第1章

音乐信息检索的产生与发展

音乐是我们社会中一个广泛的主题，每个人都会收听或者创作音乐。广义上讲，音乐信息检索[1]首要考虑的问题，就是从音乐信号、音乐的符号表示或网页资源上，提取或分析出音乐有意义的特征，用此特征为音乐建立索引，然后设计不同的查询和检索机制，如基于内容的检索、音乐推荐系统或大规模音乐数据库用户界面。音乐信息检索的目的就是从世界上海量的音乐数据中，为每个人提供检索服务。音乐信息检索联系着各种不同的音乐相关实体，为作曲家、演奏家、消费者与音乐作品、专辑、视频片段等建立了联系。这种联系的建立使得音乐信息检索作为一种新兴技术为社会带来了惊喜，尽管追溯到它的历史也只有短短二三十年，但作为一个研究领域，它始终站在技术发展的浪尖上。20世纪90年代起，随着音乐信号的压缩技术的迅速发展，个人电脑计算能力不断增强，使得用户能够在合理的时间内提取出音乐特征。同时，更加广泛的移动端音乐播放器的使用和更多音乐流媒体服务的发展，如阿里音乐、百度音乐、谷歌音乐等，使得音乐消费随时随地发生，鲜有限制。

音乐产业在经历了柱式唱片、胶片、卡带、CD时代后，迎来了全新的数字时代。数字音乐是指以数字化方式进行创作、编辑、存储，通过互联网和无线网络传播的音乐形式，主要分为在线音乐和移动音乐两大类。截至2017年6月，网络音乐用户规模达到5.24亿，较去年底增加2101万，占网民总体数量的69.8%。其中手机网络音乐用户规模达到4.89亿，较去年底增加2138万，占手机网民数量的67.6%[2]。当前主流的因特网站点存储着百万级电子音乐，使得用户查找、检索和发现相关音乐的复杂度大大提高。

目前，主流的业界系统提供的都是人工检索的方式[3]。这种检索通常是基于元数据的，例如艺术家的名字、歌曲名称等，少数表现语义特性的也就

只有音乐的风格[4-5]。对音乐集的检索往往只使用了标签或者其他上述文本信息。另外，系统也只提供最基本的音乐推荐和个性化服务，与音乐信号本身的内容无关。这些系统通过检测音乐消费信息、统计播放次数、记录用户点击率，或者其他行为信息来获取用户描述文件[6-7]。此类系统中，用户被表示成为点击率或者播放量计数值的向量。借助向量数据库，可以对相似的用户或者音乐文件进行协同推荐[8]。类似的方法还有使用基于语义标签的用户描述文件，查找相似的用户或者音乐文件。Firan[9]提出一种使用用户收听行为间接生成语义描述文件的方法，他们使用了用户收听行为和从用户个人音乐数据集中提取出来的元数据。这些标签从last.fm这种音乐服务网站获取或者从万维网上的评论、传记、日志、音乐相关的RSS种子等途径获得。文献[10-11]都使用了协同过滤完成音乐的推荐，在流行音乐数据上很有效。

然而，这种方法也存在长尾问题，例如，对不流行的音乐文件，缺乏必要的用户点击率、社会化标签及其他类型的元数据。有证据表明，使用基于内容的推荐，从音频内容本身出发，可以帮助解决该问题。只有很少的研究在用户描述模型上使用基于内容的方法或者融合的方法。这些方法存在一些缺点，其中包括它们都是单独使用音色或者节奏信息的，这些音频信息是低级别的信息，并不能直接转换成高级别的语义信息。研究表明在语义范围内的相关工作可以克服所谓的“语义鸿沟”，即在音频信号中提出的底层特征数据和人类的语义概念之间缺乏联系。

1.1　音乐信息检索历史与发展

早期的音乐检索技术的研究重点在于音乐作品的符号表示，如有结构的数字音乐曲谱表示MIDI等。随着计算能力的增强，20世纪20年代早期产生了大量的音乐信号处理技术，不但从乐谱当中，更能从音乐信号本身提取出不同的底层特征，如节奏、和弦、旋律等。正如Casey[12]所说，这些特征提取技术仍

然是当前致力于解决的问题。

同时，音乐还有一些重要的属性，如风格等[13-14]，它们不仅与音乐内容相关，也与一些社会化信息相关，如用户在互联网上贡献的标签评论信息等[15]，这些属性则需要建立用户行为模型，从中发现音乐的文化属性。因此，2000年代中期，研究工作多集中于不同数据资源的分析和发现，如网页、微博信息[16]，专辑封面图片信息，相关标签标注信息，以及为获取这些信息而设计的游戏[17]等资源。

近年来，在线资源的发展使得MIR有了一个迅速的提升，如Casey和Schedl[18]等人所述，MIR系统的模型设计和评价机制已经从系统中心转向用户为中心。在以用户为中心的模型中，新颖性、奇异性、流行性以及用户所处位置、时间等相关因素，融合着个体用户的音乐品味，从而形成音乐检索和推荐系统的结果。

伴随着技术的演化，用户为中心的策略考虑着用户对音乐感知特性的不同方面，但本质上都离不开音乐作品的相似性。比如对于音乐风格的感知，Lippens [19] 和Seyerlehner [20]都曾论述过，人们对某一音乐作品风格的判定上会达到75%~80%的一致性。类似地，在文献[21-22]中也都指出，在两首音乐作品的相似度上的感知一致性能达到80%以上。

