信息学院全日制硕士专业学位研究生
专业实践总结报告

学 号	4031442003	姓 名	宋嘉琪
年 级	2014 级	专业名称	软件工程
报告题目	面向用户曲类兴趣的个性化音乐推荐方法的研究与实现		
分 数		评卷教师签字	2017 年 2 月 2 日

以下为正文内容（2000 字左右，可附页）。

当网络逐渐成为人们获取信息以及消费的主要渠道时，信息的迅速膨胀导致了网络数据的极度冗余，致使人们无法迅速而准确的找到自己关注的重点。音乐作为传统的娱乐项目之一，也存在着一样的信息过载问题。若音乐网站建立之初只是简单地将音乐产品一一罗列，用户则需要花费大量时间浏览海量无关信息后，才可找到自己的兴趣所在。在漫长的浏览过程中，用户可能会失去兴趣进而放弃对网站的使用。因此，根据用户的历史信息等，发现用户偏好并主动为其推荐音乐将会大大减少用户手动搜索的时间，优化浏览网站的用户体验，从根本上提高用户对网站的兴趣以及关注度，从而达到提高服务质量、网站增值的目的。本文对用户曲类兴趣的个性化推荐方法展开深入研究：先利用 MapReduce 精简数据集，再利用 Apriori 算法计算歌曲间的关联规则，最后使用滑动窗口技术及艾宾浩斯遗忘曲线构造用户兴趣模型对其进行推荐。

个性化推荐

个性化推荐就是根据用户不同喜好推荐其喜欢或可能喜欢的事物。个性化推荐成功需要两个条件。第一是当出现信息过载时，因为如果用户获取过或者多次获取某个商品，那么他们便很容易知道自己偏好的事物类型；第二是用户没有明确的需求，需要根据他们的行为记录发现潜在的信息以推算出用户可能喜欢的事物。如果用户知道自己要找什么，他们只要通过搜索引擎就可以找到他们想要的东西。

基于这两个条件，个性化推荐就非常合适用于音乐类产品了。首先，乐坛每年都会有很多新歌出现，而且更新也很快，信息过载成为必然。其次，对于多数用户来说，音乐只是背景音乐，具体是哪些歌曲，对于用户而言并没有太大的影

响，只要是他们喜欢的曲类就可以。不论是用户常常听取类型的歌曲，还是找到用户从未涉及但有潜在偏好可能的歌曲，都成为用户希望获取到的重要数据，因此歌曲的个性化推荐在音乐网络电台得到了相当好的应用效果。

MapReduce 编程模型

MapReduce 由 Google 提出，是一种主要用于搜索领域的简化分布式编程模型和高效任务调度模型，同时，海量数据被两个步骤并行处理，数据处理规模大大提高。

MapReduce 由 Map 和 Reduce 两部分构成，其基本思想是将数据源分割为多个数据块，再将每个数据块进行分片处理，对每个数据片以 key/value 的形式进行 Map 处理，再以新的 key/value 输出，在 key 相同的情况下整合 value 形成数组。接着使用 Reduce 重写方法处理数据，处理结束后输出最终自定义的 key/value，并将其写入分布式文件系统（HDFS）中。

Apriori 算法

Apriori 算法是最为经典的关联规则挖掘算法，同时，许多效率较高的并行算法也都是基于 Apriori 算法改进而来。

基于两阶段频繁集思想的递推算法是 Apriori 算法的核心[32]。在该算法的运行过程中，找出所有的频繁集为首要任务，频繁集用于产生强关联规则。其查找所有频繁集的步骤如下：

（1）根据原事务集产生频繁 1 项集 L_1；

（2）根据频繁 k 项集产生第 k+1 层候选集；

（3）扫描事务集，找出第 k+1 层频繁集；

（4）循环第 2 步和第 3 步，直到 k+1 层频繁集为空；

（5）由频繁项集产生关联规则。

通过使用 Apriori 性质压缩了频繁集大小而提高挖掘的性能是 Apriori 算法的一大优点。

滑动窗口模型

在用户使用系统的过程中，其行为将被实时记录，选择试听歌曲时的每一次点击行为都会被记录下来，记录中对个性化推荐较为重要的信息即为此次点击行为所获得到的音乐对应标签以及点击事件计数。

但由于存储资源有限，而且还要避免过量无用信息冗余，因此不可以将所有信息一一记录，所以我们需要使用滑动窗口技术[35]来限定存储资源的容量。因此我们限定滑动窗口的大小固定不变，以此来记录处理涉及的数据集合。

在本文开发的系统中，滑动窗口用于存储用户最新和最偏好的曲类及其权重，并且记录用户最后一次使用系统的时间。滑动窗口中仅可容纳 20 组包含标签及其权重的键值对。用户初次使用时，窗口为空，当用户开始点击歌曲试听时，

窗口即开始记录相关的行为。当用户再次登录的时候，会调用上次退出系统时窗口中最后保存的内容，并对其进行分析，以此作为用户兴趣模型，并依据此模型进行个性化推荐。

艾宾浩斯遗忘曲线

德国心理学家赫尔曼.艾宾浩斯认为，刚刚被记忆的事情被遗忘的最快，而后，遗忘的速度将会越来越缓慢，最后趋于停止。但此时，人们只能记住原有内容的30%甚至更少。他根据实验结果绘成著名的艾宾浩斯遗忘曲线[36]，将记忆与时间联系在一起。

系统开发环境

（1）主要技术：Java + PHP + JavaScript + HTML + CSS + MYSQL

（2）软件运行环境

操作系统：Windows 7、CentOS 6.5

编译器：IntelliJ PHPStorm 2017 Ultimate、IntelliJ Idea 2017 Ultimate

数据库：MySQL 5.7 Community

数据库连接器：MySQL Workbench 5.7

浏览器：Chrome 55

虚拟机：VMWare WorkStation Pro 12.5

服务器：Apache Tomcat 9.0

分布式计算平台：Hadoop 2.6.0

（3）硬件运行环境

处理器：Intel(R) Core(TM) i5-3317U 2.60GHz

内存(RAM)：8.00GB

显卡：Intel(R) HD Graphics 4000 2GB 32bit

网卡：Realtek RTL8723A

硬盘：256GB

个性化推荐流程

系统中为用户提供了两种推荐列表，一种为“推荐歌曲”，另一种为“猜你喜欢”。

“推荐歌曲”列表中提供的是根据用户兴趣权重高低排序所获得的歌曲推荐列表；“猜你喜欢”则是根据前期准备工作中计算出的关联规则进行歌曲推荐所得到的列表。

（1）新用户的歌曲推荐

由于新的用户没有其相关偏好数据，因此本文中“推荐列表”以及“猜你喜欢”列表内容都是相同的。其内容都是包含数据库中出现频率最高的标签的音乐。

（2）“推荐列表”歌曲推荐

当用户开启系统时，系统将调用最后一次窗口中存有的数据，同时在数据库中遍历所有带有滑动窗口中包含标签的音乐。并对这些音乐进行评分。