DISCOVERING AND VISUALIZING PROTOTYPICAL ARTISTS BY WEB-BASED CO-OCCURRENCE ANALYSIS

Markus Schedl [1,2]
markus.schedl@jku.at

Peter Knees [1]
peter.knees@jku.at

Gerhard Widmer [1,2]
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

Detecting artists that can be considered as prototypes for particular genres or styles of music is an interesting task. In this paper, we present an approach that ranks 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. We demonstrate our approach on a data set containing 224 artists from 14 genres and evaluate the results using the rank correlation between the prototypicality ranking and a ranking obtained by page counts of search queries to Google that contain artist and genre. High positive rank correlations are achieved for nearly all genres of the data set. Furthermore, we elaborate a visualization method that illustrates similarities between artists using the prototypes of all genres as reference points. On the whole, we show how to create a prototypicality ranking and use it, together with a similarity matrix, to visualize a music repository.

Keywords: prototypical artist detection, visualization, asymmetric artist similarity, web mining, co-occurrence analysis

1 INTRODUCTION

Finding artists that define a music genre or style, or at least are very typical for it, is a challenging and interesting task. Information on prototypical artists may be used in various areas of application. For example, music information systems like the "All Music Guide"[1] or the "Desdichado Music Information System"[2] as well as online music stores, e.g. "Amazon"[3], could benefit considerably. For instance, information on prototypes could be exploited to support their users in finding music more efficiently.

[1] *http://www.allmusic.com*
[2] *http://www.music-i-s.com*
[3] *http://www.amazon.com*

Furthermore, prototypical artists are very useful for visualizing music repositories since they are usually well-known. Thus, also unexperienced music listeners are able to assign them to a particular genre or style of music and can use them as reference points to discover similar but less known artists. One possible way of visualizing prototypical artists and the relations to their most similar neighbors will be shown in this paper.

The approach presented here can be used to define complete rankings based on the prototypicality of artists. Such rankings enable further applications. For example, together with genre information, they can serve as a measure of the degree of artist membership in a particular genre, thus defining to which extent an artist produces music of a certain style or genre.

Prototypicality is strongly related to the topic of similarity measurement. In fact, we exploit information on co-occurrences of artist names on web pages to estimate conditional probabilities for an artist to be found on web pages of other artists. These probabilities give an asymmetric similarity matrix which is used for the calculation of a prototypicality ranking.

Using the World Wide Web for information retrieval and data mining offers the advantage of incorporating the knowledge and opinions of a large number of different people. Thus, the Internet reflects a kind of cultural knowledge that we extract and use for estimating artist similarity and, subsequently, for prototype detection. However, web-based information retrieval and data mining techniques also face some problems: First, they obviously depend on the existence of web pages dealing with the requested topic. If such web pages cannot be found, e.g. because the query for the search engine cannot be defined adequately or comprises ambiguous words, web-based data mining does not yield valuable results. For example, a search for music-related web pages that offer information about artists like "Bush", "Kiss", or "Porn" will most probably result in a large number of web pages not dealing with these artists.[4] Nevertheless, we already showed that web-based co-occurrence analysis can be used successfully for artist similarity measurement and artist-to-genre classification (Schedl et al., 2005). In this paper, we make use of the asymmetric similarity measure given by the probability estimation and show how to use

[4] To overcome this problem, we restrict the search by adding music-related keywords to the queries.

it for defining an artist prototypicality ranking.

The remainder of this paper is organized as follows. Related work is briefly summarized in Section 2. In Section 3, we present our approach to prototype detection and the performed evaluation, and we discuss the results. Section 4 describes our *"Continuous Similarity Ring (CSR)"* visualization that is used to illustrate relations between prototypical artists and their most similar neighbors. Finally, in Section 5, we summarize the work, draw conclusions, and point out possible future research directions.

2 RELATED WORK

While we could not find previous work on prototype detection for music artists, there has been some work on co-occurrence analysis in music information retrieval. One of the first publication on MIR-related co-occurrence analysis is (Pachet et al., 2001), where playlists of radio stations and databases of compilation CDs are used to detect co-occurrences between titles and between artists. In (Ellis et al., 2002; Whitman and Lawrence, 2002), first attempts to exploit the cultural knowledge offered by the World Wide Web can be found. User collections of the music sharing service "OpenNap" are analyzed to gain a similarity measure based on community metadata. The artist co-occurrences extracted from these collections are evaluated by comparison with direct subjective similarity judgments obtained via a web-based survey. In contrast to this survey of non-professionals, Cano and Koppenberger (2004) use expert opinions taken from the "All Music Guide" to create a similarity network. To this end, the "similar artists" links of 400 artists are gathered. Furthermore, co-occurrences on playlists from "The Art of the Mix"[5] are extracted and visualized as a network containing more than 48.000 artists.

Zadel and Fujinaga (2004) also investigate co-occurrences of artist names on web pages. In contrast to our work, Zadel and Fujinaga (2004) focus on the usage of web services for creating clusters of similar music artists. Starting with a seed artist, the Amazon web service "Listmania!" is used to obtain a list of potentially related artists. Based on this list, co-occurrences are derived by querying Google. Thereafter, the "relatedness" of each "Listmania!"-artist to the seed artist is calculated as the ratio between the combined page count and the minimum of the single page counts for both artists. In contrast to our co-occurrence approach, the one used in (Zadel and Fujinaga, 2004) does not yield complete similarity matrices.

In this paper, we use the same technique as described in (Schedl et al., 2005) to obtain a similarity matrix based on co-occurrences. We query Google for combinations of artist names and use the resulting page counts to estimate conditional probabilities that give an asymmetric similarity matrix.

