REVISTA DE COMPUTAÇÃO E SISTEMAS (RCS)
ISSN XXXX-XXXX
VOLUME 1 – NÚMERO 1 – ABRIL/2016
http://www.revistacomputacaoesistemas.net

Anotações Semânticas para Query-by-Humming

Danieli P. de Sousa, Luciano Silva

Faculdade de Computação e Informática – Universidade Presbiteriana Mackenzie
São Paulo – SP – Brasil
danip.sousa@gmail.com, luciano.silva@mackenzie.br

Abstract. *A difficulty reported by many people is to find the name or the artist of a song when there is no mention or reminder of its textual elements; or when these data are uncertain and do not match with the music which information are been searched. The method of Query-by-humming has been studied for this recovery purpose. This research aims to use the CALIPH (software for semantic annotation) and EMIR (a software for data recovery which are semantically referenced by use of the CALIPH) for storage and retrieval, respectively, of images that refer to the data obtained by humming.*

Resumo. *Uma dificuldade relatada por muitas pessoas é a de pesquisar o nome ou intérprete de uma música quando não há nenhuma referência ou lembrança de elementos textuais desta; ou quando estes dados são incertos e não correspondem à música sobre a qual se busca informações. O método de query-by-humming tem sido estudado com este propósito de recuperação. A presente pesquisa tem como objetivo a utilização do CALIPH (software para anotação semântica) e do EMIR (um software para recuperação de dados semanticamente referenciados pelo uso do CALIPH) para o armazenamento e a recuperação, respectivamente, de imagens que referenciem dados obtidos by humming.*

1. Introdução

Muitas vezes ouve-se uma música, mas não se sabe o nome ou o intérprete da mesma. Por essa razão, a implementação de mecanismos de consulta a bases de dados musicais que possibilitem ao usuário encontrar informações sobre determinado arquivo de áudio apenas fundamentando em um trecho ouvido do mesmo que se guardou na memória tornou-se algo tão relevante diante desse problema tanto de usuários leigos como de profissionais ligados à música (Lau *et al*, 2007).

Uma das técnicas que pode ser exploradas para a consulta em banco de dados de áudio que atenda a este requisito é a de *Query-by-Humming (QBH)*. Entretanto, por lidar com linguagem natural, esta técnica encontra dificuldades de implementação, devido a discrepâncias entre o resultado que deveria ser retornado em uma consulta e o resultado que muitas vezes é efetivamente obtido. Novas propostas de implementação de armazenamento e recuperação em sistemas *QBH* são objetivos de pesquisas dentro

da área da Computação denominada Computação Musical.

Com motivação neste cenário, propõe-se analisar a viabilidade do uso da representação semântica como aliada à implementação de *QBH*. Para tanto, serão utilizados dois softwares fundamentados no padrão *MPEG-7*, o *CALIPH* e o *EMIR*. Mais concretamente, a proposta do trabalho é que voluntários simulem trechos de músicas previamente selecionadas e que estas simulações sejam armazenadas no *CALIPH* (o qual pode ser compreendido como construtor de uma base de dados anotados semanticamente) e o *EMIR* (especialmente desenvolvido para uso conjunto com o primeiro), o qual se comporta como a interface de busca do *CALIPH*.

2. Banco de Dados Multimídia para Áudio

2.1. Conceito e Estrutura de um Banco de Dados Multimídia

Segundo Elmasri e Navathe (2007), um Banco de Dados Multimídia é uma estrutura que armazena, gerencia e possibilita a recuperação de dados multimídia (desde textos, gráficos, arquivos de áudio, imagens, ou vídeos, ou ainda a combinação de todos estes dados) no formato digital. O modo como o qual um dado multimídia é armazenado varia de acordo com a classificação do mesmo. Isto significa dizer que, para cada tipo de conteúdo multimídia que se pretende armazenar, deve-se selecionar uma técnica de representação adequada. Tão importante quanto definir a forma de representação de um dado multimídia em um banco de dados multimídia é definir os tipos de consulta que serão realizadas neste banco. Considerando-se a quantidade de dados que uma aplicação multimídia, geralmente, tem de manipular, chega-se à conclusão de que as estruturas de armazenamento destes dados são complexas, o que inclusive é um dos grandes problemas na questão do armazenamento de dados multimídia.

