，$U_{history}$ 分别为用户 $User$ 的几种特征属性，U_{age}，U_{sex}，U_{status} 表示用户情境中的用户年龄、用户性别和社会地位；$U_{history}$ 表示用户情境中的用户历史搜索情况和下载情况。对于每个 $u\in User$，每个属性都有系列的属性值与之对应。

与 $Item\ dimension$ 的定义相似，通过对任务情境的分析，对 $Item$ 的属性有以下描述：

$$Item=\langle I_{name}, I_{type}, I_{size}, I_{system}\rangle$$

I_{name}，I_{type}，I_{size}，I_{system} 分别表示歌曲的名称、类型、歌曲大小及适应的移动终端系统。

如果将 $User$ 和 $Item$ 表示为模型的两重 Dimension，即 $D1$、$D2$，那么在 $u\in User$ 及 $i\in Item$ 的情况下，U-I 模型的效用评分方程可表示为：$R(u,i)=D_1\times D_2$。

3.1.2　基于移动特征的三维模型优化

在上文对物理情境的分析中提到，移动终端用户收听音乐的时间、地点等行为习惯，带有明显的移动终端移动性、使用时间碎片化等特征。通过对用户个性化使用情境的分析，移动音乐推送的用户通常会根据自己的用户情境、任务情境、物理情境等情境因素，形成音乐收听类别偏好。综合这样的考虑，在传统二维用户_项目的偏好矩形中加入 $Dimension3=Situation$ 的维度。其三维模型表示如图 2 所示。

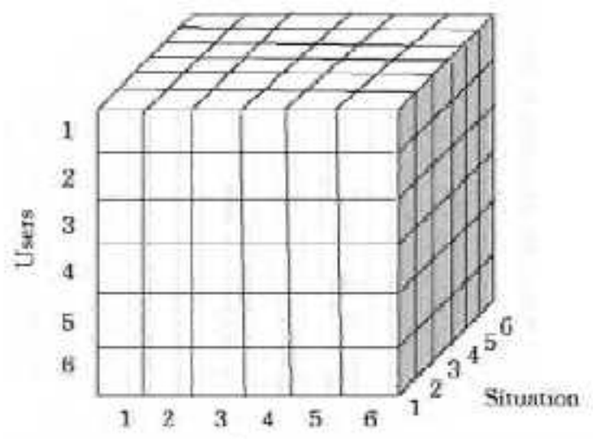


图 2　用 U-I-S 模型表示的三维用户兴趣模型

对任意 $s\in Situation$，效用评分函数表示为：$R(u,i,s)=D_1\times D_2\times D_3$。作为用户_项目二维评分系统的补充，加入移动用户特征的三维模型可以理解为：在特定的 $Situation$ 下，用户 $_u$ 对项目 $_i$ 的偏好程度。对于 $Situation$ 的具体衡量方法，将在下文算法中再加以阐述。

3.2　协同过滤算法

协同过滤算法是一种根据用户交易数据，自动推荐用户可能感兴趣的商品的算法。通过一定的信息处理，协同过滤算法在用户群的数据集中找到拥有相似兴趣的用户，综合这些用户对某商品的评价，形成系统对指定用户对此商品喜好程度的预测。简单地说，协同过滤算法通过参考与用户相似的其他用户的意见，来预测该用户可能还喜欢什么别的东西，从而进行商品推荐。

在电子商务行业，利用协同过滤算法向用户推荐商品的案例十分常见。因此，在进行移动终端的用户偏好音乐推送算法研究过程中，首先考虑协同算法过滤。

在协同过滤算法中，计算相似性是十分核心的步骤。通过相似性计算，可得到本次目标用户的最近邻居（拥有共同兴趣的用户），相似性越高，其与目标用户的相似程度越高。

在得到用户相似系数后，大多数协同过滤算法都采用平均加权策略产生后续的推荐。目标用户 $_a$ 对评分项目 $_i$ 的预测评分 P_{ai} 为：

$$P_{ai}=\overline{ra}+\frac{\sum sim(a,b)*(rb_i-\overline{rb})}{\sum|sim(a,b)|} \quad (1)$$

其中，$sim(a,b)$为用户a、b的余弦相似度，代入式(1)进行计算。该算法推测出目标用户a对项目i的评分，从而按评分排序项目，将项目评分最高的产品推荐给用户，完成协同过滤算法。

