一种基于内容及相似搜索的对等音乐文件共享系统

雷军环[1]，张光会[2,3]

(1. 长沙民政职业技术学院 软件学院，长沙 410004；2. 平顶山学院 计算机科学与技术学院，河南 平顶山 467000；3. 湖南大学 信息科学与工程学院，长沙 410082)

摘 要：提出了一个基于内容及相似搜索的对等音乐文件共享系统。该系统建立在结构化的应用层覆盖网络之上，保证了系统的可扩展性和避免了网络消息的洪泛；利用了集合对等点来完成音乐文件的元数据的注册和搜索；把音乐文件的属性名—属性值对(attribute-value pairs，AV-Pairs)通过 MFD(music file description)来表示，使系统既可以支持精确的检索，也可以完成复杂的语义相似性搜索，特别是基于内容的搜索和下载；描述了对等实体上的软件功能模块，构建了一个实际的对等音乐文件共享系统。测试结果表明，随着大量搜索请求的到达，系统保持了比较高的吞吐量，同时具有比较高的成功搜索率。

关键词：对等网络；文件共享；内容搜索；相似搜索；分布式哈希表

中图分类号：TP391 **文献标志码**：A **文章编号**：1001-3695(2012)04-1509-03

doi:10.3969/j.issn.1001-3695.2012.04.085

P2P music file sharing system supporting content and similarly searching

LEI Jun-huan[1], ZHANG Guang-hui[2,3]

(1. *School of Software, Changsha Social Work College, Changsha* 410004, *China*; 2. *School of Computer Science & Technology, Pingdingshan University, Pingdingshan Henan* 467000, *China*; 3. *School of Information Science & Engineering, Hunan University, Changsha* 410082, *China*)

Abstract: This paper presented and described a P2P audio file sharing system which provided the content and similarly searching. It used a distributed hash table(DHT) based overlay network to ensure the scalability and avoid network-wide message flooding. By using rendezvous points, achieved the music metadata registration and query. It described each file in the system by a music file description(MFD) which essentially was a set of attribute-value (AV) pairs and thus flexible searched semantics based on attribute-value pairs and automatic extraction of musical features and content-based similarity retrieval could also be accomplished. Also discussed the software architecture on a peer node subsequently. The analysis results show that the approach can maintain a high throughput and high query success rate.

Key words: P2P networks; file sharing system; content search; similarly search; distributed hash table

在音乐文件共享平台中，集中式的非结构化对等网络 Napster 由于不能抵御恶意节点的攻击而缺乏强大的鲁棒性，分布式的非结构化对等网络 Gnutella 或 KaZaA 也由于存在消息洪泛问题使系统的扩展性较差。采用结构化的分布式哈希表解决了上述问题，但是分布式哈希表往往是基于关键字的搜索与下载，缺乏更加复杂的内容分析和相似语义的分析[1,2]。所以本文提出了一种支持内容与相似语义搜索的对等音乐文件共享系统。

1 系统的体系结构

与最近的对等网络类似，本系统把内容定位和文件下载过程结合在一起进行。网络中的每个节点不仅存储需要共享的音乐文件，而且有网络中的其他音乐文件的位置信息。所以当用户在对等网络提交搜索请求时，系统会返回一组满足查询标准且可以下载的音乐文件。系统中的每个文件通过 MFD(music file description)来描述，它是一组属性名—属性值对(attribute-value pairs, AV-Pairs)。例如，演唱者 = U2，图片 = Hum，时间长短 = 3.5 min 等。这些属性可以由用户人工规定，也可以通过音乐内容来自动提取。图 1 显示了每个节点上的功能模块，音乐属性提取引擎 MFEE(music file extraction engine)是一个用来计算这些属性的组件，这些被提取出来的属性对都被用来完成基于内容的相似搜索。利用 MFD 作为参数，系统可以支持两种操作：注册和搜索。在搜索过程中，用户的查询请求被转换成合适的 MFD，这样就可以定位到那些匹配该搜索标准的节点，只要定位过程完成了，用户就可以开始下载文件。本系统关注的是高效的内容发现机制和支持多种类型的音频文件属性提取，而不是具体的下载功能本身。

在进行注册和搜索的过程中，MFD 是作为内容发现机制组件 CDS(content discovery system)的输入。CDS 模块是建立在分布式哈希表的覆盖网层面上的，如 Chord 或 Tapestry。在分布式哈希表中，每个节点负责一个区域，它通过节点 ID 表示，是一个连续的位虚拟地址空间。文件名等数据项通过属性值来与这个地址空间关联，通过对数据项进行统一的哈希函数后，那么覆盖了这些属性值的区域就存储在这些节点之上。本文使用了 CDS 内容搜索的算法，它可以把 MFDs 分发到各个节点中。分布式哈希表本身的特性可以保证路由和消息高效地

收稿日期：2011-08-24；**修回日期**：2011-10-25

作者简介：雷军环(1967-)，女，湖南永兴人，副教授，硕士，主要研究方向为计算机网络、并行分布式应用(leijunhuan@126.com)；张光会(1972-) 男，河南宝丰人，讲师，硕士，主要研究方向为数字水印、计算机网络.