MIR系统的查询系统流程图如图1.1所示。

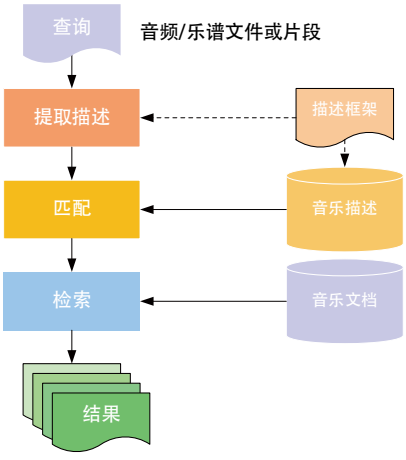


图1.1 MIR系统的查询系统流程图

1.2　音乐信息检索建模与表达

音乐作品是复杂的人类精神产物，它是诸多模型融合的产物，如音乐信号、符号表示（曲谱）、文本（歌词）、图片（音乐家的照片、专辑封面）、动作（演奏家的手势），所以它是复杂因素和谐一致的表达。Schedl等人[20]指出该模型是人对音乐作品的感知，特别是音乐相似度上的感受，主要就是受到歌词、节奏、演奏者的表现，或者用户当时的精神状态的影响。一个可计算的音乐检索系统结构可能包含的特征如图1.2所示，音乐的感知特性包括：音乐内容、音乐上下文、用户特性和用户上下文。

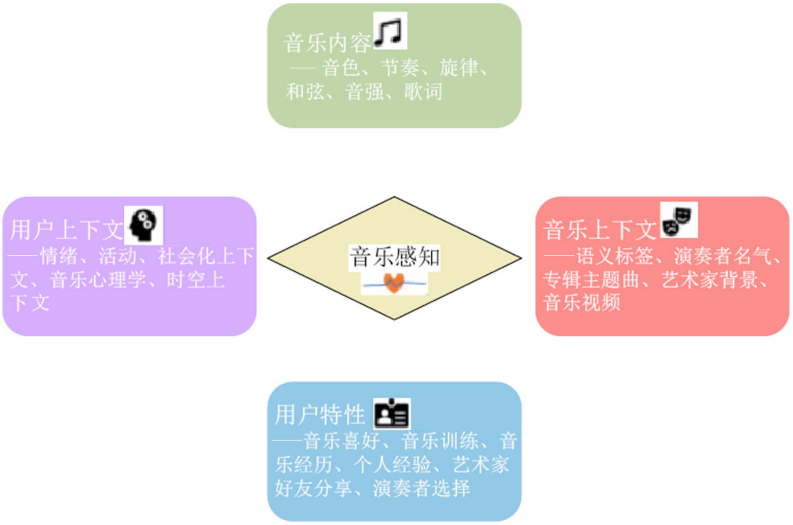


图1.2　音乐感知特性

音乐内容特征蕴含在音乐信号当中，如曲式结构、旋律和节奏。音乐上下文包含了那些不能直接从音乐信号中提取出来的信息，这些信息来自音乐片段、艺术家或者演奏者，如艺术家的文化政治背景、语义标签和专辑主打歌等。当注意力放在用户身上时，用户上下文代表着一些动态变化的因素，如用户当时所处的社会上下文、活动或者表情。相反，用户特性指的是一些不变的

或者变化缓慢的用户相关性质，如他的音乐爱好或者所受到的音乐教育，还与用户本人或他的朋友对演奏家的评价等。所以用户特性是更宽泛的长期目标，而用户上下文则受到短期的收听需求的影响。

不同类型的用户特征之间会相互影响。比如音乐相关标签的标注可以从音乐内容中建模得到，音乐风格可以从乐器的表达上得出，其他语义标签[23]如音乐作品的情感和用户心情也可以从音乐作品内容上体现出来。

理想情况下，音乐检索和推荐的方法应该融合各种类别的特征来克服所谓的“语义鸿沟”的问题。

1.3 音乐信息检索相关研究

基于内容的检索（Content-Based Retrieval，CBR）是对媒体对象的内容及上下文语义环境进行检索，如图像中的颜色、纹理、形状，视频中的镜头、场景的运动，声音中的音调、响度、音色等。基于内容的方法是从新的角度来管理多媒体数据，使得数据便于人们存取使用。不是让计算机识别和理解，而是研究人类对图片、音频内容的理解方式，从“生硬”的二进制符号串中，提取出让计算机能够进行比较和判定的特征，从而实现计算机从内容上的存取。音乐信息检索（Music Information Retrieval，MIR）是以音乐为中心的检索，利用音乐的音符和旋律等音乐特性来检索，如检索器乐、声乐作品等。音乐检索虽然可以利用文本注释，但音乐的旋律和感受并不都是可以用语言讲得清楚的。基于内容的音乐检索，通过在查询中给出示例，分析示例特征达到快速检索结果，如图1.3所示。

