



大数据在音乐推荐质量提升中的实践及应用

张玉忠,方艾,金铎,袁立宇

(中国电信股份有限公司广东研究院 广州 510630)

摘要:基于音乐推荐的应用实例,探索和实现了大数据如何提高推荐质量的过程及方法。提出通过建立基于RFM模型的用户歌曲综合评分体系,在推荐算法中引入项目稀疏度、重叠度、可信度概念作为调整因子,在混合推荐时引入飙升词、内容标签和二次规则过滤等组合方法以解决推荐系统面临的常见问题,为大数据应用提供具体的参考和指导。

关键词:大数据;RFM模型;协同过滤推荐

doi: 10.3969/j.issn.1000-0801.2014.10.008

Big Data Improving Recommendation Quality in Music Applications

Zhang Yuzhong, Fang Ai, Jin Duo, Yuan Liyu

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

Abstract: The process and methods of big data improved recommendation quality in music applications were focused. The individual music score system referring RFM model was described. The sparsity and overlapping and reliability was applied to adjust recommendation algorithm, combined soaring words and content labels and filtering rules with mixing recommendation to resolve the common problems of recommendation system.

Key words: big data, RFM model, collaborative filtering recommendation

1 引言

大数据时代下,每个人都是数据的消费者,也是数据的贡献者。IDC(International Data Corporation,国际数据公司)的调查数据显示,全球产生和复制的数据量到2015年将达到8 ZB,2020年甚至有望达到40 ZB^[1]。以具体的例子来看,百度的大数据当前虽然已经达到EB级别,但其信息量占比却不足全人类所掌握信息量的百分之一。即使在游戏、电子商务、音乐应用等垂直领域,数据量级别虽然大大减少,但数据量依然非常惊人。因此如何高效地组织和利用大数据的分析结果,挖掘大数据蕴含的潜在价值,为用户提供个性化的精确推荐服务以避免大海捞

针式的搜寻,在当前信息技术产业中掀起了一个大数据的热潮。正如《大数据时代》的作者维克托·迈尔·舍恩伯格对大数据的洞见:“大数据时代最大的转变,就是放弃对因果关系的渴求,取而代之的是关注相关关系。”

数据推荐作为大数据应用的方向之一,其目的就是通过收集和整理用户的历史行为、消费习惯、社会属性等,把看似分散的、无关的、多样性的底层数据转化为用户知识图谱和用户兴趣画像,这些综合的、格式化的知识库提供了观察用户的立体化、全方位的视角,并通过应用平台以功能组件模式向终端用户输出。这些充分组合包装的信息,尽力避免了用户和产品之间不必要的交互,进而把信息向着用户服务的“私人定制”智能化目标推进,使得信息

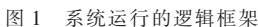


滤推荐算法在内容分类信息缺乏(如标签体系不完善)、具体内容难于量化表达(如用户音乐品味)等情境下的推荐效果较好,但是它面临的缺点也是很明显的,如系统规模过大时,系统性能会变差;项目数量变多后,会造成评分矩阵稀疏度过低、新用户的冷启动等问题^[23]。

本文以音乐产品的歌曲推荐为例,介绍大数据技术如何应用在音乐产品中,特别是如何通过多种推荐方式的组合以提高推荐质量的过程和方法。与众多关注具体单一算法的实现及准确度的推荐系统不同,本文所述的推荐系统侧重于解决准实时的在线运营产品面临的如下困难:既要考虑推荐的准确性,也要考虑算法的复杂度和高效率,从业务角度还要考虑系统维护和内容调整的方便性、实时性。因此在解决方案上采用“协同过滤+业务规则+标签内容体系等”多种模式相结合的混合推荐算法,经过一年左右的线上运营,相比于单一的推荐算法和推荐方式,本方案的推荐效果要理想很多。

系统运行的逻辑框架如图 1 所示。本推荐系统以 Cloudera 提供的 CDH 开源分布式计算框架作 ETL 处理平台,以 Mahout 工具集作为协同过滤算法的实现框架,通过

《埃森哲 2014 年技术展望》报告提出“无边界企业的崛起”概念:企业的劳动力资源不再只是企业的员工,还包括互联网上的所有用户,因为互联网时代的企业可以利用大数据技术把自己产品的用户(更广泛意义上还包含整个网络用户)视为自己的编外员工。充分发掘产品相关的“互联网用户”产生的数据价值,有助于从群体规律中提炼出用户的个性特点,以个性化和差异化的服务来帮助企业扩大用户规模,延长用户的产品生命周期。协同过滤正是上述“集体智慧”应用的典型推荐方法之一,主要分为基于用户(user-based)和基于项目(item-based)两类算法。协同过