3.3 基于终端用户音乐情境的协同过滤算法

在上述的传统协同过滤算法中，系统虽然可以根据相似性寻找到最近邻居，以得到目标用户对可能项目的评价分数，然后做出推荐，但在移动互联网音乐推送上却显示了较大的不足。如同上文分析的移动互联网音乐推送的用户情境，如果单纯从用户共同兴趣去判断其相似性，而忽略目标用户在选择歌曲时带有的自身使用习惯和特征，这样的推荐经常是不精确的。因此，下文将基于终端用户音乐情境来对协同过滤算法进行修正。

1)移动情境分析

移动用户的音乐情境，指由于用户在长期使用中形成的行为习惯，影响了用户在需求表达和评分体系中的标准。传统的协同过滤算法只考虑用户兴趣，而忽略了目标用户自身带有的这种特征。

第一，在上文音乐情境中曾提到，虽然个体用户下载和搜索歌曲的数量很多，但下载并不代表喜爱和长期地重复收听该歌曲。由于该首歌曲原本不属于自己的刚性需求和原生兴趣所在，于是短暂的试听过后便跳到下一首歌曲，等到有时间的时候再对歌曲进行收听。久而久之，这首歌曲便一直安静下去，直到某一次手机音乐的大换血——这些沉默的大多数便被直接删除。因此，一般考虑歌曲种类、宽带占有率、使用流量、使用电量、用户评分等基本情境外，考察歌曲自身在收听用户中的使用情况十分必要。

第二，每个用户收听音乐的行为习惯不同，因此造成在歌曲种类选择上的概率不同。例如，在上下班高峰时段，更多的人倾向于相对于更加柔缓的轻音乐或者古典音乐，而在周末的夜晚，摇滚、舞曲类型的音乐则更加受到人们的青睐；在用户的身份特性上，相对于中年群体而言，一般年轻群体可能对于节奏感较快的R&B拥有更高的偏好程度。与音乐类型相关联的使用行为差异引申出这样一个问题——个人由于在收听音乐总体时间、社会地位等方面的差异，选择收听不同类型音乐的概率也会出现差异。根据相似性计算，系统可能在他的最近邻居中找到评分较高的其他几类音乐进行推荐，如《青花瓷》、《发如雪》、《以父之名》。这3类音乐分别属于流行类、古典类和R&B类。如果该用户正处在下班喧闹的地铁上，显然，节奏感更加剧烈的R&B类音乐就不适用这位目标用户，那么《以父之名》就不应该出现在推荐名单中。文中所谓的“不适合”，即为由于用户的情境而导致选择某种类的概率较低。而不同种类音乐的特征，除了其原生特征外，平均收听时间、收听习惯等将由已下载用户的数据形成。

2)基于终端用户情境的协同过滤算法

①用户特征资源类别偏好

首先，需要大量基于不同类别音乐的用户情境情况，这些情况反映了在单个用户的特定情境下，对某种类音乐的偏好。在这里，定义如下的情境因素：在全部种类V的集合中，v_i代表其中某确定种类的歌曲。对于v_i来说，分别用b_1,b_2,b_3,b_4来分别表示用户收听音乐时所处位置、网络状况、所处时间段、是否是某些重大节日。在参考了一定数量的文献后，本文将采用贝叶斯方法将以上特征信息融合在传统协同过滤推荐算法中。在对所有v_i进行属性值$\langle b_1,b_2,b_3,b_4\rangle$的信息收集整理后，新添加的歌曲只需选择相应种类进入数据集，自动携带相应的特征属性。在进一步的研究中，特征属性应该随着歌曲数量和下载情况实时动态更新，以保证描述的精准性和实时性。

在如上定义后，某用户在携带自身特征属性和检索时间属性$\langle b_1,b_2,b_3,b_4\rangle$进行检索时，其选择某类型歌曲的概率用贝叶斯方法表示为：

$$P(v_i|b_1,b_2,b_3,b_4)=\frac{P(v_i)*P(b_1,b_2,b_3,b_4|v_i)}{\sum_{v_i\in V}P(v_i)*P(b_1,b_2,b_3,b_4|v_i)} \quad (2)$$

等式右边需要解决的两项：$P(v_i)$与$P(b_1,b_2,b_3,b_4|v_i)$均可由大量实验数据得到。其中概率$P(b_1,b_2,b_3,b_4|v_i)$需要分析选择v_i的用户拥有b_1,b_2,b_3,b_4的特征情况，在大量信息数据的情况下，该概率看作是准确有效的，而由于实验可采集的数据有限，因此此处认为b_1,b_2,b_3,b_4所代表的用户收听音乐时所处位置、网络状况、所处时间段、是否是某些重大节日4个因素相对独立，有如下公式：

$$P(b_1,b_2,b_3,b_4|v_i)=\prod_{j=1}^{4}P(b_j|v_i) \quad (3)$$

将式(3)代入(2)中，得到：

$$P(v_i|b_1,b_2,b_3,b_4)=\frac{P(v_i)*\prod_{j=1}^{4}P(b_j|v_i)}{\sum_{v_i\in V}P(v_i)*\prod_{j=1}^{4}P(b_j|v_i)} \quad (4)$$