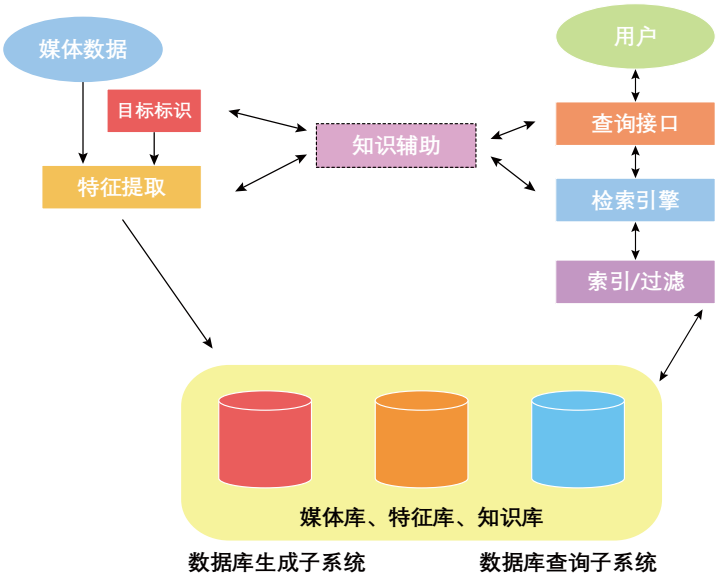


图1.3　CBR的系统结构

主要研究内容包括：

1. 音频特征的选择与提取

音频数据的特征提取和特征向量的构建，对于索引算法的设计，音频检索的效率、精度起着至关重要的作用。根据信号的物理特性，音频特征包括时域特征、频域特征、时频特征和音频片段特征。而根据感知特性又可分为时间文本特征、节奏和音高特征。从众多的特征中选择合适的特征去描述音频提供检索依据是问题解决关键。从早期英国Southampton大学的QBH系统开始[24]，大多数基于哼唱的检索系统都是利用音调，提取出旋律特征进行匹配的。而音调特征与基音频率相关，对于现在较流行的算法，无论是时域内的自相关函数还是基于滤波器的倒谱分析都存在着自身特点所带来的难以克服的缺陷。自相关算法当面对带有谐波复杂的波形时，第一个峰值可能不在整个波形的周期处出现，而是出现在谐波泛音处，导致自相关函数鲁棒性的降低和计算复杂度的增加。倒谱分析在频谱丰富的语音信号处理上滤波性较好，但面对音频的基音提取

则并不敏感。因此，基频提取算法也是MIR技术中需要迫切解决的问题，如图1.4所示。

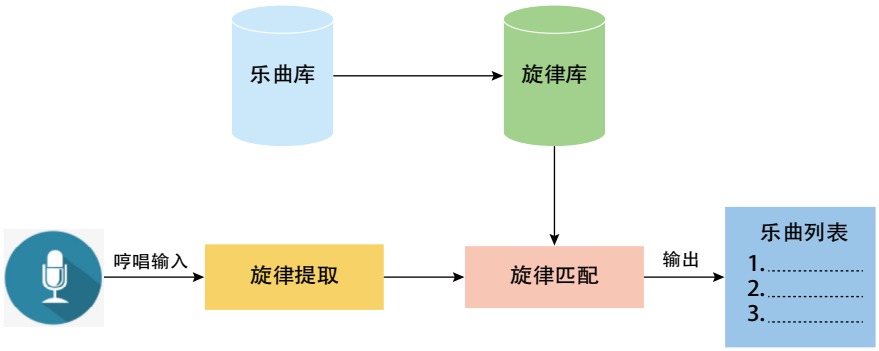


图1.4 QBH系统的工作流程

2. 特征相似度匹配算法

在广义的度量空间中，通过距离函数进行高维向量之间相似度匹配，不但能提高音频识别检索的精准率，而且在多媒体数据库、图像检索等领域都有着十分广泛的应用需求。目前经常采用的解决办法是首先对高维特征向量做降维处理，然后采用包括四叉树、k-d树、R树族等在内的主流多维索引结构，在进行相似度查询时，事先提取能够表达原始数据内容的特征向量，这些特征向量往往维数很高，而特征向量维数越高，系统的查询效率就会越低，这就是所谓的“维数危机”。通常情况下，是对降维后的向量空间建立一个索引结构以加快系统的检索速度。目前，大多数系统都使用近似字符串的匹配算法比较旋律，两个声音文件之间的距离被定义成它们的归一化描述的欧式距离。DTW算法[25]也是一种常用的方法，着重于时间规整和间距测量的概念，但对数据的可靠性没有进行有效的分析，且对连续音符的识别效果不明显。而传统的DP算法对于大型MIR系统来说速度太慢。

3. 用户接口

简便易于操作是系统中人机交互的最终目标。在MIR研究中研究者不断地

改进人机交互方式，提出了哼唱检索、音符输入以及实例检索的概念。但是这些系统的人机接口仍然存在不方便、不自然的缺点。例如，MELDEX系统[26]由于无法正确地切割音符，因此使用者必须自行留下间断或多加入“滴答”声；而SoudCompass系统[27]在使用时必须配合节拍器哼唱。所以研究者需要在用户接口方面努力，以期提供方便、快捷、直观的检索方式。现在用户接口方面需要解决的关键问题是，如何在不同的输入方式下使得系统最终获得统一的数据格式来建立索引，以及如何在多人检索的时候实现个性化定制，如图1.5所示。

