



基于标签对家庭 IPTV 业务个性化推送机制的用户体验优化研究

朱映波¹,刁建伟¹,康 波²,刘胜强²

(1. 中国电信股份有限公司数字音乐运营中心 广州 510810;

2. 中国电信股份有限公司广东研究院 广州 510630)

摘 要:爱音乐卡拉 OK 业务作为一种 IPTV 业务,现阶段基于遥控器的用户操作成本高于 PC 和智能手机,需要引入内容的个性化主动推送机制以提升用户体验。但卡拉 OK 作为一种家庭业务,具有服务器端和客户端对客户主体认定不对称的特点,依据普通的推荐算法极易导致不同家庭成员之间的错位推荐。同时,传统推荐算法本身也存在马太效应、冷启动等缺陷,这些都会导致用户体验下降。针对这些用户体验问题,从业务的用户使用场景出发,对推荐算法进行选择与改良,并结合内容标签提升推荐的针对性,以克服算法缺陷,提升用户体验。研究的首要着眼点在于用户体验效果,而非算法的复杂性。

关键词:关联规则;标签;个性化推送;层次分析法;权重;奇异值分解;独立性检验;马太效应

doi: 10.3969/j.issn.1000-0801.2014.07.018

User Experience Optimization Research on Personalized Push Approach of IPTV Service Based on Label of Contents

Zhu Yingbo¹, Diao Jianwei¹, Kang Bo², Liu Shengqiang²

(1. Digital Music Operation Center of China Telecom Co., Ltd., Guangzhou 510810, China;

2. Guangdong Research Institute of China Telecom Co., Ltd., Guangzhou 510630, China)

Abstract: iMusci Karaok is a service on IPTV developed and operated by China Telecom. It is needed to introduce active personalized push approach of content in order to improve user experience, due to higher cost of remote controller than PC and mobile phone. However, Karaok is a household service, and is ununiform to identify users on server side and terminal side, which will bring inaccurate result when normal recommendation algorithm is adopted. Meanwhile, traditional recommendation algorithm has defects of Matthew effect and cold reboot, which will decrease user experience. To tackle such problems of user experience, the recommendation algorithm was selected and optimized by combination of content label so as to increase recommendation objectively. This research is aiming at enhancing user experience effect rather than algorithm complexity.

Key words: association rule, label, personalized push, AHP, weight, SVD, independence test, Matthew effect

1 引言

爱音乐卡拉 OK 已成为中国电信 IPTV 业务中最受欢

迎且发展最快的增值业务,其遥控器的操作方式带来更高的人机交互成本,迫切需要业务系统进行主动的内容推送机制。不同于基于 PC、手机等个人终端的业务,卡拉 OK 作



为一种家庭业务,服务器端对用户主体的家庭客户身份认定与客户端即时使用者的个体用户身份之间存在明显的信息不对称,若直接使用传统的推荐算法基于家庭客户数据计算推荐规则,则当规则作用于个体的家庭成员时,往往造成即时使用者收到适合其他家庭成员的推荐内容,直接造成用户体验度降低。

另一方面,一些推荐算法,如关联规则、协同式过滤等,当用于实际业务时,本身也隐藏一些无数理逻辑错误但却不完全吻合用户体验的地方,比较典型的如内容缺少点播记录造成无法输出推送规则的“冷启动”问题、倾向于优先推荐热门内容的“马太”效应。这些都需要根据具体的业务情况进行逐一修正和处理。同时,在卡拉 OK 业务严格的内容审核过程中产生了优质的内容标签数据,具有定义准确、结构化程度高、数据完整无缺失等特点,是优化家庭客户个性化推送规则的宝贵资源。

针对以上问题和背景,设计卡拉 OK 业务的推送机制优化工作流程如下。

(1)选择关键标签:计算不同标签对用户选择歌曲的影响力,选取关键标签确保对优化工作的针对性,并提升计算效率。

(2)选择符合业务使用情景的基础推荐算法:从用户体验出发,结合客户端的推送应用场景,选择与用户行为最吻合的基础推荐算法。

(3)结合标签数据的推送机制进行优化:包括对 IPTV 个性化推荐中的家庭客户身份认定不对称性问题的解决、关联规则本身马太效应的克制、无关联规则歌曲的规则补漏。

2 关键标签的选取

针对用户选择歌曲这一特定行为,筛选对其影响显著的关键标签,以确保标签对用户歌曲选择行为的解释能

力,同时剔除冗余标签,以避免数据相关性对内容相似度算法精确性造成的不利影响。该工作分为两个阶段:计算不同标签对用户选择歌曲行为的效用大小;利用相关性检验剔除冗余标签,选出关键标签簇。

