Improving Prototypical Artist Detection by Penalizing Exorbitant Popularity

Markus Schedl [1,2] **Peter Knees** [1] **Gerhard Widmer** [1,2]
{markus.schedl, peter.knees, gerhard.widmer}@jku.at

[1] Department of Computational Perception
Johannes Kepler University (JKU)
A-4040 Linz, Austria

[2] Austrian Research Institute for Artificial Intelligence (ÖFAI)
A-1010 Vienna, Austria

Abstract. Discovering artists that can be considered as prototypes for particular genres or styles of music is a challenging and interesting task. Based on preliminary work, we elaborate an improved approach to rank artists according to their prototypicality. To calculate such a ranking, we use asymmetric similarity matrices obtained via co-occurrence analysis of artist names on web pages. In order to avoid distortions of the ranking due to ambiguous artist names, e.g. bands whose names equal common speech words (like *Kiss* or *Bush*), we introduce a penalization function. Our approach is demonstrated on a data set containing 224 artists from 14 genres.

1 Introduction and Related Work

The automatic detection of prototypical artists provides valuable information for a wide range of applications related to music information retrieval. For example, music information systems like the All Music Guide[1] as well as online music stores, e.g. Amazon[2], could benefit considerably. For instance, information on prototypes could be exploited to support users in finding music more efficiently. Furthermore, prototypical artists are very useful for visualizing music repositories since they are usually well-known. Thus, they can serve as reference points to discover similar but less known artists (for more details, see [4]).

To obtain an estimate of the prototypicality of artists, we exploit information on co-occurrences of artist names on web pages. We already showed that web-based co-occurrence analysis can be used successfully for artist similarity measurement and artist-to-genre classification [3]. In this paper, we will use the (approximated) page counts returned by Google in response to requests concerning artist names. Based on these page counts, we estimate conditional probabilities for an artist to be found on web pages of other artists. These probabilities

[1] *http://www.allmusic.com*
[2] *http://www.amazon.com*

give an asymmetric similarity matrix which is used for the calculation of prototypicality rankings within genres [4]. An emerging problem of this approach are extremely high rankings for artists whose names equal common speech words. To address this issue, we propose a modification of the ranking function.

The new contribution of this paper is the improvement of the prototypicality ranking from [4] by incorporating information on global (genre-independent) prototypicality. To this end, we penalize exorbitant and unrealistic popularity by weighting the ranking with our newly introduced *inverse overall prototypicality* measure.

2 Prototypical Artist Detection

To obtain the information used for prototypical artist detection, first we perform a co-occurrence analysis step. Given a list of artist names, we use Google to estimate the number of web pages containing each artist and each pair of artists. The only information we need for this is the page count returned by Google. This raises performance and limits web traffic. Based on the page counts, we then use relative frequencies to calculate a conditional probability matrix. Given two events a_i (artist with index i is mentioned on web page) and a_j (artist with index j is mentioned on web page), we estimate the conditional probability p_{ij} (the probability for artist j to be found on a web page that is known to contain artist i). Computing the conditional probability for every pair of artists, we obtain a similarity matrix that is obviously not symmetric. More detailed information on the calculation can be found in [3].

2.1 Prototype Detection using Backlink/Forward Link Ratios

We regard the prototypicality of a music artist as being strongly related to how often music-related web pages refer to the artist. The method used to infer prototypicality is based on an idea similar to the PageRank mechanism applied by Google where *backlinks* and *forward links* of a web page are used to measure relevancy [1]. In our approach, we call any co-occurrence of artist a and artist b (unequal to a) on a web page that is known to contain artist b a *backlink* of a (from b). A *forward link* of an artist of interest a to another artist b, in contrast, is given by any occurrence of artist b on a web page which is known to mention artist a.

For each artist a_i, we count for how many of the artists a_j from the same genre the number of backlinks of a_i exceeds the number of forward links, which gives the backlink count bl. The fl count for forward links is defined analogously. Using these counts, we calculate the ratio bl/fl. The higher this ratio, the higher the prototypicality of a_i for the respective genre. A more formal definition can be found in [4]. The intuitive assumption behind this is that frequent occurrences of an artist on other artists' web pages reflect a certain relevance. For example, pages related to the Finnish heavy metal band *Sentenced* will obviously refer more often to the well-known pioneers *Metallica* than vice versa. This indicates that *Metallica* serves as a prototype for the genre Heavy Metal.

2.2 Downranking Artists by Penalizing Exorbitant Popularity

As results in [4] show, artist names which are also used in everyday speech tend to be ranked at the top, for example, *Kiss* from the genre Hard & Heavy, *Bush*, *Hole*, and *Nirvana* from Alternative/Indie, or *Prince* and *Madonna* from Pop. The reason for this is that such words frequently occur on web pages and therefore produce a lot of backlinks for artists with such names. This kind of misleading co-occurrences are a shortcoming of web-based information retrieval methods and distort the prototypicality ranking.