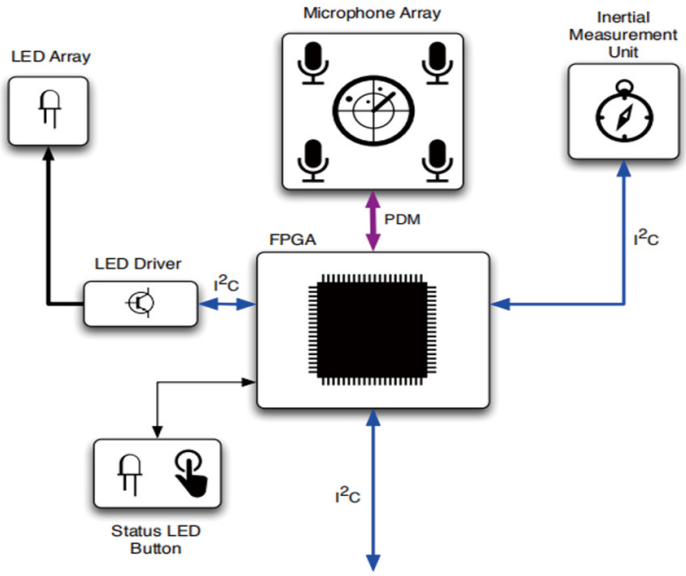


图1.5　SoundCompass的基本硬件模块

4. 实用化研究

科技的最终目的是为广大人民群众服务，而技术的实用化和产业化是实现这一目的的根本途径。对MIR的研究迄今为止还没有实现广泛的产品化，除了人机交互界面和特征匹配算法等技术层面的问题之外，还存在有实用化的领域和实现的载体等问题，这些问题的处理将直接影响此技术的应用范围。

基于内容音乐检索技术也是在网络条件下处理多媒体海量数据的一个重要技术，与图片检索、视频检索并列成为当今基于内容检索研究的热点。将音频的识别和检索技术与传统文本检索相结合可以大大提高数据检索的效率和准确率，降低检索成本。

1.4 国内外研究进展

从用户使用角度来说，音乐信息检索研究领域包含着一系列的实际应用，下文总结了音乐信息检索关键领域的国内外研究进展。

1.4.1 音乐检索

音乐检索应用帮助用户按照一定的相似度衡量规则，从大量音乐数据集中找到想要的音乐。Casey [12]和Grosche [28]提出根据两个性质对音乐检索的应用场景进行了分类。第一个性质是特异性，高级别的特异性用来标识音乐信号，而低级别的特异性用来统计计算音乐作品的相似度。第二个性质是特征粒度，也就是时间范围，大的粒度用来检索整首音乐作品，而小的粒度用来在音乐作品中定位时间点或者检索音乐片段。根据这两个性质对一些常见音乐检索任务的分类如下：

1. 音频标识符或音频指纹

这是一个高特异性低粒度的检索任务。该任务的目标是在数据库中检索或识别一个已知的音乐片段，需要有较强的鲁棒性。最著名的应用就是Wang等人提出的方法[29]，该方法已经集成到一些商业系统中，如Shazam[30]、Vericast[31]和GracenoteMusicID[32]。音频指纹技术在音乐识别和音乐作者之间版权费分配时都十分有效，如图1.6所示。

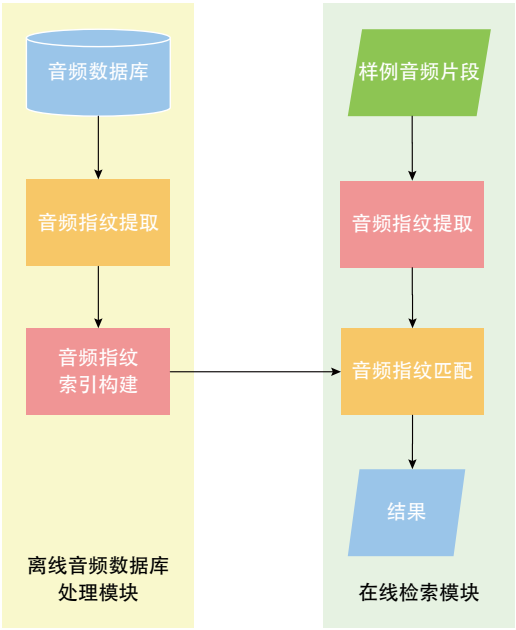


图1.6　基于音频的样例检索系统图

2. 音乐对齐、匹配和同步

这是一个音乐检索的应用场景。除了要识别出给定的音乐片段，该任务的主要目的是将两个音乐信号匹配到局部时间相关位置上，对音乐特征的鲁棒性的要求更高，需要对同一音乐片段的不同演奏进行匹配。如Dixon和Widmer [33]提出的MATCH系统和Müller [34]等人提出的系统，该系统从音频信号中提取的特征序列基础上使用动态时间扭曲算法，能够匹配不同版本的经典音乐作品。

3. 翻唱歌曲识别

这是一个低特异性检索任务。它的目标是检索同一首歌的不同版本，不同的音乐版本可能在乐器、和弦或者结构上皆有异同。Serrà等人在文献[35]中提到的版本识别系统能够描述音乐信号中的旋律或和弦，然后使用局部或全局对齐算法匹配这些描述。The Covers Project等网站系统采用上述方法识别歌曲的影响力和版本引用，如图1.7所示。

