Using Search Engine for Classification: Does It Still Work?

Sten Govaerts, Nik Corthaut, Erik Duval
Katholieke Universiteit Leuven, Dept. Computer Science
{sten.govaerts, nik.corthaut, erik.duval}@cs.kuleuven.be

Abstract

Genre classification is a key aspect of music descriptions. In 2006, Schedl et al. presented a method for genre classification through web-based co-occurrence analysis. We evaluate whether this method is still valid, given the evolution of the web search technologies. We identify some issues with page count as the main parameter for the analysis in relation with the used genre taxonomies, choice of search engine, etc. We show that the results vary over time. Depending on the required response time, we suggest two different strategies for music genre classification using this method. Finally, we discuss future work.

1. Introduction

In the rockanango project [1], we developed a music player for hotels, restaurants and bars. Specific to our approach is that a user can describe the music he wants by referring to a situation, rather than by defining the usual search criteria on artist, title, etc. Behind the scenes, we rely on almost 40 metadata elements, manually determined by music experts of Aristo Music[1]. Manual annotation is a very time-consuming and thus expensive labor. Currently, the Aristo Music database contains around 60.000 songs. We assist the experts by automating the annotation process for some of the metadata elements [2]. We focus on automatic metadata generation for elements that are most costly to annotate and of highest relevance. For this purpose, we rely on a combination of digital signal processing [3], web-based techniques [4] and external "linked data" sources [5].

In this paper, we focus on musical genre, one of the most relevant metadata elements [6]. The definition of genre is not universally agreed upon and many genre taxonomies exist [6]. Music experts rank genre as the most important parameter for selecting music when creating musical contexts (dynamic playlists based on metadata) [1]. On average, a music expert needs 33 seconds to annotate genre. As figure 1 shows, genre is the fourth most time consuming parameter to annotate. Because of the importance of the parameter and the annotation effort needed, genre is a high priority for automation. In previous work, we already automated another parameter, the origin of an artist [7].

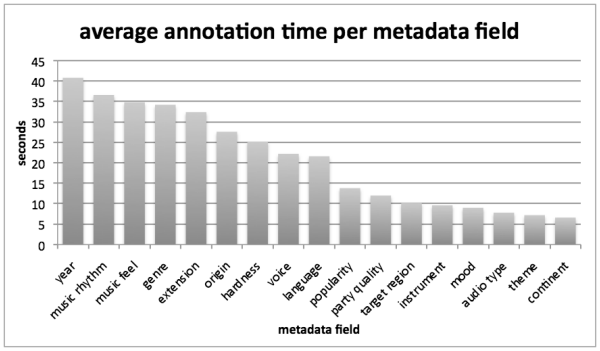

Figure 1. The average time in minutes to manually annotate a metadata field.

Web-based classification can be achieved with different techniques. Examples are the use of (public) API's[2] [7][8]. Alternatively, web pages concerning a band or a song can be analyzed [9]. This paper relies on an analysis of search engine result counts as source for classification [4], which is a very interesting approach for genre classification with good performance. This approach is simple, easy to implement and works for artists and songs that are not covered by external data sources, e.g. for artists without an artist page on Last.fm. Moreover, the technique can be applied more broadly to determine other data elements than genre, e.g. mood [10]. On the other hand, there may be legal issues with automatically querying search engines and there is a risk of becoming too dependent on third-party services that may evolve without any consideration for how we use them. This paper focuses on the performance of the approach rather than on when and how to deploy it.

[1] Aristo Music, http://www.aristomusic.com (viewed on 18 Sept. 2009)

[2] e.g. http://www.discogs.com/, http://the.echonest.com and http://last.fm (viewed on 18 Sept. 2009)

Geleijnse et al. [10] presented a very similar approach to [4] with similar results. We analyzed Schedl's results in some detail, because we initially obtained far worse results, even with the same dataset and genre taxonomy. We only reached 50% accuracy versus the 62% reported by Schedl. When we repeated the analysis, we obtained different results. We analyzed Schedl's results and not Geleijnse's, because of the larger data set. The topic of this paper is to validate whether this approach still works and analyze how it performs on different search engines.

