

Using Cultural Metadata for Artist Recommendations

Stephan Baumann
DFKI GmbH
Erwin Schrödinger Str
67663 Kaiserslautern
Stephan.Baumann@dfki.de

Oliver Hummel
DFKI GmbH
Erwin Schrödinger Str
67663 Kaiserslautern
Oliver.Hummel@dfki.de

Abstract

Our approach to generate recommendations for similar artists follows a recent tradition of authors tackling the problem not with content-based audio analysis. Following this novel procedure we rely on the acquisition, filtering and condensing of unstructured text-based information that can be found in the web. The beauty of this approach lies in the possibility to access so-called cultural metadata that is the agglomeration of several independent - originally subjective - perspectives about music.

1. Introduction

Our approach to generate recommendations for similar artists follows a recent tradition of authors [1-4] tackling the problem not from the audio-based side of content analysis. Instead we rely on the acquisition, filtering and condensing of text-based information that can be found in the web. The beauty of this approach lies in the possibility to access so-called cultural metadata which is indeed the agglomeration of several independent - originally subjective - perspectives about music, i.e. artists. The advantages of such a technique are the following:

- Incorporation of semantics: in contrast to content-based audio processing the usage of web reviews offers the possibility to access descriptive semantics about the musical work of an artist.
- Instant availability: in contrast to collaborative filtering techniques such as used by commercial portals (e.g. Amazon) we have no bootstrapping phase in web-based approaches. After the release of new songs or albums immediate access to according reviews is possible.
- Time-awareness: in contrast to the static representation of an audio-based approach the dynamics of changing cultural context, resp. artist relations are included in a web-based representation.

The paper is structured as follows: In section 2 we give a short overview of the system architecture and the embedded components, section 3 goes into the details of the feature space, vectors and similarity metrics, section 4 explains the improvements over existing approaches, section 5 comes with an evaluation of our approach

compared to other authors and different ground truth data and finally in section 6 we present some new ideas about the incorporation of relevance feedback techniques in order to make personalized recommendations. Section 7 gives some conclusions we gained from our experiments in this area.

2. Architecture

Our system consists of 3 major components: the information gathering has been realised through the Google API [5], part-of-speech tagging has been implemented by using a probabilistic open-source tagger and finally for the evaluation we relied on the datasets acquired by Whitman [3-4] and some lookups at the editor-generated similarity lists at www.allmusic.com and launch.yahoo.com. The entire system and information flow is sketched in Figure 1.

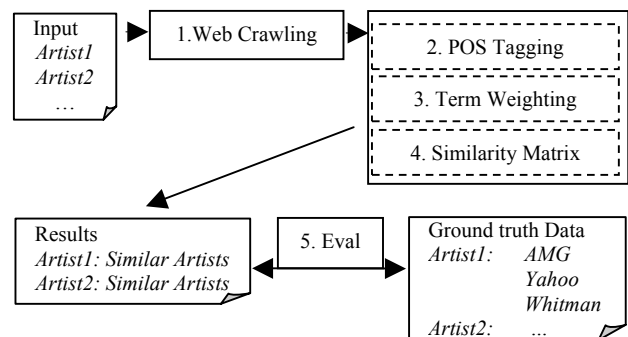


Figure 1. System architecture

Term extraction and weighting as well as the similarity computation for artists have been realised as a JAVA application which gives also access to the main tuning parameters of the system in a simple, but powerful GUI.

As a starting point Google does not allow the use of its search results for automated queries. But “Google Web APIs” enabled us to acquire the data we needed by the use of a simple and quite stable interface, without the

necessity of building an application that extracts the search results from a HTML document. The subsequently applied standard techniques from the field of Information Retrieval [6] and Computational Linguistics have been implemented in a JAVA kernel engine. Finally we used www.allmusic.com (AMG) and Yahoo Launch websites to collect their recommendations about similar artists as a kind of ground truth to evaluate our systems performance. We selected AMG and Yahoo because their approaches to construct lists for similar artists have significant differences. At AMG editors being deeply involved in the ongoing trends in music generate handcrafted contributions. Vice versa at Yahoo the recommendations are generated from end user feedback.

3. Feature Space

Our approach relies solely on the usage of textual features that are contained in the HTML documents. As a starting point we crawl about 50 pages per artist that contain reviews of the musical work and apply some filtering techniques to extract the meaningful parts out of these pages. In subsequent steps part-of-speech tagging, term weighting and the computation of artist similarities are performed. These steps are described in the following subsections.