海量原始数据的收集和处理、用户音乐评分归一化、初步歌曲推荐列表产生、内容标签体系建立和二次过滤规则等多个步骤搭建了一个实时歌曲推荐系统。

在推荐模型的算法调整和优化时采用如下方法:在用户歌曲评分上参考 RFM(指最近一次消费(recency)、消费频次(frequency)、消费金额(monetary))模型,建立考虑多个维度的综合评分;在推荐算法上,引入项目稀疏度、重叠度、可信度概念;混合推荐时引入飙升词、内容标签和二次规则过滤等方式,具体分为以下 5 个步骤。

2.1 收集大数据

用户的音乐数据从形式上来说有格式化/半格式化数据,甚至有非格式化数据;从来源来看有歌曲的浏览、试听、下载、购买、搜索、收藏等纯粹的音乐偏好行为,还有评论、分享、转发等社交性质的行为,这些信息可以从产品平台日志或通过互联网设备捕捉用户上网的 URL 清单获得。

还有一些信息如终端信息、消费能力、电信套餐及增值业务、移动支付能力等可以从电信 CRM 系统获得,其社交关系图谱等可以从电话话单、短信以及 SNS 类社交软件日志获得。

音乐的内容和标签可以从产品的字典库获得,如曲风、歌手、歌词,甚至可以依靠爬虫软件从互联网补充字典库所缺失的必要信息。

2.2 评分归一化

用户的评分数据有很多种类:如 1~5 分的 5 级数字评分表;差(差评)~好(点赞)的描述性阶梯等级,但这两种方式的评分多依赖于用户个人的主观判断,也就是说对同一首歌曲,评分为 3 分的用户 A 并不见得比评分为 2 分的用户更欣赏这一首歌曲。要解决这类问题,可以采用归一化的方式,以用户 A 对其所有歌曲的评分均值为标准,其他分数相对该均值的比值作有效评分,并二次转化为相应的可度量值。

对于没采用评分数据的系统来说,可以通过用户在浏览歌曲的相对驻留时长、对歌曲试听或下载的完整程度以及业务权重(如付费购买>免费下载>免费试听)相结合的方式作综合评价。

最终的评分结果是对上述两类评分,再做一次处理。因为用户对歌曲的偏好,单独考虑“用户评分”这个维度并不能完整表述其偏好程度,因此参考数据挖掘的 RFM 模型,以周或天为时间单位,以离当前日期的周期值作 R ,以

该周期内对同一首歌曲的试听次数作 F ,以听歌的完整程度或业务权重作 M ,构建用户对歌曲的综合评分。

2.3 相似度计算

随着推荐系统规模越来越大(以本音乐产品为例,仅版权歌曲数量就超过了 30 万首,互联网歌曲达 90 万首,更不要说淘宝这样的在线商品数超过 8 亿件的巨型系统),庞大的商品数目必然会导致两个用户之间选择的重叠非常少,因此评分矩阵的稀疏性问题也是推荐面临的难题之一。

目前解决稀疏性通常采用算法改进,以提高推荐效果,如矩阵降维、聚类等,本文引入“可信度”概念以调整用户相似度取值,具体如下:假设用户 A 和用户 B 的歌曲历史记录分别为 M 和 N ,其歌曲空间维度值为 Q ,相同歌曲数目为 P ,则取可信度为:可信度=稀疏度×重叠度= $(M/Q \times N/Q) \times (P/Q)$,因此以前的用户相似度计算也要调整为:可信度×用户相似度,按照这种修正算法,常用的 Pearson 模型可变为如下计算式:

$$D_{\text{Pearson}} = D_{\text{reliability}} \times \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)} \sqrt{E(Y^2) - E^2(Y)}} \quad (1)$$

其中, $D_{\text{reliability}}$ 是新引入的可信度值。

“稀疏度”反映了用户的向量距离远近程度,如果两个用户之间的歌曲交集差异越大,那么这两个用户的距离就越远,反映在计算式中,他们的稀疏度值接近于 0。

“重叠度”从另外一个维度反映了用户的向量距离远近程度,如果重叠度接近于 1,则表示两个向量都在某些相同的属性上具有实际值,那么得到的距离就具有较高的可靠性。

“可信度”是向量的相对稀疏度和重叠度的乘积,共同影响用户距离的计算。可信度的概念实际是对现有传统协同过滤算法中计算距离方法的一种改进,可以降低因计算距离过程中的随机性导致的距离偏差。由于在用户相似度的计算中新引入了上述“可信度”变量因子,这使得在计算用户邻居的时候更准确,降低原算法中因为缺失值导致的用户邻居发现错误的可能性,由于计算这些值都可以在很短的时间内完成,因此对原算法的性能、计算速度影响不大,都在可接受的范围内。