First, we will briefly explain the approach, and then we will elaborate on the setup of our experiments and present the observations we made. Afterwards, we analyze possible causes and strategies to cope with the problems. Finally, we conclude with future work.

2. Music genre classification by web-based co-occurrence analysis

For genre classification on artist level, Schedl et al rely on co-occurrence analysis. This means that a search engine is queried with the combination of the name of an artist and a music genre. The search engine returns a number of pages, in this paper referred to as page count, *pc*. This number is divided by the page count for the individual genres or artists names. This creates conditional probabilities, resulting in probability distributions for artist genres, e.g. the conditional probability for the artist name (*a*) to be found on a web page that mentions the genre name (*g*) can be written formally as follows $p(g|a) = pc_{a,g} / pc_a$ with pc_a the page count of the artist and $pc_{a,g}$ the page count of the combination of artist and genre. To limit the retrieved pages, schemas are introduced, which are keywords that are added to the query. We limit ourselves to the two best performing schemas for genre classification, namely "music+genre" and "music+style", hereafter named MG and MS. For an elaborate explanation we refer to [4].

To classify the genre of Bruce Springsteen for example, we execute the following queries: first the artist query "Bruce+Springsteen", then one for every genre, e.g. "Bruce+Springsteen+music+genre+blues". The page count of every genre query is divided by the page count of the artist query, and the best classification is the genre with the highest result.

We use the same ground truth data set as Schedl, referred to as C1995a. This dataset is retrieved from AllMusic and contains 1995 artists with 9 very general genres: Blues (189 artists), Country (245 artists), Electronic (98 artists), Folk (84 artists), Jazz (810 artists), Metal (263 artists), Rap (44 artists), Reggae (60 artists) and RnB (202 artists) [4].

3. Re-runs of the original experiment

3.1. Setup

Our starting point was to re-run Schedl's experiment on the same data, in order to find out whether the technique still works as well as in 2006. The experiment was carried out multiple times, in order to enable analysis over time.

As the methodology is in essence search engine independent, we also investigated how different search engines perform on the same task. Schedl used the Google API to retrieve the search result counts. This API is no longer publicly available. Therefore, we scrape the search result counts from the web page returned by the search engine. The query consists of the artist and/or the genre plus a schema. Schedl uses quotes around the artist name, while we joined the whole query with "+", like "Bruce+Springsteen+jazz+music+genre".

The experiment has been repeated 8 times over a period of 36 days, in May and June 2009, over multiple search engines: Google (www.Google.com), Yahoo! Search (search.yahoo.com) and Microsoft Live Search (search.live.com). Because the search engines employ techniques to detect software querying their services, waiting times are inserted after every query.

3.2. Observations

We will analyze the accuracy of the search engines over time and the impact of the MG and MS schemas.

The accuracy, defined here as the number of correct classifications divided by the total number of artists classified, is shown in Figure 2 for the 3 search engines and the MS and MG schemas.

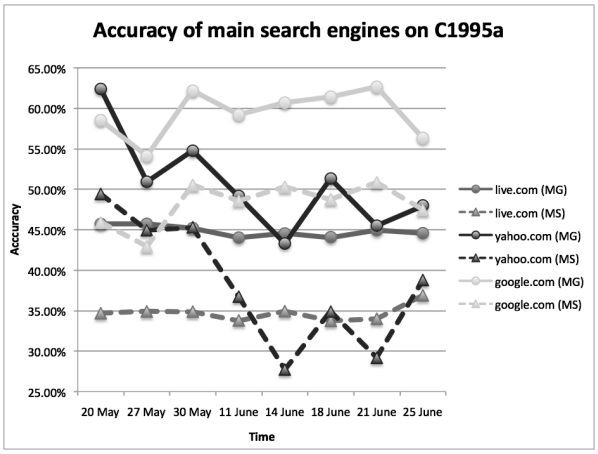

Figure 2. Accuracy of Google, Yahoo! and Live on the C1995a data set.