3.1 Part-of-Speech Tags

The part-of-speech (POS) tags offer the ability to get rid of some *language noise* that is indeed necessary for the authors to formulate their reviews. We followed the work of Whitman and used the same categories, namely nouns and adjectives. Different combinations of these categories are tested in order to achieve a maximum of performance in predicting similar artists. We chose similar to Whitman: Single occurrences of terms (n1) e.g. *self*, pairs of terms (n2), e.g. *daft punk*, single occurrences of nouns e.g. *rock*, single occurrences of adjectives e.g. *tricky*, adjective-noun pairs or adverb-adjective-noun triples, e.g. *mindless self indulgence*.

3.2 Term Weighting and Similarity Computation

We rely on the well-known TFIDF weightings [6] that can be found in standard IR (Information Retrieval) to put bias on important, resp. most-discriminating terms of a document. The computation of TFIDF is based on the number of occurrences of a term in the 50 documents that have been crawled for an artist, denoted as term frequency TF. The document frequency DF of a term counts the number of documents with occurrences of this term in the entire artist collection of size n. The product $TFIDF = TF * \log (n / DF)$ can be used as a weighting score for each

term in the collection. The representation of an artist consists of an n-dimensional vector with n = number of unique terms in the set of 50 documents and the TFIDF weightings as according values. On top of this vector representation the cosine measure has been used to compute similarity between two artists. Table 1 shows some examples of TFIDF-weighted vector representations for the *Bloodhound Gang*.

| adj Terms | TFIDF | Phrases | TFIDF |
|-----------|---------|--------------------------|---------|
| funny | 0.20463 | bad touch | 0.86982 |
| juvenile | 0.14242 | safe version | 0.80009 |
| queer | 0.12907 | fierce beer coaster | 0.40004 |
| anomalous | 0.09314 | inevitable return | 0.40004 |
| cultural | 0.08607 | great white dope | 0.40004 |
| annoying | 0.07558 | still funny | 0.40004 |
| limp | 0.07339 | mindless self indulgence | 0.40004 |
| funniest | 0.06887 | bad language | 0.36956 |
| fierce | 0.06497 | lower volume | 0.36956 |
| tricky | 0.06497 | good beat | 0.34793 |

Table 1. Examples for the *Bloodhound Gang*

4. The differences to the *Whitman* Approach

In contrast to Whitman we implemented a HTML-filtering stage to improve the quality of the text input. The first step was to remove all pages that brought more than 40 kB of parsed text. The second point was the structure of the HTML pages themselves. A lot of advertisement is placed in other table elements than the actual review and contain only a few words. Therefore our HTML just returns text from cells which contains at least one full sentence and more than 60 characters. Although our results were mostly bounded to musical reviews we got many of them where our desired artist is just mentioned somehow, e.g. we found a *Santana* review where we just wanted to have *Nickelback*, because *Nickelback's* front man is mentioned on the *Santana* review page.

| URL contains: | Title contains: | Nickelback review? |
|-------------------------|-------------------------|--------------------|
| music | music review nickelback | yes |
| music | music review | no (Santana) |
| nickelback review | review nickelback | yes |
| music review nickelback | music | yes |
| music review nickelback | music | yes |
| review | nickelback review | yes |
| review | music review | no |
| review | music review | no |
| music | music review | no |
| review | review music | no |

Table 2. Filtering with URL and title

So we started to take a closer look at the URLs and the titles of our search results (see Table 2). Based on this results we constructed another filter that combined the results of a keyword spotting in the URL, the title and first

text parts of the page. Each occurrence of the keywords *music*, *review*, *artist name* in one of the parts is scored with a weighting (i.e. the artist name gets a higher score). The total sum of these scores is thresholded to filter out the unrelated pages.