To avoid such distortions, we propose a mechanism that basically pursues the idea of the commonly used information retrieval approach $tf \times idf$ (term frequency×inverse document frequency), cf. [2]. In this approach, the importance of a term is higher if the term occurs frequently (high tf). On the other hand, a term is penalized if it occurs in many documents and hence does not contain much discriminating information (high df leads to low idf).

In our approach, we adapt this principle to penalize the prototypicality of an artist if it is high over all genres (following the naming scheme of $tf \times idf$, we could call our approach $gp \times iop$ for *genre prototypicality*×*inverse overall prototypicality*). This is reasonable, since even very popular and important artists are unlikely to be prototypes for all styles of music and related to all other artists. Furthermore, common speech words appear on artists' web pages independently of their genre. Taking a closer look of the overall bl/fl ratios from [4] supports this consideration. On a collection of 224 artists from 14 well-known genres[3], those artists whose names equal common speech words yield by far the highest overall bl/fl ratios, e.g. *Bush* (223/0), *Prince* (222/1), *Kiss* (221/2), *Madonna* (220/3), and *Nirvana* (218/5). Thus, using the information on overall (global) prototypicality, we suggest ranking of artists according to the value

$$rnk(a) = \frac{bl_{genre}(a)}{fl_{genre}(a) + 1} \cdot iop(a)^2, \tag{1}$$

where

$$iop(a) = norm(log \frac{fl_{global}(a)}{bl_{global}(a) + 1}), \tag{2}$$

and $norm$ is a function that shifts all values to the positive range by subtracting the smallest value (despite $-\infty$), replaces infinite numbers by 0, and normalizes the values by dividing by the maximum value (in the order mentioned).

To demonstrate the effects of this revised function, Table 2 shows the newly obtained rankings for genres containing artists whose names equal common speech words, as well as one ranking for a genre without such artists that remains almost unmodified (Folk). For comparing the improved prototypicality ranking with the original one as published in [4], Table 1 shows the latter for the selected genres. One can see that the artists *Bush* and *Hole* from the genre Alternative/Indie drop significantly when applying the iop weighting, which better

[3] *http://www.cp.jku.at/people/knees/artistlist224.html*

corresponds with their real importance for the genre. Also *Kiss* from Hard & Heavy drop from the top-position to the mid-range of the ranking.

Concerning evaluation, we refer to the results in [4]. Basically, we use Spearman's rank correlation between the bl/fl prototypicality ranking per genre and rankings obtained by absolute page counts for Google queries containing both artist and genre name. Comparing the original with the *iop*-weighted rankings, it can be seen that rankings and correlations remain basically unchanged for genres without ambiguous artist names. For the other genres, the evaluation method applied in [4] returns lower rank correlation values. This is no negative result, since it is the intention of our approach to overcome the susceptibility of web-based approaches for overrating of common word names.

3 Conclusions and Future Work

We presented an approach for automatic detection of prototypical artists. To this end, we use asymmetric similarity matrices gained by co-occurrence analysis of artist names on web pages. Based on these similarity matrices, we estimate a prototypicality ranking for the artists using *backlink/forward link ratios*. To overcome the problem of unjustified high rankings for artists whose names equal common speech words, we apply a penalization function that weights the local bl/fl ratios (per genre). This function is calculated using global bl/fl ratios (computed on the complete artist set).

We demonstrated our approach on a test collection containing 224 artists of 14 genres. It was shown that the introduction of a penalization function improves the ranking results obtained by the simple bl/fl prototypicality ranking in the sense that the improved rankings better reflect the real importance of artists to genres.

As for future work, it is planned to evaluate our approach on a larger artist set containing more than 950 artists from 15 genres, which is currently being compiled. This would be highly interesting since a shortcoming of the test collection used here is that most of its artists are quite popular and typical of their genre. Furthermore, we are currently elaborating on heuristics to minimize the computational complexity of the co-occurrence analysis.

4 Acknowledgments

This research is supported by the Austrian Fonds zur Förderung der Wissenschaftlichen Forschung (FWF) under project numbers L112-N04 and Y99-START, and by the EU 6th FP project SIMAC (project number 507142). The Austrian Research Institute for Artificial Intelligence is supported by the Austrian Federal Ministry for Education, Science, and Culture and by the Austrian Federal Ministry for Transport, Innovation, and Technology.