3.2.1. MG outperforms MS. The first major observation is that the accuracy with the MG schema is always better than the MS schema. We can thus confirm the results from [4] and generalize them to other search engines, than just Google. An explanation can be that style is a broader term than genre for music, including more pages of lesser relevance in the search engine queries and thus in the page counts. With a co-occurrence check on Google for "genre" and "style", we found that 25% of pages containing "style" also contain "music" as opposed to 87% of pages containing "genre" also containing "music".

3.2.2. Google outperforms Yahoo! And Live. Google outperforms Yahoo! and Live with both schemas**.**

3.2.3. Results fluctuate over time. Google fluctuates between 54% and 63% and Yahoo! fluctuates most. Microsoft Live has the most stable results over time.

3.2.4. Technical problems with Yahoo! The fluctuations are caused by partial measurements due to Yahoo! banning our batch querying process. Yahoo! never achieves to classify the whole set of artists. The maximum is 686, the minimum 347 and on average 452 artists (with a standard deviation of 100,5) are classified. The number of artists that get a classification for three different measurements does not rise above 138. This means that the results for Yahoo! are not usable for comparison.

3.2.5. Precision and recall. Figure 3 shows a motion chart[3] of the precision and recall of Google.com and Live.com search engines on the C1995a data set. Each circle displays a genre classified by a search engine. The size of the circles corresponds to the number of correctly classified artists per genre and the color of the circle visualizes the correct genre.

The jazz circles in the upper right (high precision and high recall) corner illustrate that jazz can be accurately classified, with high precision and recall, especially for Google.com.

The country circle on the lower right with a recall of 94% is for the classification of Country with Live. The precision is very low (20%), due to the fact that about 4 times more artists are classified as Country than present in the ground truth. It is clear that Country acts as an attractor. The reason is that many artist pages contain information on the country of origin. An ad hoc measurement with Google shows a 22% co-occurrence between "music+genre" and "music+genre+country". The co-occurrence analysis

[3] Gapminder, http://www.gapminder.org (viewed on 18 Sept. 2009)

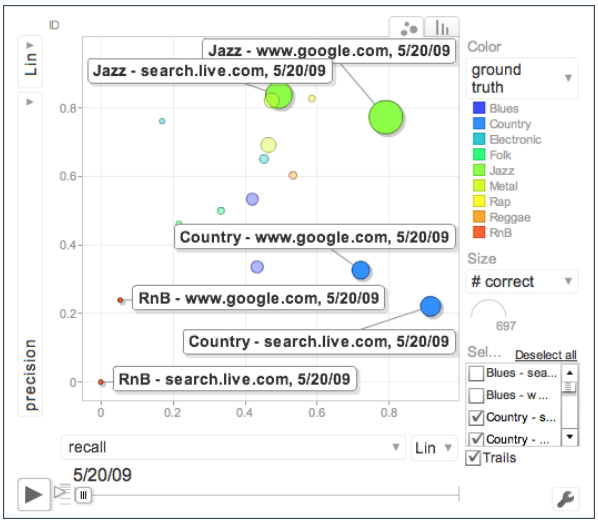

Figure 3. Motion chart of the precision and recall for Google, Yahoo! and Live on the C1995a data set.

has no good way to detect this kind of collision, apart from optimizing the genre names in the taxonomy by analyzing their behavior on the search engines.

You can experiment with the motion chart at http://hmdb.cs.kuleuven.be/muzik/gapminder.html.

As the variability of accuracy is somewhat surprising, we set up a second experiment, with a more limited data set, for fine-grained analysis of a larger number of repeated observations.

