Recuperação Inteligente de Letras de Músicas na Web

Rafael P. Ribeiro[1], **Carlos N. Silla Jr.**[1]

[1]Laboratório de Computação e Tecnologia Musical
Universidade Tecnológica Federal do Paraná (UTFPR)
Caixa Postal 86300-000 – Cornélio Procópio – PR – Brasil

{rpr.rafa,carlos.sillajr}@gmail.com

Abstract. *The task of automatic retrieval and extraction of lyrics from the web is of great importance to different Music Information Retrieval applications. Most of the existing approaches to deal with this problem rely on computation resources that often unavailable for songs which are not mainstream (usually English) songs. In this paper we present the Ethnic Lyrics Fetcher (ELF), a tool for Music Information Researchers, which has a novel lyrics extraction mechanism. To assess ELF Two experiments were performed. In the first, the novel lyrics extraction mechanism against 12 websites that have their lyrics in a well defined structure and, in addition, a comparison was made with the lyrics extraction method presented in [Geleijnse and Korst 2006]. The second experiment evaluated the performance of the ELF (Ethnic Lyrics Fetcher) as a tool for retrieving lyrics from the web. Our results show that ELF is a useful tool for users and researchers in Music Information Retrieval.*

Resumo. *A tarefa de recuperação e extração automática de letras de músicas a partir da web é de grande importância para diferentes aplicações da área de recuperação de informações musicais (Music Information Retrieval). A maior parte das abordagens existentes para lidar com este problema dependem de recursos computacionais que, muitas vezes, estão indisponíveis para músicas que não são populares ou estão em idiomas que não são o Inglês. Neste artigo, é apresentado um sistema para a recuperação automática de letras de músicas na web, denominado Ethnic Lyrics Fetcher (ELF), que possui um novo mecanismo para a detecção e extração automática de letras de músicas. Para avaliar o sistema desenvolvido foram realizados dois experimentos. No primeiro experimento, o mecanismo de extração de letras foi avaliado utilizando como base 12 websites que possuem letras de música em uma estrutura bem definida e também o método considerado o estado da arte para o problema. No segundo experimento, foi avaliado o desempenho do sistema desenvolvido como uma ferramenta de busca, identificação e extração de letras de músicas na web. A análise dos resultados experimentais obtidos mostraram que o ELF é uma ferramenta útil para auxiliar pesquisadores e usuários na recuperação de informações musicais.*

1. Introdução

As letras de músicas são utilizadas em diferentes tipos de aplicações e pesquisas na área de recuperação inteligente de informações musicais (*Music Information Retrieval* - MIR). Alguns dos exemplos de aplicações são: identificação de emoções [Hu et al. 2009, Hu and Downie 2010]; classificação automática de gêneros musicais[Mayer and Rauber 2011]; dentre outras [Baumann and Hummel 2003, Logan et al. 2004, Li and Ogihara 2004]. Além disso, existem estudos que apontam que as letras das músicas (ou pelo menos parte delas) são a segunda informação mais desejada e utilizada pelos usuários de sistemas de MIR [Downie and Cunningham 2002].

A necessidade de sistemas automatizados para recuperação e extração de letras de músicas tem sido relatada em vários trabalhos de pesquisa. Por exemplo, em [Zaanen and Kanters 2010] os autores apontam que devido a algumas limitações na forma como eles receberam o seu conjunto de dados (uma lista de títulos música e de artistas da canção, sem espaços em branco. Ex: ironmainden - fearofthedark), eles ficaram limitados a usar 5.631 letras de 10.000 músicas que eles tinham disponíveis.

Em [Shamma et al. 2005] os autores utilizaram uma biblioteca online especializada em letras de músicas (*Leo Lyrics*), porém eles mencionaram que em seus experimentos a busca direta não trabalhava com covers. Por exemplo, o sistema só iria encontrar a letra da música "Smells Like Teen Spirit"se o artista da música fornecido fosse "Nirvana". O sistema não encontraria a letra se o artista fornecido fosse "Tori Amos"(que gravou um cover da música).

Na obra de [Macrae and Dixon 2012] os autores notaram que com seu processo simples de recuperação de letras não foi possível obter corretamente todas as letras que eles precisavam. O método utilizado examinava a linha de cada página html para determinar se ela continha o texto da letra semelhante ou não e, em seguida, selecionar a mais longa sequência de linhas de textos semelhantes na página. Quaisquer letras que tinham menos de três linhas ou mais de 200 linhas foram descartadas.

