EURASIP Journal on
Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

The Ethnic Lyrics Fetcher tool

Rafael P Ribeiro, Murilo AP Almeida and Carlos N Silla Jr*

Abstract

The task of automatic retrieval and extraction of lyrics from the web is of great importance to different Music Information Retrieval applications. However, despite its importance, very little research has been carried out for this task. For this reason, in this paper, we present the Ethnic Lyrics Fetcher (ELF), a tool for Music Information Researchers, which has a novel lyrics detection and extraction mechanism. We performed two experiments to evaluate ELF's lyrics extraction mechanism and its performance as a lyrics fetcher tool. Our experimental results show that ELF performs better than the current state of the art, using website-specific lyrics fetchers or manually retrieving the lyrics from the web.

Keywords: Lyrics retrieval; Ethnic lyrics; Music information retrieval

1 Introduction

Song lyrics are used in many different types of music information retrieval (MIR) applications. Examples of MIR applications with song lyrics are Music Mood Identification [1,2], Music Genre [3] and Mood [4] Classification, Artist Similarity [5,6], Artist Style Identification [7], Enhancing Query By Humming Systems [8,9], Automatic Recognition of Lyrics in Singing [10], among others [11-13]. Furthermore, the authors of [14] have shown that lyrics (at least in part) are the second most used in a MIR system and are also the second most desired information on such system.

The need for automated lyrics retrieval and extraction systems has been reported in several research papers. For example, in [15], the authors point out that due to some limitations on how they received their dataset (a list of song artist and song titles without blank spaces, e.g. ironmainden - fearofthedark), they were limited to using 5,631 lyrics out of 10,000 songs they had available.

In [16], the authors used a specialized online lyrics library (Leo's Lyrics); however, they mentioned that in their experiments, direct searching did not always work for cover songs. For example, the system would only find the lyrics for the song 'Smells Like Teen Spirit' if the song artist provided was 'Nirvana'. It would not find the lyrics if they provided 'Tori Amos' (who has recorded a cover of the song) as the song artist.

In the work of [17], the authors noted that their simple lyrics mining procedure failed to filter out all the non-text lyrics. In their approach, they examined the line of each html page to determine whether it contained lyrics-like text or not and then selected the longest sequence of lyrics-like text lines in the page. Any lyrics that were less than three lines or over 200 lines long were discarded.

In the work of [18], the authors created a lyrics dataset manually for the task of music genre classification. However, despite having available a private collection of about 12,000 songs, they only selected (randomly) 30 to 45 songs from each of the ten music genres available because of the inexistence of reliable automated tools to retrieve the lyrics for the whole collection.

Despite the importance of song lyrics to the different types of MIR applications and systems and the problems reported by several authors [15-18], to the best of the authors' knowledge (as we shall see in Section 2), very little research has been carried out in the area of the automated lyrics retrieval from the web [19,20].

The main contributions of this paper are the Ethnic Lyrics Fetcher (ELF) System[a] (presented in Section 3) and its novel lyrics detection and extraction procedure (presented in Section 4). It should be noted that our system is named as Ethnic Lyrics Fetcher in the same sense the word 'ethnic' is used in Computational Ethnomusicology [21], i.e., our system is potentially able to retrieve lyrics for all of the world's music. In order to evaluate ELF,

*Correspondence: carlosjunior@utfpr.edu.br
Computer Music Technology Laboratory at the Federal University of Technology of Paraná, Av. Alberto Carazzai, 1.640, Cornélio Procópio 86300-000, Brazil

we have performed two experiments. In the first experiment (Section 5), we evaluate the effectiveness of ELF's novel lyrics extraction algorithm. In the second experiment (Section 6), we evaluate the effectiveness of ELF as an automated tool to retrieve lyrics from the web. Finally, in Section 7, we present the conclusions of this work and future research directions.

2 Related work

The task of automatically retrieving lyrics from the web presents two problems: The first problem is how to detect whether or not a given webpage contains the lyrics of a given song. The second problem is how to extract only the lyrics content from a webpage that contains the lyrics of a given song.

It is possible to avoid these two problems by using an online lyrics website that has an internal lyrics database such as LeosLyrics [22]. However, this approach has one main limitation, i.e., no matter how big the lyrics collection of a given website is, it will still have a limited number of song lyrics which may not contain the song lyrics the user is interested in. This is particularly true for new and/or non-mainstream songs.

Another approach that has been used by some researchers, in order to avoid these two problems, is to develop a website-specific automated lyrics fetcher, such as jLyrics [23]. This approach avoids the issue of manually retrieving and storing the desired lyrics but it is still limited to retrieving only the lyrics available in the website it was designed for. Also, in order to develop such an automated lyrics fetcher, it is necessary to know the website structure in order to correctly locate and extract the lyrics content from the website.

To the best of the authors' knowledge, the only reported studies that are related to dealing with the two problems related to the task of automatic lyrics retrieval from the web are [19,20]. The work of Knees et al. [20] is possibly the first reported work that deals with the problem of retrieving lyrics from the web. However, the main objective of their work was to use sequence alignment algorithms to produce one robust (i.e., typo-free) lyric for a given song. In [20], the authors used the Google search engine with queries containing the following information: song artist, song title, and the keyword 'lyric'. In their pre-processing phase, they removed all HTML tags and links. Since, the authors were dealing with a different task, they could retrieve a collection of webpages and use co-relation methods in order to determine which webpages might have contained the lyrics of a given song.