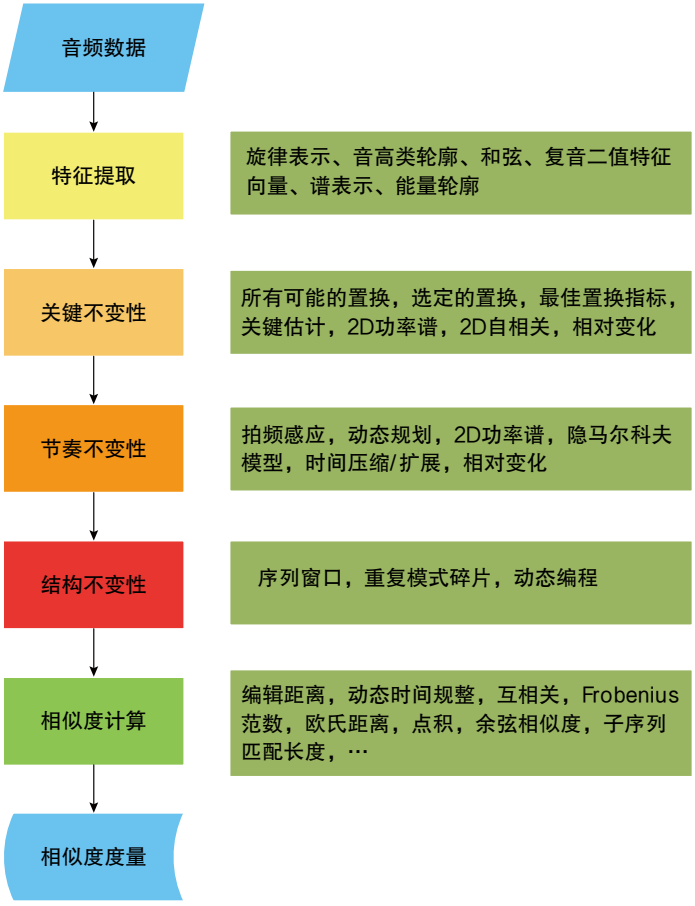


图1.7 翻唱歌曲识别系统的通用框图

4. 基于哼唱或轻拍的检索系统

此类系统的目标是使用给定的旋律或旋律作为系统输入，然后从中提取特征与音乐数据库中的文档作比较。第一个系统是Birmingham提出的MUSART[36]，如图1.8所示。

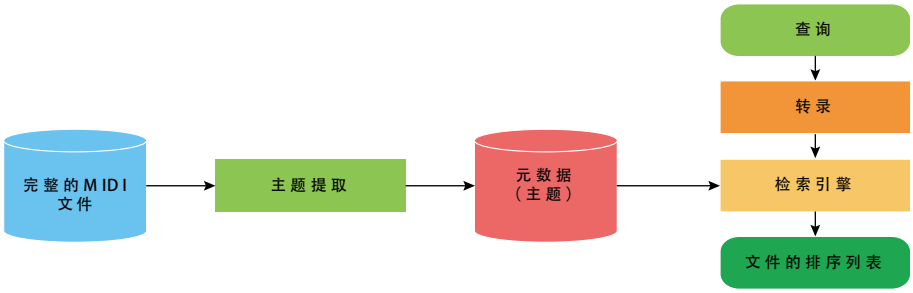


图1.8　MUSART 架构

这个任务的数据集通常是使用乐谱建立起来的，用户哼唱或轻拍的检索片段是音乐信号[37]。商业系统也是通过歌唱、哼唱或轻拍的方式来检索，其中一个典型的系统就是SoundHound[38]，它则是将用户哼唱的查询片段与之前数据库中人们哼唱的歌曲做比较，查询其中特征最相似的音乐，如图1.9所示。

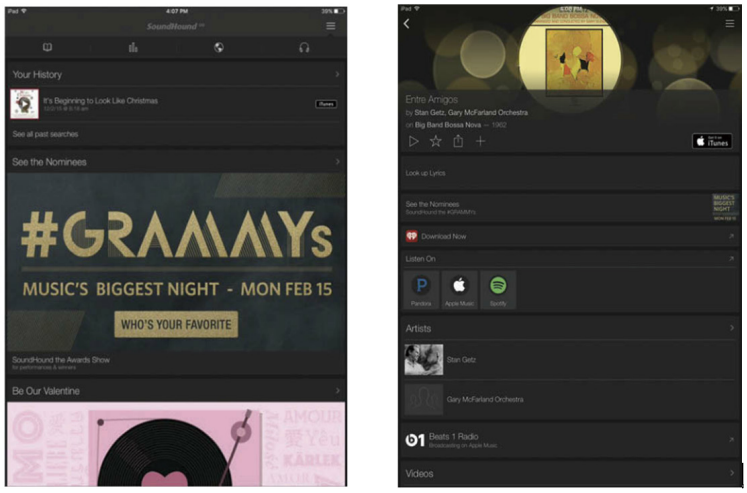


图1.9　SoundHound主界面

上述提到的应用都是将目标音乐信号（查询示例）与数据库中的音乐进行匹配，而另外一些应用，如Isaacson [39]提到的，可能按照一定的描述规则，比如“检索C调节奏为75bpm”的音乐，去完成音乐的检索。除此之外，人们还经常使用标签或语义描述（如“快乐”或“摇滚”）去查找音乐。基于语

义、标签或类别的检索系统，如Knees [40]、Turnbull [41]提出的一些方法就是从内容中估计音乐的语义标签，从而实现检索的，它们属于低特异性和高粒度的检索。

Celma等[42-43]提出的音乐搜索引擎SearchSounds，就是一个用户为中心的系统，它从用户音乐日志中找到反映音乐语义的文本，比如“温柔的吉他曲”，对音频特征进行查询扩展。Knees[44]等人提出的Gedoodle系统也是收集了网站上可编辑的元数据信息，如艺术家、专辑名等标签，扩充到相应的音频特征中去。所有补充信息作为语义融合到音乐片段中，综合得出检索结果。

