第一章 背景介绍

随着互联网技术的快速发展，信息呈指数级增长，这造成了严重的信息过载问题[1,2]。特别是随着大数据时代的来临，信息的增长速度已经远远超过人们可以理解的速度，而海量信息中更是包含着很多对人们来说无用的冗余信息以及会干扰人们抉择的错误信息。在面对大量的信息时，人们往往无所适从，进而陷入到选择悖论(Paradox of Choice)[3]之中，或无法做出合理的选择，或需要消耗很大的精力才能做出正确选择。

为了解决信息过载问题，减轻人们在做出抉择时所承受的沉重负担，信息分类、搜索引擎和推荐系统等技术应运而生。其中，信息分类通过将互联网上的各种信息分门别类的组织起来以提高人们查询的效率，比如早期的 Yahoo!和国内的 58 同城的信息分类网站。然而，信息分类技术往往依赖于人工，随着互联网上信息的增加，显然不合时宜。搜索引擎通过建立索引的方式将互联网上的网页组织起来，然后接收用户的关键字进行查询，比如 Google 和百度。显然，搜索引擎相对于信息分类技术来说已经有了很大的进步，能够适应互联网的快速发展。但是，人们很多时候要么不愿意费时费力去输入关键字，要么无法准确地用关键字去描述自己的想法，比如“今天看什么电影好呢？”“去哪儿吃饭呢？”等，搜索引擎此时就无能为力了。为了在解放人们双手的同时挖掘人们的需求，推荐系统出现了。

名称	特点	典型案例
信息分类	分门别类的组织信息	Yahoo!、58 同城
搜索引擎	根据用户输入的关键字进行查询	Google、百度
推荐系统	分析用户行为历史，主动推荐	Amazon、淘宝

推荐系统本质上是一种信息过滤系统，其通过对用户行为历史的分析挖掘出用户的行为偏好，进而帮助用户将海量信息中的无用信息过滤掉，进而将符合用户偏好的信息推荐给用户。目前，推荐系统已经在各个领域得到了广泛的应用，比如电子商务领域的亚马逊和淘宝，在线视频领域的 Netflix 和优酷，在线音乐领域的 Lastfm 和豆瓣以及个性化阅读领域 Flipboard 和无觅阅读等。随着商品推荐系统的深度整合，亚马逊在 2012 年第二财季中营收达到了 128.3 亿美元，与 2011 年同期的 99 亿美元相比大涨了 29%。图 1.1 展示了电子商务领域 Amazon

的商品推荐系统界面和视频领域 PPTV 的在线视频推荐系统界面的截图。

音乐推荐系统是推荐系统在音乐领域的应用，其旨在将人们从海量的音乐中解脱出来，通过分析用户的收听习惯以及歌曲本身的特征来为用户推荐符合口味的歌曲。音乐推荐与传统物品推荐的区别在于歌曲时长比较短、消费代价低、可重复消费以及具有强烈的情感性等特点，而人们的听歌行为往往对上下文环境往往更为敏感。比如，在不同的天气、不同的情绪以及不同的场合，人们的音乐喜好都会有所不同。因此，在做音乐推荐的时候需要充分的考虑歌曲和人们收听行为的特殊性。

为了满足人们对音乐的需求，一些个性化的音乐推荐系统被设计和实现出来，如国外的 Lastfm、Pandora、Spotify 等以及国内的豆瓣电台、虾米音乐以及 Jing 音乐等，图 1.2 展示了豆瓣电台和虾米音乐的推荐界面截图。这些音乐推荐

系统首先建立自己的曲库，然后分析歌曲的特征和用户的听歌行为习惯继而为用户做出推荐。其中，Pandora 是当今最流行的音乐推荐系统，其通过 Music Genome Project 将 400 种熟悉分配给每一首歌曲，进而按照歌曲的相似程度为用户做出推荐。从可见的资料看，Lastfm、豆瓣主要采用协同过滤的推荐算法为用户做出推荐。而虾米音乐还给出了推荐的同时还给出了做出推荐的原因，这在一定程度上提升了用户的接受程度。

尽管这些音乐推荐系统能够取得不错的推荐效果，但其仍存在以下几个问题：

1. 在对歌曲进行描述和刻画时，没有充分考虑歌曲划分标准的不确定性以及歌曲本身的多重性，导致对歌曲的描述不够全面和完整。此外，Pandora 对歌曲属性的划分往往由具有相关领域知识的专业人士进行，对专业要求过高，代价大，缺乏一定的可扩展性。

2. 没有充分考察用户所处上下文环境对用户的影响，特别是用户听歌序列对其所处上下文环境的表征作用，导致多用户行为习惯的刻画有所偏差。Lastfm、豆瓣等系统考察了用户所处的群体环境，却忽视了用户本身所处的环境，导致所做的推荐不一定符合用户当前的状态。

本文提出了一种基于多维时间序列分析的个性化音乐推荐系统框架，挖掘了歌曲划分标准的不确定性和歌曲本身的多重性，验证了用户在一定会话期内听歌行为中的序列性特征，提出了一种基于多维时间序列分析的个性化音乐推荐方法，再者本文挖掘了用户听歌行为的局部性特征和全局性特征，并将之与时序特征进行混合，最后本文实现了一个个性化音乐推荐原型系统，验证了该框架的有效性。

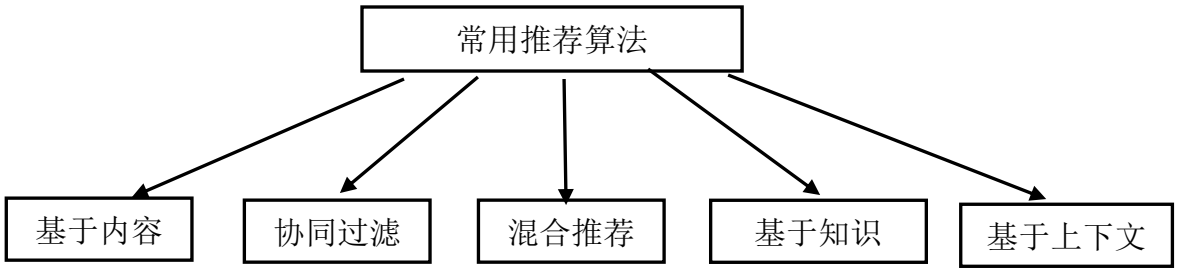
本文的组织结构如下：第 2 章介绍了推荐系统的相关工作，包括常用的推荐算法、常用的相似度计算方法、推荐系统的评测指标以及文本处理、时间序列分析的相关理论等；第 3 章描述了本文所提的一种基于多维时间序列分析的个性化音乐推荐方法，包括问题的定义、方法的执行流程等；第 4 章描述了用户听歌行为的局部性特征和全局性特征并以此为基础的个性化音乐推荐系统框架；第 5 章给出了原型系统的实现细节；第 7 章对全文进行总结并给出对未来工作的展望。

第二章　相关工作

这一章我们将介绍与本文工作相关的工作，首先我们介绍推荐系统领域常用的一些推荐算法，相似度计算方法和评估标准；其次，我们将介绍文本建模，时间序列分析以及实时流处理框架的相关工作，其中的很多方法和技术将在我们的原型系统中得到应用；最后我们讨论与本文工作相关的一些音乐推荐算法。

2.1 常用推荐算法

目前，推荐系统领域的研究工作已经取得了长足进度也出现了很多类型的推荐算法，按照检索物品方式的不同，这些算法大致可以分为基于内容的推荐、协同过滤推荐、基于知识的推荐、基于上下文的推荐以及混合推荐算法几类，本节将对这些推荐算法进行简单的介绍。



2.1.1 基于内容的推荐

基于内容的推荐(Content-based Recommender System)主要是向用户推荐与其过去喜欢的物品相似的新物品，比如用户几天前刚在亚马逊上购买了吴军博士的《数学之美》，那么系统会据此向其推荐吴军博士的另一本著作《浪潮之巅》。如文献[]所述，基于内容的推荐主要包括以下三个阶段：

1. 抽取物品的特征。比如，音乐可以有类型、创作者、创作年代等属性。

2. 抽取用户的特征。用户的特征既包括其对应的人口学特征，比如年龄、性别、职业等，也包括从其过去购买、评分和喜欢等行为中提取的特征。

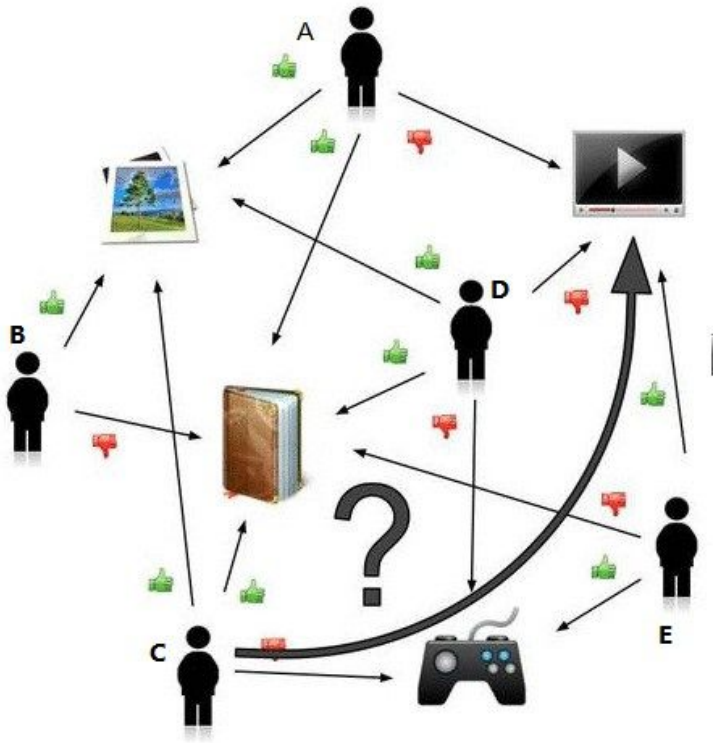
3. 根据用户的特征向其推荐与其特征匹配的新物品。比如，向青少年用户推荐新晋歌手的新歌《》，向老年用户推荐经典老歌《东方红》，向喜欢怀旧的用户推荐《海阔天空》。

基于内容的推荐往往能够为用户做出正确的推荐，而且能够在推荐的同时给

出关于推荐的解释，这有助于提升推荐的效果。但是，准确全面的抽取物品的特征往往是很困难的，一方面这需要较强的专业领域知识，比如音乐推荐系统Pandora 发动专家为音乐进行特征提取，这种方法耗时耗力，不具可扩展性。另外，物品的特征往往是不可完全列举的，这在一定程度上限制了基于内容的推荐算法的发展。

2.1.2 协同过滤的推荐

协同过滤推荐(Collabrative Filtering, CF)是当前最为流行的一类推荐算法，其挖掘用户所处的社会化环境，利用群体智能为用户做出推荐。协同过滤推荐基于这样一个假设，即如果两个用户在过去有相同的行为，那么系统认为他们在未来也会有类似的行为。当需要预测一个用户是否喜欢一个物品时，协同过滤推荐算法首先找到与当前用户喜好类似的用户，进而综合这些相似用户的喜好为用户推荐新的物品。如图所示，系统试图为用户 C 做出推荐，发现用户 D 和用户 C 都喜好图片和书籍且都不喜欢游戏，那么系统认为用户 D 和用户 C 是相似用户，参考用户 D 不喜欢视频，那么系统认为 C 也不喜欢视频。



与传统的基于内容的推荐方法相比，协同过滤推荐具有领域无关的特点，因此能够方便的推广应用到各个领域，比如音乐、电影、电子商务等。另外，协同过滤

推荐充分考虑了群体智能在推荐中的应用，因此往往能够取得比较高的推荐准确率。上文所述的协同过滤称之为基于用户的协同过滤(user-based CF)，即通过计算用户之间的相似度来进行推荐，也可以通过计算物品之间的相似度来做出推荐，这称之为基于物品的协同过滤(item-based CF)，其执行过程与 user-based CF 类似，只是考虑到物品的相对静止特点，其做推荐时不需要繁杂的线上计算，因此也得到了广泛的应用。

2.1.3 基于知识的推荐

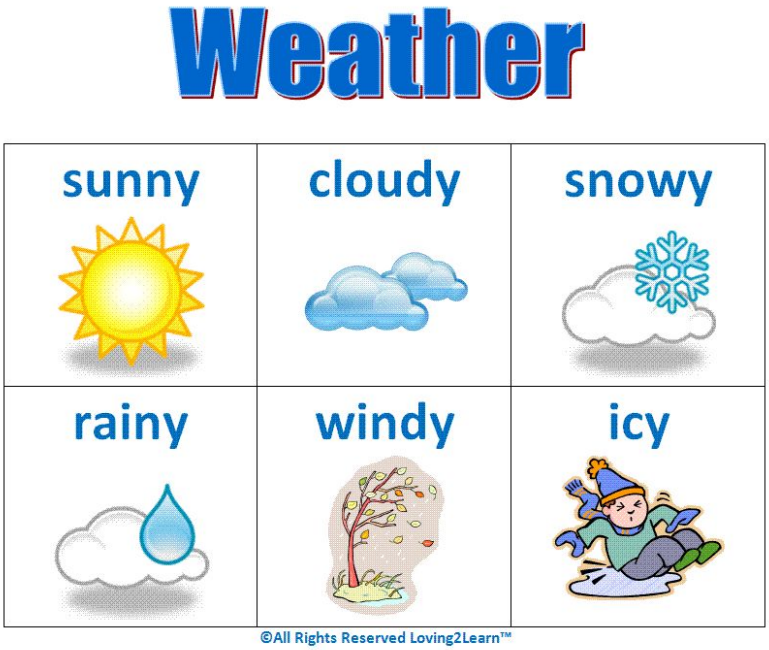
基于知识的推荐是一类特殊的推荐方法，其不依赖于用户的历史行为记录或者评分，而是依赖用户提出的需求或者一定的规则，然后通过计算物品与用户需求或者物品与规则之间的相似度来为用户做出推荐。比如，人们并不会频繁地购买房屋、汽车或者计算机，在这种情况下去维护一个用户的偏好特征就不太合适，而且人们在购买这些物品时往往会附带一些明确的需求信息，比如“汽车的最高价是 x，颜色应该是灰色的”等，这时系统只需要根据用户的需求为之做出推荐即可。基于知识的推荐一般分为基于约束的推荐和基于实例的推荐两类，这两种方法在推荐过程上比较相似：用户必须指定需求，然后系统设法找出解决方案。如果找不到解决方案，那么用户必须修改需求。此外，系统往往还会给出推荐的具体解释。这些推荐系统的不同之处在于如何使用所提供的知识：基于实例的推荐着重于根据不同的相似度度量方法检索出相似的物品，而基于约束的推荐系统依赖明确定义的推荐规则集合。基于约束的系统会在符合推荐规则的所有物品中搜索得出要推荐的物品集合，基于实例的系统会根据相似度衡量标准检索那些与特定用户需求（在预定义阈值内）相似的物品。