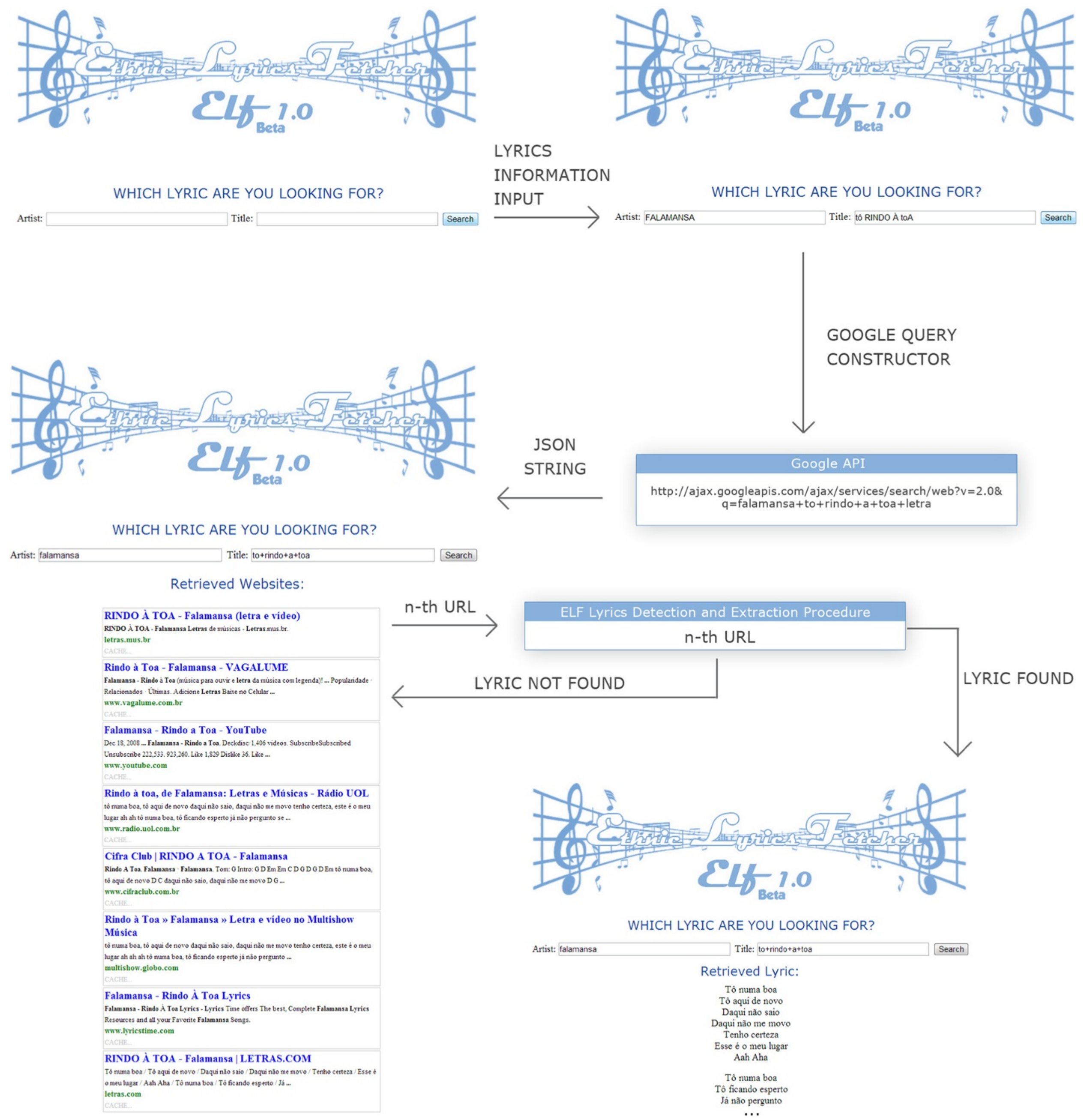
In the work of [19], their main goal is to retrieve the lyrics from a given song. In order to search for lyrics, the authors also use the Google search engine and store the first 40 retrieved results of the query. Then, for each webpage, they use a regular expression-based approach to extract the lyrics content from the webpages.

It should be noted that in both works [19,20], the authors used the same database of 258 songs, and although both papers mention the Evil Lyrics software [24], they did not make any experimental comparisons with it. The Evil Lyrics software is a plug-in used by several digital music players such as Winamp and iTunes. Although all the implementation details of the software are not available, the existing information about the software allows us to classify it as a hybrid approach to retrieve lyrics from the web. In the version of the software used in this work, the Evil Lyrics plug-in had a list of 437 websites which we assume that it knows their structure. If the lyrics are not found in one of these websites, it then searchers the web for the lyrics; however, no information is available about how this web search is performed by the software.

3 ELF system overview

Figure 1 presents an overview of the ELF system. The ELF system works as follows:

1. In the first step, the ELF system needs information about the song lyrics the user wants to retrieve. The ELF system requires that the user provides the song title information (mandatory) and song artist information (optional).
2. In the second step, the ELF system pre-processes the song title and song artist information by applying case folding (e.g.,'É' becomes 'é') and accent removal (e.g., 'é' becomes 'e').
3. In the third step, the ELF system creates a query for the search engine. In this paper, we used the Google search engine, and the search queries have been constructed by concatenating the song title and artist information and by replacing all the white spaces in the resulting string with the plus (+) symbol. We also add to the search query a special keyword which is the the word 'lyric' in the language of the lyric the user is searching for.
4. In the fourth step, the ELF system process the JSON string returned by the Google application programming interface (API) to create a list of websites which may contain the lyrics the user is interested in.
5. In the fifth step, the ELF system uses its novel lyrics detection and extraction mechanism in order to identify if a given website contains a song lyrics or not.
   If the website contains the song lyrics, the ELF system returns it to the user. Otherwise, it analyzes the next website on the list created in the fourth step.



Artist: FALAMANSA Title: tô RINDO À toA

http://ajax.googleapis.com/ajax/services/search/web?v=2.0&q=falamansa+to+rindo+a+toa+letra

Artist: falamansa Title: to+rindo+a+toa

RINDO À TOA - Falamansa (letra e vídeo)
RINDO À TOA - Falamansa Letras de músicas - Letras.mus.br.
letras.mus.br
CACHE...