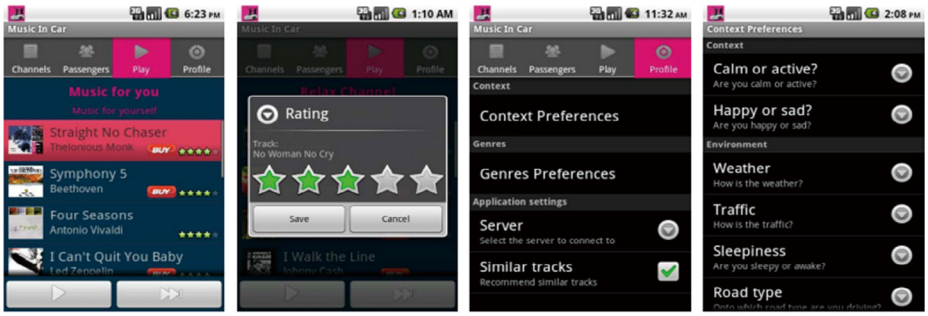
1.4.2 音乐推荐

如果说检索已知音乐是音乐信息检索的主线，那么探索和发现未知新音乐就是音乐检索的另一个方向，或者说是不可或缺的辅助。音乐推荐系统是对用户偏好进行建模，然后返回典型音乐作品列表的过程。Ricci [45]和 Celma [46]都曾论述过推荐系统需要实现的核心问题：

- 第一，准确度。即所推荐的内容要符合用户的音乐嗜好。
- 第二，多样性。与相似性相反，当推荐系统显示出一定的多样性时，用户对这样的结果会更加满意。
- 第三，透明性。当用户理解了系统是根据什么理由推荐某个音乐时，用户就会更加信任推荐系统的推荐结果。
- 第四，新颖性。用来衡量推荐结果会给用户带来多大的“惊喜”。最著名商业系统Last.fm就是基于协同推荐的，而Pandora则擅长对音乐作品进行标注。

近年来文献中提出的系统多集中在用户感知、个性化及多模式推荐上。例如：

- Baltrunaset等人[47]提出的InCarMusic音乐推荐系统是应用在汽车上（如图1.10所示）的。
- Schedl 等人[48-49]提出的位置感知相关音乐推荐技术是基于微博内容的。
- Forsblum等人[50]提出了基于位置相关的推荐系统会发现音乐节上的偶然事件。
- Wang等人[51]提出了一个结合音乐内容和用户上下文特征的统计模型来满足用户短时间的收听需要。
- Teng等人[52]设计特征感知器，收集了移动设备上的音乐收听事件，用以改进移动终端上的音乐推荐系统。



（a）拟播放的曲目　（b）给曲目评分　（c）编辑用户配置文件　（d）配置推荐系统

图1.10　InCarMusic用户界面

1.4.3　音乐播放列表生成

音乐播放列表的自动生成，又被称为“自动DJ”，可当作音乐推荐的高端应用。其目的是产生一个有序的播放列表，可能针对某类相似的作品或者某位艺人，从而为收听者提供一份有特殊意义且极具欣赏性的播放列表。这就是该任务与通常意义下的音乐推荐系统之间的主要区别，在普通的音乐推荐系统中并不在乎播放的次序。另一个区别在于音乐推荐系统更在乎如何发现新的音乐，而自动生成播放列表更倾向识别已知的材料。

Pohle等人[53]研究表明，人们评价一个自动生成的播放列表质量时，连续播放的音乐之间的相似度是非常重要的指标。如果连续播放的音乐有太高的相似度，听众就会感到十分无聊。Schedl等人[48]进一步发现了除相似度外的其他指标，包括音乐作品或者艺人的熟悉度或者说流行度、热度，是不是最新发布的，还有新颖性。这些因素都会为用户的收听增加趣味性，用户会期待这种“惊喜”，因为他被带入了一场有趣的，对未知音乐作品和艺术家的发现之旅。Zhang[54]也论述了此类模型的更多细节。

Intelligent iPod[55]就是一个典型的基于内容音乐播放列表生成系统，移动设备音乐数据集中作品的音频特征和它们之间的相似度都被提取出来，然后生成播放列表且用颜色条可视化地表现不同的音乐风格，用户可以通过滑动条轻松调节播放器，以切换自己想听的音乐。其他的商用播放列表生成系统还包括YAMAHA BODiBEAT，它集成了一套体感装置来追踪人锻炼的过程，从而生成符合个人跑步频率的音乐，如图1.11所示。

（a）Intelligent iPod　　（b）YAMAHA BODiBEAT

图1.11　基于播放列表生成系统的产品

1.4.4　音乐浏览界面

现在，音乐消费者可通过音乐流媒体网站获得数以万计的音乐。如何设计智能用户界面，使得用户能够邂逅意想不到的收听体验，已经变得越来越重

要。这些界面要支持直观的音乐数据库浏览功能，同时又可以精确地搜索到具体的曲目。以下是这类用户界面的典型示例。

1. nepTune

Knees等人[56]提出的nepTune是一个新颖的音乐数据库用户接口工。给定任意一组数字音乐文件，nepTune会创建一个虚拟的地形图，用户可在这张地形图上任意导航数据集中的音乐。系统原理在于自动提取音频中的特征，按照特征对音乐作品进行聚类。聚类的结果被用来创建三维立体“音乐岛”地形图，用户可以通过环绕立体声在地形图上自由徜徉并收听到他想要的音乐。另外，该系统还从网络上获取的文字资源中提取了知识，为地形图增加了语义信息。然后，用这些文字来描述听到的音乐，相关的图片也会出现在地形图上支持新歌曲的发现。用户能够像玩虚拟游戏一样浏览音乐，界面如图1.12所示，图中聚类后的音乐按照流行度可视化为山脉的形式，而那些流行度不高的部分被显示为海滩或者海洋。