4. Fine-grained experiments

4.1. Setup

To be able to analyze the classifications in more detail, a second data set was created, containing a random selection of 2 artists for every genre from

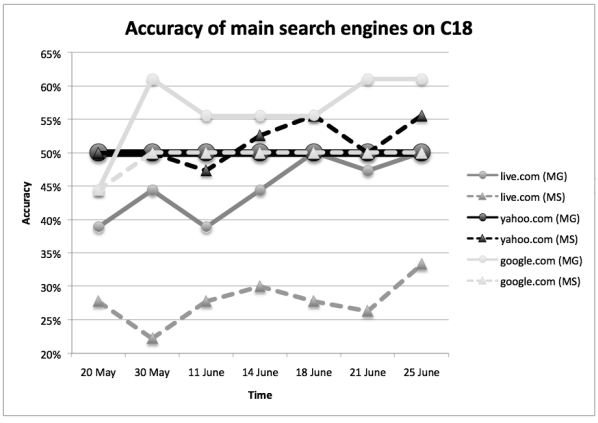

Figure 4. Accuracy on C18a at the same times as Figure 2.0

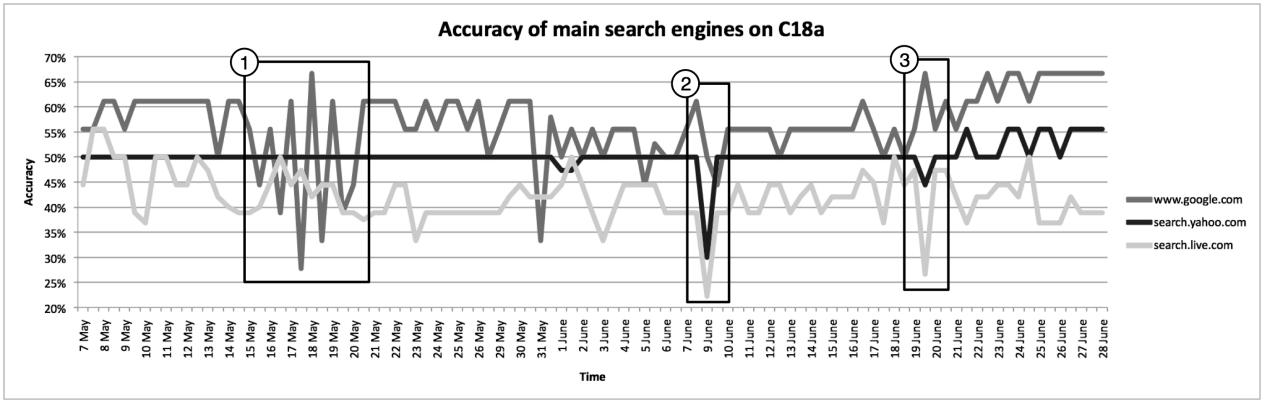

Figure 5. The accuracy of Google, Yahoo! and Live on the C18a data set.

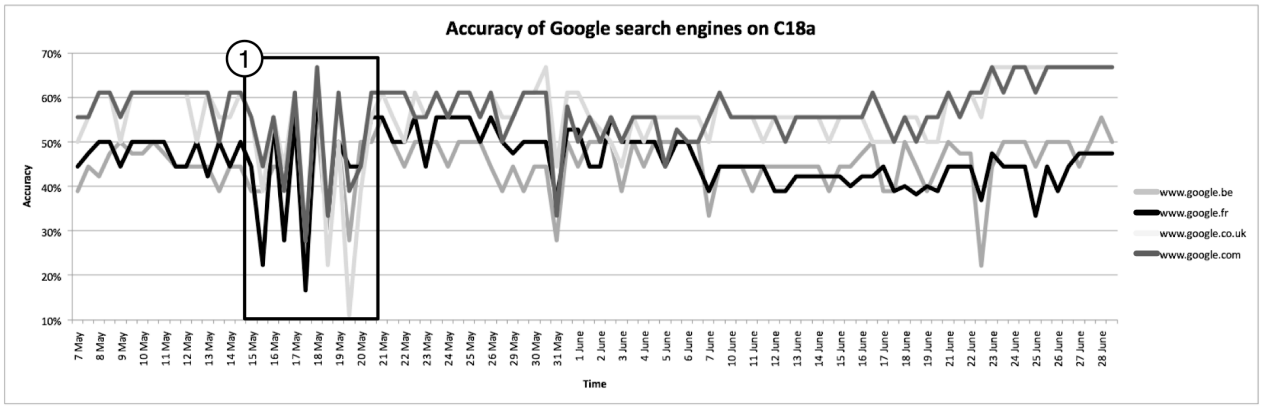

Figure 6. The accuracy of the regional Google search engines on the C18a data set.