2.1.4 基于上下文的推荐

上述的三种推荐方法主要着眼于为用户推荐与其偏好或者需求最相似的物品，却很少考虑用户所处的上下文信息，如时间、地点、天气等。换句话来说，传统的音乐推荐算法主要处理用户和物品两类实体，但并没有将它们放到一定的上下文环境中去分析。然而，很多物品对上下文环境比较敏感，比如对时间信息敏感的旅游产品，对情感信息比较敏感的电影、音乐以及对地点信息比较敏感的餐馆等，因此将上下文信息融入到为特定环境下的用户推荐物品中是非常重要的。比如，可以为一个刚刚失恋、心情低沉的用户推荐悲伤的歌曲，而对刚刚毕

业、心情舒畅的用户推荐欢快的歌曲。

基于上下文的推荐(Context-aware Recommender System,CARS)就是这样一种推荐算法，其通过一定的手段收集用户所处的上下文环境信息，进而分析用户、物品、上下文环境这三个实体，然后将与用户所处上下文环境匹配的物品推荐给用户。随着移动互联网的发展，用户所处的时间、地点、状态等信息可以比较容易地通过传感器获得，因此基于上下文的推荐获得了比较长足的发展，比如各种典型的 LBS 应用。



2.1.5 混合推荐方法

上述的几类推荐方法各有各的优势，但也各有各的劣势。混合推荐方法就是这样一类推荐方法，其将不同推荐方法综合利用，以规避单一推荐方法的缺点，充分利用单一推荐方法的优点。比如，基于内容的推荐与协调过滤推荐的混合推荐方法、基于内容的推荐与协调过滤推荐的混合推荐方法、基于上下文的推荐与协同过滤的推荐、基于知识的推荐与基于内容的推荐等。本文所述的基于多维时间序列分析的个性化音乐推荐算法本质上也是一直基于内容的推荐和基于上下文的推荐相关联的推荐方法。

名称	优点	缺点
基于内容的推荐	用户独立、可解释	领域相关

协同过滤推荐	群体智慧、领域无关	冷启动、长尾效应
基于知识的推荐	特殊应用场景	不够通用
基于上下文的推荐	考虑上下文环境的影响	上下文获取有一定难度

2.2 常用相似度衡量标准

如前节所述，无论是哪一种推荐算法，都需要进行相似度的计算，本节简单介绍一些常用的相似度度量标准。

2.1 余弦相似度，是推荐系统中计算用户与用户或者物品与物品之间相似度的常用方法，通过测量两个向量内积空间的夹角的余弦值来度量它们之间的相似性。0 度角的余弦值是 1，而其他任何角度的余弦值都不大于 1，并且其最小值是-1。如果两个向量的指向越接近，那么它们内积空间夹角的余弦值越接近于 1，即二者越相似。相反，如果两个向量的指向越相离，那么它们内积空间夹角的余弦值越接近于-1，即二者越不相似。设向量 $\mathbf{X_a}$ 和向量 $\mathbf{X_b}$ 分别表示用户 a 和用户 b 的喜好向量，其中 $\mathbf{X_a}$ 中的元素 Xa(i)表示用户 a 对编号为 i 的物品的喜好值(评分)，$\mathbf{X_b}$ 中的元素 $X_b(i)$表示用户 b 对编号为 i 的物品的喜好值(评分)。那么，用户 a 和用户 b 的相似度可以按照下式计算。其中，K 表示物品集合中物品的数目。

$$\mathrm{sim}(a,b)=\cos(\mathbf{X_a},\mathbf{X_b})=\frac{\sum_{i=1}^{K}X_a(i)X_b(i)}{\sqrt{\sum_{i=1}^{K}X_a(i)^2}\sqrt{\sum_{i=1}^{K}X_b(i)^2}}$$

2.2 Pearson 相关系数，是一个取值介于-1 和 1 之间的用于衡量两个变量线性相关程度的指标，定义为两个变量协方差和标准差的商。当两个变量的线性关系增强时，相关系数趋于 1 或-1，这种线性关系分为正相关和负相关。当一个变量增大，另一个变量也增大时，表明它们之间是正相关的，相关系数大于 0；如果一个变量增大，另一个变量却减小，表明它们之间是负相关的，相关系数小于 0；如果相关系数等于 0，表明它们之间不存在线性相关关系。Pearson 相关系数有一个非常重要的特性，即两个变量的位置和尺度变化并不会引起其取值的变化，这再一定程度解决了余弦相似度计算中不同用户之间喜好值尺度不一的问题。

$$\mathrm{sim}(a,b)=\frac{\sum_{i=1}^{K}(X_a(i)-\overline{X}_a)(X_b(i)-\overline{X}_b)}{\sqrt{\sum_{i=1}^{K}(X_a(i)-\overline{X}_a)^2\sum_{i=1}^{K}(X_b(i)-\overline{X}_b)^2}}$$

2.3 欧几里得距离，是一种非常朴素的用于计算两个点在欧几里得空间中距离的方法，引申用来计算用户或者物品之间的相似度，取值越小表示用户或物品越相似。

$$dis(a,b)=\sqrt{\sum_{i=1}^{K}(X_a(i)-X_b(i))^2}$$

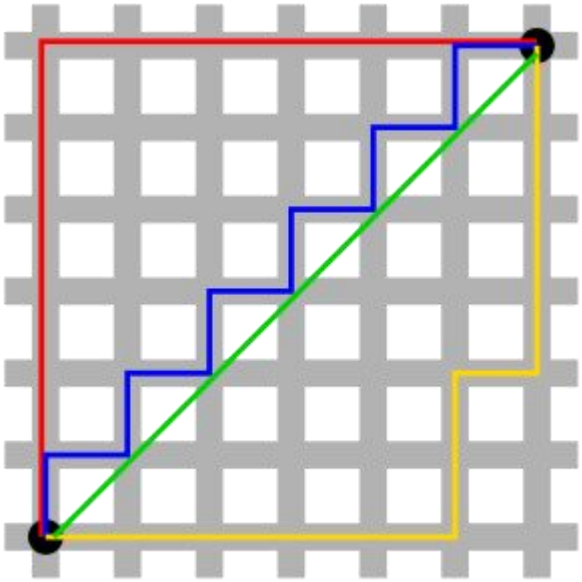
$$sim(a,b)=\frac{1}{1+dis(a,b)}$$

2.4 曼哈顿距离，用于计算两个点在标准坐标系上的绝对轴距之和，计算公式如下所示。

$$dis(a,b)=|X_a(i)-X_b(i)|^2$$

$$sim(a,b)=\frac{1}{1+dis(a,b)}$$

下图展示了欧几里得距离和曼哈顿距离在棋盘上的计算方式，其中红、蓝、黄三条线分别表示所有曼哈顿距离都拥有一样长度，而绿线表示欧几里得距离的长度。



2.5 Jaccard 指数，也称 Jaccard 相似度，是用于计算两个集合相似程度的指标。设 N(a)表示用户 a 曾经有过正反馈的物品集合，N(b)表示用户 b 曾经有过正反馈的物品集合，那么我们可以通过下式来计算这两个集合的 Jaccard 相似度并进而用之作为用户 a 和用户 b 的相似度。显然，s(a,b)的取值在 0 和 1 之间。如果 N(a)和 N(b)均为空，我们定义 s(a,b)=1。

$$s(a,b)=\frac{|N(a)\cap N(b)|}{|N(a)\cup N(b)|}$$

2.6 KL 距离，也称相对熵或 KL 散度，是两个概率分布 P 和 Q 差别的非对称性的度量，用来度量使用基于 Q 的编码来编码来自 P 的样本平均所需的额外的比特个数。如果能够把物品表示称一个概率分布，那么显然可以用 KL 距离来衡量两个物品之间的相似度。设 $P=(p_1,...,p_i,...p_k)$ 表示物品 p 对应的概率分布，物品 q 对应的概率分布用 $Q=(q_1,...,q_i,...q_k)$ 表示，那么物品 p 和 q 之间的 KL 距离可以按照下式进行计算，即按概率 P 求得的 P 和 Q 的对数差的平均值，其取值非负。KL 散度仅当概率 P 和 Q 各自总和均为 1，且对于任何 i 皆满足 $p(i)>0$ 及 $q(i)>0$ 时才有定义，若式中出现 $0\ln 0$ 的情况，其值按 0 处理。

$$dis_{KL}(p,q)=D_{KL}(P\,||\,Q)=\sum_i p(i)\ln\frac{p(i)}{q(i)}$$

由于 KL 距离是非对称的，即

$$dis_{KL}(p,q)\neq dis_{KL}(q,p)$$

然而，物品之间的相似度应该是对称的，即

$$sim(p,q)=sim(q,p)$$

因此，不能直接使用 KL 距离来计算物品之间的相似度，但我们可以使用 p 到 q 的 KL 距离与 q 到 p 的 KL 距离的平均值来作为二者的最终距离，这样就可以满足对称性的要求，如下式所示。

$$dis(p,q)=\frac{dis_{KL}(p,q)+dis_{KL}(q,p)}{2}$$

$$sim(p,q)=\frac{1}{1+dis(p,q)}$$

2.7 Hellinger 距离，是一种度量两个概率分布之间相似度的方法。与 KL 距离不同的是，其满足对称性，因此可以直接用来计算物品之间的相似度。物品 p 和物品 q 之间的 Hellinger 距离可以按照下式计算，进而按式计算二者之间的相似度。

$$dis(p,q)=\frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{k}\left(\sqrt{p(i)}-\sqrt{q(i)}\right)^2}$$

$$\mathrm{sim}(p,q)=\frac{1}{1+\mathrm{dis}(p,q)}$$

2.3 评测指标

推荐系统的研究包括评分预测问题、Top-N 推荐问题、冷启动问题、可解释性问题以及用户交互问题等多个方面。其中，评分预测问题和 Top-N 推荐问题是最广泛且最重要的研究内容，所谓评分预测即根据用户已经产生的评分记录来预测其对尚未评分物品的可能打分，而 Top-N 推荐是指为用户生成一个包含 N 个符合其偏好的物品列表，两者采用不同的指标进行评测，本节将对此作简要介绍。

2.3.1 用户满意度

用户作为推荐系统的重要参与者也是推荐系统最终的服务对象，其满意度是评测一个推荐系统优劣的最重要指标。然而，这种满意度没法离线计算，只能通过用户调查或者在线实验的方法获得。在在线系统中，用户满意度主要通过对一些用户行为的统计得到。比如，在电子商务网站中可以通过用户的实际购买情况来评判，或者通过设置“满意”“不满意”按钮进行显示的统计。更一般的情况是，我们可以使用点击率、用户停留时间和转化率等指标度量用户的满意度。

2.3.2 预测准确度

预测准确度度量一个推荐系统或者推荐算法预测用户行为的能力，该指标是最重要的推荐系统离线评测指标，从推荐系统诞生的那一天其，几乎 99%与推荐系统相关的论文都在讨论这个指标，这主要是因为该项指标可以通过离线实验计算，方便了很多学术界的研究人员研究推荐算法。

评分预测的预测准确度一般都过均方根误差(Root Mean Square Error, RMSE)和平均绝对误差(Mean Absolute Error, MAE)计算。对于测试集 T 中的一个用户 u 和物品 i，设 r_{ui} 是用户 u 对物品 i 的实际评分，而 $\hat{r}_{ui}$ 是推荐算法给出的预测评分，那么 RMSE 按照下式定义：

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{u,i\in T}(r_{ui}-\hat{r}_{ui})^2}{|T|}}$$

MAE 采用绝对值计算预测误差，其定义为：

$$\mathrm{MAE}=\frac{\sum_{u,i\in T}|r_{ui}-\hat{r}_{ui}|}{|T|}$$

Top-N 推荐一般通过准确率(precision)和召回率(recall)来度量算法的优劣。设 R(u)是根据用户在训练集上的行为给用户作出的推荐列表，而 T(u)是用户在测试集上的行为列表。那么，推荐结果的召回率定义如下:

$$\text{Recall} = \frac{\sum_{u \in U} |R(u) \cap T(u)|}{\sum_{u \in U} |T(u)|}$$

推荐结果的准确度按照如下方式定义:

$$\text{Precision} = \frac{\sum_{u \in U} |R(u) \cap T(u)|}{\sum_{u \in U} |R(u)|}$$

2.3.3 其他评测指标

除了预测准确度这一重要指标之外，推荐系统的评测还有诸如覆盖率、多样性等指标，这些指标从不同的角度看待推荐的有效性。其中，覆盖率(Coverage)描述一个推荐系统对物品长尾的发掘能力，可以简单地定义为推荐系统能够推荐出来的物品占总物品集合的比例。覆盖率越大，说明系统所能推荐的物品越广泛，也说明系统挖掘物品长尾的能力越大。假设系统的用户集合为 U，物品集合为 I，那么覆盖率可以按照下式定义:

$$\text{Coverage} = \frac{|\cup_{u \in U} R(u)|}{|I|}$$

多样性用来描述推荐列表中物品两两之间的不相似性，列表中物品两两之间的相似性越小表示推荐的多样性越大。假设 $s(i,j) \in [0,1]$ 表征物品 i 和 j 之间的相似度，那么用户 u 的推荐列表 R(u)的多样性定义如下：

$$\text{Diversity}(R(u)) = 1 - \frac{\sum_{i,j \in R(u), i \neq j} s(i,j)}{\frac{1}{2}|R(u)|(|R(u)|-1)}$$