2.4 内容分类

上述用户评分矩阵通过协同过滤算法得到推荐歌曲列表后,有时效果仍然不是很理想。其中部分原因是用户在听歌的过程中被一些外界原因干扰而出现了偏好噪音:



若用户听歌过程被打断后忘记进行“暂停”操作,但音乐产品却随机向用户播放音乐;也有可能是用户被动地接受了音乐产品的歌曲列表,而实际并没有按照自己的偏好过滤和选择。针对这种情况,分析歌曲的内容分类可以进一步提高推荐质量。

内容分类在数据挖掘领域有很多分类算法,对音乐产品来说,有基于歌词分词的文本内容分类,也有基于音符或音律的语音学分类。本文在实践中以歌词分类的方式,采用朴素贝叶斯算法,以互联网音乐分类较好的典型分类和样例歌词归类为种子,对字典库的所有歌词做了分类处理,整体效果可以达到40%~60%的匹配率。

另外一种分类是基于统计计数的分类,例如飙升词、热榜等。本文在使用这些统计数据的时候,没有采用简单的计数方式,而使用了用户归一化评分的累积方式。以飙升词为例,定时统计每个时段的统计指标后,跟上一个时段的指标值相比较,如果相对比值超过一定的阈值,则认为该指标发生了飙升,从而指导运营人员快速响应,更新产品的相关推荐列表。

2.5 混合推荐

协同过滤推荐算法在满足个性化需求方面虽然能起到较好的作用,但是在实际应用时通常需要混合的方式做歌曲推荐^[4],也就是要采用多种推荐方式相互交叉和补充,这主要是由于以下几种情况的存在。

首先,第一次使用音乐产品的用户或者信息量过少的用户,推荐系统需要处理冷启动问题。这类用户一种常用的解决办法是利用引导性的文本信息进行辅助推荐,亦可通过在首次使用或者注册时通过导航机制获得用户的关键属性信息,如年龄段、居住地/地域方言、教育程度、性别、行业等;另外音乐的标签系统(如语种、曲风、影视剧、歌手、热榜榜单、首发/经典)的广泛应用也是解决冷启动问题的可选方案,这些标签的分类整理本身就是商品内容的萃取,同时也可以视为收集用户个性化偏好的初始种子。

其次,协同过滤虽然在推荐算法上效果较好,但也会碰到多样性、新颖性和覆盖率问题。具体表现为推荐的歌曲范围太窄,特别是用户使用一段时间之后会出现重复推荐,而且容易导致曲库字典的大量歌曲成为“僵尸”,这种缺陷带来的恶果就是用户吸引力和满意度下降,甚至导致用户的留存率降低或用户流失。

再次,从业务运营的角度来看,收费歌曲、上线新歌、主推歌曲、节假日应景音乐也需要一定的人工干预

和展示途径。这些运营元素的引入,可以在一定程度上弥补单一算法带来的新颖性不够等缺陷,扩大“长尾”,获得规模效应。

3 音乐产品的实践应用及效果

利用上述的大数据平台框架及推荐处理流程,项目团队在实际的运营支撑中进行了一系列的音乐产品研发运营实践,包括建立个性化用户知识库、差异化歌曲推荐、开发和运营流程优化、用户体验及反馈分析等。

以某款手机的定制播放器为例,它是音乐运营中心与手机终端设备供应商之间约定的应用预装产品,经过一年半的运营,到2014年上半年止,该产品的注册用户数达到2500万户,日UV(unique visitor)突破100万户,产品功能包括听歌、搜索、收藏、下载、分享、订购等。这些用户产生的原始日志量每日可达50GB,有效数据超过5000万条,再加上超过2亿条CRM历史记录以及其他终端、评论、SNS数据,这些用户行为数据以日为单位,定期经过大数据平台收集、整理、更新,最终导入用户歌曲评分矩阵为推荐备用。