C1995a. This results in a set of 18 artists, named C18a and can be viewed at http://hmdb.cs.kuleuven.be/muzik/C18a.html. We classified these artists twice a day at 0:30h CET and 18:30 CET for a period of 53 days, from May 7 till 28 June. Due to technical problems with our server, we only collected 96 measurements. Moreover, in order to compare regional versions of Google and Yahoo! we added more search engines: Google.co.uk, Google.be, Google.fr, uk.search.yahoo.com and fr.search.yahoo.com. Microsoft Live Search does not provide a regional search. This results in around a 250.000 queries for the whole experiment, which illustrates why we limited our scope to 18 artists.

4.2. Observations

We analyze the accuracy of the main and regional search engines and compare this to the C1995a results.

4.2.1. Comparison with C1995a. Figure 4 shows the accuracy of the main search engines on the C18a data set, at the same points in time as Figure 2.

The accuracy is not the same as for C1995a, but the overall trends are similar. The MG schema is still more accurate, except for Yahoo! on some points. Yahoo! MG is very stable, Live is still the worst in accuracy and Google the best. We also see that the accuracy of Live and Google is a bit more variable.

4.2.2. Overall results for C18a. Figure 5 shows the accuracy of Google, Yahoo! and Live on C18a over the 99 measurements on the most accurate schema MG. Yahoo! is performing very stable in contrast to the results of Figure 2. Google is again most accurate overall and Live least. The accuracy of Google also fluctuates most, especially in box 1 in Figure 5. Microsoft launched a new search engine, Bing, (bing.com), on the 3rd of June 2009, replacing live.com. This did not lead to noticeable changes in results. From here on we will refer to the Live and Bing combination as Live.com. On 29 July 2009, Yahoo! and Microsoft announced a business collaboration to replace Yahoo! Search with Bing in the next two years[4], which means that using Yahoo! for classification will be the same as using Bing.

[4] http://www.microsoft.com/Presspass/press/2009/jul09/07-29release.mspx (viewed on 18 Sept. 2009)

4.2.3. Regional Variations. Figure 6 shows similar data the same as Figure 5, but for the regional Google search engines. Overall, Google.com performs best, closely followed by Google.co.uk. Next is Google.fr and then Google.be. The regional search engines often fluctuate in a similar manner over time. The large fluctuations in box 1 in Figure 5 and 6 are very similar for all regional engines. A possible reason why Google.fr and Google.be perform less accurate than their Anglo-Saxon siblings could be because the genres and schemas used in the queries are in English. One could also check whether French and Belgian artists are classified better or what would happen if the schemas would be translated to the French "musique+genre" or the Dutch "muziek+genre". We did not investigate this further.

4.2.4. Stability of artist classification. In order to investigate whether search engines consistently classify an artist correctly or incorrectly, Figure 7 visualizes the correct (light grey) and incorrect (dark grey) classifications for the 18 artists with the MG schema. The results are shown horizontally chronologically over the 99 measurements (equal to Figure 5 and 6) and vertically the artists are ordered to allow clustering of correct and incorrect classifications. Clearly, Yahoo! provides the most stable classifications as it consistently classifies artists correctly (light grey) or incorrectly (dark grey) at each point in time. This means that the stable overall accuracy of Yahoo! in Figure 4 can be explained by a stable classification for individual artists in the case of Yahoo!. On the other hand, Google changes the classification of individual artists most often. Artists that are often correctly classified sometimes get misclassified and vice versa. Changing from correct to incorrect occurs most, but there is no clear pattern to be observed. By contrast, Microsoft Live seems to be always struggling to classify the same one fourth of the artists and doing this more wrongly than right, hence the low accuracy. Box A in Figure 7 corresponds to 1 in Figure 5 and 6.