卡拉 OK 内容审核的标签体系共有 11 个大类 130 个标签项,表 1 是部分内容的示例。

2.1 测度不同标签对用户选择歌曲的效用值

确定标签权重的方法有定量和定性两大类。定性方法主要有 Delphi 法和 AHP(analytic hierarchy process,层次分析)法等,定量方法包括变异系数法、因子分析及响应模型判定等。定量方法相对客观,但结果的经济意义模糊,且很容易受样本数据结构的影响,并不一定反映用户的真实心理认知;而 Delphi 法更适用于战略性决策,对打分的专家的能力要求很高,且工作流程十分繁琐(需经多轮反复达成一致意见)。比较而言,集合专家意见的半定量分析法——层次分析法最适合该研究情景。在该研究中只需考虑各标签对用户选择歌曲影响力的大小,故只需建立单层模型,过程简介如下。

(1)建立权重判断矩阵

为了保证得到权重的客观性,笔者在确定权重时请到多位音乐专家及业务运营管理人员就各类标签对用户选择歌曲的影响力大小进行评判,沿用了 AHP 创始人 Saaty T L 提出的 1~9 分的比例标度法作为依据给出分值。专家团队对各标签进行两两比较,给出它们相对重要性的判断值,全部 n 个标签经过两两判定之后,形成一个比较判断矩阵 B ,如下:

$$B = (b_{ij})_{n \times n} = \begin{pmatrix} b_{11} & \cdots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{n1} & \cdots & b_{nn} \end{pmatrix} \quad (1)$$

其中, b_{ij} 为标签相对于标签的重要度。

表 1 卡拉 OK 歌曲标签表(部分)

歌曲语种	节日类型	歌曲发布年代	音乐风格 (复选)	音乐心情 (复选)	卡拉 OK 版本	歌曲嗓音类 型(复选)	歌曲节奏 类型	伴奏乐器组成	...
国语	元旦	20 世纪 50 年代 及以前	流行	开心	A 版风景	温柔	节奏动感	电子乐器为主	...
粤语	妇女节	20 世纪 60 年代	网络	甜蜜	A 版人物	甜美	节奏明快	不插电乐器为主	...
闽南语	元宵节	20 世纪 70 年代	摇滚	祝福	A 版剧情	明亮	节奏中等	电子、不插电各半	...
英语	春节	20 世纪 80 年代	电子	悠扬	A 版动画	清澈	节奏舒缓	...	...
日韩	传统情人节	20 世纪 90 年代	校园	抒情	B 版风景	吟唱	...	...	...
多语种混合	党的生日	2000 年以后	乡村	幽默	B 版人物	中性	...	...	...
...	...	...	...	...	...	...	...	...	...

(2)计算单准则条件下标签的权重系数

运用特征根法(eigenvalue method,EM)计算该权重系数,并将其归一化为各标签对用户选择歌曲的权重向量。对于判断矩阵 B , 计算满足 $BW=\lambda_{\max}W$ 的特征根与特征向量。其中, $\lambda_{\max}$ 为 B 的最大特征根, W 为对应于 $\lambda_{\max}$ 的正规化特征向量, W 的分量 W_i 即相应标签 i 单排序的权值。

(3)进行一致性检验

为了保证所得权重的合理性及正确性,在计算权重向量后,应对每个判断矩阵进行一致性检验,无法通过一致性检验的矩阵,反馈给打分者对标签分值进行适当修改。一致性检验系数 CR 为:

$$CR=\frac{CI}{RI} \tag{2}$$

其中,一致性指标 CI 可表示为:

$$CI=\frac{\lambda_{\max}-n}{n-1} \tag{3}$$

RI 为平均随机一致性指标,由 Saaty T L 给出,见表 2。

当 $CR\leq 0.1$ 时,一般认为该判断矩阵具有满意的一致性;当 $CR>0.1$ 时,则认为该判断矩阵不具有 consistency,应该调整判断值,直到通过一致性检验。