Rindo à Toa - Falamansa - VAGALUME
Falamansa - Rindo à Toa (música para ouvir e letra da música com legenda)! ... Popularidade · Relacionados · Últimas. Adicione Letras Baixe no Celular ...
www.vagalume.com.br
CACHE...

Falamansa - Rindo a Toa - YouTube
Dec 18, 2008 ... Falamansa - Rindo a Toa. Deckdisc·1,406 videos. SubscribeSubscribed Unsubscribe 222,533. 923,260. Like 1,829 Dislike 36. Like ...
www.youtube.com
CACHE...

Rindo à toa, de Falamansa: Letras e Músicas - Rádio UOL
tô numa boa, tô aqui de novo daqui não saio, daqui não me movo tenho certeza, este é o meu lugar ah ah tô numa boa, tô ficando esperto já não pergunto se ...
www.radio.uol.com.br
CACHE...

Cifra Club | RINDO A TOA - Falamansa
Rindo A Toa. Falamansa · Falamansa. Tom: G Intro: G D Em Em C D G D G D Em tô numa boa, tô aqui de novo D C daqui não saio, daqui não me movo D G ...
www.cifraclub.com.br
CACHE...

Rindo à Toa » Falamansa » Letra e vídeo no Multishow Música
tô numa boa, tô aqui de novo daqui não saio, daqui não me movo tenho certeza, este é o meu lugar ah ah ah tô numa boa, tô ficando esperto já não pergunto ...
multishow.globo.com
CACHE...

Falamansa - Rindo À Toa Lyrics
Falamansa - Rindo À Toa Lyrics - Lyrics Time offers The best, Complete Falamansa Lyrics Resources and all your Favorite Falamansa Songs.
www.lyricstime.com
CACHE...

RINDO À TOA - Falamansa | LETRAS.COM
Tô numa boa / Tô aqui de novo / Daqui não saio / Daqui não me movo / Tenho certeza / Esse é o meu lugar / Aah Aha / Tô numa boa / Tô ficando esperto / Já ...
letras.com
CACHE...

Artist: falamansa Title: to+rindo+a+toa

Tô numa boa
Tô aqui de novo
Daqui não saio
Daqui não me movo
Tenho certeza
Esse é o meu lugar
Aah Aha

Tô numa boa
Tô ficando esperto
Já não pergunto
...

Figure 1 Overview of the Ethnic Lyrics Fetcher (ELF) system.

Note that the ELF system can work with lyrics in any language that can be encoded using UTF-8.

4 ELF's novel lyrics detection and extraction procedure

The ELF system employs a novel lyric detection and extraction procedure. The usual approach behind current lyrics extraction is to assume that the lyrics of a song are somewhere within the webpage where there are ends of line tags (i.e., <BR> tags) [19]. One of the differences between ELF novel lyrics extraction procedure and the existing approaches is the way ELF uses all the html tags (including the <BR> tags) to locate the lyrics within any webpage.

Before attempting to extract the lyrics from the webpage, the system verifies if the webpage does not come from the list of websites that are known to not have lyrics.

If this is not the case, ELF's lyric extraction procedure works as follows:

- The first step is to store the retrieved webpage into a string.
- The second step is to analyze the string looking for opening html tags with the exception of '<br' (line break), '<p' (paragraph), '</' (closing tag), and '<!-' (html comment). For each opening html tag found, do the following:
  1. Create a new position in the lyric extraction vector;
  2. Add all the content of the webpage until a new opening html tag is found.
- After the webpage has been converted into the lyric extraction vector, ELF analyzes the lyric extraction vector counting the number of '<BR' tags in each position of the vector.
- In order to decide what to output, the ELF lyric extraction procedure has threshold (θ), and it considers that any position in the lyric extraction vector that has a number of '<BR' tags higher than θ is the song lyrics and output it to the user.
- If the song lyric to be shown to the user is empty, ELF considers that the lyric was not detected on the webpage given as input.

Figure 2 presents an example of the lyrics extraction procedure used by ELF with different θ values for the song 'To rindo a toa' by the Brazilian band 'Falamansa'.

5 Evaluating ELF's lyrics extraction procedure

In this section, we are interested in answering the following questions by using controlled experiments: (a) How well does ELF's lyric extraction procedure (with different θ values) performs when compared against other approaches? (b) What value should be used as the θ value of ELF's lyric extraction procedure?

5.1 Experimental settings

In order to evaluate ELF's lyric extraction procedure, we manually verified 12 different websites which have lyrics for songs of the Latin Music Database (LMD) [25]. Each of the websites has a different way of marking the song lyrics content. Table 1 presents the list of websites we used in this experiment along with specific tag delimiters each website uses to identify the song lyrics content within their webpages.

In order to evaluate ELF's lyric extraction procedure, we provide ELF with the full html page and compare its output (using different θ values) with the gold standard of each website. The gold standard of each website is obtained by creating a website-specific lyrics extractor which knows beforehand the special delimiter tags used in each website.

In order to measure how similar the output of ELF's lyric extraction procedure (using different θ values) is with that of the gold standard in each website, we use the document similarity measure from the information retrieval field, known as cosine similarity [26]. The cosine similarity measure computes how similar two documents (lyrics in our case) are. The measure provides a value between 0 and 1. The higher the value, the more similar the two documents are. We also use the method presented in [19], which is, to the best of the authors' knowledge, the reported state of the art and the lyrics obtained by using the Evil Lyrics software [24].