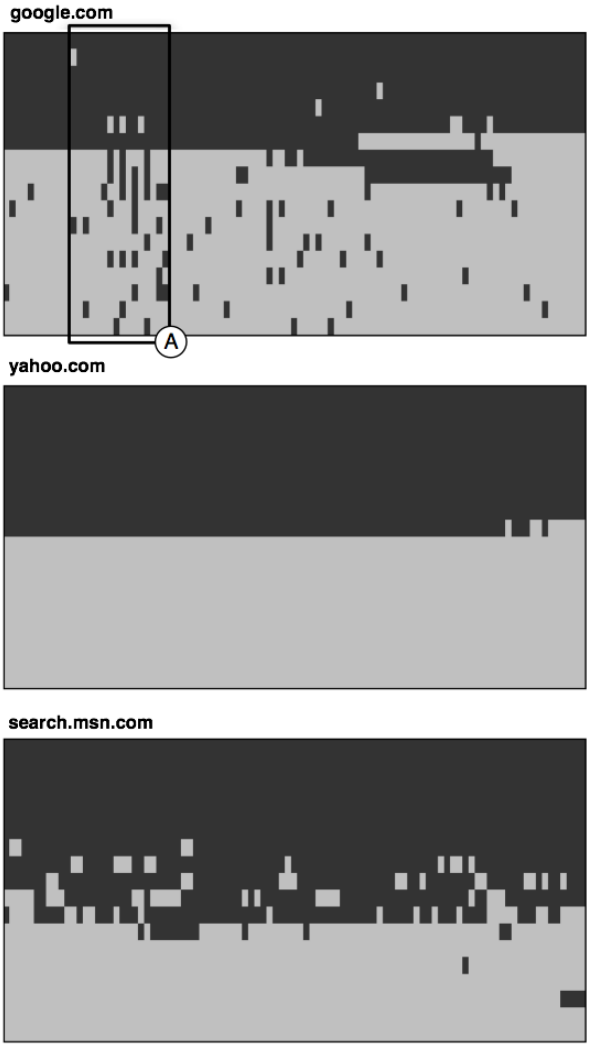

Figure 7. Classifications for artists over time ordered on number of correct classifications (light grey = correct, dark grey = incorrect)

5. Strategies

Overall, in order to obtain the most accurate results, we would prefer to use Google when it is not fluctuating, and Yahoo! when Google does fluctuate. However, we need a strategy to find out when Google is fluctuating. Analysis of figure 7 shows that in the bad day scenario different classifications occur for normally 'stable' artists. Therefore, an embedded test set can be deployed a number of times per day to verify the stability of Google. The outcome of this simple test determines the selection of the search engine.

We can detect when Google behaves sub-optimally by comparing the correctness of the test set with the average reported in Figure 5. It would be worthwhile to investigate if repeating the queries several times could help us to further increase accuracy.

6. Conclusion and future work

In conclusion, the method suggested by Schedl et al. is still a valid approach after 3 years of technological evolution. In that time span, the Web expanded from 75 million web sites in March 2006 to 238 million in June 2009[5]. The main outcome of our research so far is

[5] http://news.netcraft.com/archives/web_server_survey.html (viewed on 18 Sept. 2009)

that the accuracy of this approach varies per search engine and, more surprisingly, can also vary considerably for a specific search engine.

Unfortunately, we do not have detailed information about the internals of the different search engines to understand the origin of fluctuations in the page count, though according to [11] it could be related to propagation and partitioning of indices and query distribution. Specifically for Google, the frequent index rebuilding is often referred to as Everflux[6]. Again, we cannot provide a definitive explanation, but hope to engage in a dialogue with the search engine providers to further understand the underlying causes.

Regardless of the implementation of the search engine and based on our observations, we note that Google outperforms Yahoo!, and Live if it is not fluctuating (see figure 5). This fluctuation can be detected by deploying simple statistics. In the case of fluctuation, it is better to use another search engine.