(4)实际的计算结果

综合各专家打分,得到结果矩阵(见表 3),其中原始的 1~9 分值被平均后转化为小数。

计算出的最大特征根 $\lambda_{\max}=12.08$, 对应的特征向量为 (0.33 0.35 0.54 0.12 0.30 0.38 0.37 0.26 0.10 0.06 0.07)。一致性检验系数 $CR=0.107$,基本通过一致性检验。可以看到,对用户选择歌曲的行为而言,歌曲影响年代(0.54)、音乐风格(0.38)和音乐心情(0.37)是效用值最高的 3 个标签,在实际业务运营中必须优先考虑。

2.2 剔除冗余标签

标签体系中可能存在的高相关性会对后继的相似度计算等应用的准确性产生显著影响,并加大计算量,因此需要进行测量并剔除冗余标签。由于标签数据多为离散的列名变量,在计算相关系数时不宜采用常规的皮氏或斯氏系数,宜采用专门用于测量列名变量独立性的 χ^2 卡方检验,其统计量的计算式如下:

$$\chi^2=\sum_i\sum_j\frac{(n_{ij}-e_{ij})^2}{e_{ij}} \tag{4}$$

其中, n_{ij} 为标签 1 的第 i 水平与标签 2 的第 j 水平交

表 2 平均随机一致性指标

阶数 n	2	3	4	5	6	7	8	9	10	11
RI	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49	1.51

表 3 歌曲标签比较判断矩阵

	歌手 性别	歌曲 语种	歌曲影响 年代	歌手 人数	是否 原唱	音乐 风格	音乐 心情	歌曲节奏 类型	歌曲音调 高低	卡拉 OK 画面背景	卡拉 OK 歌曲版本
歌手性别	1	0.60	0.53	6.20	0.71	0.71	1.11	1.13	2.60	6.60	6.60
歌曲语种	1.67	1	0.53	5.40	0.71	0.73	1.11	1.13	2.60	6.60	6.60
歌曲影响年代	1.88	1.88	1	6.20	5.00	1.20	0.51	1.13	6.60	6.40	6.40
歌手人数	0.16	0.19	0.16	1	0.21	0.31	0.31	0.73	1.00	6.20	1.40
是否原唱	1.40	1.40	0.20	4.67	1	1.40	0.93	1.40	2.60	1.80	1.80
音乐风格	1.40	1.36	0.83	3.18	0.71	1	1.80	1.20	6.00	5.20	3.60
音乐心情	0.90	0.90	1.94	3.18	1.07	0.56	1	1.20	6.00	5.20	3.60
歌曲节奏类型	0.88	0.88	0.88	1.36	0.71	0.83	0.83	1	2.60	4.87	3.27
歌曲音调高低	0.38	0.38	0.15	1.00	0.38	0.17	0.17	0.38	1	1.67	3.27
卡拉 OK 画面背景	0.15	0.15	0.16	0.16	0.56	0.19	0.19	0.21	0.60	1	1
卡拉 OK 歌曲版本	0.15	0.15	0.16	0.71	0.56	0.28	0.28	0.31	0.31	1.00	1



叉的样本频数,而 e_{ij} 为对应的期望频数,计算式如下:

$$e_{ij} = \frac{n_i \times n_j}{n} \quad (5)$$

在大样本下,该统计量服从自由度为 $(r-1) \times (c-1)$ 的卡方分布,其中 r 表示标签1的水平数, c 表示标签2的水平数。通常在0.05的置信度水平下拒绝两个标签相互独立的零假设,判定二者有明显的关联性,即 $P(x \geq \chi^2) < 0.05$ 。