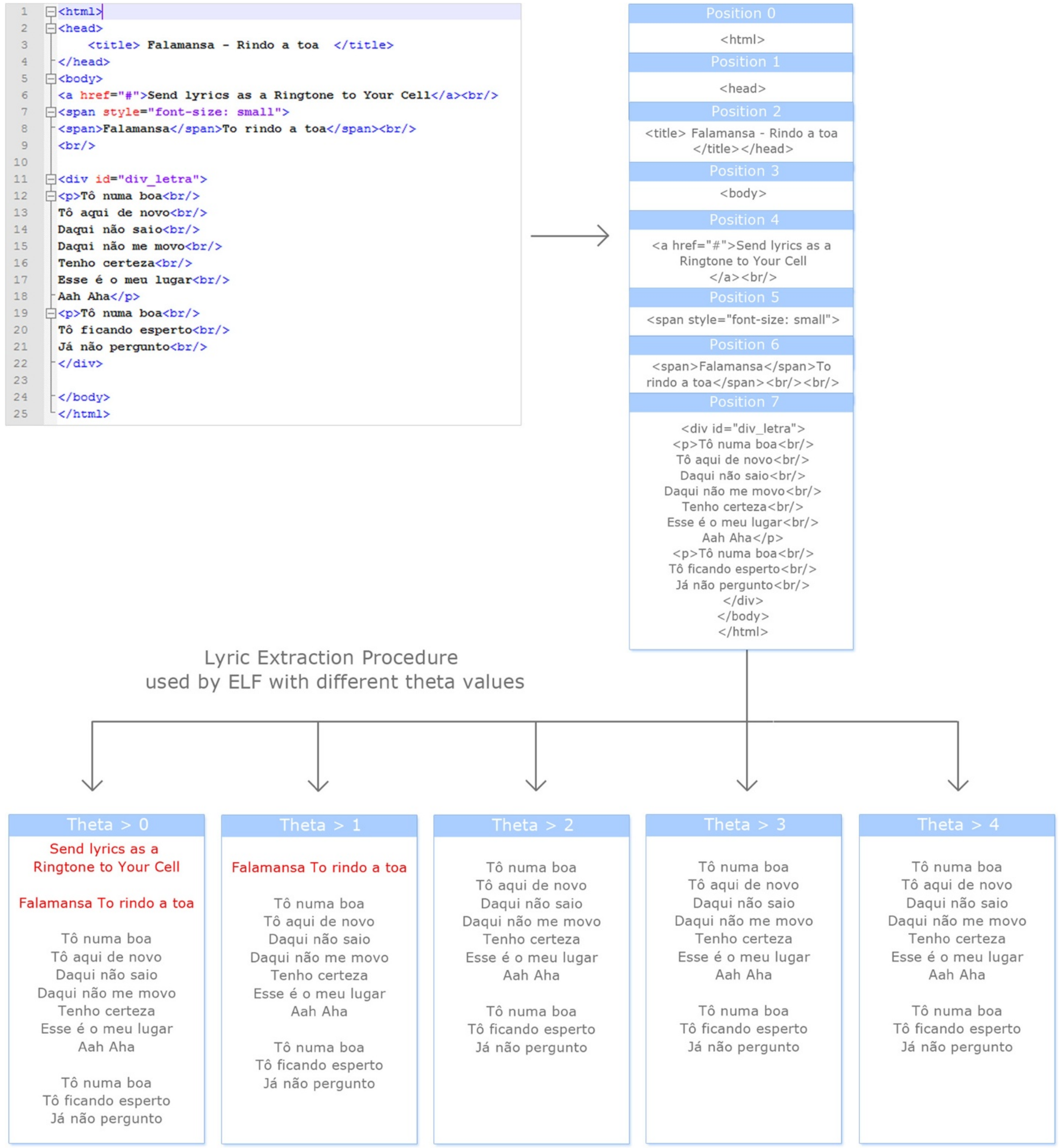
5.2 Experimental results

Table 2 presents the experimental results of the lyrics extraction mechanism used by ELF with different θ values, [19] and [24] for the 12 different websites presented in Table 1.

The analysis of Table 2 shows some interesting results. First is that for higher θ values, a higher value of overall similarity is obtained. This can be explained by the fact that a low value will simply return all the webpage content that has a <BR> tag in the same position of the lyric extraction vector.

Second, we manually inspected the lyrics with higher θ values in order to understand why the similarity was not 1.00. Interestingly, we observed that some of the websites have among their structured lyrics some kind of promotional message such as 'Ringtone - Send this ringtone to your cell phone!' which was removed by the lyric extraction procedure used by ELF. In our experiments, this happened with the lyricsmania.com, 1songlyrics.com, lyricshall.com, and lyricsreg.com websites. Another observation that we made was that some websites such as lyricsoncall.com and 6lyrics.com have in the middle of the song lyrics the name of their website. This information was also removed by the lyric extraction procedure used by ELF. In the case of the vounessa.com.br website, it included the fields composer and music genre at the end of the lyrics which were also filtered out by ELF's lyric extraction procedure.

Third, based on our experimental results, it seems that there is a trade-off between using a low value or a high value for θ. A low value might return everything, while a high value may filter out more than it should. Therefore, for the experiments in the next section, we employed the value of $\theta > 3$. Fourth, despite the value of θ used, the novel lyrics detection and extraction procedure used by ELF is better than the method



```
<html>
<head>
    <title> Falamansa - Rindo a toa  </title>
</head>
<body>
<a href="#">Send lyrics as a Ringtone to Your Cell</a><br/>
<span style="font-size: small">
<span>Falamansa</span>To rindo a toa</span><br/>
<br/>

<div id="div_letra">
<p>Tô numa boa<br/>
Tô aqui de novo<br/>
Daqui não saio<br/>
Daqui não me movo<br/>
Tenho certeza<br/>
Esse é o meu lugar<br/>
Aah Aha</p>
<p>Tô numa boa<br/>
Tô ficando esperto<br/>
Já não pergunto<br/>
</div>

</body>
</html>
```