Na obra de [Mayer et al. 2008] os autores criaram um conjunto de dados de letras manualmente para a tarefa de classificação de gêneros musicais. No entanto, apesar de haver disponível uma coleção particular de aproximadamente 12.000 músicas, foram selecionadas (aleatoriamente) de 30 a 45 músicas de cada um dos dez gêneros musicais disponíveis, devido à inexistência de ferramentas automatizadas para recuperar as letras para toda a coleção.

Apesar da importância da recuperação automatizada de letras a partir da web para diferentes tipos de aplicações e sistemas de MIR, pesquisas para resolver esse problema são quase inexistentes. As únicas exceções que os autores tem conhecimento são os trabalhos de [Geleijnse and Korst 2006, Knees et al. 2005].

Dessa forma, a principal contribuição deste trabalho é o sistema Ethnic Lyrics Fetcher (ELF)[1], apresentado na Seção 2, e seu novo procedimento de detecção e extração letras da web (apresentado na Seção 3). Para avaliar o procedimento de extração do ELF foram realizados dois experimentos. No primeiro experimento (Seção 4), foi avaliado o novo procedimento de extração de letras de músicas comparando-o com as letras recupe-

[1]Uma versão online do ELF está disponível em `http://music.cp.utfpr.edu.br/elf/`

radas de 12 websites que possuem a estrutura de página bem definida (mas diferentes entre si) para a delimitação do conteúdo com a letra da música e também contra o método considerado o estado da arte. No segundo experimento (Seção 5), o ELF foi avaliado como um sistema de recuperação inteligente de letras de músicas na web, utilizando um motor de busca e comparando seu desempenho contra os websites do primeiro experimento (com estrutura conhecida) e também com as letras recuperados manualmente. Finalmente, na Seção 6 são apresentadas as conclusões deste trabalho e futuras direções de pesquisa.

2. Visão Geral do Sistema ELF

O sistema ELF funciona do seguinte modo:

1. Na primeira etapa, o sistema ELF precisa de informações sobre a letra da música que o usuário deseja recuperar. O sistema requer que o usuário forneça as informações título (obrigatório) e artista (opcional);
2. Na segunda etapa, o sistema ELF pré-processa o título da música e do artista mantendo todas no padrão minúsculo (por exemplo, É torna-se é) e remoção dos acentos (por exemplo, é torna-se e);
3. Na terceira etapa, o sistema ELF cria uma consulta para o motor de busca. Neste trabalho foi utilizado o motor de busca Google, e as consultas foram construídas pela concatenação do título da música e nome do artista. Além disso, foi feita a substituição de todos os espaços em branco na string resultante pelo símbolo de soma (+). A palavra-chave letra também é adicionada na pesquisa (por exemplo: falamansa+xote+dos+milagres+letra);
4. Na quarta etapa, o processo do sistema ELF retorna uma string JSON, obtida da API do Google para criar uma lista de websites que podem conter a letra da música que o usuário está procurando;
5. Na quinta etapa, o sistema ELF utiliza seu novo método para detecção e extração de letras de músicas (descrito em detalhes na seção 3), a fim de identificar se um determinado website contém uma letra de música ou não. Se o website contém a letra, o ELF a retornará para o usuário. Caso contrário, ele analisa o próximo website da lista retornada pelo motor de busca.

3. Novo Procedimento de Detecção e Extração de Letras

O sistema ELF utiliza um novo procedimento de detecção e extração de letras de músicas de páginas web. A abordagem usada para extração das letras é assumir que ela está em algum lugar dentro da página web onde há marcadores de final de linha (ou seja, tags
 e <P>). Uma das diferenças entre o procedimento de extração de letras do ELF e as abordagens existentes é que o ELF utiliza todas as tags HTML (incluindo as tags de quebra de linha) para localizar as letras dentro de qualquer página da web. Antes de tentar extrair as letras da página o sistema verifica se a página não entra na lista de sites que são conhecidos por não possuírem letras, como por exemplo, Youtube, Wikipédia e Blogspot. Se este não é o caso, o procedimento de extração da letra é feito da seguinte forma:

- O primeiro passo é armazenar todo o conteúdo HTML da página web recuperada em uma string.
- O segundo passo é analisar a string com o conteúdo HTML, procurando por tags HTML de abertura (por exemplo: <div) com a exceção de "<br"(quebra de linha), "<p"(parágrafo), "</"(fecha tag) e "<!--"(comentário html).

