基于推荐技术的中国音乐数据库系统的设计

周 强[1],李 曦[2]
(1. 中国科学院 文献情报中心,北京 100190;
2. 中国科学技术大学,安徽 合肥 230026)

摘 要:目前的中国音乐网站,没有把音乐检索、音乐推荐结合起来。文中介绍的中国音乐数据库系统网站,提供了按曲名、作曲家、演奏者进行检索,在检索结果页面中,显示推荐结果,推荐结果展示分为用户收听歌曲的详细信息界面的“浏览本乐曲资料的还浏览了”、“收听本乐曲的还收听了”两部分。文中介绍了该中国音乐数据库系统网站的设计,该网站的推荐系统采用了基于物品的协同过滤算法,文中详细介绍了基于物品的协同过滤算法。接着,全面介绍了该数据库系统网站推荐系统的设计。现在,中国音乐数据库系统网站完成了演示版的开发,功能正确,系统运行稳定,用户对乐曲推荐的结果比较满意。

关键词:基于物品的协同过滤算法;乐曲;推荐;用户行为
中图分类号:TP302.1 **文献标识码**:A **文章编号**:1673-629X(2015)07-0162-04
doi:10.3969/j.issn.1673-629X.2015.07.036

Design of Database System of Chinese Music Based on Recommendation Technology

ZHOU Qiang[1], LI Xi[2]
(1. National Science Library, Chinese Academy of Sciences, Beijing 100190, China;
2. University of Science and Technology of China, Hefei 230026, China)

Abstract: Currently, the China music website, could not combine music retrieval with music recommendation. The China music database system website is introduced in this paper, provides the retrieving by song name, composer, and player. In retrieved results page, displaying recommended results, recommended results show is divided into user listening to songs of detailed information interface of "browsing this music information also browse" and "listening to this music also listen to". The design of database system of Chinese music website is introduced, the site recommendation system adopts the collaborative filtering algorithm based on item, which is described in detail. Then, provide an overview of the recommender system design of database system website. Now, the development of demo version of database system of Chinese music has been completed, with correct function and stable running, which is more satisfied by users.

Key words: collaborative filtering algorithm based on items; music; recommendation; user behavior

1 中国音乐数据库的介绍

中国音乐源远流长,风格独特,为了保存、研究经典音乐,所以创建了中国音乐数据库。

中国音乐数据库,以乐器、演奏形式分类,收藏乐曲,分为古琴、古筝、笛、箫、琵琶、二胡、高胡、葫芦丝、唢呐、管子、扬琴、柳琴、编钟、江南丝竹、民乐合奏,其中民乐合奏包括王俊雄、史志有、李志辉、李汉颖、邵容等作曲家的曲目,可以按照乐曲名、作曲家、演奏家来检索。

2 现有的音乐数据库系统简述

2.1 中国古曲网

中国古曲网是由专业的音乐人士建立,是按乐曲来索引的,其民族音乐的数量在所有网站数据库中是最多的,其民族音乐的音质是最好的,其民族音乐知识是最全面的。

收稿日期:2014-08-26 **修回日期**:2014-11-28 **网络出版时间**:2015-06-23
基金项目:国家自然科学基金资助项目(61379040)
作者简介:周 强(1971-),男,馆员,硕士,CCF会员,研究方向为网络信息系统、信息检索、推荐技术;李 曦,博士,副教授,研究方向为嵌入式系统设计、低功耗计算机系统设计、数据库应用技术、海量信息管理等。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20150623.1031.021.html

该网站于 2004 年成立,经过多年的发展,中国古曲网目前位列国内民乐网站流量第一名,古曲音乐库将近 10 000 首,民乐视频将近 2 000 个,民乐知识文章 3 000 篇,民乐曲谱 2 500 张,古典书籍 18 000 多篇。

中国古曲网,提供了乐曲分类、检索服务,但没有提供推荐曲目的服务。

2.2　Pandora 个性化音乐网络电台

Pandora 是著名的个性化音乐网络电台,其推荐算法主要基于内容,音乐家和研究人员亲自听了上万首来自不同歌手的歌,然后对歌曲的不同特性(旋律、节奏等)进行标注,接着 Pandora 根据这些标注计算歌曲的相似度,最后给用户推荐与他之前喜欢的音乐相似的音乐。