Position 0
<html>
Position 1
<head>
Position 2
<title> Falamansa - Rindo a toa </title></head>
Position 3
<body>
Position 4
<a href="#">Send lyrics as a Ringtone to Your Cell </a><br/>
Position 5
<span style="font-size: small">
Position 6
<span>Falamansa</span>To rindo a toa</span><br/><br/>
Position 7
<div id="div_letra"> <p>Tô numa boa<br/> Tô aqui de novo<br/> Daqui não saio<br/> Daqui não me movo<br/> Tenho certeza<br/> Esse é o meu lugar<br/> Aah Aha</p> <p>Tô numa boa<br/> Tô ficando esperto<br/> Já não pergunto<br/> </div> </body> </html>
Lyric Extraction Procedure used by ELF with different theta values
Theta > 0
Send lyrics as a Ringtone to Your Cell
Falamansa To rindo a toa
Tô numa boa Tô aqui de novo Daqui não saio Daqui não me movo Tenho certeza Esse é o meu lugar Aah Aha
Tô numa boa Tô ficando esperto Já não pergunto
Theta > 1
Falamansa To rindo a toa
Tô numa boa Tô aqui de novo Daqui não saio Daqui não me movo Tenho certeza Esse é o meu lugar Aah Aha
Tô numa boa Tô ficando esperto Já não pergunto
Theta > 2
Tô numa boa Tô aqui de novo Daqui não saio Daqui não me movo Tenho certeza Esse é o meu lugar Aah Aha
Tô numa boa Tô ficando esperto Já não pergunto
Theta > 3
Tô numa boa Tô aqui de novo Daqui não saio Daqui não me movo Tenho certeza Esse é o meu lugar Aah Aha
Tô numa boa Tô ficando esperto Já não pergunto
Theta > 4
Tô numa boa Tô aqui de novo Daqui não saio Daqui não me movo Tenho certeza Esse é o meu lugar Aah Aha
Tô numa boa Tô ficando esperto Já não pergunto

Figure 2 An example of the lyric extraction procedure used by ELF with different theta values.

proposed in [19] or the method used by the Evil Lyrics software.

6 Evaluating ELF as a lyrics retrieval system

In this section, we are interested in evaluating how well ELF performs as an online lyrics retrieval system. That is, we want to evaluate how well it performs by using a search engine API and its lyrics detection mechanism.

6.1 Experimental settings

In order to evaluate ELF as a online lyrics retrieval system, we have provided ELF with the song artist and song title information for all the songs available in the LMD [25]. The motivation behind using the LMD is that it contains songs with lyrics in English, Spanish, Brazilian Portuguese, and Spanglish (i.e., lyrics with words in English and Spanish).

Table 1 Lyrics websites and their lyrics identifier tags

Website	Lyrics-content starting tag	Lyrics-content enclosing tag
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>`

In this experiment, we have used the Google API as ELF's search engine for querying possible websites that may contain the lyrics for the song the user is looking for. As a baseline for ELF, we have implemented 12 website-specific lyrics fetchers (those listed in Table 1) that have a website-specific URL encoding and the lyrics-specific information. Table 3 presents the website structure for each of the 12 lyrics websites.

We also evaluate two versions of the ELF system. One version uses the song title and song artist information, referred to as $\mathrm{ELF}_{\mathrm{ST+A}}$, and the other version uses only the song title information, referred to as $\mathrm{ELF}_{\mathrm{ST}}$.

It should be noted that in the experiments reported in this section, the following websites have been listed in ELF as non-lyrics websites: 'blogspot', 'wikipedia', and 'youtube'.

Additionally, in order to compare ELF with the current state-of-the-art technology in online lyrics fetching, we also present the performance of the Evil Lyrics software [24]. To make a fair comparison with ELF, we have also run the experiments with Evil Lyrics using song title and song artist information (referred to as $\mathrm{EvilLyrics}_{\mathrm{ST+A}}$) or using only song title information (referred to as $\mathrm{EvilLyrics}_{\mathrm{ST}}$).

6.2 Experimental results

Let us now evaluate ELF as a lyrics retrieval system by comparing the total number of lyrics retrieved by it with the other approaches. Table 4 presents the experimental results (broken down by music genre) when using $\mathrm{ELF}_{\mathrm{ST+A}}$, $\mathrm{ELF}_{\mathrm{ST}}$, the other 12 lyrics websites, and the two configurations of the Evil Lyrics software to retrieve lyrics for the Latin Music Database.

Table 2 Evaluation of ELF's and the lyrics extraction approaches used in [19] and [24]

Website	ELF with $\theta > 0$	ELF with $\theta > 1$	ELF with $\theta > 2$	ELF with $\theta > 3$	ELF with $\theta > 4$	[19]	[24]
letras.mus.br	0.9226	1.0	1.0	0.9998	0.9998	0.7199	0.3556
vagalume.com.br	0.9151	0.9191	1.0	1.0	1.0	0.7202	0.3434
songlyrics.com	0.9753	0.9761	0.9767	0.9780	0.9780	0.6389	0.3753
6lyrics.com	0.9141	0.9800	0.9972	0.9972	0.9966	0.5808	0.3721
vounessa.com.br	0.9996	0.9996	0.9995	0.9996	0.8960	0.6379	0.3409
lyricmania.com	0.9552	0.9617	0.9618	0.9617	0.9617	0.6295	0.3703
lyricshall.com	0.9702	0.9706	0.9706	0.9706	0.9162	0.8781	0.1893
lyricsoncall.com	0.9668	0.9666	0.9666	0.9787	0.9787	0.5932	0.3232
lyricspedia.com	0.9974	0.9967	0.9920	0.9920	0.9905	0.6725	0.3756
lyricsreg.com	0.9533	0.9686	0.9689	0.9730	0.9730	0.7305	0.3474
moron.nl	0.8847	0.9890	1.0	1.0	1.0	0.8088	0.3907
themusic-world.com	0.9905	0.9890	0.9922	0.9923	0.9922	0.7096	0.3646
Average	0.9537	0.9764	0.9855	0.9869	0.9736	0.6933	0.3457

Table 3 Fixed structure of the 12 lyrics websites used as baselines

Website	Structure
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/$song title$`

In Table 4, from row 1 to row 12, we present the results of using the individual lyrics websites and their lyrics page structure to retrieve the song lyrics from the LMD. The analysis of these results shows that the Brazilian websites vagalume.com.br (1,569 lyrics retrieved) and vounessa.com.br (1,944 lyrics retrieved) contain the highest number of lyrics for songs from the LMD among the individual websites. To some extent, this is an expected result since six out of the 10 music genres in the LMD are Brazilian music genres.