而推荐系统的整体多样性可以定义为所有用户推荐列表多样性的平均值：

$$\text{Diversity} = \frac{1}{|U|} \sum_{u \in U} \text{Diversity}(R(u))$$

2.4 实时性流计算框架

尽管在摩尔定律的影响下单机的处理能力已经得到极大的提升，然而其在应对大数据时代产生的海量数据时非常吃力。为了解决大数据处理和分析的问题，Google 提出了一种分布式的计算模型，即 MapReduce，使得由一般机器组成的集群可以完成大规模的计算工作。在 Google 工作的启发下，Apache 于 2005 年开发了目前广泛得到应用的分布式应计算框架 Hadoop。然而，Hadoop 天然适用于批处理的工作，对于一些实时性要求较高的流计算有所欠缺。为此，S4、Storm、Puma 等实时性流计算框架如雨后春笋般出现了，其中 Storm 是由 Twitter 发布的一款开源的分布式实时计算框架，目前已经得到广泛的应用，本节将对此进行简单介绍。Storm 主要适用于流数据处理和分布式远程过程调用(Distributed Remote Procedure Call, DRPC)两种场景。其中，在流数据处理场景中，Storm 可以用来处理源源不断流进来的消息，处理之后将结果写入到某个存储中去。此外，由于 storm 的处理组件是分布式的，而且处理延迟极低，所以可以作为一个通用的分布式远程过程调用框架来使用。本节下面将简要介绍 Storm 框架的基本概念和其在分布式过程调用场景中的使用。

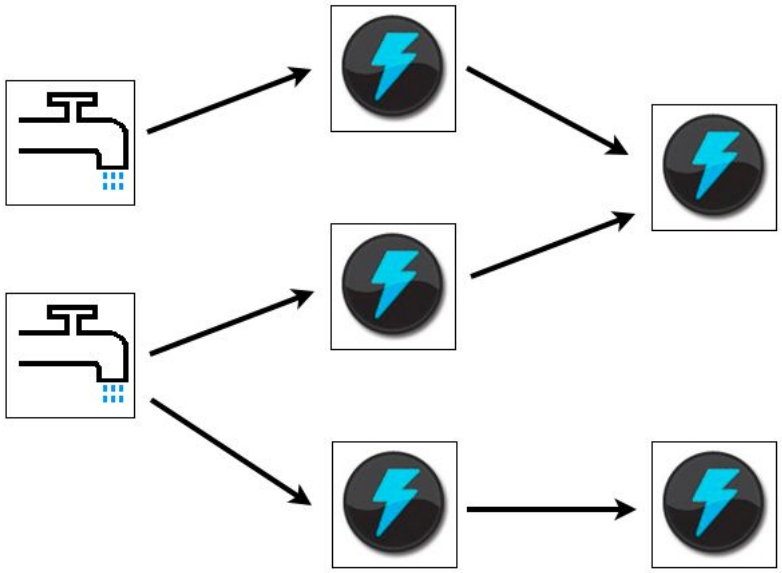
2.4.1 Storm 基本概念

Storm 集群由一个主节点和多个工作节点组成。主节点运行了一个名为“Nimbus”的守护进程，用于分配代码、布置任务及故障检测。每个工作节 点都运行了一个名为“Supervisor”的守护进程，用于监听工作，开始并终止工作进程。Nimbus 和 Supervisor 都能快速失败，而且是无状态的，这样一来它们就变得十分健壮，两者的协调工作是由 Apache ZooKeeper 来完成的。

在 Storm 中，一个实时应用的计算任务被打包作为 Topology 发布，这同 Hadoop 的 MapReduce 任务相似，但有一点不同的是：在 Hadoop 中，MapReduce 任务最终会执行完成后结束；而在 Storm 中，Topology 任务一旦提交后永远不会结束,除非你显示去停止任务。计算任务 Topology 是由不同的 Spouts 和 Bolts 通过数据流连接起来的拓扑图。其中，Spout 是 Storm 中的消息源，用于为 Topology 生产消息（数据），一般是从外部数据源（如 Message Queue、RDBMS、NoSQL、Realtime Log）不间断地读取数据并发送给 Topology 消息 (tuple 元组)；Bolt 是 Storm 中的消息处理者，用于为 Topology 进行消息的处理，Bolt 可以执

行过滤，聚合，查询数据库等操作，而且可以一级一级的进行处理。Topology 中每一个计算组件（Spout 和 Bolt）都有一个并行执行度，在创建 Topology 时可以进行指定，Storm 会在集群内分配对应并行度个数的线程来同时执行这一组件。下图是 Twitter Storm 官方给出的一个典型的 Topology 示意图，其中水龙头表示用以生产数据的 Spout 单元，闪电表示用户处理数据 Bolt 单元。

除此之外，Storm 还有 Stream、Stream Grouping、Task、Worker 等关键概念。其中，Stream 表示被处理的数据，Stream Grouping 表示 Bolt 接收什么样的数据作为输入数据，Task 表示运行于 Spout 和 Bolt 中的线程，Worker 表示运行这些线程的进程。



2.4.2 DRPC

分布式 RPC（Distributed Remote Procedure Call，DRPC）用于对 Storm 上大量的函数调用进行并行计算过程。对于每一次函数调用，Storm 集群上运行的拓扑接收调用函数的参数信息作为输入流，并将计算结果作为输出流发射出去。DRPC 通过 DRPC Server 来实现，DRPC Server 的整体工作过程如下：

(1)接收到一个 RPC 调用请求；

(2)发送请求到 Storm 上的拓扑；

(3)从 Storm 上接收计算结果；

(4)将计算结果返回给客户端。

DRPC 的工作流大致如图所示：

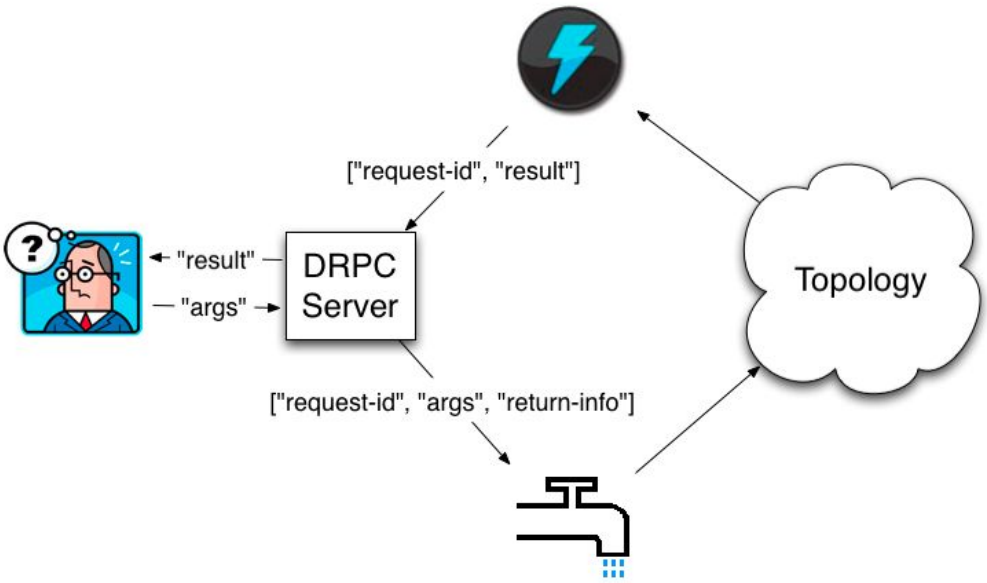
（1）Client 向 DRPC Server 发送被调用执行的 DRPC 函数名称及参数。

（2）Storm 上的 topology 通过 DRPCSpout 实现这一函数，从 DPRC Server 接收到函数调用流；

（3）DRPC Server 会为每次函数调用生成唯一的 id；

（4)Storm 上运行的 topology 开始计算结果，最后通过一个 ReturnResults 的 Bolt 连接到 DRPC Server，发送指定 id 的计算结果；

（5）DRPC Server 通过使用之前为每个函数调用生成的 id，将结果关联到对应的发起调用的 client，将计算结果返回给 client。



2.5 文本建模

为了全面地描述歌曲特征，本文将使用文本分析的方法对表征歌曲的文本信息进行分析。因此，本节简单介绍一些基本的文本建模方法。

2.5.1 向量空间模型

计算机不具备人脑的结构，无法理解自然语言，所以需要首先将无结构的自然语言文本转化为计算机可计算的特征文本。为此，Salton 等人在 20 世纪 70 年代提出了向量空间模型(Vector Space Model, VSM)，并成功地应用于著名的 SMART 文本检索系统。向量空间模型首先将每一个文档看做一个词袋(Bag of Words)，进而将文档表示成一个向量，向量的每一维表示一个词项，而向量每一维的取值表示该词项在文档中的权重。对于文档集合 D 中编号为 j 的文档 d，可以将之表示成一个 t 维的向量 $d_j = (w_{1,j}, w_{2,j}, ..., w_{t,j})$，其中 t 表示词项的数目，

$w_{i,j}(1 \le i \le t)$ 表示第 i 个词项在文档 d 中的权重。在对文本进行建模的过程中，词的选取及权重的计算有以下几种典型方式：

（1）布尔模型。这是最为简单直观的一种计算词项权重的方法，即将词项在文档中是否出现作为其权重，如果词项在文档中出现那么将其权重记为 1，否则记为 0。虽然这种方法比较简单，但是它没有体现词语在文档中出现的频率。一般来讲，词语在文档中出现的越多，说明它对该篇文档的重要性越大（“的”、“得”、“地”、“是”等停用词除外）。

（2）词频模型。与布尔模型不同的是，词频(Term Frequency, TF)模型统计词项在文当中出现的次数，然后得到词项的频率，并将之作为词项的权重。词项 t_i 相对于文档 d_j 的词频表示成 $tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$，这里 $n_{i,j}$ 表示该词项在该文档中出现的次数。这突出了词频对词项重要性的影响，能够取得比布尔模型较好的效果。但是，词语的重要性不仅随着它在文档中出现的次数成正比增加，而且可能会随着它在语料库中出现的频率成反比下降。也就是说，一个词语在整个语料库中出现得越频繁，则它对于文档的重要性越低，对文档的区分度量越差。

（3）词频-逆向文本频率模型。词频-逆向文本频率(Term Frequency - Inverse Document Frequency，TF-IDF)是对上述 TF 模型的补充，词项的重要性随着其在特定文档中出现次数的增加而增强，但同时随着其在全体文本中出现次数的增加而减弱，即该模型认为对区别文档最有意义的词语应该是那些在文档中出现频率高、而在整个语料库中的其他文档中出现频率少的词语。词项 t_i 相对于文档 d_j 的 TF-IDF 取值表示为 $tfidf_{i,j} = tf_{i,j} \times idf_i$。这里，$idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$ 表示逆向文本频率，其中|D|表示文档集合中的文件总数，$|\{j : t_i \in d_j\}|$ 表示文档集合中包含词项 t_i 的文件数目。TF-IDF 结构简单，容易理解，被广泛应用。但是，其无法准确捕捉文档内部与文档间的统计特征，也不能解决同义词和多义词的问题，因此精确度不是很高。

2.5.2 隐含狄利克雷分配模型

为了解决同义词和多义词的问题，从而进一步地挖掘文本的语义信息，Blei

等人于 2003 年提出了隐含狄利克雷分配模型(Latent Dirichlet Allocation，LDA)。LDA 是一种典型的词袋模型，它认为一篇文档是由一组词构成的一个集合，词与词之间没有顺序以及先后的关系。一篇文档可以包含多个主题，文档中每一个词都由其中的一个主题生成。进一步地可以认为 LDA 是一种主题模型，它将文档看做是一组隐含主题的概率分布。这里，主题表示一个概念、一个方面，表现为一系列相关的单词，是这些单词的条件概率。形象来说，主题就是一个桶，里面装了出现概率较高的单词，这些单词与这个主题有很强的相关性。这样，LDA 模型便通过隐含主题将文本与词项联系起来，从而达到降维的目的。LDA 是一种生成模型，一篇文档按照如下所示的规则生成：

1. 假设有两种类型的桶，一种是文档-主题桶，桶里的每一个球代表一个主题；另一种桶是主题-词汇桶，桶中的每一个球代表一个词汇。
2. 文档的生成过程就是不断从桶中取球的过程，每一次先从文档-主题桶中取出球，得到该球代表的主题编号 z。
3. 从编号为 z 的主题-词汇桶中取球，得到一个词汇。
4. 不断重复 2,3 两步，即可生成一篇文档。

在 LDA 模型中，记文档-主题的概率分布为多项式分布$\vec{\theta}$，主题-词汇的概率分布为多项式分布$\vec{\varphi}$。此外，LDA 模型认为$\vec{\theta}$和$\vec{\varphi}$是模型中的参数，且二者都是随机变量，考虑到$\vec{\theta}$和$\vec{\varphi}$都是多项式分布，模型选择狄利克雷(Dirichlet)分布作为其先验分布。在确定了这些分布之后，LDA 下面需要做的就是估计这些分布的参数，如下所示的吉布斯取样（Gibbs Sampling）是目前比较流行的采用方法：

1. 首先对所有文档中的所有词遍历一遍，为其都随机分配一个主题，即$z_{m,n}=k\sim Mult(1/K)$，其中 m 表示第 m 篇文档，n 表示文档中的第 n 个词，k 表示主题，K 表示主题的总数，之后将$n^{(k)}_m$、n_m、$n^{(t)}_k$、n_k都加1, 它们分别表示在第 m 篇文档中主题 k 出现的次数、第 m 篇文档中主题数量的和、主题 k 对应的词 t 的次数，k 主题对应的总词数。

2. 对第1篇文档中的所有词进行遍历，假如当前文档中的词 t 对应主题为 k，则将$n^{(k)}_m$、n_m、$n^{(t)}_k$、n_k都减1, 即先拿出当前词，之后根据 LDA 中 Topic Sample

的概率分布取样出新的主题，再将$n^{(k)}_m$、n_m、$n^{(t)}_k$、n_k都加1。其中，α和β为对应的Dirichlet分布的参数，V为词汇总数。

$$p(z_i = k \mid z_{-i}, w) \propto \frac{(n^{(t)}_{k,-i} + \beta_t)(n^{(k)}_{m,-i} + \alpha_k))}{\sum_{t=1}^{V}(n^{(t)}_{k,-i} + \beta_t)}$$