以音乐心情为例,计算各心情标签(共计17个)两两比较的卡方独立检验统计量的截尾概率,最终获得表4的矩阵(截尾概率越小,代表对应的两项标签关联性越高)。

表4中大量的数值很小,提示很多标签之间是非独立的。结合专家意见进行整理以后,获得如下几个关键的标签集中代表音乐心情,见表5。

由此获得最能代表歌曲嗓音类型特点的6个变量,从而降低后继推送规则计算中的计算量。类似地,诸如音乐风格等有复选结构的标签项也采用相同的方式进行处理。

3 符合业务使用情景的个性化推荐算法的选择

个性化推荐算法是个性化推荐机制的核心,目前主流的方法有基于内容相似度的推荐和基于行为相似用户的推荐,后者又可分为协同式过滤和关联规则,两种算法简要介绍如下。

3.1 典型关联规则算法

关联规则挖掘就是发现数据集中项集之间有趣的关联或相互联系,是数据挖掘领域的一个重要分支。设 $I=\{i_1,$

表4 歌曲标签卡方独立检验结果矩阵

Obs	标签	粗犷	沧桑	磁性	低沉	高亢	浑厚	空灵	铿锵	明亮	清澈	沙哑	甜美	温柔	细腻	慵懒	悠扬	中性
1	粗犷	0	0	0.1%	0	12.9%	0	95.6%	0	0.1%	0	0	0	0	7.5%	62.7%	0	7.2%
2	沧桑	0	0	0	0	78.6%	56.7%	25.0%	51.2%	0	0	0	0	0	48.0%	31.8%	0	24.6%
3	磁性	0.1%	0	0	0	0	0.4%	0	5.3%	0	0	0.1%	0	0	40.4%	59.3%	0	0
4	低沉	0	0	0	0	0	3.0%	77.7%	65.0%	0	0	0	0	0	45.2%	45.8%	0	39.2%
5	高亢	12.9%	78.6%	0	0	0	0	0	0	0	0	0.3%	5.5%	0	0	78.1%	0.4%	0
6	浑厚	0	56.7%	0.4%	3.0%	0	0	18.7%	0	0.5%	0	43.5%	0	0	2.6%	76.8%	0	49.8%
7	空灵	95.6%	25.0%	0	77.7%	0	18.7%	0	12.3%	96.7%	19.3%	23.8%	0.3%	0	3.6%	0	39.9%	5.8%
8	铿锵	0	51.2%	5.3%	65.0%	0	0	12.3%	0	97.9%	2.9%	1.8%	0	0	0.8%	46.9%	0	29.3%
9	明亮	0.1%	0	0	0	0	0.5%	96.7%	97.9%	0	0	0	0	0	18.6%	1.4%	2.4%	0
10	清澈	0	0	0	0	0	0	19.3%	2.9%	0	0	0	0	0	2.8%	1.7%	0	0
11	沙哑	0	0	0.1%	0	0.3%	43.5%	23.8%	1.8%	0	0	0	0	0	41.7%	29.4%	0	54.9%
12	甜美	0	0	0	0	5.5%	0	0.3%	0	0	0	0	0	0	21.6%	31.4%	0	0
13	温柔	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.3%	0	0
14	细腻	7.5%	48.0%	40.4%	45.2%	0	2.6%	3.6%	0.8%	18.6%	2.8%	41.7%	21.6%	0	0	38.8%	0	60.6%
15	慵懒	62.7%	31.8%	59.3%	45.8%	78.1%	76.8%	0	46.9%	1.4%	1.7%	29.4%	31.4%	0.3%	38.8%	0	1.8%	88.5%
16	悠扬	0	0	0	0	0.4%	0	39.9%	0	2.4%	0	0	0	0	0	1.8%	0	0
17	中性	7.2%	24.6%	0	39.2%	0	49.8%	5.8%	29.3%	0	0	54.9%	0	0	60.6%	88.5%	0	0

表 5 关键歌曲标签的卡方独立检验显著性水平矩阵

Obs	标签	沧桑	空灵	铿锵	细腻	慵懒	中性
1	沧桑	0	25.0%	51.2%	48.0%	31.8%	24.6%
2	空灵	25.0%	0	12.3%	3.6%	0	5.8%
3	铿锵	51.2%	12.3%	0	0.8%	46.9%	29.3%
4	细腻	48.0%	3.6%	0.8%	0	38.8%	60.6%
5	慵懒	31.8%	0	46.9%	38.8%	0	88.5%
6	中性	24.6%	5.8%	29.3%	60.6%	88.5%	0

$i_2, \dots, i_m$ 是项的集合, D 是数据库事务的集合, 每个事务 T 是不同项的集合, 使得 $T \subseteq I$ 。设 A 是一个项集, 事务 T 包含 A , 当且仅当 $A \subseteq T$ 。关联规则是形如 $A \Rightarrow B$ 的蕴含式, 其中 $A \subset I, B \subset I$, 并且 $A \cap B = \emptyset$ 。规则 A, B 在事务集 D 中成立, 具有支持度 $support$ 和置信度 $confidence, support = (A \Rightarrow B) = P(A \cap B); confidence = (A \Rightarrow B) = P(B|A)$ 。同时满足最小支持度阈值 (min_sup) 和最小置信度 (min_conf) 的规则, 就可作为输出的强关联规则。