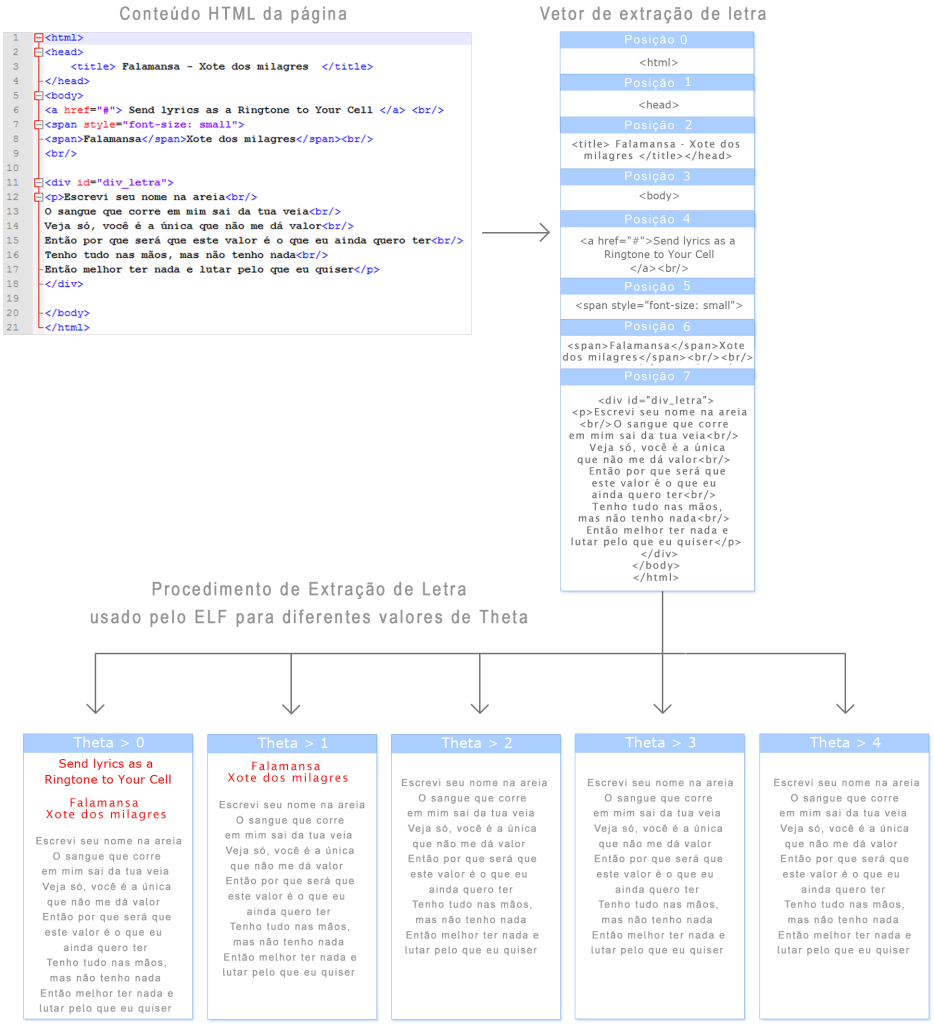

Figura 1. Um exemplo do procedimento de extração de letra utilizado pelo ELF com diferentes valores de theta (θ)

- Para cada tag HTML de abertura, faça:
    1. Crie uma nova posição no vetor de extração da letra;
    2. Adicione todo o conteúdo da página até que uma nova tag HTML de abertura seja encontrada.
- Após converter a página Web no vetor de extração de letras, o ELF analisa o vetor contando o número de ocorrências de `<br` e `<p` tags em cada posição do vetor.
- Para decidir qual conteúdo será extraído como sendo parte da letra da música (e exibido ao usuário), o ELF seleciona as posições do vetor de extração de letras que tenham o número de tags de quebra de linha superior ao parâmetro θ (constante que indica a quantidade mínima de tags de quebra de linha).
- A Figura 1 apresenta um exemplo do procedimento de extração de letras usado pelo ELF para diferentes valores de θ com a música “Xote dos Milagres”da Banda “Falamansa”.
- Se a letra da música que será exibida ao usuário estiver vazia, o ELF considera que a letra não foi detectada no website.