Diante desta problemática, o uso padrões de auxílio à descrição e compressão de dados multimídia, como o MP3 (*MPEG 1 Layer-3*) para áudio, *JPEG (Joint Photographic Experts Group)* para imagens e o *MPEG (Moving Picture Experts Group)* para áudio e vídeo tem sido fundamentais para a tentativa de lidar com este problema. A seguir, será tratada a maneira pela qual um conteúdo de áudio pode ser representado em um banco de dados multimídia.

2.2. Representações de Arquivos de Áudio em um Banco de Dados Multimídia

De acordo com Kosh e Döller (2005), a tarefa de representar dados multimídia pode ser subdivida em dois tipos: a representação de alto nível e a representação de baixo nível. A representação de baixo nível compreende a maneira como o dado multimídia será reconhecido e inserido no banco. Em soluções de mercado como o *Oracle 10g* e *IBM DB2*, objetos como áudio são definidos como um tipo de dados específico, denominado *BLOB* (*Binary Large Objects*). Desse modo, a manipulação de conteúdos multimídia era generalizada, sendo o *BLOB*, portanto, um objeto de representação geral, que utiliza um vetor de *bytes* para o armazenamento dos dados multimídia no banco. Entretanto, esta representação de baixo nível não disponibiliza mecanismos necessários para a realização da consulta e recuperação dos dados de modo eficiente (ou seja, o retorno bem sucedido da busca), bem como a maneira como a consulta será realizada para que este requisito tenha sido atendido. Neste cenário, faz-se necessária uma representação

de alto nível do áudio, para que esta realize uma intermediação entre o objeto *BLOB* e a aplicação que utilizará o banco de dados multimídia.

Esta representação denomina-se representação semântica. Ela permite um tratamento diferenciado dos arquivos de áudio (e de outros tipos de conteúdo multimídia), por tornar a própria representação do conteúdo uma ferramenta de auxílio à sua consulta, porém utilizando características encontradas por meio da análise do mesmo. Ferramentas têm sido criadas com o objetivo de tornar essa representação eficiente a ponto de otimizar os resultados de consulta.
A seguir, serão introduzidos dois importantes padrões de descrição de conteúdo multimídia para a realização de anotação semântica em dados: o *MPEG-7* e o *MPEG-21*.

2.3. Padrões de Descrição de Conteúdo Multimídia MPEG-7 E MPEG-21

Os padrões *MPEG-7* e *MPEG-21* foram desenvolvidos por um grupo de pesquisa ligado à *ISO (International Standard Organization)*, denominado *Moving Pictures Experts Group (MPEG)*. O padrão *MPEG-7* é também denominado Interface de Descrição de Conteúdo Multimídia. A realização de uma descrição em *MPEG-7* depende de um conjunto de metadados, somados à estrutura e conexões existentes entre os mesmos. Este padrão é totalmente centralizado na descrição do conteúdo, sendo as definições no que se referem aos mecanismos de busca a serem utilizados com o auxílio deste ou os modos pelos quais este pode ser aplicado a um sistema multimídia, tarefas não definidas em seu escopo. Este padrão é constituído por três elementos principais: Descritores (D), Esquemas de Descrição (ED) e Linguagem de Definição de Descrição (LDD), além da definição de uma outra representação, a representação binária (BiM), conforme mostrado na Figura 1, a seguir:

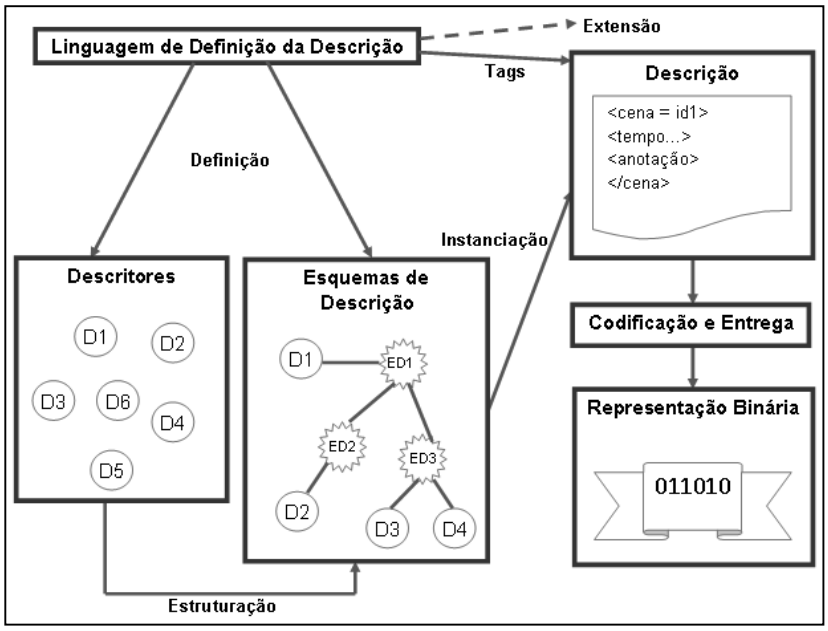

Figure 1. Principais elementos do padrão MPEG-7 [adaptada, Sonera (2003)]

A descrição em MPEG-7, em uma de suas formas de representação, utiliza-se da chamada Linguagem de Definição da Descrição (LDD), ou Linguagem de Definição da Descrição. Uma LDD emprega a sintaxe da *XML* (*eXtensible Markup Language*) para originar novos descritores e esquemas de descrição. A *XML* proporciona o uso de marcadores definíveis, as quais tornam esse tipo de linguagem favorável à criação de esquemas de descrição. É importante ressaltar que a validação de um documento de descrição nesta linguagem é realizada com fundamentação em uma gramática denominada DTD (*Document Type Definition*), conforme mencionado em Sonera (2003). O padrão *MPEG-7*, juntamente com os padrões *MPEG-1*, *MPEG-2*, *MPEG-4*, foi a base para o desenvolvimento do padrão *MPEG-21*. O padrão *MPEG-21* surgiu da necessidade de uma normalização mais abrangente no sentido de construção de um verdadeiro *framework* multimídia que pudesse facilitar a manipulação da crescente quantidade de dados multimídia inseridos em grande parte das aplicações distribuídas (sobretudo as aplicações para a Internet).

Sua composição engloba sete seguintes áreas fundamentais:

- Declaração do Item Digital;
- Descrição e Identificação do Item Digital;
- Uso e Manuseio de Conteúdo;
- Proteção e Gerenciamento da Propriedade Intelectual;
- Redes e Terminais;
- Representação de Conteúdo e Relatório de Eventos.

Para representá-lo, é utilizada o que se denomina Linguagem de Declaração de Item Digital (LDID), a qual por sua vez é definida em esquemas *XML* (assim como a Linguagem de Definição da Descrição, LDD, do padrão *MPEG-7*). A Declaração do Item Digital (DID) consiste justamente no documento *XML* que se refere a um Item Digital, de acordo com os padrões acordados na LDID. Pode-se perceber que a maneira como o *MPEG-21* realiza a representação dos dados multimídia é similar a que ocorre no *MPEG-7*.

3. *Query-by-humming*

Query-by-humming (QBH), de acordo com Tripathy *et al*. (2009), é um termo que pode ser compreendido como "conceito de interação na qual a identificação de uma canção tem de ser revelada rapidamente e ordenadamente, a partir de uma dada entrada de áudio cantada, utilizando um grande banco de dados de melodias conhecidas".Uma aplicação multimídia *QBH*, portanto, caracteriza-se por possibilitar ao usuário buscar informações como nome, gênero, autores e intérpretes, além do próprio arquivo de áudio correspondente a uma música, sem que este saiba palavras que existam na composição da mesma. Como não há uma tradução técnica oficial para o referido termo, em Português pode-se compreendê-lo como "consulta pelo assovio ou solfejar". Além de identificar os componentes de um sistema *QBH,* é necessário compreender como estes interagem para alcançar a finalidade proposta por uma aplicação deste tipo. O

componente **Entrada de Áudio** pode ser também denominado áudio *humming*, considerando que este é o arquivo que contém a gravação da melodia produzida pelo usuário do sistema na realização da consulta. As **Entradas do Banco de Dados Multimídia** são as canções previamente armazenadas no banco de dados multimídia do sistema.