请音乐家和研究人员亲自聆听音乐,来进行分类,需要有非常大的费用支出。

2.3　Last. fm 音乐电台

Last. fm 是世界上最大的社会音乐平台,在这里网友可以寻找、收听、谈论自己喜欢的音乐。Last. fm 一直致力于营造更民主的音乐文化:每个人均可按自己的方式收听自己想听的音乐,而无需他人为其做出选择。网站使用十二种语言为全球听众提供音乐服务。

Last. fm 记录了所有用户的听歌记录以及用户对歌曲的反馈,在这一基础上计算出不同用户在歌曲上的喜好相似度,从而给用户推荐与他有相似听歌爱好的其他用户喜欢的歌曲。这种方式比较好,文中设计的中国音乐数据库系统网站就采用了这种方法。

3　中国音乐数据库系统设计

中国音乐数据库系统,包括了搜索、推荐系统。文中重点介绍推荐系统[1-2],推荐算法采用基于物品的协同过滤算法。

数据存储在关系数据库中,文中采用 MySQL 数据库。为了加快检索速度,提高系统性能,创建了 Solr 索引。检索时,从 Solr 索引中读取数据。

Web Server 采用 Tomcat。

中国音乐数据库系统网站架构如图 1 所示。

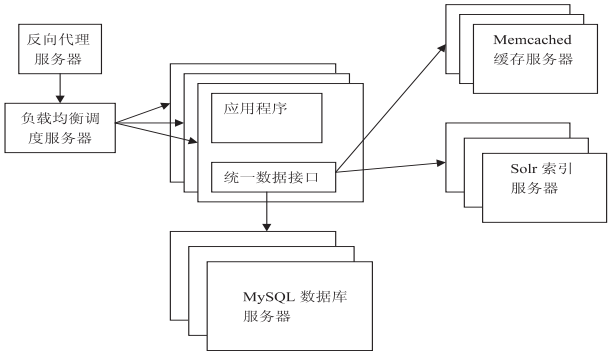

图 1　中国音乐数据库系统网站架构

在系统的最前端,配置了 Squid 进行反向代理,Squid 缓存中如果有相应资源,则直接返回客户端,提高了系统性能。

采用了 DNS 轮转技术来进行负载均衡,部署了三套系统,包括 MySQL 数据库、Solr 索引、Web 应用程序。每套系统各自分担三分之一的访问压力,提高了系统性能。如果其中一套系统宕机了,还有两套系统可用,提高了系统的可用性。

同时,采用了 Memcached 分布式缓存,提高了系统性能。Memcached 采用集中式的缓存集群管理,缓存系统部署在一组专门的服务器上,缓存集群规模可以很容易地实现扩容,具有良好的可伸缩性。

Web 应用程序中,设计开发了统一数据接口,统一访问 MySQL 数据库、Memcached 缓存、Solr 索引。

中国音乐数据库系统如图 2 所示。

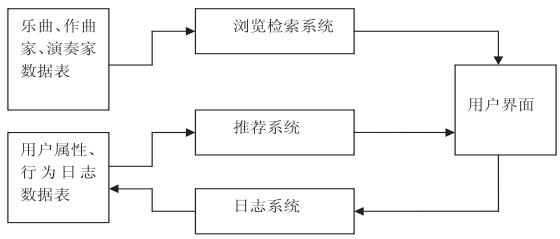

图 2　中国音乐数据库系统

乐曲数据表,字段有乐曲标识、乐曲名、演奏乐器、演奏形式、作曲家、演奏家、乐曲描述、此曲的养生功能、此曲的心情分类。

作曲家数据表,字段有作曲家名字、性别、出生日期、代表作品、作曲家简介。

演奏家数据表,字段有演奏家名字、性别、出生日期、代表作品、演奏家简介。

乐曲、作曲家、演奏家数据表提供数据给浏览检索系统,为了加快检索速度,创建了 Solr 索引,从 Solr 索引中检索数据。

系统的用户包括未注册用户、注册用户。用户属性表记录注册用户的数据,包括用户标识、名字、性别、年龄、专业、学历、居住地等;未注册用户,在日志中记录用户上网计算机的 IP 地址。