3. 重复步骤2直至遍历所有文档。

4. 输出LDA模型中的参数$\vec{\theta}$和$\vec{\varphi}$。

$$\varphi_{k,t} = \frac{n^{(t)}_k + \beta_t}{\sum_{t=1}^{V} n^{(t)}_k + \beta_t}$$

$$\theta_{m,k} = \frac{n^{(k)}_m + \alpha_k}{\sum_{k=1}^{K} n^{(k)}_m + \alpha_k}$$

在本文后续实验中，我们将分别以TF-IDF为代表的向量空间模型和以LDA为代表的主题模型对歌曲对应的文档进行建模，发现LDA能够获得较高的推荐准确率，因此我们将采用LDA作为我们主要的文本建模方法。

2.6 时间序列预测

歌曲时间短、消费代价低的特点决定了其能够较容易的形成序列，且是这种序列有严格的时间顺序，本文工作将通过对用户在会话期内所收听歌曲形成的时间序列的分析来预测用户的接下来的行为。因此，本节将简单介绍一些常用的时间序列预测模型。

2.6.1 简单平均法

简单平均法是以观察期内时间序列的各期数据（观察变量）的平均数作为下期预测值，按照采用的平均方法又可以分为算术平均法、加权平均法和几何平均法三类。其中，算术平均法以观察变量的算术平均数作为下期预测值，加权平均法以观察变量的加权算术平均数作为下期的预测值，而几何平均法是以观察变量的几何平均数作为下期的预测值。简单平均法比较简单、直观，但其预测误差一般偏高。

2.6.2 移动平均法

移动平均法是在简单平均法的基础上发展起来的，其将简单平均法改进为分段平均，并且按照时间序列数据点的顺序，逐点推移。根据时间序列逐项移动，依次计算包含一定项数的平均数，形成平均数时间序列，并据此对预测对象进行预测。移动平均可以消除或减少时间序列数据受偶然性因素干扰而产生的随机变动影响。按照移动的次数又可以分为一次移动平均法和二次移动平均法两类。其中，一次移动平均法是依次取时间序列的 n 个观察值进行平均，并依次移动，得出一个平均序列，并且以最近 n 个观察值的平均数作为预测值的预测方法。二次移动平均就是对时间序列的一次移动平均值再次进行第二次移动平均,就是利用一次移动平均值和二次移动平均值的滞后偏差的演变规律，建立线性方程进行预测的方法。二次移动平均法与一次移动平均法相比，其优点是大大减少了滞后偏差，使预测准确性提高。

2.6.3 指数平滑法

指数平滑法是由移动平均法改进而来的，是一种特殊的加权移动平均法。这种方法既有移动平均法的长处，又可以减少历史数据的数量。首先，它把过去的数据全部加以利用。其次，它利用平滑系数加以区分 ，使得近期数据比远期数据对预测值影响更大。它特别适合用于观察值有有长期趋势和季节变动，必须经常预测的情况。按照平滑的次数可以可分为一次指数平滑法和多次指数平滑法。其中，一次平滑法是计算时间序列的一次指数平滑值，以当前观察期的一次指数平滑值为基础，确定下期预测值。与二次移动平均法类似，二次指数平滑法就是对时间序列的一次指数平滑值再次进行指数平滑。

2.6.4 差分整合移动平均自回归模型

差分整合移动平均自回归模型(Autoregressive Integrated Moving Average model,ARIMA)又称为 Box-Jenkins 方法，是一种由 Box 和 Jenkins 于 1970 年提出的一种时间序列预测方法，目前已经得到广泛的应用。在模型 ARIMA(p,d,q)中，“AR”表示自回归模型，p 为自回归阶数，“MA”表示滑动平均模型，q 表示滑动平均的阶数，d 表示将时间序列转化为平稳时间序列所作的差分次数，即差分阶数。显然，ARIMA 首先需要进行 d 次差分从而将非平稳序列转化为平稳序列,然后利用自回归模型和滑动平均模型对转换后的平稳序列进行预

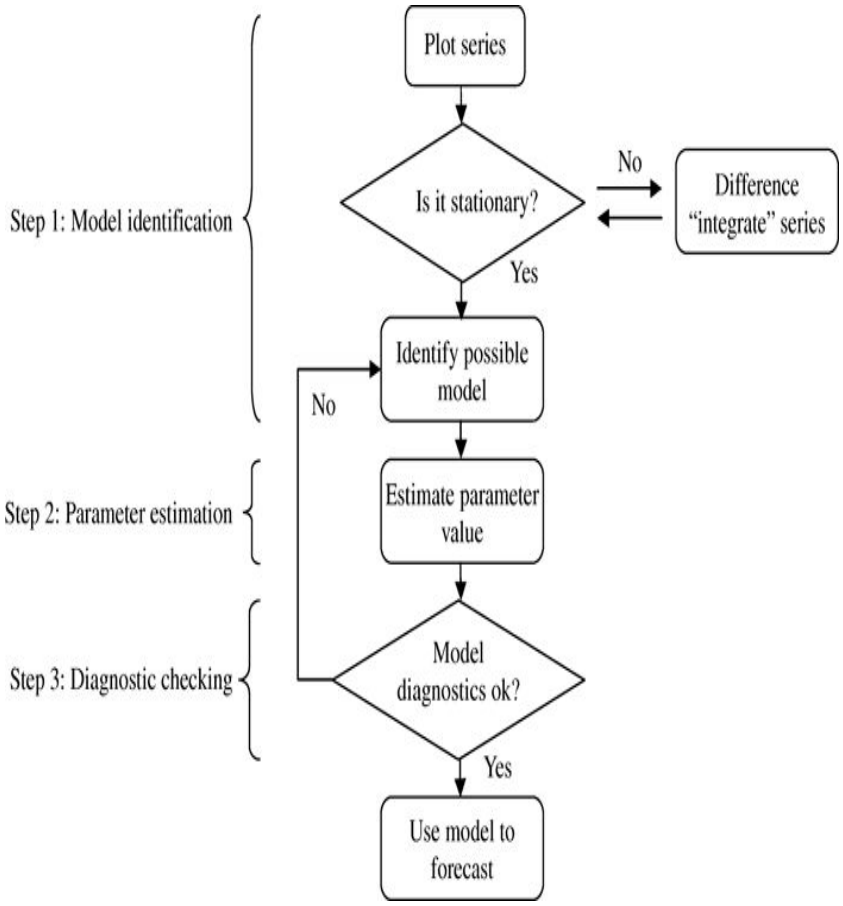
测。进一步的，模型 ARIMA(p,d,q)可以表示成如下三个式子。

$$\Phi(B)(1-B)^d y_t = \delta + \Theta(B)\varepsilon_t$$

$$\Phi(B) = 1 - \sum_{i=1}^{p} \varphi_i B^i$$

$$\Theta(B) = 1 + \sum_{i=1}^{q} \theta_i B^i$$

其中，y_t 为时间序列 Y 的第 t 个取值,B 是滞后算子，φ 和 θ 为模型参数,它们可以通过最小二乘等方法获得。



ARIMA 的执行流程如图所示，主要分为模型识别(Model Identification)、参数估计(Parameter Estimation)和诊断检测(Diagnostic Checking)三个阶段。其中,模型识别阶段主要完成检测序列是否平稳的工作。如果序列不平稳，那么则通过差分的方法将序列转化为平稳序列并给出差分阶数 d。在此基础上，识别序列适用的可能模型，如自回归模型或滑动平均模型或者二者的混合。而参数估计阶段主要完成模型参数的估计工作，即通过最小二乘法估计参数 φ 和 θ。诊断

检测阶段用以检测所给出的模型及参数是否符合条件，如果符合则选用此模型进行预测，否则重新识别模型。

ARIMA 模型将应用在本文后续的工作中，已完成对用户行为序列的分析和预测。后续实验结果显示，使用 ARIMA 模型进行时间序列预测，能够得到较好的效果。

2.7 音乐推荐算法

文献[]给出了一种简单朴素的音乐推荐算法，其引导用户通过歌曲名、曲作者姓名以及歌词等文本信息去检索歌曲。文献[]首先抽取歌曲的声音的音色、节奏等声学特征，然后通过比较用户收听歌曲的声学特征与曲库中歌曲声学特征为用户做出推荐。文献[]是典型的基于内容的音乐推荐，能够快速地为用户推荐符合偏好的歌曲，但它们具有较强的领域相关特点，可扩展性不强而且耗时耗力。文献[]给出了一种基于群体行为的推荐方法，即首先分析用户的收听行为然后找到与当前用户相似的用户并进而根据相似用户的行为为用户做出推荐，这是典型的协同过滤推荐。这种方法充分利用了用户的从众心理，能够得到较好的推荐效果。同时，其具有领域无关性，解决了基于内容的推荐所面临的问题。然而，这种推荐方法容易遇到冷启动和长尾效应等问题，一方面无法很好地为新用户做出推荐，另一方面往往推荐的是一些比较流行的歌曲而导致长尾歌曲无法被推荐。考虑到歌曲具有强烈的情感色彩，文献给出了一种基于情绪的音乐推荐算法，首先引导用户选择自己当前的情绪状态，然后从曲库中选择与用户选择情绪相近的歌曲。这种推荐方法能够很好地为用户做出符合当前情绪状态的推荐，但是情绪的描述和定义往往是很困难的，因此这种推荐的结果有时不太准确。此外，其需要用户的参与，在一定程度上增加了用户的成本。

上述给出的几种音乐推荐算法从基本属性、声学特征、情绪等不同侧面对歌曲进行了描述和刻画，进而根据用户的喜好检索歌曲并作出最终推荐。然而，这些算法往往只考虑了歌曲一方面的特征而忽视了其他方面的特征，比如[]考虑了声学特征却忽略了情绪特征，而[]考虑了情绪特征却忽略了声学特征。此外，歌曲还具有风格、年代、语言、场合等不同的侧面，这些都没有得到很好的体现。另外，一首歌曲往往并不是确定地属于某一种类别，而是以不同的概率隶属于不同的类别。比如，张雨生的歌曲“大海”即属于“经典”也属于“怀旧”还属于

“流行”，而上述的几种算法都没有对歌曲的这种特质进行体现。为了较为全面地描述歌曲特征并体现歌曲隶属类别的不确定性，文献[]充分利用了社会化标签系统的成果，首先利用用户对歌曲所打的标签构造歌曲对应的文档，继而通过主题模型建模的方法将歌曲表示成一组隐含主题的概率分布。

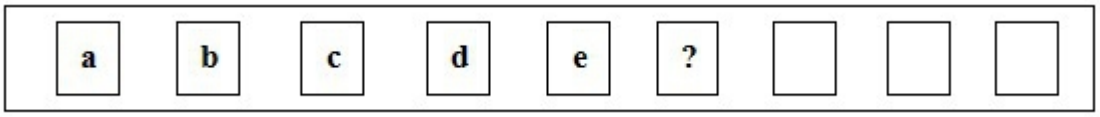
歌曲具有很强的情感色彩，用户收听歌曲的行为与其所处的天气、位置、场合等上下文环境有十分密切的关系，但直接获取这些上下文环境往往比较困难或者代价比较大。歌曲具有时间短、消费代价低的特点，因此用户往往会一次性地收听多收歌曲，而这些歌曲就天然形成了一个序列。文献[][]认为歌曲序列能够在一定程度上代表用户所处的上下文，进而通过对序列的分析来预测用户接下来的行为。它们首先利用[]中的思想将歌曲表征成若干隐含主题的分布，然后选择隶属概率较大的几个主题作为显著主题来表征歌曲，这样便可以把歌曲序列表示成显著主题序列。得到显著主题序列数据库后，[][]分布通过模式匹配和马尔科夫链进行分析并预测用户可能收听的下一首歌曲所述的显著主题并将该主题中的显著歌曲推荐给用户。实验表明，文献[][]能够获得比较好的推荐效果。然而，文献[][]同样需要参考其他用户的行为，荣誉遇到冷启动问题和长尾效应。此外，它们只考虑了显著主题的贡献而忽略了其他主题的影响。最后，它们是通过定性的方法预测用户可能收听的下一首歌曲的主题，没有充分考虑用户行为的时序性。本文将给出一种基于多维时序分析的音乐推荐方法，该方法也利用文献[]中的思想将歌曲表示成若干隐含主题的分布，然后利用时间序列预测的方法逐一预测每一个主题隶属度序列的取值。这样我们便可以得到用户可能收听的下一首歌曲的概率分布，通过计算相似度我们便可以得到候选的推荐列表。

第三章 一种基于多维时间序列分析的音乐推荐方法

本章将提出一种基于多维时间序列分析的音乐推荐方法。首先，我们将介绍本文所研究问题的相关描述和定义。其次，我们给出一种基于多维时间序列分析的音乐推荐方法，包括隐含主题的抽取、多维时间序列的构造及分析预测、最终推荐结果的生成等内容。最后，我们介绍基于所提方法所进行的实验，包括实验的设计、结果和分析。

4.1 问题描述和定义

音乐推荐的目的在于通过对音乐本身属性和用户行为习惯的分析，帮助用户过滤掉不必要的信息，并最终为用户推荐符合其喜好的音乐作品。换句话来说，就是解决如何在已知歌曲特征以及用户之前所收听歌曲的情况下准确地预测用户可能收听的下一首歌曲的问题(注:本文所考察的是用户在一定会话周期内的行为)，如图 1 所示[14]。



为了更好地描述该问题，我们首先定义用户集和歌曲集，如下所示：

用户集 C: 所有用户的集合,如式(1)所示,其中 m 为用户的数目,即 m=|C|.

$$C = \{c_1, c_2, \ldots, c_m\}$$

歌曲集 S: 所有可以推荐给用户的物品(这里就是歌曲，本文不加区分地使用“歌曲”和“物品”)的集合，如式(2)所示，其中 v 为所有可推荐歌曲的数目，即 v=|S|。为方便起见，我们认为用户收听的歌曲一定在“可推荐歌曲”中。

$$S = \{s_1, s_2, \ldots, s_v\}$$

对于给定的一个用户 c，音乐推荐系统的目标就是为这个用户推荐其可能喜欢的下一首歌曲。为了衡量用户对歌曲的喜欢程度，我们定义效用函数如下：

效用函数 u(c,s): 表征歌曲 s 对用户 c 的推荐度(如歌曲 s 符合用户 c 喜好的程度、歌曲 s 与预测歌曲的相似度等)。

效用函数反映了用户对某首歌曲的喜爱程度，其值越大表明喜欢程度越大。如前所述,本文的方法建立在隐含主题分类和用户在一定会话周期内听歌序列的基础上,我们下面进一步地定义歌曲对应的隐含主题集合以及用户所收听歌曲对应的序列。