4. Avaliação do Procedimento de Extração de Letras do ELF

Nesta seção, foram respondidas às seguintes perguntas, usando experimentos controlados: (a) O quão bem o procedimento de extração de letras do ELF funciona (variando-se o parâmetro θ) quando comparado com as letras extraídas de diferentes websites com estrutura da página claramente definida? (b) Até que ponto o sistema ELF funciona como um sistema de recuperação de letras quando comparada com outras abordagens? Em todos os experimentos foi utilizada a Latin Music Database (LMD) [Silla Jr. et al. 2008].

Tabela 1. Websites de Letras de músicas e os rótulos (tags) utilizados.

Website	Tag para início da letra	Tag para fim da letra
1songlyrics.com	`<div id="wrapper">`	`<p><b>`
6lyrics.com	`<span id="ctl00_ContentPlaceHolder1_lbl Content" style="display: block">`	`</span></p>`
lyricmania.com	`<div id="wrapper">`	`<div id="wrapper2">`
www.lyricsoncall.com	`<p><strong>`	`<div id="info">`
www.moron.nl	`<td colspan=2 style="line -height:15px; size:14px;">`	`</td>`
vounessa.com.br	`<div class="bloco-c">`	`<!-- Fim div bloco c -->`
letras.mus.br	`<div id="main_cnt">`	`</div>`
www.vagalume.com.br	`<div id=lyr_original class="left originalOnly">`	`</div>`
Lyricshall.com	`<div class="correct-button">`	`<h3 class="smalltitle">`
Lyricspedia.com	`<!-- AddThis Button END -->`	`</a></b></div>`
Lyricsreg.com	`</div><div style="text -align:center;">`	`<g:plusone>`
Themusic-world.com	`<BLOCKQUOTE>`	`</BLOCKQUOTE>`

4.1. Experimentos

Para avaliar o procedimento de extração de letras do ELF foram verificados manualmente doze websites diferentes que têm letras para as músicas da Latin Music Database [Silla Jr. et al. 2008]. Cada um dos websites tem uma maneira distinta para delimitar seu conteúdo com as letras de músicas. A Tabela 1 apresenta a lista com os websites utilizados neste experimento, bem como as respectivas tags que delimitam as letras de música dentro das páginas.

Para avaliar procedimento de extração de letras do ELF, foram comparados a saída de seu mecanismo (usando diferentes valores de θ) com a extração utilizando os delimitadores padrão de cada website. Foi utilizado um extrator de letra específico, que sabe quais são os rótulos específicos de cada website.

A fim de mensurar o quão semelhante é a saída do procedimento de extração do ELF (usando diferentes valores de θ) com o padrão de delimitadores em cada website, foi utilizada a medida de similaridade de documentos a partir da área de recuperação de informação, conhecida como similaridade co-seno [Salton and Buckley 1998]. A medida de similaridade co-seno calcula o quão semelhantes são dois documentos (letras de música neste caso). A medida fornece um valor entre 0 e 1. Quanto maior for o valor da medida entre dois documentos, mais semelhantes eles são.

4.2. Resultados

A Tabela 2 apresenta os resultados experimentais do mecanismo de extração de letras do ELF com diferentes valores de θ e para os doze websites apresentados na Tabela 1. A análise da Tabela 2 apresenta alguns resultados interessantes. Em primeiro lugar, é que, quanto maior o valor de θ, obtém-se um valor mais alto de similaridade geral. Isto pode ser explicado pelo fato de que um menor valor para θ simplesmente retornará todo o conteúdo da página web que contém pelo menos uma tag `<br>` no vetor de extração.

Tabela 2. Avaliação dos mecanismos de extração de letras de música do ELF e do estado da arte.