关联规则的挖掘过程分为两步: 生成所有频繁项集, 项集的出现频率不小于 min_sup ; 由频繁项集产生强关联规则, 规则必须同时满足 min_sup 和 min_conf 。

关联规则的一个关键问题是对生成规则的选取, 主要通过各种对规则重要性的度量如支持度 $support$ 和置信度 $confidence$, 选择最优规则。

3.2 典型协同式过滤算法

协同过滤算法是个性化推荐中应用最广泛的方法, 基于邻居用户的兴趣爱好预测目标用户的兴趣偏好。先使用统计技术 k -NN (nearest-neighbor, 最近邻) 算法, 寻找与目标用户有相同喜好的邻居, 然后根据目标用户邻居的偏好产生向目标用户的推荐。算法包括如下 3 个阶段。

(1) 表示

输入数据通常可以表示为一个 $m \times n$ 的矩阵形式, 即项目评估矩阵 R , m 是用户数, n 是项数, r_{ij} 是第 i 个用户对第 j 项的评估值, 这样的评估值可以有几个等级。矩阵 R 可表示为:

$$R = \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{m1} & \cdots & r_{mn} \end{pmatrix} \quad (6)$$

(2) 邻居形成

基于协同过滤技术, 推荐系统的核心是为当前用户寻找最相似的“最近邻居”集, 即对一个用户, 要产生一个根据相似度大小排列的“邻居”集合 $U = \{u_1, u_2, \dots, u_s\}$, u 不属于

U , 从 u_1 到 u_s , 按与 u 的相似度 $sim(u, u_i)$ 从大到小排列。用户之间的相似性测度可使用 Pearson 相关度方法和目前常用的夹角余弦计算方法。确定邻居用户有两种方法: 一是根据预先确定的相似性阈值, 选择相关性大于阈值的作为邻居用户; 二是根据预先确定的邻居数 k , 选择相关性最大的前 k 个用户作为邻居用户。

(3) 产生推荐

“最近邻居”集产生后, 可计算用户对任意项的兴趣度和推荐集。设定用户 u 和相应的已选项集 I_u , 则任意项 t 对用户 u 的效用值 $u_{u,t}$ 计算如下:

$$u_{u,t} = r_u + \frac{\sum_{i=1}^k [sim(u, u_i) \times (r_{i,t} - r_i)]}{\sum_{i=1}^k sim(u, u_i)} \quad (7)$$

其中, r_u 是用户 u 对所有项的平均评价分值, u_i 是“最近邻居”集中的用户, $sim(u, u_i)$ 是用户 u 和用户 u_i 之间的相似度, $r_{i,t}$ 是用户 u_i 对项 t 的评估值, r_i 是用户 u_i 对所有项的平均评价分值。按照效用测度值的高低即可产生用户的推荐集。

3.3 基于用户体验和使用场景选择基础推荐算法

卡拉 OK 业务作为一种电视业务, 只能使用遥控器进行人机交互, 导致用户交互成本较高。对于需要用户打分的协同式过滤算法来说, 要用户对大多数歌曲进行打分显然会增大用户的操作成本, 影响用户操作的流畅性进而降低用户体验, 也更容易造成用户对项目评估矩阵 R 的高度稀疏, 进而影响推荐的效果。相比之下, 关联推荐由于无需用户主动打分的环节, 实施的难度相对较小, 但同样存在“冷启动”的问题, 即内容在无点播记录的情况下, 无法完成推送规则的计算, 为弥补关联规则的这一漏洞, 可使用基于内容相似度的推送机制, 本文建立的关键标签选取机制已经为相似度计算提供了良好的数据基础。

在卡拉 OK 业务中选择关联规则而非协同式过滤的另一个重要原因是基于用户使用场景——为降低用户操作



成本,提升用户交互流畅性,最好的推荐时机是在用户点选或播放某首歌时,这不同于以商品销售为目的的电商网站(综合用户所有的购买记录再向其推荐最吻合的货品),优先考虑的是人机交互操作的流畅性,对算法的要求是与当前内容的高关联性。从这个场景出发,关联规则无疑比普通的协同式过滤更为适用。