②基于终端用户情境的协同过滤算法

结合用户根据自身情境对类别偏好，融合传统协同过滤算法，参考文献后，将式(4)与式(1)进行相乘处理：

$$\begin{aligned}P_{ai}^{*}&=P(v_i|b_1,b_2,b_3,b_4)*P_{ai}\\&=\frac{P(v_i)*\prod_{j=1}^{4}P(b_j|v_i)}{\sum_{v_i\in V}P(v_i)*\prod_{j=1}^{4}P(b_j|v_i)}*(\overline{ra}+\\&\quad\frac{\sum sim(a,b)*(rb_i-\overline{rb})}{\sum|sim(a,b)|})\end{aligned} \quad (5)$$

等式分为两部分理解，左侧P_{ai}^{*}为参考了用户习惯特征对类别偏好影响因素后的预测评分；右侧分别为传统协同过滤算法下的用户预测评分，和目标用户确定$\langle b1,b2,b3,b4\rangle$属性值对类别$v_i$的偏好情况。对于推荐列表中同一类别的歌曲来说，其偏好程度相同。经过修正后，可根据最后总分在原推荐列表中筛选更符合目标用户使用习惯的歌曲进行推荐，以提高推荐精准性。整体流程步骤如图3所示。

图3 基于终端用户情境的协同过滤算法操作步骤

3)基于朴素贝叶斯分类算法的模型构造

在上文所阐述的算法中提到，修正协同过滤算法的表达式，需要通过输入大量数据的训练集，形成贝叶斯模型，也就是形成成熟的类别集合，待分类元素进入模型后，即可判断其

最大概率所属类别，从而对样本也就是目标用户的移动类别偏好进行判断。在这里，应用了贝叶斯模型中最为简单的分类算法——朴素贝叶斯，对以上流程进行完善。

①朴素贝叶斯分类算法概述

对于朴素贝叶斯分类算法，目前的算法定义和方法较明确：对于给出的待分类项，求解在此项出现的条件下各个类别出现的概率，认为此分类项属于概率最大的类别。对于目标用户的类别偏好问题，朴素贝叶斯分类算法可以描述为：在能够捕捉到目标用户移动音乐情境的前提下，判断该用户选择哪种音乐类别的概率最大，则将该用户分类到其下的类别集合中。

为便于对算法的应用进行理解，表5列出了朴素贝叶斯分类算法常用的几个公式定义。

表5 朴素贝叶斯分类算法的表达式及含义

表达式	含义
$X=\{a_1,a_2,a_3,\cdots,a_m\}$	X为待分类项，a_i为X的特征属性
$C=\{y_1,y_2,y_3,\cdots,y_n\}$	C为类别集合，y_i为各个类别
$P(y_1\mid x)*P(y_2\mid x)\cdots$	在已知x的情况下，类别为y_i的概率
$P(a_1\mid y_1)*P(a_m\mid y_n)$	在各个已知类别下，各个特征属性的概率

在上一节的算法展开中，已经提出将$P(y_i\mid x)$按照朴素贝叶斯分类模型进行拆解。在各个特征属性为独立的情况下，可以得到如下表达：

$$P(y_i\mid x)=\frac{P(x\mid y_i)*P(y_i)}{P(x)} \tag{6}$$

要得到最符合待分类项的分类概率，只需要计算等式右边的最大值即可。对于$P(x)$，集合C中的所有分类项都是一样的，即为常数，则只需要计算分子的最大值。由于此处默认的特征属性独立，因此有：

$$P(x\mid y_i)*P(y_i)=P(y_i)*\prod_{j=1}^{m}P(a_j\mid y_i) \tag{7}$$

因此得出，要得到最终待分类项的结果，只需要计算各个类别的$P(y_i)*\prod_{j=1}^{m}P(a_j\mid y_i)$数据即可，其最大值项即为分类结果。