Website	$\theta > 0$	$\theta > 1$	$\theta > 2$	$\theta > 3$	$\theta > 4$	Estado da arte [Geleijnse and Korst 2006]
letras.mus.br	0.9226	1.0	1.0	0.9998	0.9998	0.7199
vagalume.com.br	0.9151	0.9191	1.0	1.0	1.0	0.7202
songlyrics.com	0.9753	0.9761	0.9767	0.9780	0.9780	0.6389
6lyrics.com	0.9141	0.9800	0.9972	0.9972	0.9966	0.5808
vounessa.com.br	0.9996	0.9996	0.9995	0.9996	0.8960	0.6379
lyricmania.com	0.9552	0.9617	0.9618	0.9617	0.9617	0.6295
lyricshall.com	0.9702	0.9706	0.9706	0.9706	0.9162	0.8781
lyricsoncall.com	0.9668	0.9666	0.9666	0.9787	0.9787	0.5932
lyricspedia.com	0.9974	0.9967	0.9920	0.9920	0.9905	0.6725
lyricsreg.com	0.9533	0.9686	0.9689	0.9730	0.9730	0.7305
moron.nl	0.8847	0.9890	1.0	1.0	1.0	0.8088
themusic-world.com	0.9905	0.9890	0.9922	0.9923	0.9922	0.7096
Média	0.9537	0.9764	0.9855	0.9869	0.9736	0.6933

Em segundo lugar, as letras com maiores valores de similaridade foram inspecionadas manualmente a fim de compreender por que a semelhança não era 1.0. Curiosamente, foi observado que alguns dos websites têm entre a estrutura de suas letras algum tipo de mensagem promocional, como "Ringtone - Enviar esta ringtone para o seu celular", que foi removido pelo procedimento de extração do ELF. Nos experimentos, foi possível notar isto nos websites: lyricsmania.com, 1songlyrics.com, lyricshall.com e lyricsreg.com. Outra observação que pode ser feita é que alguns websites, como o lyricsoncall.com e 6lyrics.com têm em meio das letras das músicas, a URL da página. Esta informação também foi removida pelo procedimento de extração usado pelo ELF. O website vounessa.com.br incluiu os campos Compositor e gênero musical no final das letras que também foram filtrados pelo procedimento de extração do ELF.

Em terceiro lugar, com base nos resultados experimentais, é possível perceber que existe um trade-off entre a utilização de um valor baixo ou alto para θ. Um valor baixo pode retornar todo o conteúdo como resultado, enquanto um valor alto pode filtrar mais do que deveria. Portanto, para os experimentos na próxima seção foi utilizado o valor de θ maior do que 3.

5. Avaliando o ELF Como um Sistema Para Recuperar Letras da Web

Nesta seção, será avaliada a performance do ELF como um sistema para recuperação de letras de música online. Para isto, é utilizada uma API como ferramenta de busca e o mecanismo de extração.

5.1. Experimentos

Para avaliar o ELF como um sistema de recuperação de letras da Web, foram fornecidos o artista e o título para todas as músicas disponíveis na LMD. A motivação por trás da utilização da LMD são as músicas com letras em Inglês, Espanhol, Português e "Spanglish"(Letras com trechos em Espanhol e outros em Inglês).

Neste experimento foi utilizada a API do Google como motor de busca no ELF para consultar possíveis websites que possam conter as letras de música que o usuário está procurando. Como base de comparação para o ELF foi utilizada a implementação de um coletor de letras para 12 (doze) websites específicos (os mesmos websites apresentados na Tabela 1), que têm uma codificação de URL específica e informações das letras bem delimitadas. A Tabela 3 apresenta a estrutura da URL específica de cada website. Também foram avaliadas duas versões do sistema ELF. Uma usando o nome da música e do artista, referido como ELF$_{ST+A}$ e outro usando apenas as informações do nome da

Tabela 3. Estrutura fixa dos sites de letras usados como base

Website	Estrutura
letras.mus.br	http://letras.mus.br/$song artist$/$song title$
vagalume.com.br	http://www.vagalume.com.br/$song artist$/$song title$.html
moron.nl	http://www.moron.nl/lyrics/$song artist$/$song title$-lyrics.html
lyricsoncall.com	http://www.lyricsoncall.com/lyrics/$song artist$/$song title$-lyrics.html
lyricsreg.com	http://www.lyricsreg.com/lyrics/$song artist$/$song title$
vounessa.com.br	http://www.vounessa.com.br/musicas/$song title$
6lyrics.com	http://www.6lyrics.com/$song artist$-lyrics-artista.aspx
lyricshall.com	http://www.lyricshall.com/lyrics/$song artist$/$song title$
lyricmania.com	http://www.lyricmania.com/$song artist$-$song title$-lyric.html
1songlyrics.com	http://www.1songlyrics.com/$first letter of song artist$/$song artist$song title$.html
lyricspedia.com	http://www.lyricspedia.com/$song artist$/$song title$-lyrics/
themusic-world.com	http://www.themusic-world.com/$song artist$/lyrics/$song title$

música, conhecido como ELF$_{ST}$. Note-se que nos experimentos relatados nesta seção, os seguintes websites foram listados como websites que não possuem letras e, portanto, são ignorados: “blogspot”, “wikipedia”e “youtube”.