当用户登录后,首页显示的是一系列通过复杂、大规模、实时的算法得到的差异化信息,同时还有一些热榜、专辑介绍等信息。以图2为例,当用户搜索并播放歌曲《高山流水》时,在用户界面的左上部以列表的形式显示经过混合推荐处理过的Top10歌曲,而在左下部显示推荐的少量音乐专辑及图片,在歌曲的播放过程中通过开关切换来控制是否显示源于互联网的歌词或者歌手图片(这些信息是爬虫定期更新和运营人员人工整理的);右半部分的用户



图2 用户界面示例

界面以 tab bar(选项卡)切换视图的形式显示 Top10 的热榜/新榜/飙升榜 3 类不同维度的歌曲以满足新用户的选择。整体来看,在这个用户界面设计里,个性化推荐歌曲、热榜歌曲、业务运营所需的人工干预歌曲都体现了出来,在一定程度上解决了冷启动和新颖性不足等缺点,可以满足不同用户的音乐品味,收到了较好的效果。

4 结束语

由于 Hadoop 等开源项目的发展及应用,特别是 CDH 等开源平台及生态系统的稳定以及相关工具如 Hive、Impala、Machine Learning 等的广泛使用,解决了早期的推荐系统在数据量过大时面临的内存、CPU 等计算瓶颈的难题。基于分布式并行计算的算法改良可以充分发挥云计算的特点,提高大规模数据分析的效率,推动基于大数据的推荐系统的发展^[5];同时由于分布式计算架构也降低了大数据处理的预算成本和技术门槛,IT 部门可以较高的性价比进行大数据项目的研发和实践,近来以 Spark 为基础的基于内存的迭代式计算框架与以 MapReduce 为基础的文件流模式的计算框架互为补充,也为机器学习(ML)和数据挖掘(DW)提供了技术保障。但就目前而言,大数据的应用及产生的效益还处于起步阶段,缺乏成熟的大规模的商业模式,信息内容最终还是需要转化为以结构化信息为主,未来随着挖掘工具和算法的进一步发展和成熟,大数据在应用智能化和商业模式创新等方面必定会起到更大作用。

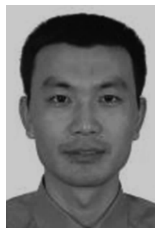
未来,互联网推荐类应用可以通过跨域行业应用数据和互联网海量数据的深度整合,建立多维的、综合的统一用户视图,勾勒一个比较全面的用户画像。这种 360 度全方位的客户洞察系统的建立,可以打造一个面向用户本身及其衍生的社会网络、生活消费等多层面的个性化知识库,通过深入分析用户在不同业务中的行为及关系并挖掘

其中的潜在用户价值,可以进一步促进不同行业领域的跨域业务协同发展,同时也会为用户带来更好的体验,形成一个良性循环的生态圈。

参考文献

- 1 大数据就在你身边. <http://info.secu.he360.com/2014/01/201838764677.shtml>, 2014
- 2 大数据应用之个性化推荐的十大挑战. <http://www.kddchina.com/article-49-1.html>, 2014
- 3 个性化推荐十大挑战. <http://blog.sciencenet.cn/blog-3075-588779.html>, 2014
- 4 张瑶,陈维斌,傅顺开. 协同过滤推荐研究综述. 微型机与应用, 2013(6)
- 5 李改,潘嵘,李章凤等. 基于大数据集的协同过滤算法的并行化研究. 计算机工程与设计, 2012, 33(6)

[作者简介]



张玉忠,男,中国电信股份有限公司广东研究院工程师,主要研究方向为数据挖掘、云计算。

方艾,男,中国电信股份有限公司广东研究院工程师,主要研究方向为数据挖掘、云计算。

金铎,男,中国电信股份有限公司广东研究院高级工程师,主要研究方向为电子商务、电子支付。

袁立宇,男,中国电信股份有限公司广东研究院工程师,主要研究方向为软件开发。

(收稿日期:2014-09-23)

· 简讯 ·

2014 年 8 月份通信业经济运行情况

2014 年 8 月,我国 3 家基础电信企业通信业务保持稳定发展。

电信业务收入同比增速持续回落。8 月,3 家基础电信企业电信业务总量完成 1 528.1 亿元,同比增长 14.7%;电信业务收入完成 998.7 亿元,按可比口径测算同比增长 3.3%,比上月同比增长回落 1.3 个百分点。

1-8 月,电信业务总量完成 11 660.2 亿元,同比增长 15.0%;电信业务收入累计完成 7 924.1 亿元,按可比口径测算同比增长 6.2%,比 1-7 月同比增长回落 0.4 个百分点。

(来源:工业和信息化部运行监测协调局)