基于上文基本分类算法的理论，可以总结出朴素贝叶斯分类算法的流程步骤（见图4），以便于在实际问题应用中逐步解决问题。

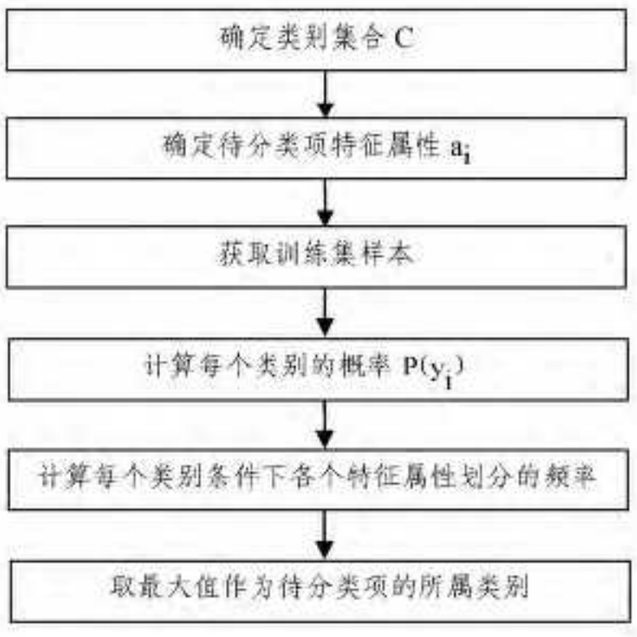


图4 朴素贝叶斯算法的流程步骤

②基于朴素贝叶斯分类算法的模型构造

在以上步骤的指导下，对本文中的“目标用户移动应用类别偏好”问题进行算法应用。首先，对计算过程中需要使用到的表达式进行定义和解释，如表6所列。

表6 本文应用中的算法表达式及含义

表达式	含义
$X=\{b_1,b_2,b_3,b_4\}$	X为目标用户，b_i为X的特征属性
$V=\{v_1,v_2,v_3,\cdots,v_n\}$	V为类别集合，v_i为歌曲各个类别
$P(v_1\mid x)*P(v_2\mid x)\cdots$	在已知目标用户情境情况下，选择类别v_i的概率；
$P(b_1\mid v_1)*P(b_3\mid y_n)$	在各个已知类别下，各个特征属性的概率

第一，确定歌曲类别集合V。

首先，需要将数量众多的歌曲集中地划分为几种歌曲分类，在qq音乐、酷狗音乐、百度音乐等不同的网络音乐库中，音乐类别的基本情况如图5所示。根据这些分类所含歌曲的数量、收听情况等因素，进行应用类别集合V的确定。

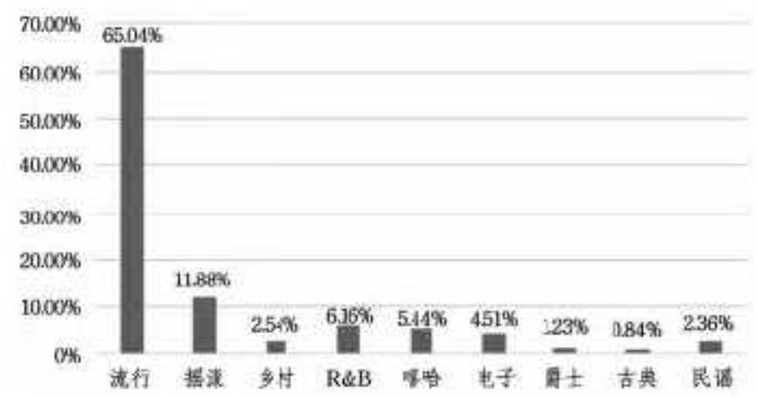


图5 音乐库中音乐分类及数量占比

根据ISO及安卓平台的音乐分类及数量占比情况，可看出流行、摇滚、电子、R&B、嘻哈这几类音乐在音乐市场中的数量比例较大，是用户下载较为普遍的音乐类型。因此，类别集合V有如下定义：

$V=\{$流行，摇滚，R&B，嘻哈，电子$\}$

第二，确定目标用户的移动音乐情境属性b_i。

在实际的情境中，可以用于描述用户收听歌曲习惯的因素有很多，特征属性的分类标准也比较详细，这里根据上文所分析的用户特征，选取比较有代表性的几种特征来收集数据和计算，如表7所列。

表7 特征属性及结果集

字母	含义	结果集
b_1	用户收听音乐时所处位置	{交通工具，休闲场所，住宅，工作环境，夜店迪吧}
b_2	网络状况	{网络状况极差，网络状况较差，网络状况一般，网络状况较好，网络状况极好}
b_3	所处时间段	{上班高峰，午餐时间，下午茶时间，下班高峰，下班之后}
b_4	是否是某些重大节日	{是，否}