5.2. Resultados

A Tabela 4 apresenta os resultados experimentais (separadas por gênero musical) ao usar ELF_{ST}, o ELF_{ST+A} e outras 14 abordagens para recuperar as letras das músicas da LMD. As análises destes resultados mostram que os websites brasileiros: vagalume.com.br (1569 letras) e vounessa.com.br (1944 letras) contêm o maior número de letras para canções da LMD entre os websites individuais. Este resultado já era esperado, afinal 6 dos 10 gêneros musicais na LMD são de gêneros musicais Brasileiros.

Para saber se os websites individuais podem ser combinados em um sistema de recuperação de letras, a fim de obter um resultado melhor foi feita a combinação como é apresentada na linha 13 da Tabela 4, onde 2.480 letras das músicas da LMD foram recuperadas. Deve notar-se que, nestes resultados (individuais ou utilizando a abordagem combinada) as páginas da web podem ser extraídas usando a estrutura de página conhecida e podem conter algum conteúdo adicional além das letras de música (por exemplo, algum tipo de publicidade).

Na linha quinze da Tabela 4 são apresentados os resultados utilizando o ELF_{ST+A}. Ou seja, utilizando as informações sobre o título da música e o artista. O ELF_{ST+A} recuperou 2.847 letras de músicas da LMD. Estas letras foram obtidas automaticamente pelo sistema. Para avaliar se todas as letras recuperadas eram letras reais, cada uma delas foi inspecionada manualmente. Após esta análise, foi verificado que 204 das 2.847 letras recuperadas não eram letras de músicas, mas algum outro tipo de informação, como biografias de artistas. Ainda assim, obteve-se um total de 2.643 letras corretas recuperadas automaticamente pelo ELF_{ST+A}. Outro aspecto observado é que as letras recuperadas pelo ELF_{ST+A} foram extraídas de 157 websites diferentes. Os cinco principais websites utilizados pelo ELF_{ST+A} foram: letras.mus.br (1.490 letras), www.todotango.com (279 letras), www.musica.com (224 letras), www.vagalume.com.br (91 letras) e www.mp3lyrics.org (76 letras). É interesse observar que os autores do trabalho só conheciam dois destes websites antes de realizar os experimentos.

Na linha dezesseis da Tabela 4 são apresentados os resultados para a utilização do ELF_{ST}. Ou seja, utilizando apenas o título da música. Esta versão do ELF recuperou 3.116 letras da LMD. Para avaliar se todas as letras recuperadas eram letras de fato letras das músicas desejadas, cada uma delas foi inspecionada manualmente. Após essa análise foi verificado que apenas 60 das 3.116 letras recuperadas não eram letras de mu-

sicas. Portanto, o ELF_{ST} recuperou com sucesso 3.056 letras utilizando 110 websites diferentes. Os cinco principais websites utilizados foram: letras.mus.br (1.454 letras), www.musica.com (760 letras), www.todotango.com (274 letras), www.vagalume.com.br (114 letras) e www.lyricsmode.com (38 letras).

Na linha dezessete da Tabela 4 são apresentados os resultados das letras pesquisadas e recuperados manualmente por membros do laboratório de computação e tecnologia musical. Este processo levou à recuperação de 2.503 letras de músicas da LMD utilizando o título da música e do artista usando a ferramenta de busca Google.

Na linha dezoito da Tabela 4, é apresentado o número total de músicas da LMD por gênero e na linha 19 é apresentado o número de músicas instrumentais contidas na LMD.