主题集 T：由所有隐含主题组成的集合，如式(3)所示，其中 K 为隐含主题的数目。

$$T = (t_1, t_2, ..., t_K)$$

事件 $e(c,\tau,s)$：表示用户 c 在时刻 τ 收听了歌曲 $s \in S$，显然 s 可由 c 和 τ 唯一决定，因此 $e(c,\tau,s)$ 可简化为 $e(c,\tau)$。

序列 Q(c)：代表用户 c 在一定会话周期内的所有听歌事件按时间先后顺序排列而成的列表，如式(4)所示，其中 n 为用户 c 所收听或喜欢的歌曲数目，τ 为事件发生的时间，且 $\tau_1 < \tau_2 < ... < \tau_i < \tau_{i+1} < ... < \tau_n (1 \le i \le n)$。

$$Q(c) = < e(c,\tau_1), e(c,\tau_2), ..., e(c,\tau_n) >$$

会话 Session:本文所考察的歌曲序列均是用户在一个会话周期之内的行为，即用户连续不中断的收听行为。我们定义会话为满足式(5)式的序列 Q(c)。其中，ε 为最长时间间隔，本文将其设为 480 秒。方便起见，本文依然用 Q(c)来代表用户 c 的一个特定的会话。

$$|\tau_i - \tau_{i-1}| \le \varepsilon \quad (1 < \tau \le n)$$

目标歌曲：符合用户喜好或者用户接下来可能收听的歌曲。

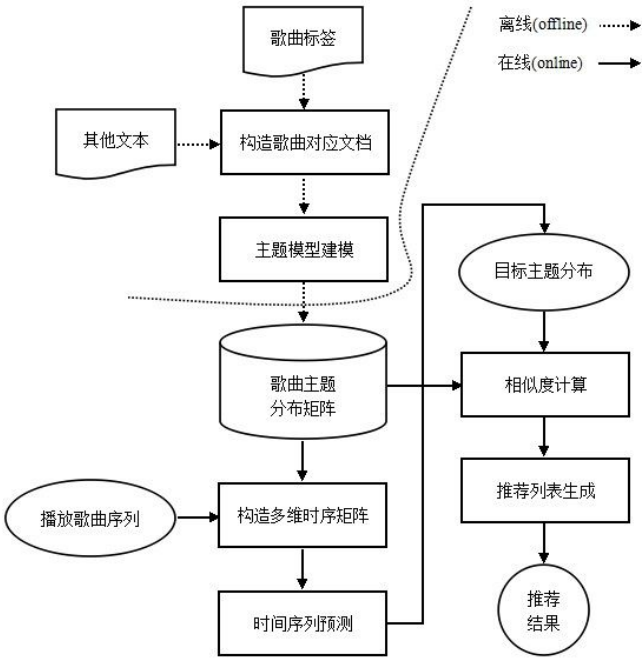
由之前的分析可知，本文所提方法的本质输入是歌曲对应的隐含主题 T 和用户某个会话下的收听序列 Q(c)，最终目标是为用户推荐其最可能收听的下一首歌曲。也就是说，在已知主题集 T 和用户 c 收听序列 Q(c)的情况下从歌曲集 S 中找出那些对用户的效用函数取值最大的目标歌曲 N(c)并推荐给用户，如式(6)所示。

$$N(c) = \underset{s \in S}{\operatorname{argmax}}\, u(c,s)$$

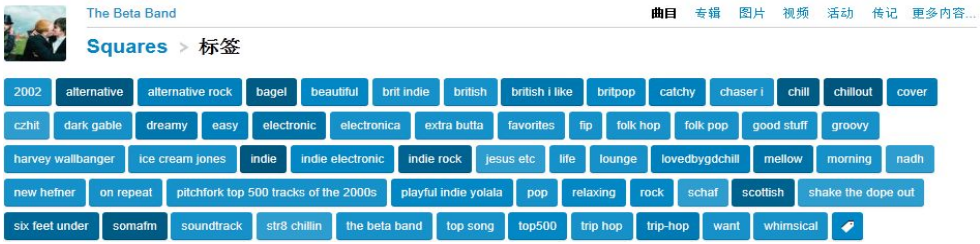
4.2 算法框架

针对上文所定义的音乐推荐问题，本文提出一种基于多维时间序列分析的音乐推荐方法，该方法按如图 2 所示的流程工作。其中，虚线箭头表示离线处理模块，实线箭头表示在线处理模块。首先，为了向主题模型建模方法提供输入，本文方法将定期从音乐网站上抓取以用户标签为主的歌曲信息，然后将这些标签信息按照一定的方式组合成文档，那么这些文档将和歌曲集 S 中的歌曲一一对应。随后，我们对这些文档组成的文档集进行主题模型建模，从而抽取出歌曲中包含

的隐含主题。这样，每一首歌曲可以用一个多维的表示隐含主题权重的向量表示，向量中的每一维取值表征该主题对歌曲内容的贡献程度。另外，由于用户听歌事件之间有严格的时间先后关系，我们将用户所听歌曲构成的序列建模为一个以歌曲对应的每一维隐含主题权重为变量的多维时间序列。进一步地，通过对该多维时间序列分析和预测，我们得到用户下面可能收听歌曲所对应的隐含主题权重的向量。最后，我们将曲库中与预测歌曲最相似的歌曲推荐给用户。在下面的章节中，我们将对框架中的主要步骤进行一一介绍。



4.3 主题模型建模



如前所述，为了更好地刻画歌曲的特征，本文在将歌曲映射成一个个文档的基础上通过主题模型建模来对歌曲进行描述和刻画。隐含狄利克雷分配(Latent Dirichlet Allocation,LDA)模型[15]是当前最具代表性也是最流行的概率主题模型，已经在文本挖掘、信息处理、多文档摘要等领域得到了广泛的应用[16,17]，其能够将文档映射到若干隐含主题的分布，而这个分布可以用一个主题权重向量表示，向量的每一维表征了该维主题对文档内容的贡献程度。通过 LDA 主题模型

建模，我们不但可以抽象出文档包含的隐含主题集合 T，而且能够以量化的方式表达不同文档之间的距离和相似度。本节详细介绍利用 LDA 主题模型对歌曲建模的过程。

首先，我们从 Lastfm(http://last.fm)、豆瓣(www.douban.com)等音乐网站上抓取用户对歌曲所标注的标签,这些标签对歌曲内容的描述比较全面，既包含了歌曲的名称、曲作者信息、专辑信息、发行年代等基本信息，还包含了歌曲所表达主题、歌曲类型、用户心情、适合场合等扩展信息。以 The Beta Band 的 Squares 为例，用户为其标注的标签有表征年代的“2002”/“2000s”、表征类型的“indie rock”/“folk pop”、表征用户感受的“beautiful”/“want”等，如图 3 所示。在得到歌曲对应的标签信息后，我们将这些标签按照一定的方式进行组织，从而生成歌曲对应的文本文档。为了减少噪音，我们只利用那些被多数人使用的标签来完成歌曲对应文档的构造，具体来说我们只考虑被标记次数大于 10 的标签。这样，每一首歌曲 s 都将对应一篇文档 d，而歌曲集 S 将对应到一个文档集 D。最后，我们对文档集 D 进行 LDA 主题模型建模，从而得到包含 K 个用来全面刻画歌曲特征的隐含主题集合 T。同时，对于歌曲集 S 中的任意歌曲 s，我们能够得到其对应的隐含主题权重的向量，如式(7)所示。

$$\mathbf{s} = (w_1, w_2, \ldots, w_i, \ldots, w_K)$$

其中，当 $w_i = 0$ 时表示歌曲完全不属于该隐含主题代表的类别，即该隐含主题对歌曲的内容完全没有贡献；当 $w_i = 1$ 时表示歌曲完全属于该隐含主题代表的类别；当 $0 < w_i < 1$ 时表示歌曲在一定程度上隶属于该隐含主题代表的类别。

表 1 展现了歌曲 Squares 的若干显著主题及其隶属于这些主题的概率。由表 1 可以看出，歌曲 Squares 隶属于主题 508 的概率为 0.215，这说明主题 508 对该歌曲贡献度为 0.215。类似的，主题 75 对歌曲的贡献度为 0.028，这说明主题 75 对该歌曲的贡献度为 0.028。因此，我们认为歌曲能够以更大的概率划分到主题 508 中。如前所述,我们不去也无法描述这些主题具体是什么含义，而我们正是利用这一点来回避确定性分类带来的弊端。

主题 id	508	313	296	106	231	100	307	354	8	75

隶属度	0.215	0.186	0.134	0.084	0.072	0.072	0.057	0.030	0.030	0.028

4.4 时间序列的构造和预测

传统的基于“最相似”假设的推荐方法只考虑用户当前的收听习惯，较少考虑用户行为所形成的序列性趋势，而本文所提出的基于多维时间序列分析的音乐推荐方法则对此进行了考虑。在通过 LDA 主题模型对歌曲进行建模而得到歌曲对应的主题权重向量后，我们需要做的就是获取用户对应的在一定会话周期内收听的歌曲构成的序列并构造其对应的时间序列，然后使用相应的时间序列分析方法进行预测。

设用户 c 当前会话周期内收听的歌曲序列长度为 n，为了预测其可能收听的下一首歌曲，我们需要获取该歌曲对应的主题权重向量，如式(8)所示。

$$\mathbf{N}(c) = (w_1(n+1), w_2(n+1), ..., w_i(n+1), ..., w_K(n+1)) \tag{8}$$

其中，$w_i(j)$ 表示第 j 首歌曲隶属于第 i 个隐含主题的概率($1<i<K$)。换句话说，我们需要准确地预测对应的主题权重向量中每一维的值，从而能够在歌曲集 S 中找到最匹配的歌曲推荐给用户。由前文可知，用户 c 的听歌事件按照事件发生的先后顺序构成了一个与时间相关的序列 Q(c)，且每一个听歌事件中对应的歌曲可以表示成如式(7)所示的 K 维隐含主题权重向量，那么我们将 Q(c)按照式(7)展开，可以得到如式(9)所示的矩阵 M(c)。

$$\mathbf{M}(c) = \begin{bmatrix} w_1(1) & w_1(2) & \cdots & w_1(n) \\ w_2(1) & w_2(2) & \cdots & w_2(n) \\ \vdots & \vdots & \ddots & \vdots \\ w_K(1) & w_K(2) & \cdots & w_K(n) \end{bmatrix} \tag{9}$$

其中，M(c)的行向量表示用户 c 对应的事件序列 Q(c)在各个隐含主题上的展开，而矩阵的列向量即为用户 c 收听事件序列中每一个事件对应的歌曲的隐含主题权重向量。考察如式(10)所示的 M(c)的第 i 个行向量对应的序列:

$$q_i = < w_i(1), w_i(2), ..., w_i(n) > \tag{10}$$

其中，$w_i(j)$ 的含义如前所述。显然，w_i 是一个随着时间而变化的实数变量，因此可以将 q_i 看作一个以 w_i 为时间变量的长度为 n 的时间序列。进一步地，我们可以利用时间序列分析和预测的方法对该时间序列进行分析和预测，从而可以估算第 i 个隐含主题对用户可能收听的下一首歌曲的贡献程度，即 N(c)对应的主

题权重向量中第 i 维元素值的估计值。

我们可以将 M(c)看做是一个多维的时间序列，其每一个行向量对应一个维度，共计 K 个维度。然后，我们对这 K 个维度的时间序列逐一利用 ARIMA 模型进行分析并预测该维度的估计值。最终我们可以获得如式(14)所示各个维度的估计值。显然，该向量可以看作是用户下面可能收听的歌曲在 K 个主题上的近似分布。为简便起见，我们将该分布对应的歌曲记为$\hat{s}$。

$$\hat{\mathbf{N}}(c) = (\hat{w}_1(n+1), \hat{w}_2(n+1), ..., \hat{w}_i(n+1), ..., \hat{w}_K(n+1)) \tag{14}$$

4.5 歌曲相似度计算及其推荐

通过 LDA 主题建模我们得到了如式(7)所示歌曲集 S 中歌曲对应的主题权重向量，考虑到向量中每一维度的值表征歌曲隶属于某一隐含主题的概率，我们将这些主题权重向量看作是离散的概率分布。进一步地，通过对多维时间序列的分析和预测，我们得到了如式(14)所示用户 c 可能收听的第(n+1)首歌曲对应的主题权重向量的估计向量，即一个估计概率分布。这样，我们便可以计算歌曲集中的歌曲对应的概率分布与这个估计概率分布的距离。显然，如果 s 与$\hat{s}$距离足够小，那么我们就可以将歌曲 s 推荐给用户 c。因此，我们可以用 s 与$\hat{s}$的距离的倒数表示歌曲集中任一歌曲 s 对用户 c 的推荐度，即上文定义的效用函数 u(c,s)，如式(15)所示。

$$u(c,s) = \frac{1}{dis(s,\hat{s})} \tag{15}$$

因为我们是通过主题模型建模将歌曲表示成离散的概率分布，所以我们可以使用 KL 距离(Kullback-Leibler Divergence)[19]以及 Hellinger 距离[20]等来度量两首歌曲之间的距离。考虑到 KL 距离不具有对称性，本文采用 Hellinger 距离来度量歌曲之间的距离。给定两个离散的概率分布 $P = (p_1, p_2, ..., p_k)$ 和 $Q = (q_1, q_2, ..., q_k)$，它们的 Hellinger 距离如式(16)所示。

$$H(P,Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{k} (\sqrt{p_i} - \sqrt{q_i})^2} \tag{16}$$

因此，我们可以定义歌曲之间的 Hellinger 距离，如式(17)所示。

$$dis(s_i,s_j)=\frac{1}{\sqrt{2}}\sqrt{\sum_{k=1}^{K}(\sqrt{w_{ik}}-\sqrt{w_{jk}})^2}$$