日志系统对于推荐系统是很重要的。

浏览检索日志数据表,字段有用户标识、上网计算机的 IP 地址、用户浏览时间、浏览的乐曲标识。

收听日志数据表,字段有用户标识、上网计算机的 IP 地址、用户收听时间、收听的乐曲标识。

用户通过用户界面进行浏览检索时,把用户的行为写入了浏览检索日志数据表;用户通过用户界面收听乐曲时,把用户的行为写入了收听日志数据表。

文中的推荐系统是根据用户的日志来计算并返回推荐结果,推荐结果展示是用户收听歌曲的详细信息

界面的“浏览本乐曲资料的还浏览了”、“收听本乐曲的还收听了”两个模块部分。

4 推荐算法简述

中国音乐的乐曲非常多,从大量乐曲中找到自己感兴趣的乐曲是一件相对困难的事情。推荐系统的任务是联系用户和信息,一方面帮助用户发现对自己有价值的信息,另一方面让信息能够展现在对它感兴趣的用户面前。

推荐算法[1-7]采用基于物品的协同过滤算法,该算法主要分为两步:

(1)计算物品之间的相关度;

(2)根据物品的相关度和用户的历史行为给用户生成推荐列表。

物品相似度的公式:

$$w_{ij} = |N(i)| \cap |N(j)| / \sqrt{|N(i)| \, |N(j)|}$$

其中,$N(i)$ 是喜欢物品 i 的用户数;$N(j)$ 是喜欢物品 j 的用户数。

计算物品相似度的步骤如下:

(1)建立用户-物品倒排表(即每个用户建立一个他喜欢的物品的列表);

(2)对于每个用户,把他物品列表中的物品两两在共现矩阵 $\boldsymbol{C}$ 中加 1,其中 $C[i][j]$ 记录了同时喜欢物品 i 和物品 j 的用户数;

(3)把矩阵 $\boldsymbol{C}$ 归一化,得到物品之间的余弦相似度矩阵 $\boldsymbol{W}$。

在得到物品之间的相似度后,通过如下公式计算用户 u 对一个物品的兴趣:

$$P_{uj} = \sum_{i \in N(u) \cap S(i,K)} w_{ji} r_{ui}$$

其中,$N(u)$ 是用户喜欢的物品集合;$S(i,K)$ 是和物品 i 最相似的 K 个物品的集合;w_{ji} 是物品 j 和 i 的相似度;r_{ui} 是用户 u 对物品 i 的兴趣。