Tabela 4. Número de letras recuperados por abordagens para cada gênero musical na LMD. A = Axé; Ba = Bachata; Bo = Bolero; F = Forró; G = Gaúcha; M = Merengue; P= Pagode; Sa = Salsa; Se = Sertaneja; T = Tango

Linha #	Contexto	A	Ba	Bo	F	G	M	T	P	Sa	Se	Total
1	1songlyrics.com	28	94	71	7	0	53	24	2	102	0	381
2	6lyrics.com	17	107	112	2	0	70	101	9	96	2	516
3	Letras.mus	132	98	127	229	126	39	168	168	140	14	1.241
4	Lyricmania.com	21	98	72	8	0	52	23	3	109	0	386
5	Lyricshall.com	0	0	2	0	0	0	0	0	0	0	2
6	Lyricsoncall.com	46	80	67	10	6	42	23	36	85	27	422
7	Lyricspedia.com	13	43	61	11	0	35	22	1	77	0	263
8	Lyricsreg.com	11	23	18	1	0	6	4	0	32	0	95
9	Moron.nl	0	12	6	1	0	3	0	0	7	0	29
10	Themusic-world.com	0	23	66	1	0	18	0	0	35	0	143
11	Vagalume.com.br	204	125	136	208	105	68	113	224	129	257	1.569
12	Vounessa.com.br	247	210	239	259	179	171	266	0	230	143	1.944
13	Os 12 websites	268	238	251	277	208	185	288	237	253	275	2.480
14	ELF_{ST+A}	298	247	233	307	272	227	386	301	267	309	2.847
15	ELF_{ST}	303	303	301	313	300	291	397	301	299	308	3.116
16	Recuperadas Manualmente	292	251	268	296	178	101	336	299	172	310	2.503
17	LMD Total	304	308	302	315	306	307	404	301	303	310	3.160
18	LMD Instrumentais	0	0	0	6	24	0	54	2	0	0	86

Tabela 5. Avaliação das diferentes abordagens para as Letras Recuperadas da Web.

Abordagens	Precisão	Recall	F-measure
1songlyrics.com	1	0.120569620	0.215193448
6lyrics.com	1	0.163291139	0.280739935
Letras.mus	1	0.392721519	0.563962736
Lyricmania.com	1	0.122151899	0.217710096
Lyricshall.com	1	0.000632911	0.001265022
Lyricsoncall.com	1	0.133544304	0.235622557
Lyricspedia.com	1	0.083227848	0.153666375
Lyricsreg.com	1	0.030063291	0.058371736
Moron.nl	1	0.009177215	0.018187520
Themusic-world.com	1	0.045253165	0.086587950
Vagalume.com.br	1	0.496518987	0.663565236
Vounessa.com.br	1	0.615189873	0.761755486
Os 12 websites	1	0.784810127	0.879432624
ELF_{ST+A}	0.928345627	0.900949367	0.914442348
ELF_{ST}	0.980744544	0.986075949	0.983403021
Recuperadas Manualmente	1	0.792405063	0.884180791

Os resultados sob a recuperação das informações, as métricas de precisão, Recall e F-Measure são apresentados na Tabela 5. A análise da Tabela 5 apresenta alguns resultados interessantes. Em primeiro lugar, deve notar-se que, para os websites específicos, embora a sua precisão seja sempre 100%, existe a necessidade de se conhecer, previamente, a estrutura do site e de suas páginas para buscar corretamente as letras, o que é inviável na prática.

Em segundo lugar, com a exceção dos websites vagalume.com.br e vounessa.com.br, todos os websites específicos, possuem valores baixos de recall, pois re-

cuperaram apenas uma pequena parte do número de letras solicitadas.

Em terceiro lugar, ambas as versões do sistema Ethnic Lyrics Fetcher alcançaram as taxas mais elevadas de F-Measure, sendo 91,44% para o ELF_{ST+A} e 98,34% para o ELF_{ST} . A diferença entre o desempenho de ELF_{ST+A} e o ELF_{ST} está no fato de que, na música latina, muitos autores fazem covers de canções interpretadas por artistas conhecidos. Por este motivo, em muitos casos, quando a música na LMD é cantada por uma banda pouco famosa o ELF_{ST+A} pode não encontrar as letras, mesmo que a letra esteja amplamente disponível sob o nome do artista famoso. Em quarto lugar, o F-measure alcançado pelo ELF_{ST} é maior do que pesquisar as letras manualmente (88,41%) ou usar todas os doze websites de letras juntos (87,94%). Isso demonstra a viabilidade do uso do sistema ELF para diferentes aplicações e pesquisas da área de MIR.

6. Conclusões

A tarefa de recuperação e extração automática de letras a partir da web não é trivial. Apesar de sua importância para diversas aplicações em Music Information Retrieval, muito pouca pesquisa tem sido realizada até hoje. Neste artigo foi apresentado o sistema Ethnic Lyrics Fetcher (ELF), que possui um novo procedimento de detecção e extração de letras de música de páginas na web.