其中，$dis(s_i,s_j)$ 为歌曲 s_i 和 s_j 对应的主题概率分布之间的 Hellinger 距离，K 为隐含主题数目，w_{ik} 为歌曲 s_i 在第 k 个隐含主题上的概率。显然，当两首歌曲越相似，那么其对应的距离越小。相反的，若两首歌曲越不相似，其主题概率分布对应的距离越大。

最后，我们根据效用函数计算 S 中所有歌曲的效用值，然后根据效用值对歌曲排序，并将排名最高的 N 首推荐给用户，这样我们就为用户 c 推荐了一个长度为 N 的歌曲列表供用户选择。

4.6 实验设计和结果

本节将给出我们实验的设计思路、实验结果以及结果分析。

4.6.1 实验数据收集

通过对基于多维时间序列分析的音乐推荐算法的分析，可以看出我们需要的数据集主要包括包含标签文本信息的歌曲数据集以及用户在一定会话周期内所收听的歌曲序列的数据集。虽然 Berenzweig 等人于 2003 年从 Art of the Mix(http://www.artofthemix.org/)上抓取了播放列表数据集[23]，但是其存在如下三个问题：

(1) 缺少歌曲对应的标签等文本信息。

(2) 歌曲名称经过处理，无法与 Last.fm 对应，导致可用数据较少。

(3) 给出的播放列表意义不清，没有时间信息，无法确定是用户在一个会话周期内的行为，可能跨越多个会话周期。

为此，我们从 Last.fm 上重新爬取了一个数据集。该数据集既包含歌曲的基本信息(包括标签等文本信息)也包含用户的基本信息(包括用户在一定会话周期内所收听的歌曲列表)。为了消除噪音，我们只选用包含歌曲数目多于 10 首的列表，出现频次大于 10 的标签以及可用标签大于 4 的歌曲。该数据集的统计信息如下表所示，目前该数据集已经发布在 http://lastfmseq.sinaapp.com/上，而其使用说明如附录 1 所示。

听歌事件	34930(个)
不同歌曲	24992(首)
不同列表	1530(个)

不同歌手	5479(位)
最小长度	10
最大长度	30
平均长度	22.83

4.6.2 实验评测标准

对于用户 c，我们通过对其收听记录所形成的有序列表进行分析，为其生成一个包含 N 首歌曲的推荐歌曲列表，如果这 N 首歌曲中包含用户真实收听的下一首歌曲，那么我们认为这个对于用户 c 的推荐是有效的。显然，类似音乐推荐这种为用户推荐一组物品供选择的问题是典型的 Top-N 推荐问题(Top-N Recommendation, TNR)。由文献[1][21]可知，召回率(recall)和准确率(precision)是衡量一个 Top-N 推荐算法优劣的重要标准，我们这里也用这两种标准来评测本文提出的方法。记 R(c)是根据用户在训练集上的行为给用户推荐的歌曲列表，而 T(c)是用户在测试集上的行为列表，那么表征“检索出的相关文档数和文档库中所有的相关文档数的比率”的召回率 Recall 的定义如式(18)所示：

$$\mathrm{Recall}=\frac{\sum_{c\in C}|R(c)\cap T(c)|}{\sum_{c\in C}|T(c)|}$$

可以看出，召回率表征用户真实收听的歌曲被推荐的数目与用户真实收听歌曲总数的比率。因为在音乐推荐系统中，用户 c 同一时刻在测试集上只会收听一首歌曲，即|T(c)|=1。因此，我们将式(18)简化为如式(19)所示的命中率(hit ratio)：

$$h(N)=\frac{\sum_{c\in C}\mathrm{hit}}{|C|}$$

其中，N 为推荐系统为用户推荐的歌曲数目，hit 表示用户实际收听的歌曲是否在推荐列表中，若在则为 1，否则为 0。如果 hit 为 1，我们称之为“命中一次”。记用户 c 实际收听的歌曲为 s，则 hit 可表示为如式(20)：

$$\mathrm{hit}=\begin{cases}1 & ,s\in R(c)\\ 0 & ,s\notin R(c)\end{cases}$$

准确率表征了“检索出的相关文档数和系统所有检索到的文件总数的比率”，即用户真实收听的歌曲被推荐的数目与被推荐的歌曲总数的比率，其定义如式(21)所示：

$$\mathrm{precision}=\frac{\sum_{c\in C}|R(c)\cap T(c)|}{\sum_{c\in C}|R(c)|}$$

考虑到$|T(c)|=1$，式(21)可简化为式(22)：

$$precision = \frac{\sum_{c \in C} hit}{\sum_{c \in C} |R(c)|}$$

考虑到召回率和精确度此消彼长的关系，文献[21]中使用 F1-Score 对模型进行评估，F1-Score 的取值越大那么模型对应的综合效果越好，反之越差。F1-Score 可以用式(23)表示：

$$F = 2 * \frac{recall * precision}{recall + precision}$$

在如上所述的评测标准中，召回率和准确率的在很大程度上取决于如式(19)所示的命中率。如果歌曲未被用户喜欢或收听，那么就认为该歌曲未命中。然而，文献[22]指出没有明显的证据表明未被评分的物品对用户来说是完全否定的。也就是说，即使歌曲未命中，也不代表用户不喜欢该歌曲。假设用户 u 真实收听的下一首歌曲为s_{next}，如果系统为其推荐了列表$L1 = \{s_{11}, s_{51}, s_{31}, s_{next}, s_{21}, s_{41}\}$，那么我们认为该推荐是有效的，因为目标歌曲在推荐列表中。相反，如果系统为其推荐列表$L2 = \{s_1, s_5, s_3, s_6, s_2, s_4\}$，由于其中未包含$s_{next}$，我们认为该推荐是无效的。然而，如果$s_{next}$与 L2 中的歌曲相似度很高，用户显然也会喜欢该列表。那么，认为列表 L2 无效就不够合理。为了解决这种矛盾，我们可以考虑使用推荐偏差的大小来衡量算法的优劣。也就是说，推荐偏差越小，算法越好，反之越差。这里的偏差可以用歌曲对应的主题概率分布的 Hellinger 距离来表示。如果一个推荐列表中歌曲与目标歌曲的相似度较高，那么该列表中歌曲与目标歌曲的距离就应该较小，即推荐偏差较小，这时即使目标歌曲不在该列表中，我们也应该认为该列表合理。而如果一个推荐列表中的歌曲与目标歌曲的相似度整体偏低，导致推荐偏差较大，即使其中包含目标歌曲，我们也应该降低该列表被认可的权重。推荐系统主要包含评分预测和 Top-N 推荐两类问题，在评分预测中我们常常使用均方根误差(RMSE)和平均绝对误差(MAE)来衡量算法的优劣，这里我们将其借鉴到音乐推荐的问题中并用以衡量不同算法的优劣，其定义如式(24)(25)所示。

$$RMSE = \sqrt{\frac{\sum_{c \in C} e(c)^2}{|C|}}$$

$$MAE = \frac{\sum_{c \in C} |e(c)|}{|C|}$$

其中，C 为测试用户集，|C|为用户数，e(c)为向用户推荐的结果的误差。考虑到我们为用户推荐的是一个列表，我们将 e(c)看做是列表中所有歌曲与目标歌曲的平均距离。如式(26)所示。

$$e(c)=\frac{\sum_{s\in R(c)} dis(s,s_{next})}{|R(c)|}$$

4.6.3 实验设置

为了客观地衡量本文所述方法的效果，我们首先将所有数据随机地分为 10 份并将其中的 9 份作为训练集，剩余的 1 份为测试集，然后进行 10 轮交叉实验。最后，我们将 10 轮实验的结果进行平均，从而得到最终的实验结果。其中，本文将隐含主题数目设为 30。本文实验实在 Dell Optiplex74 的台式机上进行，操作系统为 Ubuntu12.04，CPU 为 Intel 酷睿 2 E6300，内存大小为 2G，硬盘空间 160G。实验所用编程语言为 Python2.7。目前，实验源码和实验结果已经上传至 https://github.com/wwssttt/GitRepo/tree/master/Python/experiment，具体的使用说明如附录 2 所示。

4.6.4 实验结果与分析

4 展示了当被推荐歌曲数目 N 由 1 到 100 的增长过程中，不同推荐算法的召回率的变化情况，这里包括基于用户的最近邻算法(UserKNN)[7]、基于模式挖掘的推荐算法(PatternMining)[4]、基于 1 阶马尔科夫链的推荐算法(1st-Markov)[4,12]、基于 3 阶马尔科夫链的推荐算法(3rd-Markov)[4,12]以及本文提出的基于多维时间序列分析的音乐推荐方法(MTSA)。其中，横坐标表示被推荐歌曲的数目，纵坐标表示算法的召回率。由图 4 可以看出，代表本文所述方法的曲线与其他曲线能够明显分开且位于其他曲线之上，表明本文所提方法能够获得比其他同类工作更好的召回率且提升效果比较明显。此外，随着被推荐歌曲数目的增加，本文所述方法的召回率也同时提升且呈逐渐上升趋势。

考虑到基于用户的最近邻算法(UserKNN)在同类工作中的召回率最高，本文考察基于多维时间序列分析的音乐推荐方法相较于 UserKNN 在综合指标 F1-Score 上的提升效果(倍数)，如图 5 所示。其中，横坐标表示被推荐歌曲的数目，纵坐标表示算法 F1-Score 的提升倍数。如果纵坐标取值大于零，表明本文所述方法的 F1-Score 相较于 UserKNN 算法有所提升；如果纵坐标取值为零，表明效果没有提升；如果纵坐标取值小于零，表明本文所述方法不但没有提高

F1-Score 而且还有所下降。由图 5 可以看出，无论被推荐歌曲数目为何值，纵坐标取值总是大于零，说明本文所述方法能够提升推荐的 F1-Score 且提升幅度在 80%以上。随着推荐列表长度的增长，提升的幅度也继续增长(可达 150%)。

图 6 展示了随着推荐列表长度的增加，几种不同的音乐推荐算法的均方根误差(如左图 a 所示)和平均绝对误差(如右图 b 所示)的变化趋势。其中，横坐标表示被推荐歌曲的数目，纵坐标表示不同推荐算法的误差。由图可以直观地看出，本文所提方法的推荐误差较之其他几种算法都比较小。随着推荐列表长度的增加，列表中无效的歌曲增多，使得推荐误差有所上升，但这种上升幅度也是是非常小的。

综合以上实验结果可以看出，无论是从命中率的角度去考察算法的优劣，还是从误差的角度去考察算法的优劣，本文所述的基于多维时间序列分析的音乐推荐算法都能够取得比较好的效果。这验证了本文所提方法的合理性，说明在音乐推荐系统中使用隐含特征的方法刻画歌曲并结合对用户在一个会话周期内行为趋势的预测能够提高推荐的效果。

第四章 一个基于用户行为三元特征的混合推荐框架

第三章给出了一种基于多维时间序列分析的个性化音乐推荐方法，同时通过实验验证了其在一定条件下的有效性和合理性。然而，上文所给的推荐方法仍然存在一些问题。本章首先将对上文所提的方法进行分析，并在此基础上给出用户听歌行为的三元特征，进而我们给出一个基于用户行为三元特征的混合推荐框架。最后，我们通过实验验证该混合推荐框架能够得到比单一的基于多维时间序列推荐较优的效果。

4.1 问题分析

如第三章所述，基于多维时间序列分析的个性化音乐推荐算法主要是通过对用户在会话期内行为的分析来预测用户可能收听的下一首歌曲。显然，这要求会话期有一定的长度，这样才能够获得比较准确的预测。那么，第三章所述方法对于会话刚开始阶段的预测效果如何尚不可知。这里，我们从常识假设其对于会话刚开始阶段的推荐效果不如会话已经形成规模时。因为，会话刚开始时，歌曲序列果断，序列包含的信息较少，不足以得到高效的推荐。本节将通过实验来验证我们的直观假设。

图中横坐标表示为用户推荐的列表长度，纵坐标表示推荐的命中率。由图可以看出，尽管随着推荐列表长度的增加推荐的效果也有所增加，但推荐的命中率明显低于会话形成时，甚至低于“最相似”这种最基本的推荐。推荐的准确度、F1 值、均方误差、绝对误差也能获得类似的结果，这里不一一列出。

由上面的结果可以看出，本节开始提出的假设是成立的，即第三章所述方法在会话刚形成时预测效果不佳。

除了上述问题之外，还存在预测误差偏大的问题。分析图可知，尽管该方法能够得到较低的预测误差，但误差的绝对值比较高。

综上所述，三种所述的方法有两个主要的问题需要解决：其一是对会话刚形成时的处理，其二是降低预测的绝对误差。

为了解决这两个显著问题，本章将给出一种基于用户行为三元特征的混合推荐框架。该框架再利用用户行为的序列性特征的基础上，进一步地挖掘用户行为的其他特征。本章下面将首先介绍用户行为的局部性特征和全局性特征，然后给出完整的混合推荐框架，最后通过实验验证框架的有效性。

4.2 用户行为的三元特征

为了更好地为用户做出推荐，我们对用户行为特征进行深度挖掘并提出用户行为的三元特征，即用户行为具有时序特征、局部特征和全局特征。其中，时序特征即用户收听行为具有时序性，通过对收听时序的分析可以在一定程度上预测用户接下来的行为，这在上文提出的基于多维时间序列分析的个性化音乐推荐算法中能够得到证明。本节将给出对用户收听行为局部特征和全局特征的描述并简单介绍其背后蕴含的心理学原理。

4.2.1 局部特征

众所周知，程序的执行具有局部性原理，即在一段时间内，程序的执行仅限于程序中的某一部分，这包括空间局部性和时间局部性。时间局部性是指如果程序中的某条指令一旦执行，则不久之后该指令可能再次被执行；如果某数据被访问，则不久之后该数据可能再次被访问。空间局部性是指一旦程序访问了某个存储单元，则不久之后其附近的存储单元也将被访问。考察空间局部性，这里我们将“用户”看做“程序”，将“收听”看做“访问”，而将曲库中的歌曲按照相似度大小依次排开即越相似的歌曲离得越近并将“歌曲”看做“存储单元”，那么我们便可以将用户收听某一首歌曲的行为看做是程序访问某个存储单元的行为。类似的，我们认为用户收听歌曲的行为也具有局部性原理，即如果用户在某一时刻收听了某首歌曲，那么其在不久之后将会继续收听与该歌曲类似的歌曲，我们将这种特点称之为用户行为的局部特征。进一步地，我们做出如下假设：