Como mencionado em Unal *et al*. (2008), para que ambas as entradas sejam eficientemente comparadas, estas deverão ser convertidas em um formato similar, o qual é a **Representação Melódica**. Com fundamento nesta representação, o mecanismo de **Comparação** busca as similaridades entre os arquivos, e ao encontrar, dentre os armazenados no banco de dados, o que mais apresente similaridades com o áudio *humming* pelo usuário e realiza a recuperação do arquivo em seu formato original, juntamente com as respectivas informações associadas ao mesmo. Entretanto, um dos fatores limitantes a resultados totalmente corretos nesta busca são as imperfeições e variações presentes no áudio *hummed*.

4. Proposta

Serão apresentadas as ferramentas *CALIPH* e EMIR, ambas voltadas para uso com imagens e fortemente relacionadas ao padrão MPEG-7.

Como descrito em Lux (2009), o *CALIPH* ("*Common and Light Weight Photo Annotation*") é uma ferramenta fundamentada na linguagem Java, cujo objetivo é realizar, sobretudo em arquivos de imagens, a tarefa de anotação semântica e de extração de metadados contidos nestes. Para a manipulação dos metadados, o *CALIPH* utiliza-se do padrão *MPEG-7*. Esta ferramenta livre possibilita a realização manual ou automática da anotação semântica. Neste trabalho, será priorizada a realização de anotação manual, a qual inclusive pode ser realizada de diferentes formas, de acordo com o objetivo específico que se pretende alcançar ou pela forma como será realizada a busca do arquivo em que foi efetuada a anotação.

EMIR (*Experimental Metadata based Image Retrieval*), de acordo com Lux (2009), é uma ferramenta desenvolvida em Java para ser utilizada em conjunto com o *software CALIPH*, possibilitando a realização de buscas em arquivos anotados semanticamente pelo mesmo, conforme mencionado em item anterior deste mesmo capítulo. Esta ferramenta tem seu funcionamento fundamentado nos descritores *MPEG-7* criados pelo *CALIPH*, e possibilita a recuperação das imagens por intermédio de texto (com a utilização de palavras-chave dos descritores), por consulta a nós de grafos semânticos armazenados ou ainda por similaridade das demais imagens do repositório com uma imagem previamente selecionada.

A implementação deste trabalho experimental fundamentou-se nas seguintes etapas:

1. Definição de uma listagem de músicas para a realização da anotação semântica; nesta etapa, foram selecionadas 10 músicas (cujos títulos não apresentam relação direta com palavras do refrão destas ou que não apresentam "letra cantada").

2. Obtenção dos dados a serem anotados por intermédio do software de edição de áudio Audacity;

Nesta etapa, iniciou-se a utilização do Audacity, um software livre que se destina à captação, edição, gravação e análise de arquivos de áudio. Cada uma das músicas teve um intervalo selecionado, correspondente ao seu refrão ou à parte mais propícia a ser lembrada pela maioria dos usuários no momento de realização de uma consulta.

Após este procedimento, é realizada a análise do trecho de áudio. A análise é fundamentada na frequência atingida a cada variação de decibéis da música. O *Audacity* retorna o resultado desta análise graficamente, como é representado pela Figura 2, da próxima página. A realização da anotação semântica com o *CALIPH* utilizará, além da imagem deste gráfico, os dados referentes ao mesmo, contidos em um arquivo de texto simples, gerado pelo próprio *Audacity*, e denominado espectro de frequência.