第三，选取训练样本。

根据用户已有数据，获取不同v_i的特征属性情况，得到如下值，此部分数据将在实验验证阶段获取。

$P(b_i\mid v_j)(1\leqslant i\leqslant 3;1\leqslant j\leqslant 7)$

选取用户在音乐平台上的数据，则得到$P(v_i)$的相关数据如下：

$P(v_1)=25181/38717=0.6504$

$P(v_2)=4599/38717=0.1188$

$P(v_3)=2385/38717=0.0616$

$P(v_4)=2105/38717=0.0544$

$P(v_5)=1748/38717=0.0451$

第四，根据训练集样本，计算每个类别下的特征属性条件概率。

4 实验验证

本文根据用户在移动终端收听音乐的特点，修正了协同

过滤算法，形成了新的用户兴趣推荐模型。下面将用真实的数据评估文中所提出的协同过滤修正算法进行歌曲推荐的准确度，即推荐的质量。根据对参考文献的阅读，一般协同过滤算法的评价引用准确度(Precision)作为考核推荐结果的标准，而本文将通过比较一般协同过滤算法和基于移动终端特征修正过后的协同过滤算法的准确度高低来评判算法的正确性和有效性。

对于数据集的获取，通过综合考虑我国几大主要音乐平台的音乐种类、音乐歌手、音乐下载量等因素，考虑歌曲150首，分为5类：流行、摇滚、乡村、R&B、嘻哈，每类30首。实验中将对算法中总结出的移动终端情境——用户所处位置、网络状况、所处时间段、是否是重大节日4个因素进行考察。本实验的测试者共30人，均是19～25岁的学生。实验共有两部分，第一部分是通过对测试者进行实验，获取贝叶斯训练集；第二阶段是通过对比两算法的准确度，评估推荐效果。

(1)训练集生成

在本阶段，向30位参与测试的实验者每人提供30首待评分的歌曲(来自不同类别)，并通过系统设置保证每人所评价的音乐列表有至少50%的重复。为得到传统协同过滤算法的数据，将要求测试者对每个待评价的歌曲按照“10分制”进行评分，为保证数据的真实性，要求每个测试者在30首歌曲中，相同评分数量不得超过6个。即使选择音乐平台中下载数量靠前的歌曲作为数据源，仍存在部分实验者从未接触过目标歌曲的情况，在这种情况下，实验者可选择不评分。为避免实验数据的稀疏性，实验要求每个实验者至少完成50%的歌曲评分，将未达标的测试者数据删除，直至合格数据达到30份，停止收集评分。为收集对应用户的歌曲收听情境，分别让使用IOS及安卓的测试者安装speedtest及GPS定位，这两款软件可以在手机后台实时捕捉到所有本机收听音乐时的使用情况，包括用户所处位置、网络状况、所处时间段、是否是重大节日等。以两周为时间跨度，对该软件收集到的数据进行记录、统计和整理。

(2)算法推荐歌曲效果评估

在本阶段，首先将上一阶段所获得的训练集信息整理为贝叶斯公式中描述的相应情境对应类别选择概率的形式。然后，让测试者提供自己在移动端收听音乐的特征信息$\langle b_1, b_2, b_3, b_4\rangle$，根据训练集确定其类别偏好，并根据其音乐评分情况选择相似邻居，根据式(5)进行推荐得分计算，得分最高的TOP-10音乐确定为新算法为其推荐的偏好音乐。用户通过给推荐列表里的音乐评分，来提供偏好是否准确的数据信息。通过对参考文献的学习，大多数协同过滤推荐算法都通过计算准确率或召回率来衡量推荐效果，在这里选择使用准确率来进行效果考核。

(3)实验设计

上文中提到，本实验将选择准确率来考核推荐名单是否准确。准确率是指推荐列表中被用户标注为喜欢的项占该列表项总数的比例。准确率的计算方法表述如下：

$$Precision=|A\cap B|/|B| \tag{8}$$

其中，A表示用户标注为“喜欢”的项集合；B表示推荐列表中的所有的项集合；$|A\cap B|$为推荐准确的用户偏好的音乐项数量；$|B|$表示列表中所有项的总数量。准确率越高，证明推荐列表中符合用户偏好的音乐数量越多，反之亦然。在本实验中，将通过原协同过滤推荐算法、基于移动终端特征的推荐算法两种方式，分别给出两种推荐名单，每份推荐名单给出10首音乐，其中评分靠前的7首音乐推荐给用户。由于本实验涉及的实验样本较少，采用原有的准确率计算方法表达数据差异不大，因此采取变形的准确率计算方法：

$$Ru=\frac{\mathrm{abs}(Ra_j-Rb_j)}{Size(R)} \tag{9}$$

因此，在数据收集过程中，测试者将根据名单，给出心理预期下的歌曲偏好排名。在这种准确率的计算方法下，可以更精准地衡量根据该用户收听音乐特征修正的名单是否真正符合了用户的需求。