1.用户的状态是稳定的

2.用户在短期内所听歌曲的特征保持不变或仅有细微变化

$$\hat{w}(\mathrm{n}+1,\mathrm{i}) = w(\mathrm{n},\mathrm{i})(1 \le \mathrm{i} \le \mathrm{K})$$

我们这种假设的合理性除了可由程序执行的局部性原理类比获得之外，也可以由心理学的相关理论导出，我们这里介绍海德的平衡理论和注意的稳定性特征。

4.2.2 全局特征

上节介绍了本文基于心理学分析的基础上给出的用户行为具有局部特征的假设，本节将介绍用户行为的全局特征。心理学理论指出，人们的性格具有一定

的倾向性，而用户长期的行为能够在一定程度上反映这种倾向性。为此，本文给出如下的用户听歌行为的全局特征假设：

用户长期收听的歌曲能够反映用户对歌曲的整体偏好

简单起见，我们仅用歌曲特征的平均值来标准用户对歌曲的偏好，那么上述假设可以具体描述为：

用户对歌曲的整体偏好可由其所收听的所有歌曲的特征的平均值表示。

$$\hat{w}(n+1,i)=\frac{\sum_{j=1}^{n} w(j,i)}{n}(1\le i\le K)$$

4.3 基于用户行为三元特征的混合推荐框架

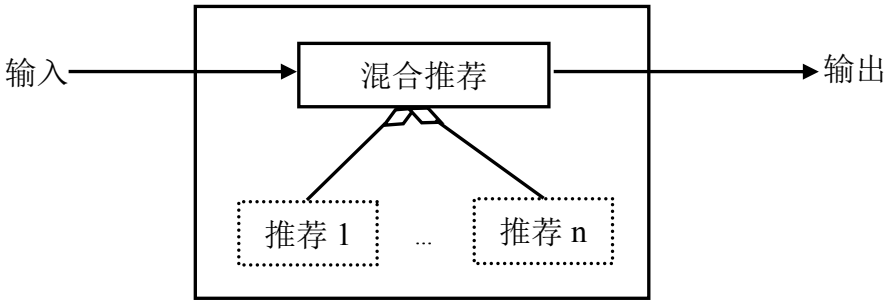
上文分别给出了用户听歌行为的局部性特征和全局性特征的假设，加上上一章描述的用户听歌行为的时序性特征，这些就构成了用户听歌行为的三元特征。为了描述方便，我们用 $\hat{w}_{sequential}(n+1,i)$ 表示根据时序特征预测的用户可能收听的下一首歌曲在编号为 i 的隐含主题上的权重，$\hat{w}_{local}(n+1,i)$ 表示根据时序特征预测的用户可能收听的下一首歌曲在编号为 i 的隐含主题上的权重，$\hat{w}_{global}(n+1,i)$ 表示根据时序特征预测的用户可能收听的下一首歌曲在编号为 i 的隐含主题上的权重，而 $\hat{w}(n+1,i)$ 表示用户可能收听的下一首歌曲在编号为 i 的隐含主题上的权重。

由前文可知，单纯的基于时序特征的推荐存在预测误差偏大以及冷启动处理不佳的问题，本节给出基于三元特征的混合推荐框架，该混合推荐框架综合考虑用户行为的局部特征、全局特征和时序特征。混合推荐的主要特征就是混合设计，发挥各个独立算法的优势同时规避各个独立算法的缺点。尽管 Burke 的分类方法区分出了七种不同的混合策略，但从更综合的角度来看这七种策略可以概括为三种基本设计思想：整体式混合设计、并行式混合设计和流水线式混合设计，下文将对这三种设计思想进行简单介绍并给出本文所提混合推荐框架所采用的设计思想。

4.3.1 整体式混合设计

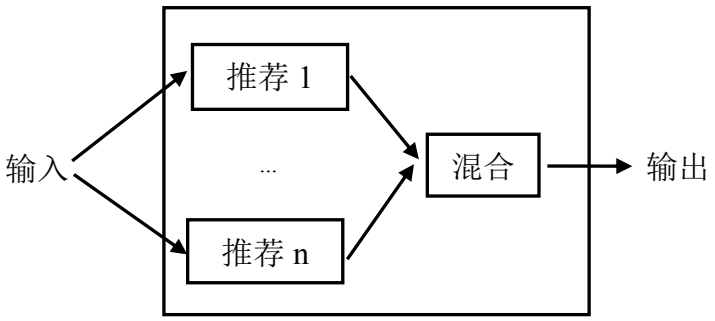
所谓整体式设计即将几种不同推荐策略整合到一个算法中实现的混合设

计，实际上确实所有推荐策略都在发挥作用，如下图所示。举例来说，可以通过基于局部特征的推荐得到用户可能收听的下一首歌曲在 K 各隐含主题上的概率分布，而同样可以通过基于全局特征和基于时序特征的推荐得到用户可能收听的下一首歌曲在 K 各隐含主题上的概率分布。整体式推荐的一个简单实现就是在得到三个概率分布的基础上，混合得到最终的概率分布并以此来产生推荐列表。



4.3.2 并行式混合设计

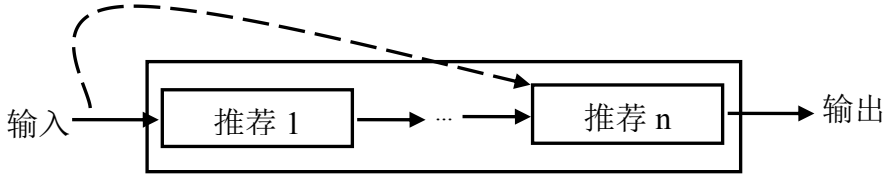
所谓并行式混合设计就是不同的推荐算法根据输入并行运行分布产生推荐列表，最后将这些推荐列表混合产生最终的推荐列表推荐给用户的设计思路，如图所示。举例来说，并行式混合设计需要分别利用局部特征、全局特征和时序特征分别为用户生成一个推荐列表，然后将这三个列表进行混合以生成最终的推荐结果。至于混合的策略，可以采用加权、交叉以及切换等进行混合。交叉式混合推荐在在用户交互界面层面上将不同的推荐结果组合到一起，各种方法所得到的结果一同被呈现。加权式混合推荐关键字是加权，通过计算多个推荐结果分数的加权之和将其组织到一起。切换式混合需要根据用户记录或推荐结果的质量来决定哪种情况下应用什么推荐系统。



4.3.3 流水线式混合设计

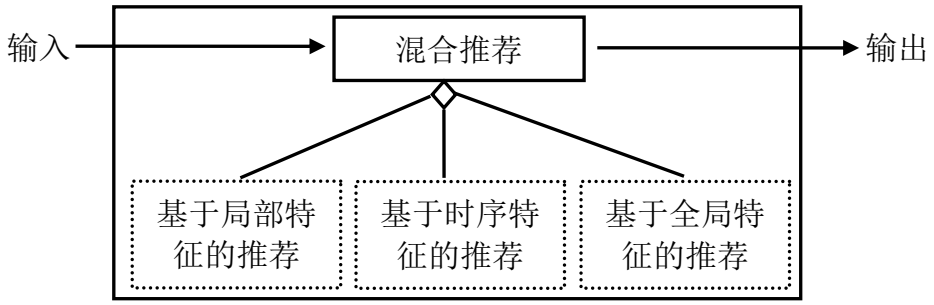
如下图所示，流水线式混合设计将多个推荐策略按照流水线架构连接起来，前一

个推荐系统的输出变成后一个推荐系统的输入部分。当然，后面的推荐单元也可以选择使用部分原始输入数据。



4.3.4 混合方案的选取和分析

由上面的介绍可知，整体式混合设计实际上是一种推荐算法，其在推荐内部对不同推荐策略获取的特征进行混合，然后基于此做出最终推荐列表；并行式混合设计和流水线式混合设计都需要至少实现两个不同的推荐算法，其中并行式混合设计将不同推荐算法获取的最终推荐结果加以混合，而流水线式混合则将推荐分为不同的阶段然后在不同的阶段使用不同的推荐策略进行推荐并最终得到推荐结果。显然，本文所述的音乐推荐不具有明确的阶段划分，因此使用流水线式的混合设计不太合适，而整体式混合设计和并行式混合设计均可应用，只是二者混合的阶段不一致。由于整体式的混合设计思路是从内部对不同推荐策略的结果进行混合，我们认为这样的早期混合能够得到比较符合用户偏好的特征，因此我们主要考察这种混合设计方法，如图所示。



首先，我们考察将时序特征和局部特征进行混合，混合方法如下所示：

$$\hat{w}_{\text{seq\_local}}(n+1,i)=\alpha\hat{w}_{\text{sequential}}(n+1,i)+(1-\alpha)\hat{w}_{\text{local}}(n+1,i)$$

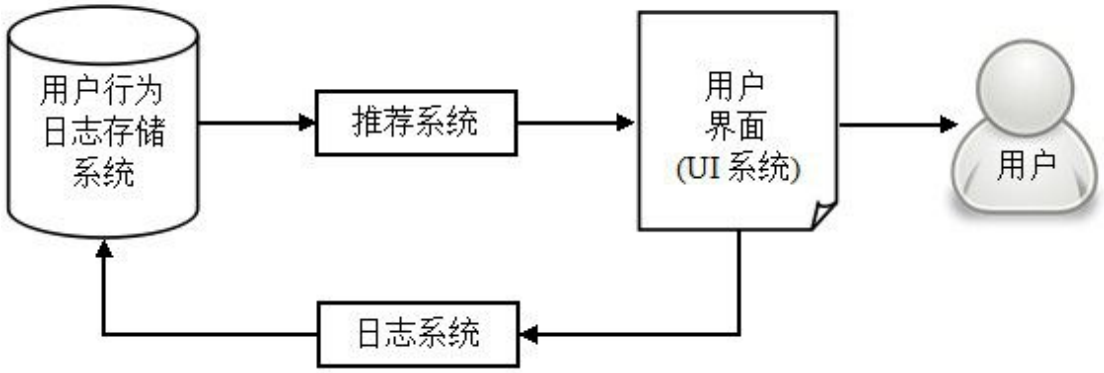
$$\hat{w}_{\text{seq\_global}}(n+1,i)=\alpha\hat{w}_{\text{sequential}}(n+1,i)+(1-\alpha)\hat{w}_{\text{global}}(n+1,i)$$

$$\hat{w}_{\text{all}}(n+1,i)=\alpha\hat{w}_{\text{sequential}}(n+1,i)+\beta(1-\alpha)\hat{w}_{\text{local}}(n+1,i)+(1-\beta)(1-\alpha)\hat{w}_{\text{global}}(n+1,i)$$

第五章　系统实现

前述章节介绍了本文所述的基于多为时间序列的个性化音乐推荐方法和基于用户三元特种的混合推荐框架，并从实验的角度验证了本文所提方法和框架的有效性。为了进一步验证本文所提方法和框架的可行性，本文实现了一个个性化音乐推荐原型系统，本章将详细介绍该系统的实现细节。

5.1 系统架构



尽管优秀的推荐算法能够为用户推荐合理的结果，但只靠推荐很难构成一个完整的可用系统。要构建一个可用的推荐系统，比如一个电影推荐网站，就需要考虑推荐与系统其他组件的关系，只有这样才能最终实现推荐的价值。项亮在文献中给出了如上图所示的一般意义上推荐系统和网站其他系统之间的关系。首先，基本上所有的网站都配有一个用户界面，即 UI 系统，该系统主要用来向用户展示页面效果以及与用户进行交互。其次，网站往往还会配置日志系统，该系统主要用来将用户在用户界面上的各种有效行为记录下来并保存到对应的日志存储系统中。需要注意的是，这里的日志存储系统既可以是数据库，也可以说是缓存，还可以是文件系统。推荐系统作为一个网站的核心，其主要作用是分析存储在日志存储系统中的用户行为历史，在此基础上生成该用户对应的推荐列表并将结果直接展示到用户界面上以供用户体验。由上图可知，推荐系统要想将强大的作用发挥出来，还要依赖于用户的历史行为信息和用户界面。