4 内容推送机制的优化

4.1 对家庭客户身份认定不对称的解决方案

卡拉 OK 是基于 IPTV 的家庭业务,用户唯一标识是宽带 ID 代表的家庭客户,而家庭客户的使用数据涵盖了家庭不同成员的点播记录,尤其是老中青三代和男女家庭成员在内容偏好、点播行为上的差异显著,而即时使用业务并感知推荐效果的又是作为个体的家庭成员。因此,基于家庭客户数据计算的个性化推荐规则极易将不同成员偏好的歌曲错位推荐,致使即时使用业务的家庭成员接收到不需要的歌曲推荐,如老年人点唱红歌之后却收到儿歌的推荐,或儿童点唱儿歌之后被推荐流行歌曲等大相径庭的推荐效果,严重影响用户体验感知和操作的流畅性。

由于服务器端很难直接认定业务即时使用者的家庭成员身份,一个替代方法是先选取影响用户选择权重大的标签对歌曲进行分隔,并藉此对用户点播记录进行分类处理,再在各个分类数据集上逐个运行关联规则,从而获取该分类的个性化推送规则,避免差异太大的歌曲之间的错位推荐。

对分隔标签个数的选择是此处的一个难点,如果分隔标签太多,则造成内容被过度切割,每一类的记录数过少而不能产生足够的推送规则;如果分隔标签太少,则无法将用户作为个体的音乐偏好切割开来。在实践工作中,筛选了影响年代、音乐风格、音乐心情等对用户影响力最大的关键标签维度,基于这些维度完成了歌曲点播记录的切分。从实际效果看,基于关键标签切分数据集计算的关联规则能更好地迎合不同家庭成员的个性化需求。

4.2 对推送规则中马太效应的遏制

内容点播业务的数据具有典型的长尾特征——少量歌曲占据绝大部分点播量,而大量歌曲则点播量低下,构成数据分布的长尾,卡拉 OK 的歌曲点播量分布如图 1 所示。

这种长尾的分布很容易带来的一个缺陷是马太效应——热门愈热,冷门愈冷。由关联规则的置信度计算公式 $P(B/A) = \frac{P(AB)}{P(A)}$ 可知,歌曲 B 越热门,同其他歌曲 A 一起

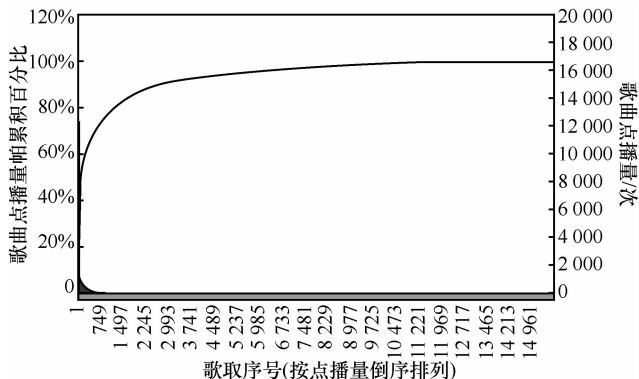


图 1 卡拉 OK 歌曲点播量帕累托累进曲线

被点播的概率 $P(AB)$ 就越高,分子 $P(AB)$ 就会越发逼近分母 $P(A)$,造成置信度 $P(B/A)$ 虚高,尤其对点播量并不高的歌曲 A 更是如此,从而占据歌曲 A 推荐列表的前列,进而为热门歌曲带来更多的点播量,加剧马太效应的发生,少量热门歌曲 B 长期霸占用户视听感官,带来用户获取不到足够内容资源与内容资源利用效率低下的矛盾。另一方面,如果关联规则的前项歌曲 A 够热门($P(A)$ 够大),则与其他歌曲的交叉集(分子 $P(AB)$)也必须够大才能确保置信度够高,因此热门歌曲的关联对象自然也是比较热门的歌曲,适合将热门歌曲单独进行关联规则的计算。针对这种情况,必须对关联规则的计算进行适当的调整,以避免马太效应的持续扩大。研究方法如下。

(1) 利用点播量的累积帕累托曲线,根据帕累托经典的二八原则(或设定歌曲对总点播量的贡献阈值),作为划分热门与非热门歌曲的依据。选取 2012 年 12 月的歌曲点播数据计算其累计概率,发现点播量最高的 1 697 首歌曲占据总点播量的 80%,而这只占总歌曲数量(15 697 首)的 7.4%,可将其定义为热门歌曲与非热门歌曲的拐点。