5. Benchmarking

We were interested to see if we could reach similar performance as being presented by Whitman and if our improvements in the early acquisition stage had an impact on the overall performance of the approach. In addition to the 414 artists used by Whitman we had a second set from earlier experiments using audio-based similarity recommendations [7]. It consists of 50 different artists that could be grouped by expert reviewing along an axis of genre similarity. This set was used for a coarse visual inspection with graphical similarity matrices. Furthermore we considered two major additional aspects:

- Predicting symmetric vs. non-symmetric artist similarity
- The influence of expert vs. end-user ground truth data

5.1 Ground truth: AMG evaluation

Whitman et. al calculated the accuracy of their system predicting the proportion of artists as known to be similar from the ground truth to the same number of randomly chosen artists. They compute the mean of the similarity scores of five recommended artists from AMG. They did the same for five randomly chosen artists from the entire set. If the mean score from the AMG set is higher than the random mean score the system has the ability to predict the AMG recommendation. We considered this accuracy metric as valuable for comparing systems to each other. Although we were able to reach in this benchmark an accuracy of almost 90% with our n1-data, we are far away from e.g. finding an AMG recommendation on all the first places of an artist’s similarity list. Table 3 illuminates this dilemma closer:

| | Accuracy | Top 1 Hits | Top 3 Hits | Top 5 Hits |
|------------|------------|------------|------------|------------|
| n1 | 89% (78%) | 37 | 111 | 169 |
| n2 | 83% (80%) | 43 | 96 | 144 |
| Adjectives | 79% (69%) | 41 | 92 | 125 |
| Phrases | 75% (82%) | 42 | 86 | 125 |
| Nouns | 84% (n.a.) | 33 | 99 | 152 |

Table 3. Results (Whitman results).

The table shows the values reached with the plain TFIDF scoring. We were able to outperform even Whitman’s best result with his Gaussian weighting function in place without any changes on our TFIDF metric. And especially our values of the n1 terms and the adjectives are much better, which is due to the fact that our filter is indeed able

to reduce the noise in the documents (e.g. link lists etc.). Another interesting fact we could discover here was that the nouns, which are usually much noisier than other parts of speech, delivered also quite useful results.

We decided to add the “roots and influences” as well as the “followers” from AMG to the groundtruth data. Now we got 2247 recommendations that appear in our set, instead of 1421 before. There are 378 artists with at least one found in the similarity recommendation of AMG. Table 4 shows the improvement of the top hits with the larger data.

| | Accuracy | Top 1 Hits | Top 3 Hits | Top 5 Hits |
|------------|----------|------------|------------|------------|
| n1 | 91% | 70 | 182 | 263 |
| n2 | 84% | 63 | 138 | 203 |
| Adjectives | 83% | 59 | 129 | 177 |
| Phrases | 78% | 58 | 126 | 174 |
| Nouns | 89% | 67 | 164 | 243 |

Table 4. Performance for increased ground truth.

5.2 Ground truth: Yahoo Evaluation

Yahoo Launch offers up to forty artists that *fans of ... also tend to like*. Here the users have the possibility to rate an artist directly. Yahoo seems to collect its data with collaborative filtering techniques, but does not explain how this is done internally. Because of this we can only speculate if their similarity list contains a ranking or not. But we noticed during our work that the data definitely changes over time, the next table shows the changes from December 2002 to February 2003 in comparison to Whitman’s OpenNap data from august 2001. This reflects our initial assumption that cultural recommendations need to be time-aware.

| OpenNap (08/01) | Yahoo (12/02) | Yahoo (02/03) |
|-----------------|-------------------|-----------------|
| Culture Beat | Depeche Mode | Depeche Mode |
| Thompson Twins | Madonna | Madonna |
| New Order | Sting | Eurythmics |
| Blondie | *NSYNC | The Cure |
| Erasure | Duran Duran | Enya |
| Duran Duran | Dido | Sade |
| Roxette | Alanis Morissette | Sting |
| Eurythmics | Nelly Furtado | Pink |
| Ace of Base | Tears For Fears | U2 |
| Wham | INXS | Phil Collins |
| Depeche Mode | Foreigner | The Cranberries |
| A-Ha | Sade | Duran Duran |

Table 5. Dynamics of similarity (*Pet Shop Boys*)

Finally we used the actual Yahoo data set to compare expert vs end-user artist recommendations. In comparison to the original AMG lists accuracy is slightly higher (92% vs. 89% for n1). As a conclusion we see that cultural recommendations based on textual web sources are able to predict both *expert and end-user perception* of similarity.

Furthermore the results show that the fine-grained expert recommendations (pure similarity, roots, follower) found at AMG are kind of subsets of a coarse mainstream-like set of recommendations as found at Yahoo.