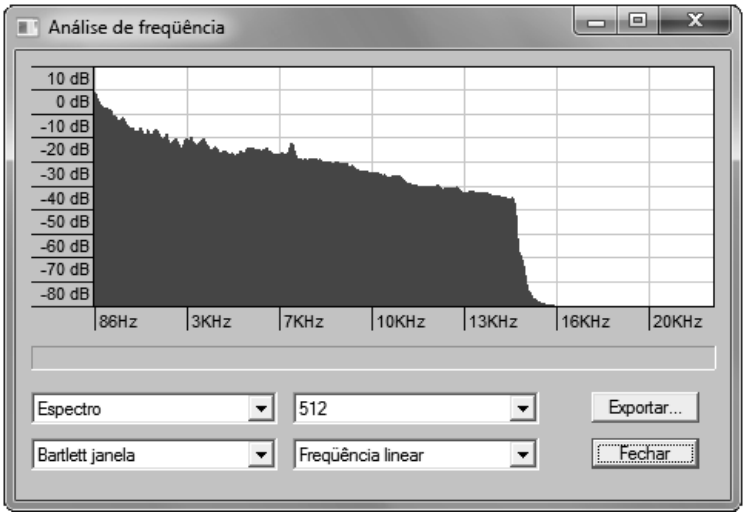

Figure 2. Gráfico gerado pelo *Audacity*, após a análise do intervalo selecionado de "*Bitter Sweet Symphony*".

3. Realização da anotação semântica das músicas pertencentes à listagem por intermédio do CALIPH;

O CALIPH possui a interface desenvolvida propriamente para anotação semântica em arquivos de imagem. Ele é preparado para identificar, dentro de um diretório selecionado, somente arquivos com extensões de imagem. Considerando que a proposta de análise deste trabalho refere-se a arquivos de áudio, foi utilizada uma imagem para representar cada um destes, para que os mesmos pudessem ser anotados semanticamente. Para aumentar a eficiência da anotação semântica, foi utilizado como imagem, para cada um dos trechos, o seu respectivo gráfico de análise de frequência (gerado pelo Audacity).

4. Coleta do assovio de voluntários simulando um trecho musical referenciado na base de dados criada com o CALIPH;

A coleta de assovios de voluntários ocorreu de duas maneiras: presencialmente, com a utilização de um microfone acoplado ao fone de ouvido (*headset*) e por meio da

ferramenta de comunicação *on-line Skype*, com a utilização do meio de captação de voz pessoal de cada voluntário.

Para cada uma das músicas selecionadas, foi criado um arquivo denominado "Trecho nº de identificação da música", contendo somente o intervalo correspondente ao refrão ou a alguma parte considerada marcante desta. A seguir, é apresentada a Tabela 1, próxima página, com as músicas e a nomenclatura de seus respectivos trechos, bem como a duração de cada um deles.

Tabela 1. Listagem das músicas e dos tempos de seus respectivos trechos selecionados para serem assoviados

Música	**Trecho**	**Duração (em segundos)**
Bitter Sweet Symphony	Trecho 1	23 s
Blue Monday	Trecho 2	17 s
Enjoy the Silence	Trecho 3	16 s
Entre a cruz e a espada	Trecho 4	20 s
L'Aurora	Trecho 5	19 s
Losing my Religion	Trecho 6	16 s
Not Exactly	Trecho 7	19 s
The Rockafeller Skank	Trecho 8	12 s
Uma Brasileira	Trecho 9	20 s
Unchained Melody	Trecho 10	20 s

O experimento foi realizado com cada participante individualmente. Foram, no total, 8 participantes (4 via *Skype* e 4 presencialmente) , os quais somaram 21 amostras de assovios. A coleta de amostras realizada pelo Skype ocorreu do seguinte modo:

- Era enviado ao participante um arquivo compactado, denominado "Trechos.rar.
- O participante era instruído a descompactar o arquivo, ouvir os dez trechos das músicas por uma vez, e selecionar o número de trechos que gostaria de assoviar.
- Após a escolha de quais trechos seriam assoviados, o participante poderia ouvi-los novamente, antes de iniciar o assovio correspondente ao trecho, pelo número de vezes que considerasse necessário.
- A gravação do assovio era então iniciada com um aplicativo adicional do *Skype,* denominado *Pamela Call Recorder*.

Presencialmente, o áudio era captado por um microfone acoplado a *headset* e gravado pelo *Audacity*.

5. A anotação dos arquivos de áudio captados dos participantes foi realizada seguindo a mesma sequência da anotação dos trechos das músicas.