(2) 利用拐点分割歌曲点播记录为热门歌曲和非热门歌曲,在已经计算出的关联推荐规则中,从规则后项上区分热门歌曲和非热门歌曲分割歌曲的关联推荐规则集合,使非热门歌曲的关联推荐规则的后项优先推荐其他非热门歌曲,而热门歌曲的关联推荐规则优先推荐其他热门歌曲。

4.3 无关联规则歌曲的规则补缺

无关联规则的情形产生于以下几种情况:

- 歌曲无点播次数——冷门歌曲和新歌;
- 歌曲关联规则的置信度和支持度达不到设定的阈值;

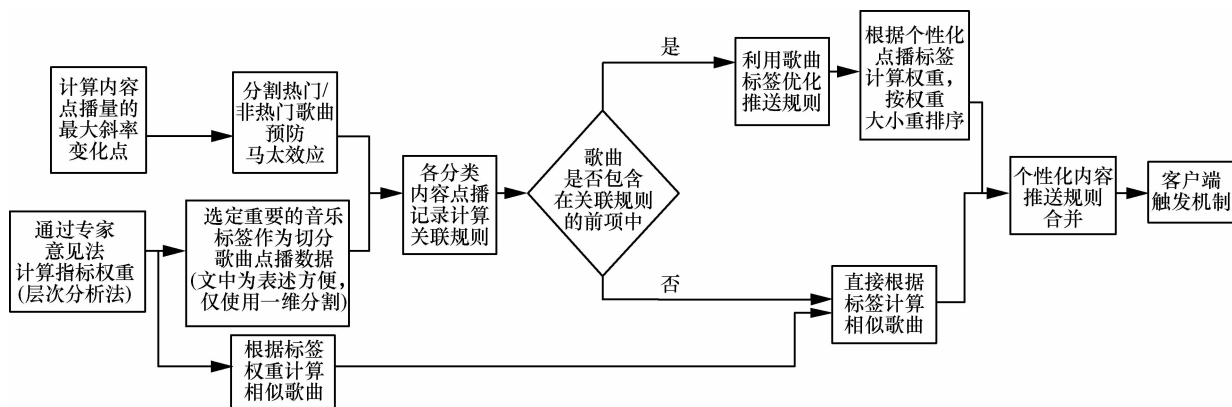


图 2 卡拉 OK 内容推送机制流程

- 规则后项数量达不到客户端页面预留的推荐内容个数。

对于这些歌曲,如果不提供推送内容或统一提供相同的推荐内容,都难免影响用户体验,因此针对这部分歌曲,可通过歌曲标签数据计算发掘与其风格相似的歌曲并进行补充推送。在实践中,筛选了关键标签并剔除了相关性高的冗余标签,为距离的测量提供了良好的数据基础。目前常用的相似度计算式有如下几种。

皮尔逊相关系数:

$$P = \frac{Cov(i, j)}{\sigma_i \times \sigma_j} \quad (8)$$

夹角余弦:

$$C = \frac{|i, j|}{|i| \times |j|} \quad (9)$$

欧式距离:

$$O = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (10)$$

由于标签统一为 0-1 型数据,无需数值的归一化处理,故直接使用欧式距离进行计算,结合前文计算出的标签影响力权重,加权的欧式距离公式如下:

$$O = \sqrt{\sum_{i=1}^n w_i \times (x_i - y_i)^2} \quad (11)$$

5 内容推送机制设计思路

结合业务处理流程和各个难点问题的解决,设计出卡拉 OK 内容推送机制流程,如图 2 所示。

在这一过程中,最基础的环节是分类内容点播记录的关联规则计算以及基于标签的歌曲相似度计算,其余环节是基于用户体验对规则计算结果的优化。该机制采用离线

计算的方式,由独立于系统的外部服务器定期更新,不会对现网运营系统造成负担。

6 针对家庭客户的智能推荐的前景展望

中国电信爱音乐卡拉 OK 业务作为首个由基地集约运营的全国性 IPTV 增值业务,采用本智能推荐算法后,明显提升了用户对推荐内容的点播次数,业务发展呈现良好的态势。从长期来看,技术的发展和用户行为习惯的变化将推进个性化推送机制的持续完善,主要的改进方向集中在以下 3 个方面。

(1) 对家庭用户细分的进一步思考

该研究中对家庭成员点播记录的区隔是采用了重点标签的交叉分割,但统一划分标准的选取可能带来对小众用户需求不同程度的忽视,后续研究中可考虑采用无监督的用户点播内容标签数据的高维聚类方法,以获取更为全面的家庭客户分类结果。