传递。每个节点维持一个本地 MFD 的数据库,这些数据是通过 CDS 来分配的。当一个查找请求到达时,节点将检查它的本地 MFD 数据库,返回可以匹配查找标准的 MFD 集合,然后查询初始化模块就从拥有目标音乐文件的节点处开始下载实际的音乐文件。

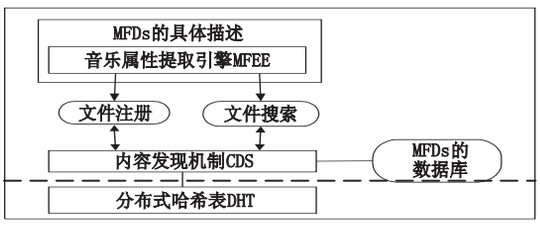

图1 对等实体上的软件功能模块

2 文件属性的提取

通过 PCM(pulse code modulated)或压缩格式(MP3-MPEG)的声音文件作为 MFEE 组件的输入,MFEE 将其输出为一个属性向量,这个属性向量用于基于内容的检索,它把音乐文件的 AV-Pairs 表示出特定的性质。本文参考了文献[3]中提出的音乐文件的类型分类方法。各种不同类型的文件信息通过属性向量清晰地表示出来,在实现内容搜索时,系统可以支持各种不同的查询要求。例如,用户可以按照播放时间长短查询,或者通过音乐文件的大小来搜索,也可以把这些条件结合起来搜索。

使用标准的线性量子化和常规化方法把那些连续属性的动态范围转换成分离的属性值,这样就可以完成基于属性名—属性值对的搜索。线性量子化方法的运用使得那些属性分布的统计不会改变。在本文中,每个属性被量子化为若干个分离的值。实验中发现,自动的原始连续属性的分类和量子化方法的分类在测试结果上没有太大的区别,并不影响系统性能。通过文献[3]中提出的属性集和数据集的方法,利用非量子化方法和量子化方法取得了精确的结果。运用 MFEE 组件的功能,配合人工的有注释的元数据,这样就可以形成 MFD 的文件描述,这就是属性名—属性值对的描述,它可以支持手动的有注释的属性与基于内容的属性。

为了定义一个查询请求,MFD 会形成文件搜索的标准,特殊情况下,MFEE 模块也用来产生一个查找的 MFD。任何一个 MFD 的子集都可以用来作为查找的具体要求,带有 AV-Pairs 属性对的 MFD 既可以支持传统的基于关键字的查找,也可以支持标准更加复杂的查找。例如下面这些复杂的查找例子:Search for {artist = U2}; Search for {artist = U2, year = 1985, tempo = 80, beats-per-minutes = 100}; Search for {10 most similar to x. mp3(基于内容的搜索)} Search for {10 most similar to x. mp3, artist = U2}。为了完成上面的两个基于内容的相似搜索,查找目标中有 x. mp3 的 10 个最相似的文件,它首先通过 MFEE 模块转换成数字化的属性,描述了音乐的内容,用来完成基于内容的相似搜索。

3 可扩展的内容提取

3.1 音乐文件描述符的注册

与集中式的对等网络不同,本文的 CDS 内容检索机制采用了一个可扩展的集合点 RP(rendezvous points)机制来完成注册和查找服务。这种对等网络是结构化的,可用来高效地表示属性名—属性值对以完成查找与检索。为了注册一个 MFD,CDS 针对每个 AV-Pairs 在其节点 ID 上使用了一个统一的哈希函数,如 SHA-1。MFD 把哈希值发送给对等网络中的这些节点,然后这些节点就作为 MFD 的集合点。在接收到一个 MFD 后,节点将把 MFD 插入到其相应的数据库,然后每个节点为 AV-Pairs 完成查找,可以映射到目标节点。MFD 是一个二维的描述,水平方向表示其第一个属性,垂直方向表示其第二个属性,它也可以是多维的。

实际的系统能够处理多维的数据。基本上二维的节点就包括一个行和列的属性名—属性值 AV-Pairs 的表格。很明显,这个结果在应用层覆盖网络是强鲁棒性的,因为它可以在每个节点的本地数据库之间传递。由于一个 MFD 中的 AV-Pairs 数目是较小的,一个 MFD 集合点中的尺寸也是较小的,所以注册过程是很高效的。不同的 MFD 具有不同的集合点组,这样就可以很自然地分离系统的注册负载,而且 MFD 的 AV-Pairs 注册后的子集也可以被继续搜索,这样就可以完成更多语义相似的检索。