However, we were also interested in knowing if these websites could be combined into one lyrics retrieval system in order to obtain a higher number of lyrics. The result of this combination is shown in Table 4 row 13. The combined approached of the individual websites retrieved 2,480 lyrics for the songs in the LMD. It should be noted

Table 4 Number of lyrics retrieved by the different approaches for each music genre in the LMD

Row number	Approach	A	Ba	Bo	F	G	M	P	Sa	Se	T	Total
1	1songlyrics.com	28	94	71	7	0	53	2	102	0	24	381
2	6lyrics.com	17	107	112	2	0	70	9	96	2	101	516
3	letras.mus.br	132	98	127	229	126	39	168	140	14	168	1,241
4	lyricmania.com	21	98	72	8	0	52	3	109	0	23	386
5	lyricshall.com	0	0	2	0	0	0	0	0	0	0	2
6	lyricsoncall.com	46	80	67	10	6	42	36	85	27	23	422
7	lyricspedia.com	13	43	61	11	0	35	1	77	0	22	263
8	lyricsreg.com	11	23	18	1	0	6	0	32	0	4	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	224	129	257	113	1,569
12	Vounessa.com.br	247	210	239	259	179	171	0	230	143	266	1,944
13	All 12 websites	268	238	251	277	208	185	237	253	275	288	2,480
14	Manually retrieved	292	251	268	296	178	101	299	172	310	336	2,503
15	EvilLyrics$_{ST+A}$	289	283	275	311	272	292	288	257	298	298	2,863
16	ELF$_{ST+A}$	298	247	233	307	272	227	301	267	309	386	2,847
17	EvilLyrics$_{ST}$	294	272	292	289	298	289	278	255	289	379	2,935
18	ELF$_{ST}$	303	303	301	313	300	291	301	299	308	397	3,116
19	LMD total	304	308	302	315	306	307	301	303	310	404	3,160
20	LMD instrumental songs	0	0	0	6	24	0	2	0	0	54	86

A, Axé; Ba, Bachata; Bo, Bolero; F, Forró; G, Gaúcha; M, Merengue; P, Pagode; Sa, Salsa; Se, Sertaneja; T, Tango.

that among these results (individual or the combined approach), the webpages would be extracted using the known webpage structure and, as discussed in section 5.2, may contain some additional content besides the song lyrics (such as some type of advertisement).

In Table 4 row 14, we present the results of the lyrics being manually searched (using the Google search engine) and retrieved by the authors of this paper. It should be noted that the authors are familiar with all the languages present in the songs of the LMD and that the method used for manually gathering the lyrics consisted on, first, using well-known websites for the particular music genres presented in the LMD such as letras.mus.br and then, on a second stage, the authors searched for lyrics using song title and artist information using Google search engine. This process has led to the retrieval of 2,503 lyrics of songs from the LMD.

In Table 4 row 15, we present the results for the Evil Lyrics software using song artist and song title information (EvilLyrics$_{\mathrm{ST+A}}$). This version of Evil Lyrics has retrieved 2,863 lyrics for songs from the LMD. In order to evaluate if all the lyrics retrieved were the actual lyrics of the songs, we have manually inspected each one of them. After this analysis, we verified that 422 out of the 2,847 retrieved lyrics were not actual lyrics but some other type of information such as artist's biographies and discographies. Therefore, in total, the EvilLyrics$_{\mathrm{ST+A}}$ correctly retrieved 2,425 lyrics.

In Table 4 row 16, we present the results for ELF$_{\mathrm{ST+A}}$ (with $\theta > 3$). This version of ELF has managed to retrieve 2,847 lyrics from songs in the LMD. This is a very interesting result, as these lyrics were retrieved automatically by the system. In order to evaluate if all the lyrics retrieved were the actual lyrics of the songs, we have manually inspected each one of them. After this analysis, we verified that 204 out of the 2,847 retrieved lyrics were not actual lyrics but some other type of information. Therefore, in total, the ELF$_{\mathrm{ST+A}}$ correctly retrieved 2,643 lyrics.

Another aspect that we analyzed was which lyrics websites were being used by ELF$_{\mathrm{ST+A}}$ to retrieve lyrics. In total, the ELF$_{\mathrm{ST+A}}$ has gathered lyrics from 157 different websites. Due to space limitations, it is not viable to list all the websites here, but it is interesting to note that the top five websites used by ELF$_{\mathrm{ST+A}}$ were letras.mus.br (1,490 lyrics), www.todotango.com (279 lyrics), www.musica.com (224 lyrics), www.vagalume.com.br (91 lyrics), and www.mp3lyrics.org (76 lyrics). Out of these websites, only two of them were known to the authors of this paper before these experiments.

In Table 4 row 17, we present the results for the Evil Lyrics software using only song title information (EvilLyrics$_{\mathrm{ST}}$). This version of Evil Lyrics has retrieved 2,935 lyrics for songs in the LMD. In order to evaluate if all the lyrics retrieved were the actual lyrics of the songs, we have manually inspected each one of them. After this analysis, we verified that 215 out of the 2,935 retrieved lyrics were not actual lyrics but some other type of information. Therefore, in total, the EvilLyrics$_{\mathrm{ST}}$ correctly retrieved 2,720 lyrics.

In Table 4 row 18, we present the results for using ELF$_{\mathrm{ST}}$ (with $\theta > 3$), i.e., using only song title information. This version of ELF has managed to retrieve 3,116 lyrics for songs in the LMD. As with the ELF$_{\mathrm{ST+A}}$, we manually verified if all the lyrics were the actual lyrics of the songs. Our analysis has shown that for ELF$_{\mathrm{ST}}$, only 60 of the retrieved lyrics were not the actual lyrics of the songs. Therefore, the ELF$_{\mathrm{ST}}$ has successfully retrieved 3,056 lyrics of the LMD using 110 different websites. The top five websites used were letras.mus.br (1,454 lyrics), www.musica.com (760 lyrics), www.todotango.com (274 lyrics), www.vagalume.com.br (114 lyrics), and www.lyricsmode.com (38 lyrics).