5. Conclusões e Trabalhos Futuros

Dentre os 21 assovios coletados, observou-se que a variação da relevância da consulta está fortemente conectada ao tipo de música assoviada e à maneira como o voluntário identifica o elemento principal da música. Nas imagens geradas pela análise dos trechos propostos para assovio foi realizada a anotação semântica com o *CALIPH*, contudo estas não foram indexadas na base de dados do *EMIR* para as consultas, devido a incompatibilidades com uma amostragem efetuada somente por assovios. Estas incompatibilidades devem-se ao conjunto de instrumentos existentes e de elementos vocais adicionados eletronicamente nas músicas que não podem ser reproduzidos com um assovio. Este fator, no entanto, não representa a inviabilidade do método *QBH* ser realizado com o auxílio de um *software* de anotação semântica no processo de análise das consultas efetuadas. Foi identificada a necessidade de maior quantidade de amostras para cada música armazenada ou referenciada em um sistema com esta abordagem.

Como trabalhos futuros, podem ser propostos: a busca por um modo de otimização dos índices de relevância das consultas, com a utilização do padrão MPEG-21 para a produção de novos descritores; a ampliação da base de dados do *CALIPH*, sem interferência nos resultados satisfatórios obtidos; a realização de um paralelo entre uma nova base de dados anotada semanticamente por meio de outras características dos arquivos de áudio analisados; e a associação de outras opções de anotação do *CALIPH* e de outras opções de recuperação do EMIR para as amostras apresentadas .

Referências Bibliográficas

AUDACITY (2003). Software de análise sonora. Revista Sonora, 4 (2), São Paulo: Abril.

BARRIGTON, Luke; CHAN, Antoni, TURNBULL, Douglas; LANCKRIET, Gert (2007) Audio information retrieval using semantic similarity. Proceedings of the IEEE ICASSP, 32, p. II-725–II-728.

DJERABA, Chabane; SEBE, Nicu; LEW, Michael S. (2005) Systems and Architectures for Multimedia Information Retrieval. ACM Multimedia Systems Journal. 10 (6), p. 457-463.

ELMASRI, Ramez; NAVATHE, Sham (2007). Fundamentals of Database Systems. Boston: Pearson/Addison Wesley.

FOOTE, Jonathan T. (1999) An overview of audio information retrieval. Multimedia Systems. 7 (1). 1, p. 2-11, ACM Press/Springer-Verlag.

GAGLIARDI, Isabella; PAGLIARULO, Patrizia (2005) Audio Information Retrieval in Hypermedia Environment. In: ACM Conference on Hypertext and Hypermedia, 6, Salzburg.

ISO/IEC TR 21000-1:2004 (2004). "Information technology – Multimedia framework (MPEG-21) – Part 1: Vision, Technologies and Strategy", New York: ISO.

KOSCH, Harald; DÖLLER, Mario (2005). Multimedia Database Systems: Where are we now? In: IASTED - INTERNATIONAL CONFERENCE ON DATABASES AND APPLICATIONS, Innsbruck.

LEW, Michael S; SEBE, Nicu; DJERABA, Chabane; JAIN, Ramesh (2006). Content-based Multimedia Information Retrieval: State-of-the-art and Challenges. ACM Transactions on Multimedia Computing, Communication, and Applications. 2(1) p. 1-19.

LAU, Edmond; DING, Annie; ON, Calvin (2010). MusicDB: A Query by Humming System. Massachusetts, 2007. Disponível em: <http://people.csail.mit.edu/edmond/projects/musicdb/musicdb.pdf>. Acesso em: 07 Nov. 2010.

PEREIRA, Fernando; KOENEN, Rob (2010). MPEG: Context, Goals and Working Methodologies. Disponível em: <http://media.wiley.com/product_data/excerpt/18/04700101/0470010118.pdf> . Acesso em: 02 Set.2010.

SONERA (s.d.) (2010). MPEG-7 White Paper,. Disponível em: <http://www.medialab.sonera.fi/workspace/MPEG7WhitePaper.pdf>. Acesso em: 02 Set.2010.

TZANETAKIS, George; COOK, Perry (2000). Audio information retrieval (AIR) tools. Proceedings of the Annual International Symposium on Music Information Retrieval (ISMIR 2000), Los Angeles, p. 135-144.