(4)数据处理

为方便描述统计结果，对b_1, b_2, b_3, b_4的区间分类进行定义，如表8所列。

表8 移动特征参数简称对应表

特征参数\简称	a	b	c	d	e
b_1	交通工具	休闲场所	住宅	工作环境	夜店迪吧
b_2	网络状况极差	网络状况较差	网络状况一般	网络状况较好	网络状况极好
b_3	上班高峰	午餐时间	下午茶时间	下班高峰	下班之后
b_4	是	否			

(5)协同过滤算法计算

在所有收集到的30位测试者的数据中，将挑选5位用户作为待推荐测试者，其在协同过滤推荐算法及贝叶斯算法中所需的各项参数情况提取如表9所列。

表9 5位测试者的移动应用情境情况

测试者\参数	平均得分	b_1	b_2	b_3	b_4
1	7.00	d	d	b	a
2	7.33	b	c	c	b
3	7.39	a	b	d	a
4	6.84	c	c	d	a
5	6.07	e	b	e	b

计算贝叶斯算法在各类别偏好下，各移动特征的占比情况。此处，将测试者在所有音乐的评分情况分类后的平均数据作为该用户对该类别的评分。综合真实实验中所有类别的得分情况，将7分作为偏好水平线，即若某首音乐的得分超过7分，则将其作为用户偏好的一类音乐。

根据测试模板，对用户评分信息收集后进行预处理，得到的数据集合如图6所示。

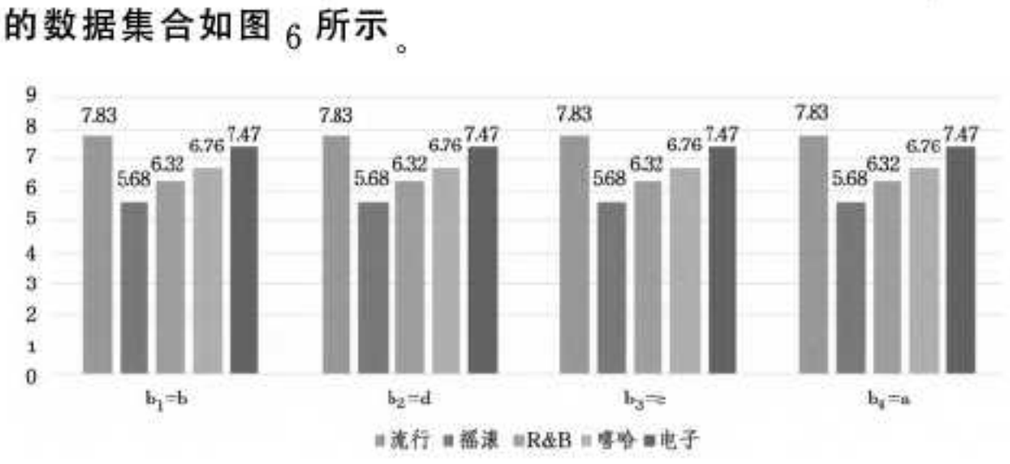


图6 测试者1各音乐类别平均得分数据情况

该数据表明，在带有(b,d,c,a)移动应用特征下的用户，对流行和电子类型的音乐偏好程度高(评分>7)。对形式相同的30份数据做同样的处理，获取各类别偏好测试者带有不同应用特征的概率(见图7)，即获取$P(b_i|v_i)$，其中$v_i\in V, b_i\in B$。