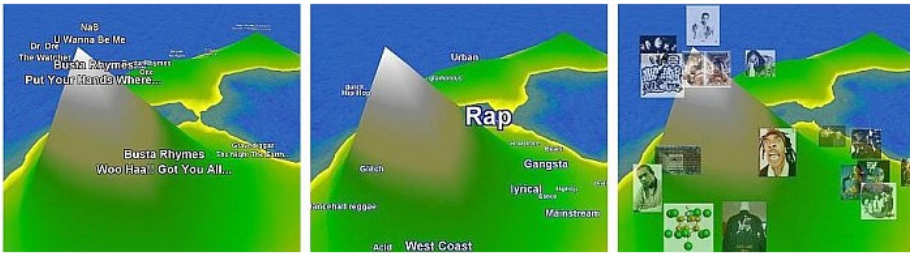


图1.12　nepTune智能音乐导航界面

2. 三维音乐数据集浏览界面

另一个类似的三维音乐数据集浏览界面是由Lübbers和Jarke [57]提出的。与nepTune音乐岛的比喻不同，该系统将音乐作品聚类后可视化为山谷，而那些数据集中表现特殊的稀疏作品表现为山峰，同时该系统还支持用户对地形图的编辑和变形，如图1.13所示。

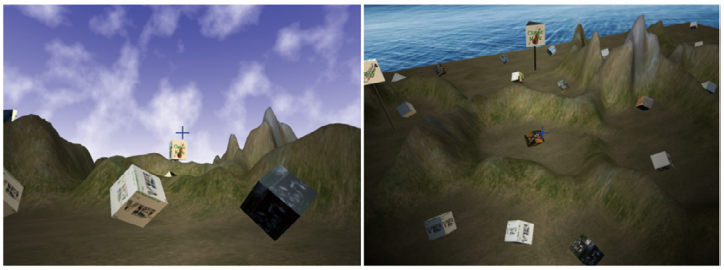


图1.13　Lübbers和Jarke提出的三维音乐数据集浏览界面

3. Musicream

Goto [58]提出的Musicream是另一发现未知音乐，音乐偶然性捕捉的例子。该系统使用水龙头的比喻，界面中有一组彩色的水龙头，每一个水龙头代表音乐的不同风格，当用户打开这个虚拟控制条时，相应的水龙头就会创建一个歌曲流。用户可以截取并播放音乐，或者把它们取回并创建一个播放列表。当用这种方式创建列表时，不相似的歌曲很容易区分出去，而相似的歌曲也易于结合在列表中，如图1.14所示。

（a）音乐光盘流媒体功能　（b）基于相似度粘贴功能

（c）元播放列表功能　（d）时间机器的功能

图1.14　Musicream的四个功能

4. Songrium

Songrium是一个网页上的应用，它能够丰富音乐收听过程。该应用由日本产业技术综合研究所（AIST）开发。如Hamasaki和Goto在文献[59]中所示，Songrium提供了浏览歌曲的一种不同方式，基于音频信号相似度创建了一幅音乐“星空图”，一首歌曲和它的衍生作品像太阳系的结构似的展示出来，而通过歌曲之间的交集发现新的音乐，音乐星空图的界面如图1.15所示。

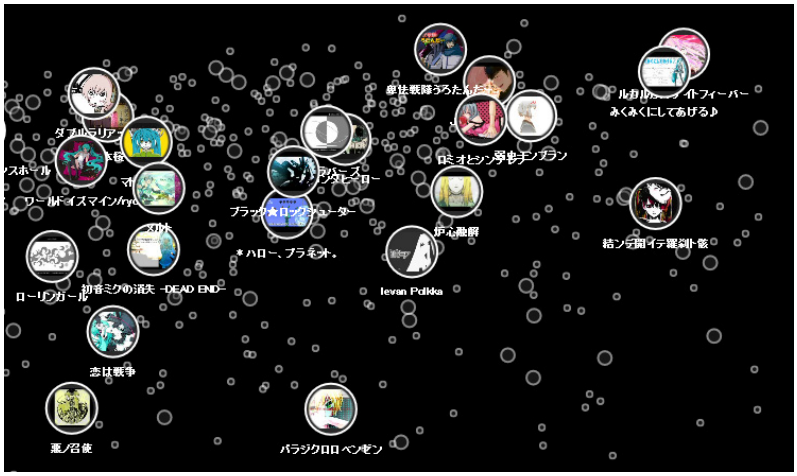


图1.15　Songrium音乐导航应用

1.4.5　其他检索应用

除了基本的检索场景，音乐检索技术还包含了其他一些检索应用。在可计算的应用理论中有一种应用就是音乐内容描述技术，使用大规模数据库做比较研究或者建立专家系统。另外，一些音乐创作应用也能够从音乐检索技术中获益，例如通过“音频拼接”方式，目标音乐被分析后，从小片段中提取音频描述文件，然后这些片段由大的音乐数据库中得到的新的相似片段替换。这些应用在MIReS项目[60]“音乐信息研究路标”中详细论述。

Downie[61-62]、Lee[63-64]和Bainbridge[65-66]等人在文献中都曾总结了音乐检索中丰

富的研究领域。音乐信息检索的研究者已在这些领域做了大量的工作，这些研究为商业应用提供了基础。这些典型的研究方向及研究任务如表1.1所示。研究的起点则是从音乐内容和相关上下文中提取出有意义的特征，然后特征被用来计算两个音乐作品之间的相似度，或者按照情感、乐器或风格等不同的标准进行分类。