这些计算是在离线的情况下进行的。

这样,可以构造物品—物品的倒排索引。

5 推荐系统设计

文中物品指的是乐曲。

图 3 是推荐系统[8-19]的架构图,说明了数据的流向。

用户的特征包括两种:一种是从用户的注册信息中提取出来的,即用户的人口统计学特征;另一种是从用户的行为中计算出来。

一个特征向量由特征以及特征的权重组成,在计算时需要考虑以下因素。

(1)用户行为分为浏览歌曲、在线收听音乐两种,其中在线收听音乐的权重大。

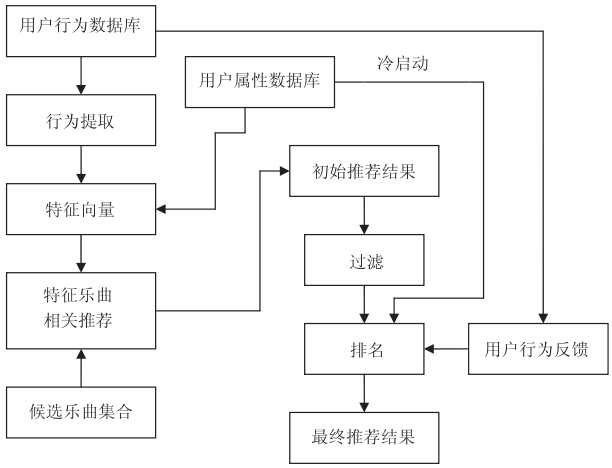

图 3 推荐系统设计

(2)用户行为产生的时间,用户近期的浏览、收听行为比较重要。

(3)用户行为的次数,用户会听一首乐曲很多次,用户对同一乐曲的收听次数反映了用户对乐曲的兴趣,收听次数多的乐曲对应的特征权重高。

(4)乐曲的热门程度,如果用户对一个热门乐曲产生了行为,有可能是跟风,可能对该乐曲没有太大兴趣,因此,对于不热门乐曲的权重高。

在得到用户的特征向量后,可以根据离线的相关表得到初始的乐曲推荐列表,其存储格式如下所示:

特征 ID、乐曲 ID、乐曲名字、心情分类、权重

在得到初步的推荐列表后,需要过滤掉不符合要求的乐曲—用户已经产生行为的乐曲。

经过过滤后的推荐结果,如果对它们进行一些排名,则可以更好地提升用户满意度。

一般排名模块包括很多不同的子模块,文中采用了按新颖度排名的模块,给用户推荐不知道的乐曲。

最后生成最终推荐列表,这推荐列表是离线生成的。为了加快查询速度,把这些推荐列表的数据写入 Solr 索引中,从 Solr 索引中检索数据。

初始的用户,没有用户行为,是冷启动的问题,这样,可以根据用户属性,来推荐一些乐曲。

6 运行结果

现在,该音乐数据库系统网站完成了演示版的开发,功能正确,系统运行稳定。

7 结束语

文中首先介绍了中国音乐数据库的需求,接着简述了现有的音乐数据库系统。在此基础上,介绍了中国音乐数据库系统的设计,并说明了基于物品的协同过滤算法,然后说明了推荐系统的设计方案。

参考文献:

[1] Degemmis M, Lops P, Semeraro G. A content-collaborative recommender that exploits WordNet-based user profiles for neighborhood formation[J]. User Modeling and User-Adapted Interaction,2007,17(3):217-255.

[2] Girardi R, Marinho L B. A domain model of Web recommender systems based on usage mining and collaborative filtering [J]. Requirements Engineering,2007,12(1):23-40.

[3] Han Jiawei, Pei Jian, Yin Yiwen, et al. Mining frequent patterns without candidate generation:a frequent-pattern tree approach[J]. Data Mining and Knowledge Discovery, 2004, 8(1):53-87.

[4] Goldberg K, Roeder T, Gupta D, et al. Eigentaste:a constant time collaborative filtering algorithm[J]. Information Retrieval,2001,4(2):133-151.

[5] 杨　博,赵鹏飞. 推荐算法综述[J]. 山西大学学报:自然科学版,2011,34(3):337-350.

[6] 马宏伟,张光卫,李　鹏. 协同过滤推荐算法综述[J]. 小型微型计算机系统,2009,30(7):1282-1288.

[7] 李　聪,梁昌勇,马　丽. 基于领域最近邻的协同过滤推荐算法[J]. 计算机研究与发展,2008,45(9):1532-1538.

[8] 王国霞,刘贺平. 个性化推荐系统综述[J]. 计算机工程与应用,2012,48(7):66-76.

[9] 许海玲,吴　潇,李晓东,等. 互联网推荐系统比较研究[J]. 软件学报,2009,20(2):350-362.

[10] 李亚楠,王　斌,李锦涛. 搜索引擎查询推荐技术综述[J]. 中文信息学报,2010,24(6):75-84.

[11] 刘建国,周　涛,汪秉宏. 个性化推荐系统的研究进展[J]. 自然科学进展,2009,19(1):1-15.

[12] 刘　鲁,任晓丽. 推荐系统研究进展及展望[J]. 信息系统学报,2008(1):82-90.

[13] 孙雨生,董　慧. 基于语义网格的数字图书馆个性化推荐研究—体系结构与总体框架[J]. 情报理论与实践,2009,32(6):63-66.

[14] 陈定权,朱维凤. 关联规则与图书馆书目推荐[J]. 情报理论与实践,2009,32(6):81-84.