Pop		*Hard & Heavy*		*Alternative/Indie*		*Folk*	
artist ranking	**bl/fl**	**artist ranking**	**bl/fl**	**artist ranking**	**bl/fl**	**artist ranking**	**bl/fl**
Prince	15:0	Kiss	15:0	Bush	15:0	Bob Dylan	15:0
Madonna	14:1	Metallica	14:1	Hole	14:1	Donovan	14:1
Britney Spears	13:2	Slayer	13:2	Nirvana	13:2	Leonard Cohen	13:2
Michael Jackson	12:3	AC/DC	12:3	Beck	12:3	Joni Mitchell	12:3
Avril Lavigne	10:5	Iron Maiden	11:4	Radiohead	11:4	Cat Stevens	11:4
Janet Jackson	10:5	Anthrax	10:5	Sonic Youth	10:5	John Denver	10:5
Jennifer Lopez	9:6	Black Sabbath	9:6	Pearl Jam	9:6	Joan Baez	9:6
Christina Aguilera	8:7	Def Leppard	8:7	Weezer	8:7	Tracy Chapman	8:7
Robbie Williams	7:8	Deep Purple	7:8	Smashing Pumpkins	7:8	Pete Seeger	7:8
ABBA	6:9	Megadeth	6:9	Depeche Mode	6:9	Don McLean	6:9
Justin Timberlake	4:11	Pantera	5:10	Foo Fighters	5:10	Townes van Zandt	5:10
N'Sync	4:11	Alice Cooper	4:11	The Smiths	3:12	Suzanne Vega	4:11
Shakira	3:12	Judas Priest	3:12	Alice in Chains	3:12	Crosby Stills & Nash	3:12
Spice Girls	3:12	Sepultura	0.75	Belle and Sebastian	3:12	Tim Buckley	2:13
O-Town	1:14	Skid Row	1:14	Jane's Addiction	1:14	Steeleye Span	1:14
Nelly Furtado	1:14	Queensryche	0:15	Echo and the Bunnymen	0:15	Woodie Guthrie	0:15

Table 1. Artist ranking according to original prototypicality for selected genres of the test set. Moreover, the *backlink/forward link (bl/fl) ratio* within the genre is shown for every artist.

Pop			*Hard & Heavy*			*Alternative/Indie*			*Folk*		
artist ranking	**rnk**	**iop**	**artist ranking**	**rnk**	**iop**	**artist ranking**	**rnk**	**iop**	**artist ranking**	**rnk**	**iop**
Britney Spears	0.20	0.22	Metallica	0.44	0.25	Radiohead	0.18	0.29	Bob Dylan	0.82	0.23
Madonna	0.20	0.17	Slayer	0.38	0.29	Weezer	0.17	0.41	Leonard Cohen	0.70	0.40
Janet Jackson	0.19	0.33	Iron Maiden	0.33	0.39	Beck	0.16	0.23	Donovan	0.65	0.30
Avril Lavigne	0.17	0.32	AC/DC	0.30	0.32	Pearl Jam	0.16	0.36	Joni Mitchell	0.60	0.45
Jennifer Lopez	0.14	0.33	Anthrax	0.24	0.38	Nirvana	0.14	0.18	Cat Stevens	0.55	0.50
Michael Jackson	0.14	0.22	Black Sabbath	0.19	0.38	Sm. Pumpkins	0.13	0.40	John Denver	0.46	0.53
Christina Aguilera	0.12	0.35	Def Leppard	0.17	0.41	Sonic Youth	0.13	0.27	Joan Baez	0.38	0.54
Robbie Williams	0.10	0.36	Kiss	0.16	0.10	Hole	0.12	0.13	Tracy Chapman	0.33	0.57
ABBA	0.09	0.38	Deep Purple	0.14	0.42	Depeche Mode	0.10	0.41	Pete Seeger	0.26	0.57
N'Sync	0.08	0.50	Megadeth	0.11	0.43	Foo Fighters	0.09	0.44	Don McLean	0.20	0.58
Prince	0.06	0.06	Pantera	0.09	0.44	Belle & Sebastian	0.08	0.57	Townes van Z.	0.16	0.60
Justin Timberlake	0.05	0.40	Alice Cooper	0.07	0.45	Alice in Chains	0.06	0.49	Suzanne Vega	0.13	0.61
Spice Girls	0.05	0.48	Judas Priest	0.05	0.46	The Smiths	0.05	0.46	Crosby S. & N.	0.10	0.66
Shakira	0.05	0.46	Sepultura	0.04	0.53	Jane's Addiction	0.02	0.58	Tim Buckley	0.07	0.71
O-Town	0.03	0.64	Skid Row	0.02	0.54	Bush	0.00	0.00	Steeleye Span	0.04	0.75
Nelly Furtado	0.02	0.50	Queensryche	0.00	0.55	Echo and the B.	0.00	0.69	Woodie Guthrie	0.00	1.00

Table 2. Artist ranking according to the improved prototypicality for selected genres of the test set. Moreover, the *ranking value (rnk)* and the value of the *inverse overall prototypicality (iop)* is shown for every artist.

References

1. L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. In *Proc. of the Annual Meeting of the American Society for Information Science*, pages 161–172, January 1998.
2. G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.
3. M. Schedl, P. Knees, and G. Widmer. A Web-Based Approach to Assessing Artist Similarity using Co-Occurrences. In *Proc. of the 4th International Workshop on Content-Based Multimedia Indexing*, June 2005.
4. M. Schedl, P. Knees, and G. Widmer. Discovering and Visualizing Prototypical Artists by Web-based Co-Occurrence Analysis. 2005. submitted.