In Table 4 row 19, we present the total number of songs available within each music genre of the LMD, and in row 20, we present the number of instrumental songs in the LMD. After the analysis of this information, it becomes clear that the maximum number of lyrics that could be retrieved correctly for the songs in the LMD is 3,074.

Let us now analyze the results under the information retrieval metrics of precision, recall, and F-measure presented in Table 5. The analysis of Table 5 shows some interesting results. First, it should be noted that for the

Table 5 Evaluation of the different approaches for retrieving lyrics from the web

Approach	Precision	Recall	*F*-measure
lyricshall.com	1	0.000632911	0.001265022
moron.nl	1	0.009177215	0.018187520
lyricsreg.com	1	0.030063291	0.058371736
themusic-world.com	1	0.045253165	0.086587950
lyricspedia.com	1	0.083227848	0.153666375
1songlyrics.com	1	0.120569620	0.215193448
lyricmania.com	1	0.122151899	0.217710096
lyricsoncall.com	1	0.133544304	0.235622557
6lyrics.com	1	0.163291139	0.280739935
letras.mus.br	1	0.392721519	0.563962736
vagalume.com.br	1	0.496518987	0.663565236
vounessa.com.br	1	0.615189873	0.761755486
EvilLyrics$_{\mathrm{ST+A}}$	0.847013622	0.906012658	0.875520318
All 12 websites	1	0.784810127	0.879432624
Manually retrieved	1	0.792405063	0.884180791
ELF$_{\mathrm{ST+A}}$	0.928345627	0.900949367	0.914442348
EvilLyrics$_{\mathrm{ST}}$	0.926746167	0.928797468	0.927770684
ELF$_{\mathrm{ST}}$	0.980744544	0.986075949	0.983403021

specific websites, although their precision is always 1.0, there is a need to know their website structure and their webpage structure in order to correctly fetch lyrics which is unfeasible in practice.

Second, with the exception of the vagalume.com.br and vounessa.com.br websites, all the specific websites provided low recall values as they only retrieved a small portion of the number of lyrics requested. Third, the Evil Lyrics software using both song artist and song title information achieved an F-measure similar to the baseline 12-website combined approach and the manually retrieval of the songs. This might have happened because the Evil Lyrics software is a hybrid approach to web-based lyrics retrieval, since it has a list of websites which we assume it knows their structure. In the version of the software used in the experiments, the websites letras.mus.br and www.vagalume.com.br were present in the list of known websites. For this reason, we believe that it might have used this websites first when searching for the lyrics, since it already knew their structure, hence its similar performance with the 12-website combined approach.

Fourth, both versions of the Ethnic Lyrics Fetcher system have achieved high F-measure rates, being 0.9144 for the $\mathrm{ELF}_{\mathrm{ST+A}}$ and 0.9834 for the $\mathrm{ELF}_{\mathrm{ST}}$. The difference between the performance of $\mathrm{ELF}_{\mathrm{ST+A}}$ and the $\mathrm{ELF}_{\mathrm{ST}}$ lies in the fact that in Latin music, many authors perform covers of songs performed by well-known artists. Therefore, in many cases when the song in the LMD is sung by a less famous band, the $\mathrm{ELF}_{\mathrm{ST+A}}$ may not find the lyrics, even though the lyrics is widely available under the famous artist name. This might also explain why the $\mathrm{EvilLyrics}_{\mathrm{ST}}$ was the second best performing approach in our experiments.

Fifth, the F-measure achieved by the $\mathrm{ELF}_{\mathrm{ST}}$ is higher than those of all the other approaches: manually searching for the lyrics (0.8841), using all the 13 lyrics websites together (0.8794), or using the $\mathrm{EvilLyrics}_{\mathrm{ST}}$ (0.9277). This demonstrates the feasibility of using the ELF system to automatically retrieve lyrics for different music information retrieval tasks and research experiments.

7 Conclusions

The task of automatic lyrics retrieval and extraction from the web is a non-trivial task. Despite its importance to several music information retrieval applications, very little research has been carried out so far. For this reason, in this paper, we have presented the ELF tool and its novel lyrics detection and extraction mechanism.

In order to evaluate ELF's novel lyric extraction method from any webpage, we have evaluated its performance against 12 lyrics websites with known structure for delimiting the lyrics content and also against the method proposed by [19]. Our experimental results show that ELF's novel lyrics extraction procedure is better than the method prosed by [19] and that it should be used with a θ value higher than 3 ($\theta > 3$).

We have also evaluated ELF as a lyrics retrieval system. In our experiments, we compared its performance with 12 site-specific lyrics fetchers, the manual retrieval of the lyrics, and also with two versions of the Evil Lyrics software [24]. In these experiments, two versions of ELF were used: the first version ($\mathrm{ELF}_{\mathrm{ST+A}}$) uses both the song artist and song title information, while the second version ($\mathrm{ELF}_{\mathrm{ST}}$) uses only the song title information. Our experiments have shown that both versions of ELF can be used as lyrics retrieval tools, but that the $\mathrm{ELF}_{\mathrm{ST}}$ has a better performance (measured by the F-measure) because it is capable of correctly fetching lyrics for cover songs. Furthermore, $\mathrm{ELF}_{\mathrm{ST}}$ has outperformed all the other approaches, even the Evil Lyrics software which is currently one of the most well-known and used plug-ins in several digital music player softwares. As future research, we intend to perform experiments with the Ethnic Lyrics Fetcher system on other music databases with songs in other languages.

Endnote

[a] Available online at http://music.cp.utfpr.edu.br/elf/.