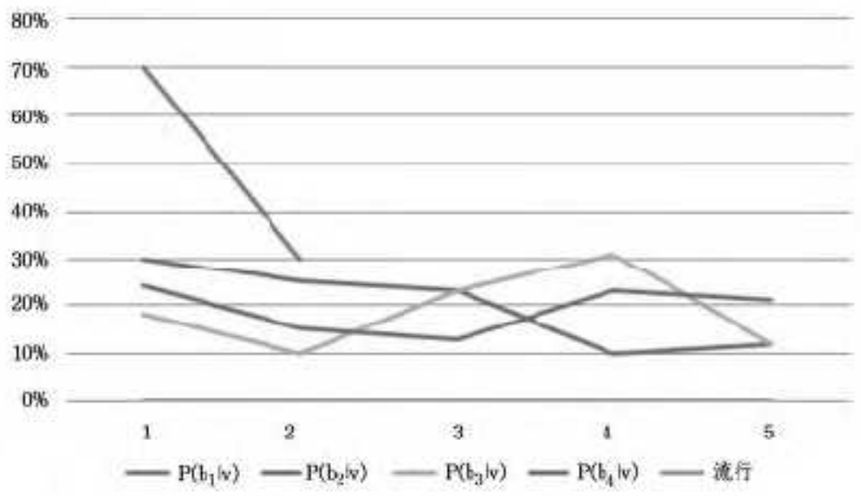


图 7　各类别偏好测试者中带有不同应用特征的概率分布

(6)实验结果

按以上方法处理数据，根据贝叶斯数据集计算修正系数(以测试者 1 为例)，数据的散点图分布如图 8 所示。

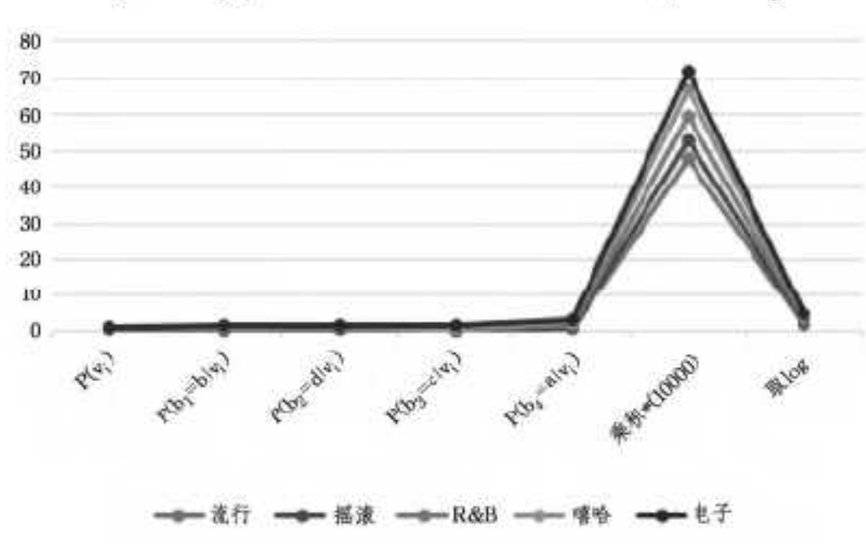


图 8　修正系数的散点图分布

根据协同过滤算法得原始推荐分数及音乐(以测试者 1 为例)，如图 9 所示。

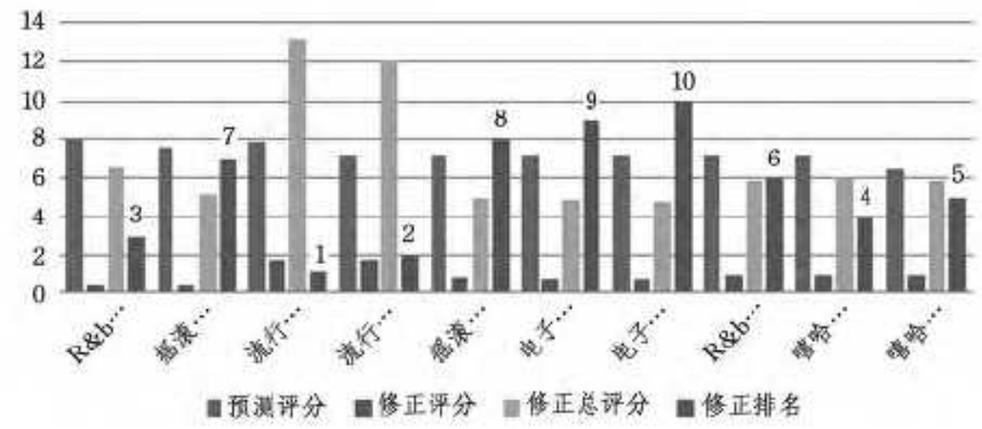
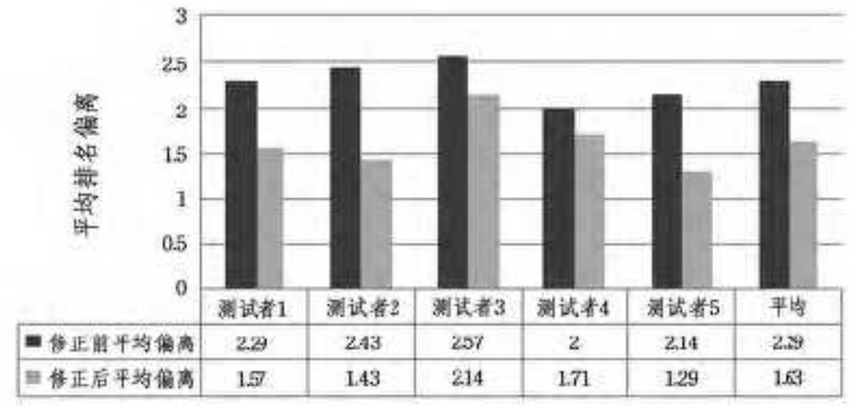