3.2 文件的搜索过程

文件的搜索过程分为两种情形,即精确搜索和相似搜索。在精确搜索中,用户主要是搜索匹配所有的 AV-Pairs 属性对的 MFDs,每个在 MFD 中、但不在查询中的确定的 AV-Pairs 将被忽视。假设搜索是 Q:$\{a_1 = U_1, a_2 = U_2, \cdots, a_m = U_m\}$,因为匹配了上面的搜索 Q 的 MFDs 都已经在集合点中进行了注册,那么 CDS 将发送一个单独的查找消息到能够解析该搜索的节点中。为了完成高效的搜索,CDS 将在数据库中选择一个具有最小 MFD 的节点。只要接收到搜索,节点将进行数据库中所有的实体完成属性对的比较,发现最合适的 MFDs。例如 beartist = U_2, year = 1985, tempo = 100 bpm, 那么意味着匹配的 MFDs 必须同时满足上面的三个条件。最可能包括了 U_2 的所有的节点将具有最小的本地 MFD 数据库中的值,在满足了这个条件后,它将可以联系到后面的同时满足 year 和 tempo 的节点。

在相似搜索中,用户将试图发现那些具有类似属性向量的音乐文件,如用户想查找 10 个与 x. mp3 类似的文件,向量为 $\{f_1 = v_1, f_2 = v_2, \cdots, f_m = v_m\}$。它的搜索过程与上面类似,CDS 可能选择一个 AV-Pairs 属性对,发送这个搜索请求给节点,节点将计算在数据库中的查找向量和每个 MFD 的距离,距离的定义可以是 $d(f,f') = |v_1 - v_1'| + |v_2 - v_2'| + \cdots + |v_m - v_m'|$ 的形式。距离的计算可能还有其他的定义方式,如在某些文献上用到了余弦距离。最后将返回 10 个距离最短的 MFD 给用户。还有其他的更加复杂的相似搜索,这些都需要建立更加复杂的数据结构。

由于在对等网络中使用到了集合点,网络消息的洪泛现象可以避免,实际上一个搜索过程中可能有很多满足查找条件的 AV-Pairs 的 MFDs,所以某些节点会出现负载过大的现象。对于 CDS 可以通过文献[4]中的方法完成分布式的动态负载均衡。

4 系统测试与性能分析

4.1 测试环境的构建

本节对系统进行了测试与性能分析。MFEE 组件参考了文献[5]中的方法,它是一个针对音频文件分析的开放的软件框架。通过事件驱动的模拟器来进行模拟[4]。在实验中建立了一个具有 10 000 个节点的对等网络,每个节点相互连接的

网卡为1 Mbps。大约有30个基于内容的音乐文件的属性被收集起来,这些属性都是从5 000多个音乐文件中人工提取出来的。图2显示了在这些文件中的AV-Pairs的分布情况,一共有2 178个不同的AV-Pairs,它们的分布状态是有一定规律的,大部分普通的AV-Pairs出现在53%的MFDs中,41个AV-Pairs出现在一个MFD中。

4.2　结果与讨论

对于注册了的搜索请求,通过把5 000个文件复制了20次,系统产生了100 000个文件的MFDs,并把它们随机地分配到对等网络的节点上。每个节点注册这些文件并且把它们分配到系统中,普通的AV-Pairs注册导致了多个分区。对于搜索请求的产生,形成了100 000个搜索按照图2中的分布方式分配给AV-Pairs。每个搜索对应了一个特殊的音乐文件的属性。这样设置是为了模仿用户的行为,让其可以通过提交音乐复制来查找到类似的音乐。有超过10%的最常见的MFD处于查找之中,大部分的MFDs只是处于一部分查找之中。查找初始化器随机地分发到所有的节点之中。为了简化,只有精确的匹配被返回。当节点的连接百分比达到了100%后,一个节点可以拒绝一个查找并且返回失效的结果。测试了两个场景,在查找请求按照正常的泊松分布达到的情况下,图3比较了查询成功的比率。

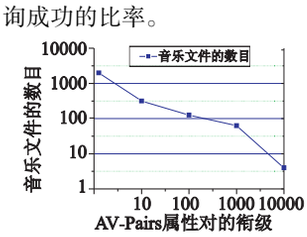

图2　音乐文件的数目随着AV-Pairs变化情况

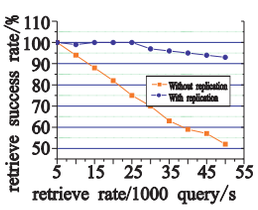

图3　系统的测试结果

在第一个场景中,当达到了连接百分比后,一个节点将拒绝一个新到达的查找,也不备份它的内容到其他节点上。由于每个查找的CDS具有30个候选的AV-Pairs,所以查找的负载在没有备份的情况下呈线性变化。结果表明,系统在高负载情况下可以获得一个高的成功率。在每秒10 000个查找情况下,成功率是90%以上。尽管这样,随着查找数目的增加,节点相应的常见的查找将处于饱和状态,查找的成功率将低于90%。在第二个场景中,通过使用动态的副本备份机制,那些具有高负载的节点将备份它们的数据库到其他节点上,这样就可以驱散那些集中的查找负载。通过结果分析,如果带有副本备份,系统将维持一个比较高的查找成功率,可以达到95%。两个场景下,系统都维持了一个高的吞吐量。实验结果表明,该方法是一种灵活、高效的音乐文件内容相似检索系统。