(2) 对基于用户行为和内容标签的推送规则的进一步融合

该研究虽然分离了热门与非热门歌曲,降低了马太效应的影响,但完全将热门歌曲与非热门歌曲隔离,并非十全十美的设计。一个优化的设想是,结合歌曲之间的用户点播关联度和内容本身的标签相似性,提供既有用户行为关联,也具有歌曲相似性的推送机制,并在这两种方式之间取得一定的平衡。

(3) 推送机制的进一步个性化

每个用户的内容偏好是不同的,但如果考虑每个用户的个性化偏好,势必导致计算成本的大量增加。一个可以替代的方法是在已经计算出的推送规则中,根据每个用户自身的标签偏好(单用户标签分布重心(均值)),二次计算



与每首被推荐歌曲之间的空间距离,继而按照这个距离对推荐歌曲进行重新排序。

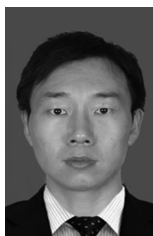
参考文献

- 1 覃亮,王喜成.层次分析法在制造业电子商务网站评价中的应用.桂林电子工业学院学报,2006(1):74~75
- 2 余力,刘鲁.电子商务个性化推荐研究.计算机集成制造系统,2004,10(10)
- 3 奉国和,梁晓婷.协同过滤推荐研究综述.图书情报工作,2011,55(16)
- 4 陈萌,杨成,王欢等.交互式电视中个性化推荐系统的研究.电视技术,2012,36(14)
- 5 徐淮杰,张二芬.基于关联规则与奇异值分解的音乐推荐系统.电子设计工程,2013,21(1)

[作者简介]



朱映波,男,博士,中国电信股份有限公司数字音乐运营中心高级工程师、总经理,在多媒体及互联网应用、电信增值业务领域有丰富的开发、管理和运营经验,主要研究方向为移动互联网音乐相关技术和业务。



刁建伟,男,中国电信股份有限公司数字音乐运营中心中级工程师、市场营销部总监,主要研究方向为互联网下音乐业务的发展。



康波,男,中国电信股份有限公司广东研究院中级经济师、客户研究经理,主要研究方向为数据挖掘、用户体验。



刘胜强,男,中国电信股份有限公司广东研究院中级经济师、消费者实验室主任,主要研究方向为用户体验、数据挖掘。

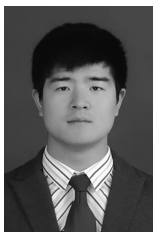
(收稿日期:2014-05-20)

(上接第83页)

- 6 3GPP TS36.401. E-UTRAN Architecture Description, 2010
- 7 李艳,张治中. LTE 网络 S1AP 监测方案的研究与实现. 电信科学, 2013, 29(1): 31~38
- 8 张应. 基于综合特征的 P2P 流量识别与控制系统研究. 复旦大学硕士学位论文, 2009
- 9 Chen H, Hu Z, Ye Z, *et al.* A new model for P2P traffic identification based on DPI and DFI. Information Engineering and Computer Science, 2009(ICIECS 2009), Wuhan, China, 2009
- 10 刘建明. 垂直搜索引擎中的主题爬虫技术研究. 广东工业大学硕士学位论文, 2013
- 11 Chaudhary A, Sardana A. Software based implementation methodologies for deep packet inspection. ICISA 2011, Jeju Island, Korean, 2011
- 12 Guo L, Wang Y, Yao Q, *et al.* A fast regular expression matching algorithm for deep packet inspection. ICITIS 2010, Beijing, China, 2010
- 13 Yu F, Chen Z, Diao Y, *et al.* Fast and memory-efficient regular expression matching for deep packet inspection. Proceedings of the 2006 ACM/IEEE Symposium on

Architecture for Networking and Communications Systems, San Jose, California, USA, 2006

[作者简介]



王建,男,重庆邮电大学通信网与测试技术重点实验室硕士生,主要研究方向为宽带通信网测试技术。

张治中,男,重庆邮电大学通信网与测试技术重点实验室博士生导师、教授,主要研究方向为通信网测试技术、宽带信息网络、NGN 等。

骆云龙,男,重庆邮电大学通信网与测试技术重点实验室硕士生,主要研究方向为宽带通信网测试技术。

(收稿日期:2014-06-10)