6. Objective vs. subjective evaluation

The proposed approach seems to have an upper bound for predicting similar artists as we can see from the results presented in section 5. AMG as well as Yahoo recommendations could be predicted within a range of 80-90% accuracy in contrast to a random prediction. But the recall is still very low. A real breakthrough may be realized only by concentrating on subjective evaluations, resp. incorporating relevance feedback of the users. We found some interesting examples indicating such hypothesis when looking manually at the result lists. We listed some of these examples in the following table. The left hand side shows the Top 12 artists in our n2 similarity list for the *Beastie Boys*. The right hand side presents all the artists that are recommended by AMG.

| Position | Artist | Score | Artist | Score |
|----------|---------------|----------|---------------------|----------|
| 1 | Jamiroquai | 0.204484 | House of Pain | 0.204263 |
| 2 | House of Pain | 0.204263 | Kid Rock | 0.069803 |
| 3 | Offspring | 0.154857 | Cypress Hill | 0.018370 |
| 4 | Run-DMC | 0.147424 | Rage Against the... | 0.009184 |
| 5 | Corrs | 0.137784 | Bad Brains | 0.008918 |
| 6 | Papa Roach | 0.126888 | Beck | 0.008712 |
| 7 | TLC | 0.123031 | Santana | 0.008702 |
| 8 | DMX | 0.118420 | Stevie Wonder | 0.004436 |
| 9 | Janis Joplin | 0.087388 | Led Zeppelin | 0.004216 |
| 10 | Limp Bizkit | 0.084984 | 311 | 0.004024 |
| 11 | Talking Heads | 0.076104 | Everlast | 0.003557 |
| 12 | Kid Rock | 0.069803 | LL Cool J | 0.002259 |

Table 6. Automatic vs. AMG

6.1 Website Collection

In order to make a more thorough evaluation of such possibilities we plan to set up a website presenting the top 10 recommended artists to a set of test persons with different cultural and educational background. In [8] a powerful user model has been presented which should be fed by a web-based interview for each user. Similar work has been done in the past at research (www.musicseer.com) and industrial affiliations (www.MoodLogic.com).

6.2 Relevance Feedback

The abovementioned setup will enable us to provide immediate refined result lists by using relevance feedback processing. We made some initial experiments using the classic term reweighting as proposed by Rocchio [6].

The formula allows for incorporation of the most-interesting terms of an artist representation and it puts a penalty on the terms of the non-relevant artists. The modified artist vector is simply calculated by adding the sum of the artists' vectors being judged as relevant to the original vector that is a *positive* feedback strategy. In the most general approach, the *standard Rocchio feedback*, additionally the sum of the non-relevant artist vectors is subtracted. Both operations may be weighted, traditionally the sum of positive feedback is weighted by 1 and the weight for the sum of negative feedback is smaller.

7. Conclusion

The approach described in this paper has essential potential as a bottom-up recommendation engine for similar artists. By incorporation of heuristic and structural filtering methods we could improve slightly over related work in this field. But from our perspective the main challenge for the MIR community is the transition from the system-centric development of MIR applications to a user-centric development. First experiments with relevance feedback techniques confirm this hypothesis. Therefore our future work will focus on the integration of relevance feedback and the investigation of learning similarity metrics.

8. References

- [1] Pachet F., Westermann G., Laigre D., Musical Data Mining for Electronic Music Distribution, Proceedings of First Int. Conf. on WEB Delivering of Music (Wedelmusic 2001), pp 12-19, Florence, Italy, November 2001
- [2] Pachet F., Aucouturier J., Representing Musical Genre: A State of the Art, Journal of New Music Research 2002
- [3] Whitman, B., Lawrence S., Inferring Descriptions and Similarity for Music from Community Metadata, Proceedings of the 2002 ICMC. pp 591-598, Göteborg, Sweden, 16-21 Sep.2002
- [4] Whitman, B., Smaragdis P., Combining Musical and Cultural Features for Intelligent Style Detection, Proceedings of the Third ISMIR, 13-17 October 2002, Paris, France
- [5] Baeza-Yates R., Ribeiro-Neto B., Modern Information Retrieval, Addison-Wesley Publishing Company, 1999
- [6] Baumann, S., Klüter A., Using natural language and audio analysis for a human-oriented MIR system, Proceedings of the Second Int. Conf. on WEB Delivering of Music (Wedelmusic 2002), Darmstadt, Germany, December 9-11, 2002
- [7] Chai W., Vercoe B., Using User Models in Music Information Retrieval Systems, Proceedings of the First ISMIR 2000, Plymouth, USA, 15-17 October 2000