表1.1 音乐检索相关研究方向及研究任务

研究方向	研究任务
特征提取	音色描述 [67–68]
	音乐改变及旋律提取 [69–70]
	端点检测 [71]、节拍追踪 [72] 和节奏估计 [73]
	音色 [74]、半音 [75–76]、音符估计 [77]
	曲式分析 [78]、音符分割 [79] 和音乐摘要 [80]
相似度	相似度度量 [81–82]
	翻唱歌曲识别 [83–84]
	哼唱检索 [85–86]
分类	情感识别 [87–88]
	风格分类 [89–90]
	乐器分类 [91]
	作曲家、艺术家和演唱者识别 [92]
	自动化标签标记 [93–94]
其他应用	音频指纹 [95–96]
	基于内容检索 [97]
	音乐推荐 [98–99]
	播放列表生成 [100–101]
	音频乐谱对齐及音乐同步 [102–103]
	歌曲 / 艺术家流行度估计 [104]
	音乐可视化 [105]
	音乐浏览互动 [106]
	用户交互互动 [107]
	个性化、上下文感知和适应系统 [108]

1.5　研究思路

1.5.1　框架

音乐检索技术离不开对音乐内容的描述、音乐内在特征的提取，以及音乐所表达情感等语义的理解。

首先，从本质内容出发，研究了音乐信号波形中固有的、客观的特征，如音乐的音高、节拍、音色等，提取了音乐波形低层的特征，对这些特征进行存储、比较和变换，进一步得到音乐旋律等较高层次的内容特征，用它表达音乐内容，描述检索条件，得出检索结果。

然后，对整首音乐的低层特征进行了提取和表达，提取出一首音乐中最关键的部分特征，挖掘出了不同音乐之间潜在语义上的相似性，获得了在音乐风格等部分语义上相似的歌曲。

最后，进一步探索了音乐所表达的语义，将音乐的语义作为特征向量，研究了音乐内容与语义之间的相互关系，从而得出在语义上相似的歌曲，以达到从人的主观感受出发，检索人们“想”要听的音乐。

本书研究框架如图1.16所示。

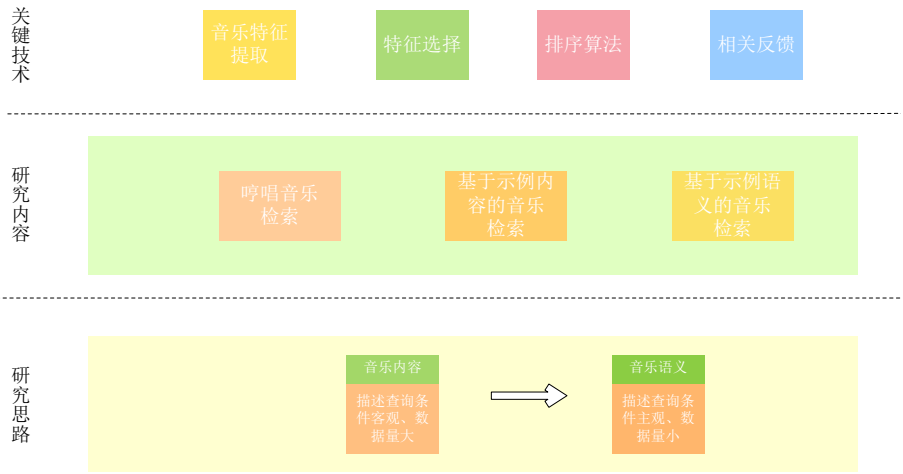


图1.16　框架

1.5.2 研发思路

基于内容的检索和基于语义的音乐检索是不同的两种检索应用。基于内容的可以以哼唱检索、音乐片段检索等为检索条件。而基于语义的检索，是从人们使用音乐的场景出发，获取用户主观感受，分析用户使用上下文，更符合用户的检索意图。使用语义去检索音乐，便于新音乐的发现和推荐。利用当前用户的播放和检索内容，得到用户收听习惯和乐于接受的检索条件，为用户提供那些他们原本未知的陌生音乐，更利于音乐产业的推广和发展。

音乐的内容是客观的，音频信号是短时信号量，因此对音乐内容的分析要经过分帧、加窗和预处理，得到的内容描述特征数据量是非常庞大的。例如一首三分钟的歌曲的wave文件大约为十几MB，从中提出的特征文件也约有3MB。如果将音乐内容作为查询条件，那么查询条件的分析、处理和检索时间较长。而音乐的语义是客观的，如果能够将内容映射到相应的语义上，那么语义描述词数量有限，常用到的只有几十个文本词，若建立语义检索向量，则向量中字符个数也就几十个。这种检索方式中，查询条件的分析、处理和检索过程耗费的时间较之音乐内容要少得多。若离开音乐物理信号载体，语义检索则毫无客观性，语义的感受因人而异，无法客观评价语义检索的结果。

基于内容和语义的音乐检索研究是相辅相成、密切相关的，基于内容的检索、基于语义的检索是更符合用户需求的检索形式，是检索的更高要求和目标。因此，本书的思路是以音乐内容为起点，以客观的信号为基础，研究从内容到语义之间的联系，逐步从内容的检索，过渡到更符合用户主观感受的语义检索，更加高效、精确、丰富地提供检索结果。