[15] 李树青,徐　侠,许敏佳. 基于读者借阅二分网络的图书可推荐质量测度方法及个性化图书推荐服务[J]. 中国图书馆学报,2013(3):83-95.

[16] 王永固,邱飞岳,赵建龙,等. 基于协同过滤技术的学习资源个性化推荐研究[J]. 远程教育杂志,2011(3):66-71.

[17] 范　旭. 以豆瓣网和中国国家图书馆为案例的网上书目推荐系统研究[J]. 图书馆学研究,2008(8):44-48.

[18] 丁　雪. 基于数据挖掘的图书智能推荐系统研究[J]. 情报理论与实践,2010,33(5):107-110.

[19] 王　义,马尚才. 基于用户行为的个性化推荐系统的设计与应用[J]. 计算机系统应用,2010,19(8):29-33.

(上接第 161 页)

粮食应急的工作流程,根据粮食应急过程的特点,研发了一套集数据采集、智能预案管理、预警、应急指挥整个流程管理于一体的供省、市、县级政府使用的粮食保障决策指挥系统,用于在突发事件发生后,有效地解决粮食应急保障问题。本系统已投入正式使用,经过应用演练,各个功能模块运行稳定,有效地提高了粮食应急保障水平。

参考文献:

[1] 北京市粮食调控局. 北京市粮食供给应急预案[EB/OL]. 2011. http://www.bj-yj.gov.cn/yjya/bsyj/shaq/t1094341.html.

[2] 上海市粮食局. 上海市粮食应急预案[EB/OL]. 2006. http://www.shang-hai.gov.cn/shanghai/node2314/node26533/node26534/node26541/node26544/node26548/u21ai203375.html.

[3] 广东省粮食局. 广东省粮食应急预案[EB/OL]. 2007. http://www.gdemo.gov.cn/yasz/yjya/zxya/shaqlya/200712/t20071202_36150.htm.

[4] 陈　军,李　杰,南立波. GIS 发展应用综述[C]//第十四届全国青年通信学术会议论文集. 北京:电子工业出版社,2009:500-505.

[5] Google Developers. Google maps JavaScript API v3 X[EB/OL]. 2013. https://dev-elopers.google.com/maps/documentation/javascript/tutorial.

[6] 刘美生. 全球定位系统及其应用综述(二)—GPS[J]. 中国测试技术,2006,32(6):5-11.

[7] 李　刚. 疯狂 Android 讲义[M]. 北京:电子工业出版社,2013.

[8] 工业和信息化部电信研究院. 移动互联网白皮书[EB/OL]. 2011. http://www.catr.cn/kxyj/qwfb/bps/201212/t20121204_898927.html.

[9] 陈霄凯,刘明辉. 基于 Google Maps 的通用定位服务平台的开发研究[J]. 计算机技术与发展,2011,21(11):215-218.

[10] 岳　泉. 基于谷歌地图的车辆远程监控系统的设计与实现[D]. 武汉:武汉理工大学,2013.

[11] 徐光侠,封　雷,涂　演,等. 基于 Android 和 Google Maps 的生活辅助系统的设计与实现[J]. 重庆邮电大学学报:自然科学版,2012,24(2):242-247.

[12] 刘胜前,陈立定. 基于 Android 平台的车辆导航系统设计与实现[J]. 自动化与仪表,2012,27(4):1-4.

[13] 崔　洋,贺亚茹. MySQL 数据库应用从入门到精通[M]. 北京:中国铁道出版社,2012:381-387.

[14] 布里泰恩. Tomcat 权威指南[M]. 北京:中国电力出版社,2009:93-100.

基于推荐技术的中国音乐数据库系统的设计

作者： 周强， 李曦， ZHOU Qiang， LI Xi

作者单位： 周强,ZHOU Qiang(中国科学院 文献情报中心,北京,100190)， 李曦,LI Xi(中国科学技术大学,安徽 合肥,230026)

刊名： 计算机技术与发展 ISTIC

英文刊名： Computer Technology and Development

年，卷(期)： 2015(7)