5　与相关工作比较

关于对等网络的音乐格式文件的搜索系统,目前国内外有很多相关的研究。本文的研究主要关注相似搜索与基于内容的检索,它支持各种类型的音频文件的分析[6,7]。文献[8~10]中也提出了更加复杂的音乐文件的搜索方法,包括基于本文分析和目录分类的复杂搜索,这些研究都关注属性的提取和类型的支持。

在体系结构上,Chord等人提出了基于分布式哈希表的对等网络音乐文件搜索,解决了一些可扩展性问题,但是缺乏内容搜索与相似搜索机制。本文中的CDS组件模块也建立在分布式哈希表层上,但其包括了精确搜索和相似检索等各种复杂的查找。把音乐文件的检索建立在分布式哈希表层之上的思想由文献[11]提出,但它也存在一些可扩展性问题。文献[12]通过改进,增加了基于语义网的音乐文件检索,还有其他的利用混合式非结构化的JXTA来完成基于内容的搜索。这些项目在使用对等网络时,都把焦点集中在集合点和提取属性完成内容及相似检索上,而不是网络体系结构本身[4]。本文是通过提出集合点的方式完成属性的提取和相似搜索,构造了一个真实的对等网络音乐文件搜索系统,取得了有价值的实验结果。

6　结束语

本文提出与分析了一个支持内容及相似搜索的音乐文件搜索系统,通过属性名—属性值对AV-Pairs来描述音乐文件的MFDs,支持属性的自动提取和复杂类型的音频文件搜索;通过集合点来完成音乐文件的注册和检索过程,能够保证系统的可扩展性和避免网络的消息洪泛;通过构建实际的对等网络文件共享系统对本文的方法进行了测试,测试结果表明,随着不同数目的搜索请求到达后,系统具有高吞吐量、高成功搜索率。下一步的工作是进行更加复杂的属性名—属性值对的分析,让系统支持更多音频类型的音乐文件检索。

参考文献:

[1] ZHAO B Y, HUANG Ling, STRIBLING J, *et al*. Tapestry: a resilient global-scale overlay for service deployment[J]. IEEE Journal on Selected Areas in Communications, 2004, 22(1): 41-53.

[2] STOICA I, MORRIS R, KARGER D, *et al*. Chord: a scalable peer-to-peer lookup service for Internet applications[C]//Proc of SIGCOMM. 2001: 149-160.

[3] TZANETAKIS G, COOK P. Musical genre classification of audio signals[J]. IEEE Trans on Speech and Audio Processing, 2002, 10(5): 293-302.

[4] GAO J, TZANETAKIS G, STEENKISTE P. Content based retrieval of music in scalable peer-to-peer networks[C]//Proc of International Conference on Multimedia and Expo. 2003.

[5] TZANETAKIS G, COOK P. MARSYAS: a framework for audio analysis[J]. Organized Sound, 2000, 4(3): 169-175.

[6] FOOTE J, MATHEW C. Audio retrieval by rhythmic similarity[C]// Proc of International Conference on Music Information Retrieval. 2002: 265-266.

[7] BAUMANN S. Music similarity analysis in a P2P environment[C]// Proc of the 4th European Workshop on ImageAnalysis for Multimedia Interactive Services. 2003: 314-319.

[8] FUTRELLE J, DOWNIE S J. Interdisciplinary communities and research issues in music information retrieval[C]//Proc of International Conference on Music Information Retrieval. 2002, 215-221.

[9] PACHET F. Content management for electronic music distribution: the real issues[J]. Communications of ACM, 2003, 46(4): 71-75.

[10] WHITMAN B, SMARAGDIS P. Combining musical and cultural features for intelligent style detection[C]//Proc of International Conference on Music Information Retrieval. 2002: 47-52.

[11] WANG C, LI J, SHI S. A kind of content-based music information retrieval method in a peer-to-peer environment[C]//Proc of International Conference on Music Information Retrieval. 2002: 178-186.

[12] BAUMANN S, KLUTER A. Super-convenience for non-musicians: querying MP3 and the semantic Web[C]//Proc of International Conference on Music Information Retrieval. 2002: 297-298.