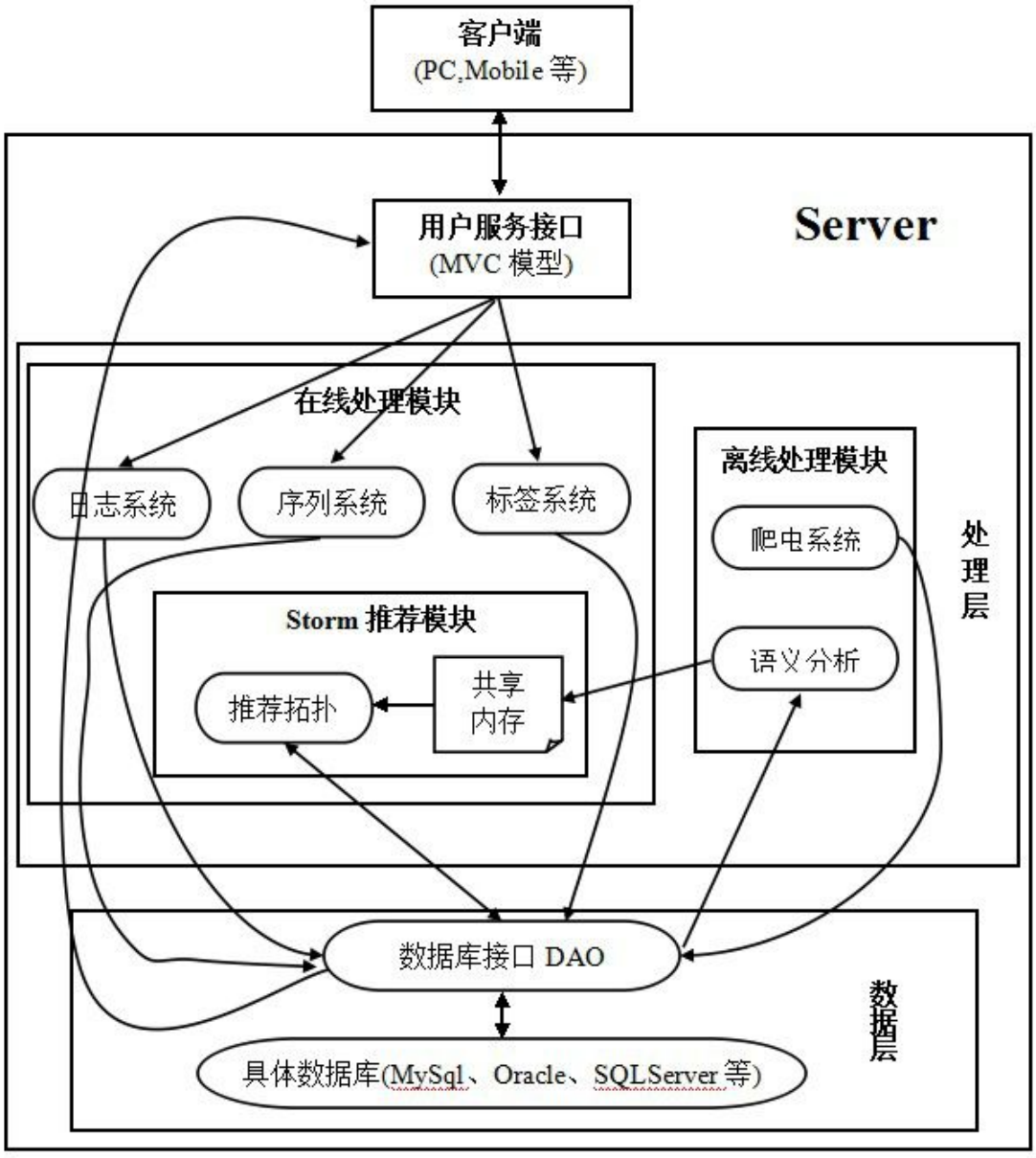
上图虽然给出了一般意义上的推荐系统的架构，给本文所实现的系统原型带来了很大启发，但这种架构显然过于简单，不能直接用于本文的系统原型。在认真分析本文所提推荐框架组成的基础上，本文按照下图对系统原型进行设计。

本文所实现的原型系统主要包括客户端、用户服务接口、处理层和数据层构

成。

客户端对应于上图中的 UI 系统，旨在为用户提供一个可以收听推荐歌曲并产生交互的窗口，具体到实现上既可以使用 Web，也可以是 Mobile 还可以是普通的窗口客户端。

用户服务接口用于将用户在客户端上的行为传递给后续模块进行处理以及从将推荐的结果返回给客户端。



处理层主要用于完成相关数据处理，其包括在线处理模块和离线处理模块。其中，离线处理模块主要包括从百度音乐、豆瓣音乐、虾米音乐、Lastfm 等数据源爬取歌曲基本属性数据和标签数据的爬虫系统对歌曲对应的文档集合进行语

义分析以获取每一首歌曲对应的隐含主题分布的语义分析模块。需要强调的是，这里离线的含义是指该模块的工作与用户行为无关，在具体实现上可以设定一定的时间间隔或周期执行一次。

在线处理模块主要用于实时记录和处理用户的行为并为之生成最终推荐结果，其主要包括日志系统、序列生成系统、标签系统以及 Storm 推荐引擎几个部分组成。日志系统对应于文献中的日志系统，用于将用户的行为记录到数据存储系统中以待后续分析。标签系统主要用于记录用户对当前所听歌曲所标注的标签，其与离线模块中的爬虫系统一起生成最终的歌曲标签信息，一定周期后可供语义分析模块处理。序列生成系统用于根据用户在当前会话期内的行为生成对应的歌曲序列。Storm 推荐引擎模块的主要作用是使用 Storm 对用户当前会话期的歌曲序列进行实时分析并产生最终的推荐结果。

数据层主要完成用户属性、歌曲属性以及用户行为的存储功能，其包括数据库接口 DAO 以及具体数据库两部分。其中，数据库接口提供供其他层次模块调用的方法以避免直接操作数据库，增强了独立性。具体数据库即真实的数据存储引擎，既可以是 Mysql，也可以是 Oracle，当然也可以是 SQLServer。

用户产生一个积极行为到系统为其推荐歌曲的过程如下所示：

1. 用户在客户端产生“即将收听完当前歌曲”的行为。
2. 用户服务接口接收用户的当前行为状态并将该状态传递给在线处理模块中的日志系统和序列生成系统。
3. 日志系统将用户当前行为状态通过数据库接口 DAO 记录到数据库中。
4. 序列生成系统通过数据库接口 DAO 读取日志数据库中的用户行为，构建其在当前会话期的收听序列并将该序列传递给 Storm 推荐引擎。
5. 推荐引擎对用户当前会话期的收听序列进行分析和处理，生成推荐列表并通过数据库接口保存到数据库。
6. 用户服务接口通过数据库接口从数据库中读取推荐列表。
7. 用户服务接口将推荐列表展示给用户。

用户对歌曲打标签的执行过程如下所示：

1. 用户在客户端选定一首歌曲。
2. 用户为选定歌曲打标签。

3. 用户服务接口接收用户的打标签行为以及打标签的对象和标签内容并将它们传递给在线处理模块中的标签系统。

4. 标签系统对标签进行分析和处理。

5. 标签系统将处理过的内容通过数据库接口 DAO 保存到数据库中。

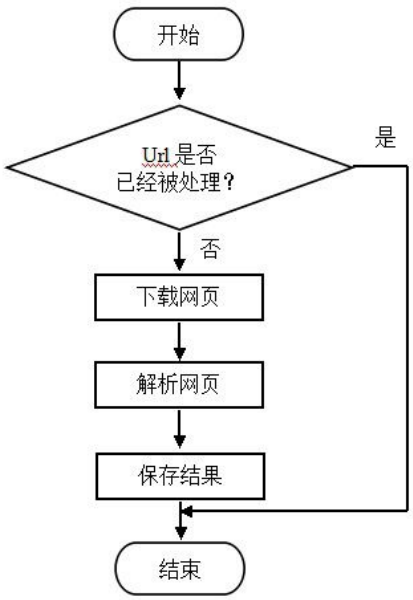
类似的，可以给出用户其他行为的处理过程。

5.2 离线处理模块

如上所述，离线处理模块所做的工作主要包括爬虫系统和语义分析系统两部分组成，这些工作都是独立于用户进行的，所以称之为离线。本节将详细介绍上述框架中的离线处理模块中设计的一些技术和工具，而作为算法核心的语义分析模块将是介绍的重点内容。

5.2.1 爬虫系统

网络爬虫是一种能够按照一定地规则自动地抓取互联网上的网页并对网页内容进行解析的网络机器人，又称之为网络蜘蛛。本文所实现的音乐网络爬虫首先从豆瓣、虾米、百度等音乐网站上抓取相关歌曲网页，然后对网页的 HTML 进行分析和解析，进而得到歌曲的名称、创作者、发行时间、歌词等基本属性以及用户对歌曲所打的标签内容，最后将这些内容进行整理并保存到数据库中。对于一个待抓取的歌曲页面 URL，本文按照如下流程图进行工作：



网络爬虫目前已经广泛应用在数据挖掘、搜索引擎、信息检索、推荐系统等领域，同时也出现了很多网络爬虫框架以简化爬虫的实现。其中，Scrapy 是一种纯

python 实现且构建于异步框架 twisted 之上爬虫框架，其用户只需要定制开发几个模块就可以轻松的实现一个爬虫，用来抓取网页内容以及各种图片，非常之方便。要想创建一个网络爬虫，人们只需要执行命令 scrapy startproject projectname 就可以得到如下图所示的项目目录，可见生成的目录包含若干文件，即 Scrapy 的模块。用户只需要在对应的文件中实现相应文件即可。其中，items.py 文件对应于 Scrapy 中的项模块，用于定义抓取结果中单个项所需要包含的所有内容，如歌曲的名称、创作者、发行时间等；piplines.py 对应于 Scrapy 中的管道模块，定义如何对抓取到的内容进行再处理，例如输出文件、写入数据库等；spider 目录下存放写好的爬虫实际抓取逻辑。

```
projectname
|---   projectname
|      |---   __init__.py
|      |---   items.py
|      |---   pipelines.py
|      |---   settings.py
|      |---   spider
|              |---   __init__.py
|---   scrapy.cfg
```

右上可知，在抓取指定网页内容之后需要对网页进行解析，这里我们使用 Python 语言的第三方包 BeautifulSoup 来解析下载下来的歌曲网页对应的 HTML 文件。BeautifulSoup 使用起来非常简单，可以非常容易地完成对 HTML 文件的解析，同时它也支持按照不同的条件来查找相关元素，比如按标签查找、按属性查找、按名称查找、按结构查找等。

5.2 结巴分词工具

由信息检索等相关知识可知，要对文本进行分析，往往首先需要进行分词，即将连续的字序列按照一定的规范重新组合成词序列的过程。比如将句子“南京市长江三桥”分成“南京/市长/江三桥”或者“南京市/长江/三桥”这样的词汇序列。目前，学术界和工业界也出现了众多成熟的分词工具，比如基于词频词典的机械中文分词引擎 SCWS、中科院的汉语词法分析系统 ICTCLAS(Institute of Computing Technology, Chinese Lexical Analysis System)、基于 HTTP 协议的开源中文分词系统 HTTPCWS 以及支持多种分词模式的结巴分词等。考虑到各工具

的效率、可用性、精度以及源码获取难易程度，本章所述原型系统采用结巴分词作为分词工具。

结巴分词是一个基于 Python 语言开发的开放源代码的中文分词工具，其由百度(Baidu Inc.)的 Sun Junyi 开发并发布在 Github 上，其目标是“做最好的 Python 中文分词组件”。结巴分词自从 2012 年 10 月 7 日被发布到 pypi 以来不断地被改进和完善，目前已经发布的最新版本为 0.32。其所采用的算法如下所示：

(1)基于 Trie 树结构实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图（DAG)。

(2)采用了动态规划查找最大概率路径, 找出基于词频的最大切分组合。

(3)对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法。

总得看来，结巴分词之所以能够被广泛关注和使用，主要是因为其具有以下特点：

(1)安装简单。用户可以直接通过 easy_install jieba 或者 pip install jieba 进行安装，就行安装普通的 Python 包一样。

(2)使用简单。用户只需使用一句代码即可实现分词操作，如用户通过如下代码即可将句子“南京市长江三桥”分词若干词汇序列并将词汇保存到列表中。

seg_list = jieba.cut("他来到了网易杭研大厦")；

(3)支持多种分词模式。结巴分词支持三种分词模式，即精确模式、全模式和搜索引擎模式。其中，精确模式试图将句子最精确地切开，适合文本分析；全模式，把句子中所有的可以成词的词语都扫描出来, 速度非常快，但是不能解决歧义；搜索引擎模式，在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。

(4)支持多种操作。除了基本的分词之外，结巴分词还支持添加自定义词典、关键词提取、词性标注、并行分词、繁体分词以及 Tokenize 等功能。

(5)支持多语言。除了最基础的 Python 语言版本外，目前结巴分词还支持 Java、C++以及 Node.js 三种语言且源码均已发布到 Github 上。

对于歌曲对应的文档，我们使用结巴分词工具进行分词，得到的结果为：

5.3 Gensim 软件包

为了得到每一首歌曲在隐含主题上的概率分布，我们使用以 LDA 为代表的主题模型分析方法对每首歌曲对应的文档进行分析，这里的文档是由用户为歌曲标记的标签构成。LingPipe 和 Mallet 都是非常优秀的自然语言处理软件包，但考虑到它们比较复杂且对 LDA 的实现欠佳，我们选择使用另一个优秀的软件包 Gensim。Gensim 最初是作为一组被用在捷克数学文献存取网站 dml.cz 中的 Python 脚本的集合而出现，而其功能只是简单地根据给定的文档来生成一组近似的文档，Gensim 正是“Generate Similar”的简称。为了使用隐含语义的方法对文档分析，作者于2010年将其扩展为一个 Python 包，随后作者于2011年开始使用 Github 来管理源代码并于2013年设计了 Gensim 独特的 Logo 和网站。Gensim 可以非常方便地实现主题模型，正如其介绍中所说—“为人类而设计的主题模型开发包”，其主要具有以下特点：

1.可扩展性。Gensim 通过使用增量式的在线训练方法可以处理大规模的语料库，从而不需要将所有语料一次性装入内存，降低了内存的负担，增强了可扩展性。
2.平台无关性。Gensim 是纯 Python 实现，可以运行在 Linux、Windows、OS X 以及其他支持 Python 和 Numpy 的平台上。
3.鲁棒性。Gensim 已经被很多个人和组织应用在各种系统中超过四年，早已过了一个开源项目的“妈妈，我发布了一个脚本”的初始阶段。
4.开源性。Gensim 开放源代码，其使用 GNU LGPL 许可证，允许个人和商业机构使用和修改该项目。
5.高效性。Gensim 中的各种算法都是使用经过优化的方法进行实现的，使得算法的效率较高；另外，Gensim 实现了一些算法的分布式版本，使得算法可以并行执行或者在集群上执行，进一步增加算法的执行效率。
6.Gensim 包含对一些常用数据格式的高效内存实现方式，同时支持不同数据格式之间的转换。
7.Gensim 除了可以快速地执行主题模型建模，还提供了快速计算文档相似性的方法。

下面简单介绍以下使用 Gensim 软件包进行 LDA 建模的方法和流程：

1. 准备语料库，这里就是需要进行主题模型建模的文档集合。
2. 对文档集合中的每一篇文档进行分词并利用分词的结果构造词典，同时可以得到每个词或者词组在词典中的编号。
3. 词典生成好之后就生成语料库，语料库中的每一个语料与文档集合中的每一篇文档一一对应，而语料的表示形式即是文档的向量空间模型，即词典中的某个词或词组在该文档中出现的次数。
4. 将上述向量空间模型表示的语料库转换成 TF-IDF 模型表示的语料库，即此时得到的语料库可以表征每一个词或者词组的重要程度。
5. 进行 LDA 主题模型建模，得到建模结果。

5.4 序列生成系统