Para avaliar o método de extração do ELF a partir de qualquer website, foi avaliado o seu desempenho contra 12 websites que possuem as letras em uma estrutura conhecida e bem delimitada. Os resultados experimentais mostram que o novo procedimento de extração do sistema deve ser usado com valor θ maior que três.

O sistema desenvolvido também foi avaliado como um sistema de recuperação de letras de músicas da web. Nos experimentos seu desempenho foi comparado com doze extratores desenvolvidos para websites específicos que contém letras de músicas e também com a recuperação manual das letras. Além disso, foram utilizadas duas versões do sistema, uma usando tanto informação do artista quanto do título da música, denominado ELF_{ST+A} e a outra versão utilizando apenas informações do título da música, denominado ELF_{ST}. Os experimentos realizados mostraram que ambas as versões do sistema podem ser utilizadas para a recuperação automática de letras de músicas na web, entretanto o ELF_{ST} possui um melhor desempenho.

Como trabalhos futuros serão realizados experimentos com o sistema desenvolvido em diferentes bases de dados musicais, visando trabalhar com outros idiomas.

Referências

Baumann, S. and Hummel, O. (2003). Using cultural metadata for artist recommendations. In *Proc. of the 3rd Int. Conf. on Web Delivering of Music*, pages 138–141.

Downie, J. S. and Cunningham, S. J. (2002). Toward a theory of music information retrieval queries: System design implications. In *Proc. of the 3rd Int. Society for Music Information Retrieval Conference (ISMIR)*, pages 299–300.

Geleijnse, G. and Korst, J. (2006). Efficient lyrics extraction from the web. In *Proc. of the 7th Int. Society for Music Information Retrieval Conference (ISMIR)*, pages 371–372.

Hu, X. and Downie, J. S. (2010). When lyrics outperform audio for music mood classification: A feature analysis. In *Proc. of the 11th Int. Society for Music Information Retrieval Conference (ISMIR)*, pages 619–624.

Hu, X., Downie, J. S., and Ehmann, A. F. (2009). Lyric text mining in music mood classification. In *Proc. of the 10th Int. Society for Music Information Retrieval Conference (ISMIR)*, pages 411–416.

Knees, P., Schedl, M., and Widmer, G. (2005). Multiple lyrics alignment: Automatic retrieval of song lyrics. In *Proc. of the 6th Int. Society for Music Information Retrieval Conference (ISMIR)*, pages 564–569.

Li, T. and Ogihara, M. (2004). Music artist style identification by semi–supervised learning from both lyrics and content. In *Proc. of the 12th annual ACM Int. Conf. on Multimedia*, pages 364–367.

Logan, B., Kositsky, A., and Moreno, P. (2004). Semantic analysis of song lyrics. In *Proc. of the IEEE Int. Conf. on Multimedia and Expo*, pages 827–830.

Macrae, R. and Dixon, S. (2012). Ranking lyrics for online search. In *Proc. of the 13th Int. Society for Music Information Retrieval Conference (ISMIR)*, pages 362–366.

Mayer, R., Neumayer, R., and Rauber, A. (2008). Rhyme and style features for musical genre classification by song lyrics. In *In Proc. of the 9th Int. Conf. on Music Information Retrieval (ISMIR)*, pages 337–342.

Mayer, R. and Rauber, A. (2011). Musical genre classification by ensembles of audio and lyrics features. In *Proc. of the 12th Int. Society for Music Information Retrieval Conference (ISMIR)*, pages 675–680.

Salton, G. and Buckley, C. (1998). Term weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Shamma, D. A., Pardo, B., and Hammond, K. J. (2005). Musicstory: a personalized music video creator. In *Proc. of the 13th Annual ACM Int. Conf. on Multimedia*, pages 563–566.

Silla Jr., C. N., Koerich, A. L., and Kaestner, C. A. A. (2008). The latin music database. In *Proc. of the 9th Int. Conf. on Music Information Retrieval (ISMIR)*, pages 451–456.

Zaanen, M. V. and Kanters, P. (2010). Automatic mood classification using tf*idf based on lyrics. In *Proc. of the 11th Int. Society for Music Information Retrieval Conference (ISMIR)*, pages 75–80.