Further research can help to understand the differences between localized versions of the search engines. In our tests, the Anglo-Saxon version performed better (see Figure 6). Accuracy may depend on the characteristics of the music collection – google.fr may perform better in the classification of French chanson, for instance.

Since Yahoo! Search will cease to exist, research on other alternatives (e.g. Ask.com and Lycos.com) or search engines tailored to music (e.g. AllMusic) could help to further understand the available options.

The used taxonomy is a very important factor in achieving high accuracy. The current approach is pretty naïve and causes problems for some genres, like RnB and Country. The taxonomy should be run in a small test run to check if conflicts occur, e.g. terms that attract irrelevant search results. This should also be checked in combination with the chosen schemas.

A generalization of the approach analyzed here can be used for classifications other than genre. A generic implementation has been added to our metadata generation framework [12] and we plan to analyze this in detail in the future.

7. Acknowledgement

We gratefully acknowledge the support of IWT Vlaanderen (project MuziK, grant IWT 080226) and Aristo Music NV for the metadata annotation time study and their musical expertise. We also acknowledge the implementation work of our master thesis student, Benedikt Raes.

[6] http://www.buzzle.com/articles/Google-everflux-just-what-we-needed-Google-on-speed.html (viewed on 18 Sept. 2009)

8. References

[1] S. Govaerts, N. Corthaut, E. Duval, "Moody Tunes: The rockanango Project", *Proceedings of the 7th International Conference on Music Information Retrieval*, Victoria, Canada, 2006, pp. 308-313.

[2] S. Govaerts, N. Corthaut, E. Duval, "Mood-ex-Machina: Towards Automation of Moody Tunes", *Proceedings of the 8th International Conference on Music Information Retrieval*, Vienna, Austria, 2007, pp. 347-350.

[3] G. Tzanetakis, P. Cook, "Musical Genre Classification of Audio Signals", *IEEE Transactions on Speech & Audio Proceedings*, Jul. 2002, 10(5), pp. 293-302.

[4] M. Schedl, T. Pohle, P. Knees, G. Widmer, "Assigning and Visualizing Music Genres by Web-based Co-occurrence Analysis", *Proceedings of the 7th International Conference on Music Info. Retr.*, Victoria, Canada, 2006, pp. 260-265.

[5] J. Bergstra, A. Lacoste, D. Eck, "Predicting Genre Labels for Artists Using FreeDB", *Proceedings of the 7th International Conference on Music Information Retrieval*, Victoria, Canada, 2006, pp. 85-88.

[6] J. J. Aucouturier, F. Pachet, "Representing Musical Genre: A State of the Art", *Journal of New Music Research*, 2005, pp. 83-93.

[7] S. Govaerts, E. Duval, "A Web-based Approach to determine the Origin of an Artist", *Proceedings of the 10th International Conference on Music Information Retrieval*, Kobe, Japan, 2009.

[8] P. Lamere, "Social tagging and music information retrieval", *Journal of New Music Research*, 37(2), 2009, pp. 101–114.

[9] P. Knees, E. Pampalk, G. Widmer, "Artist Classification with Web-based Data", *Proceedings of the 5th International Conference on Music Information Retrieval*, Barcelona, Spain, 2004, pp. 517-524.

[10] G. Geleijnse, J. Korst, "Web-based Artist Categorization", *Proceedings of the 7th International Conference on Music Information Retrieval*, Victoria, Canada, 2006, pp. 266-271.

[11] R. Baeza-Yates, C. Castillo, F. Junqueira, V. Plachouras, and F. Silvestri, "Challenges in Distributed Information retrieval", *International Conference on Data Engineering (ICDE)*, IEEE CS Press, Istanbul, Turkey, April 2007.

[12] K. Cardinaels, M. Meire, E. Duval, "Automating Metadata Generation: the Simple Indexing Interface", *Proceedings of the 14th International Conference on World Wide Web*, WWW 2005, Chiba, Japan, 2005, pp. 548-556.