图 9　协同过滤推荐后的分布

(7)精确性对比

分别将两份名单中排名靠前的 7 首音乐推荐给测试者，测试者根据偏好情况，给出 1 到 7 的偏好排名。按平均排名偏离水平方法比较两种名单的排名准确度，5 位测试者的数据如图 10 所示，其中第六列为前 5 列的平均值。



	测试者1	测试者2	测试者3	测试者4	测试者5	平均
修正前平均偏离	2.29	2.43	2.57	2	2.14	2.29
修正后平均偏离	1.57	1.43	2.14	1.71	1.29	1.63

图 10　综合推荐排名与原推荐排名准确度的比较

(8)实验总结

通过实验可看出，修正后的排名提高了推荐音乐排名的精准性，说明在综合了实验者的移动应用情境后，排名更接近测试者的真实偏好。综合考虑移动应用特征后，对原推荐排名进行调整，将更符合用户的真实偏好，接近用户情境。

结束语　本文定位于根据移动端用户音乐行为特征，完善其在音乐市场得到的音乐推荐准确性。在传统的协同过滤算法中，两维用户_项目模型(U-I 模型)为普遍认可的算法模型。通过寻找相似邻居的项目评分，确定目标用户在 U-I 中的位置，进行应用推荐。在该模型中加入 *Situation*(情境)维度，在用户进入模型时，首先寻找其相似情境，确定 *Situation* 维度的值，再在该情境下寻找相似邻居。第三维情境模式，包括所处位置、网络状况、所处时间段、是否是某些重大节日。将这 4 种元素通过贝叶斯算法聚合为用户对某种应用的类别偏好，由此将 *Situation* 更加详细地描述为类别情境。添加类别情境的优势在于，更加精准地判断某个相似邻居是否为情境相同的相似邻居。

参考文献

[1] 莫同，李伟平. 一种情境感知服务系统框架[J]. 计算机学报，2010，33(11)：2084-2092

[2] 周玲元，段隆振. 移动情境感知服务系统研究[J]. 图书馆学研究，2014，36(8)：36-44

[3] 徐步刊，周兴社，等. 一种场景驱动的情境感知计算框架[J]. 计算机科学，2012，39(3)：. 216-222

[4] 张静. 基于情境感知的自适应个性化知识服务研究[J]. 情报科学，2011，29(11)：1658-1661

[5] 叶剑，李锦涛，等. 分布式情境感知分层模型设计与角色分析[J]. 电子学报，2012，8(8)：1572-1575

[6] 於志勇，周兴社，王海鹏，等. 动态上下文知识的获取与共享[J]. 计算机科学，2009，36(9)：218-223

(上接第 502 页)

[4] Yamaguchi Y, Amagasa T, Kitagawa H. Tag-based User Topic Discovery Using Twitter Lists[C]// Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining. 2011：13-20

[5] Ma Y, Zeng Y, Ren X, et al. User Interests Modeling based on multi-source personal information fusion and semantic reasoning [C]// Proceedings of the 7th international conference on active media technology. 2011：195-204

[6] Chen J, Nairn R, Nelson L, et al. Short and tweet：experiments on recommending content from information streams[C]// Proceedings of the 28th International Conference on Human Factors in Computing Systems. 2010：1185-1194

[7] Wen J, Lim E P, Jiang J, et al. TwitterRank：finding topic-sensitive influential twitterers[C]// Proceedings of the third ACM International Conference on Web Search and Data Mining. 2010：261-270

[8] 宋巍，张宇，谢毓彬，等. 基于微博分类的用户兴趣识别[J]. 智能计算机与应用，2013，3(4)：80-83

[9] Chen J, Nairn R, Nelson L. Short and tweet：experiments on recommending content from information streams[C]// Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2010：1185-1194

[10] Zhao Hua, Zeng Qing-tian. Micro-blog Hot Event Detection Based on Dynamic Event Model[J]. Lecture Notes in Artificial Intelligence，2013，8041：161-172

[11] http://wordpress.org/plugins/cardoza-3d-tag-cloud/