Abbreviations

API: Application Programming Interface; ELF: Ethnic Lyrics Fetcher; $\mathrm{ELF}_{\mathrm{ST}}$: Ethnic Lyrics Fetcher using only the song title information; $\mathrm{ELF}_{\mathrm{ST+A}}$: Ethnic Lyrics Fetcher using the song title and artist information; MIR: music information retrieval.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements
We thank the anonymous reviewers for their insightful and valuable feedback on the earlier versions of this manuscript.

Received: 18 September 2013 Accepted: 14 June 2014
Published online: 04 September 2014

References

1. X Hu, JS Downie, AF Ehmann, Lyric text mining in music mood classification, in *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)* (Kobe, Japan, 26–30 Oct 2009), pp. 411–416
2. X Hu, JS Downie, When lyrics outperform audio for music mood classification: a feature analysis, in *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)* (Utrecht, Netherlands, 9–13 Aug 2010), pp. 619–624
3. R Mayer, A Rauber, Musical genre classification by ensembles of audio and lyrics features, in *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)* (Miami, USA, 24–28 Oct 2011), pp. 675–680
4. B den Brinker, R van Dinther, J Skowronek, Expressed music mood classification compared with valence and arousal ratings. EURASIP J. Audio Speech Music Process. **2012**, Article ID 24 (2012)
5. S Baumann, O Hummel, Using cultural metadata for artist recommendations, in *Proceedings of the 3rd International Conference on Web Delivering of Music* (Leeds, UK, 15–17 Sept 2003), pp. 138–141
6. B Logan, A Kositsky, P Moreno, Semantic analysis of song lyrics, in *Proceedings of the IEEE International Conference on Multimedia and Expo* (Taipei, Taiwan, 27–30 June 2004), pp. 827–830

7. T Li, M Ogihara, Music artist style identification by semi-supervised learning from both lyrics and content, in *Proceedings of the 12th Annual ACM International Conference on Multimedia* (New York, USA, 10–16 Oct 2004), pp. 364–367
8. Z Guo, Q Wang, G Liu, J Guo, Y Lu, A music retrieval system using melody and lyric, in *Proceedings of the IEEE International Conference on Multimedia and Expo Workshops* (Melbourne, Australia, 9–13 July 2012), pp. 343–348
9. C-C Wang, J-SR Jang, W Wang, An improved query by singing/humming system using melody and lyrics information, in *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)* (Utrecht, Netherlands, 9–13 Aug 2010), pp. 45–50
10. A Mesaros, T Virtanen, Automatic recognition of lyrics in singing. EURASIP J. Audio Speech Music Process. **2010**, Article ID 546047 (2010)
11. S Funasawa, H Ishizaki, K Hoashi, Y Takishima, J Katto, Automated music slideshow generation using web images based on lyrics, in *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)* (Utrecht, Netherlands, 9–13 Aug 2010), pp. 63–68
12. JPG Mahedero, A Martinez, P Cano, M Koppenberger, F Gouyon, Natural language processing of lyrics, in *Proceedings of the 13th Annual ACM International Conference on Multimedia* (Singapore, 6–11 Nov 2005), pp. 475–478
13. T O'Hara, N Schuler, Y Lu, DE Tamir, Inferring chord sequence meanings via lyrics: process and evaluation, in *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)* (Porto, Portugal, 8–12 Oct 2012), pp. 463–468
14. JS Downie, SJ Cunningham, Toward a theory of music information retrieval queries: system design implications, in *Proceedings of the 3rd International Conference on Music Information Retrieval* (Paris, France, 13–17 Oct 2002), pp. 299–300
15. MV Zaanen, P Kanters, Automatic mood classification using tf*idf based on lyrics, in *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)* (Utrecht, Netherlands, 9–13 Aug 2010), pp. 75–80
16. DA Shamma, B Pardo, KJ Hammond, Musicstory: a personalized music video creator, in *Proceedings of the 13th Annual ACM International Conference on Multimedia* (Singapore, 6–11 Nov 2005), pp. 563–566
17. R Macrae, S Dixon, Ranking lyrics for online search, in *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)* (Porto, Portugal, 8–12 Oct 2012), pp. 362–366
18. R Mayer, R Neumayer, A Rauber, Rhyme and style features for musical genre classification by song lyrics, in *Proceedings of the 9th International Society for Music Information Retrieval Conferece (ISMIR)* (Philadelphia, USA, 14–18 Sept 2008), pp. 337–342
19. G Geleijnse, J Korst, Efficient lyrics extraction from the web, in *Proceedings of the 7th International Society for Music Information Retrieval Conference (ISMIR)* (Victoria, Canada, 8–12 Oct 2006), pp. 371–372
20. P Knees, M Schedl, G Widmer, Multiple lyrics alignment: automatic retrieval of song lyrics, in *Proceedings of the 6th International Society for Music Information Retrieval Conference (ISMIR)* (London, UK, 11–15 Sept 2005), pp. 564–569
21. G Tzanetakis, A Kapur, WA Schloss, M Wright, Computational ethnomusicology. J. Interdiscip. Music Stud. **1**(2), 1–24 (2007)
22. Leo's Lyrics, http://www.leoslyrics.com/. Accessed 14 Oct 2013
23. jLyrics, http://jmir.sourceforge.net/jLyrics.html. Accessed 14 Oct 2013
24. Evil Lyrics, http://www.evillabs.sk/evillyrics/. Accessed 10 Dec 2013
25. CN Silla Jr, AL Koerich, CAA Kaestner, The Latin Music Database, in *Proceedings of the 9th International Conference on Music Information Retrieval* (Philadelphia, USA, 14–18 Sept 2008), pp. 451–456
26. G Salton, C Buckley, Term weighting approaches in automatic text retrieval. Inf. Process. Manag. **24**(5), 513–523 (1998)

doi:10.1186/s13636-014-0027-4
Cite this article as: Ribeiro *et al.*: The Ethnic Lyrics Fetcher tool. *EURASIP Journal on Audio, Speech, and Music Processing* 2014 **2014**:27.