Similarity measures based on subjective or cultural opinions are in general asymmetric. For example, it is more natural to say that the Finnish heavy metal band "Sentenced" sounds like the well-known pioneers "Metallica" than vice versa. This can be explained by the fact that "Metallica" serves as a prototype for the genre heavy metal. Ellis et al. (2002) regard this asymmetry as a problem since it undermines a Euclidean model of similarity. In fact, nearly all of the cited publications dealing with co-occurrence-based similarity measurement consider the asymmetry a shortcoming and perform operations to symmetrize the similarity matrices. To contrast, in this paper, we describe a prototype detection approach that capitalizes on asymmetric similarity matrices.

[5] *http://www.artofthemix.org*

3 PROTOTYPE DETECTION

In the following, we sketch how we use co-occurrence analysis to define an asymmetric similarity measure. To this end, we apply the same technique as in (Schedl et al., 2005). Based on this similarity measure, we then elaborate our novel method for calculating the prototypicality ranking.

3.1 Methodology

3.1.1 Co-Occurrence Analysis

Given a list of artist names, we use Google to estimate the number of web pages containing each artist and each pair of artists. Since we are not interested in the content of the found web pages, but only in their number, the search is restricted to display only the top-ranked page. In fact, the only information we use is the page count that is returned by Google. This raises performance and limits web traffic.

Addressing the issue of finding only music-related web pages, we add additional keywords to the Google search query. More precisely, we use the scheme ***"artist1" ["artist2"]*+music+review** to form queries. This scheme, already used in (Whitman and Lawrence, 2002), proved to yield good results for classification tasks (Knees et al., 2004; Schedl et al., 2005). Furthermore, it performed slightly better than ***"artist1" ["artist2"]*+music+genre+style** in first experiments of prototype detection.

The outcome of the querying procedure is a symmetric matrix C, where element c_{ij} gives the number of web pages containing the artist with index i together with the one indexed by j. The values of the diagonal elements c_{ii} show the total number of web pages containing artist i. Based on the page count matrix C, we then use relative frequencies to calculate a conditional probability matrix P as follows. 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) as shown in Formula 1.

$$p\left(a_i \wedge a_j \mid a_i\right) = \frac{c_{ij}}{c_{ii}} \tag{1}$$

P gives a similarity matrix that is obviously not symmetric. It can be symmetrized and used, for example, for classifying new artists into a given genre taxonomy, e.g. (Schedl et al., 2005), for generating playlists with similar pieces of music, e.g. (Aucouturier and Pachet, 2002; Logan, 2002), or for visualizing music repositories, e.g.

(Pampalk et al., 2003; Schedl, 2003). In contrast, we can also benefit from the asymmetry of P and use it for prototype detection as described in the following.

3.1.2 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 and build a model upon this consideration.

Our approach is based on an idea similar to the "PageRank Citation Ranking" (Page et al., 1998) used by Google. Like Page et al. (1998), we use information about the number of *backlinks* and *forward links* of a web page. Page et al. (1998) define a forward link of a web page w as a link that is placed on w and links to another web page. A backlink of a web page w, in contrast, is defined as a link on any web page other than w that links to w.

Since we investigate co-occurrences rather than links, we slightly modify the above definitions. In our model for prototypicality ranking, we calculate the number of backlinks of an artist of interest a by focusing a and counting how many web pages that are known to mention another artist also mention artist a. Thus, 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.

Using these definitions, we create a model for prototypicality ranking. To obtain the ranking of an artist of interest a_i, we investigate, for each artist tuple $(a_i,\, a_{j,\, j \neq i})$, whether the number of backlinks of a_i from a_j exceeds the number of forward links of a_i to a_j. We count, for how many of the artists a_j from the same genre as a_i this is the case. The larger this count, the more often artist a_i is mentioned in the context of another artist from the same genre and thus, the higher the prototypicality of a_i for the respective genre.

More formally, using the similarity matrix P, we define a ranking function r for each artist a_i (i is the index of the artist in P) as shown in Formula 2. Here, n is the total number of artists and $bwl\,(i,j)$ and $fwl\,(i,j)$ are functions that return boolean values, cf. Formulas 3 and 4 respectively. These functions use the estimated conditional probabilities as already defined in Formula 1.

$$r\,(a_i) = \frac{\sum_{j=1}^{n,\, j \neq i} bwl(i,j)}{\sum_{j=1}^{n,\, j \neq i} fwl(i,j)} \tag{2}$$

$$bwl\,(i,j) = \begin{cases} 1 & \text{if } p_{ij} < p_{ji} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$fwl\,(i,j) = \begin{cases} 1 & \text{if } p_{ij} \geq p_{ji} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

More precisely, $bwl\,(i,j)$ gives the value 1 if artist a_i has more backlinks from artist a_j (relative to the total number of web pages mentioning artist a_i) than forward links to artist a_j (relative to the total number of web pages mentioning artist a_j). Analogously, $fwl\,(i,j)$ returns the value 1 if artist a_i has more forward links to artist a_j (relative to the total number of web pages mentioning artist a_i) than backlinks from artist a_j (relative to the total number of web pages mentioning artist a_j).

We call $r\,(a_i)$ the *backlink/forward link (bl/fl) ratio* of artist a_i since it counts how often the relative frequency of backlinks for a_i exceeds the relative frequency of its forward links and relates these two counts. Our assumption is that $r\,(a_i)$ measures the prototypicality of artist a_i since its value is the higher the more web pages of other artists mention artist a_i.

3.2 Evaluation

We applied the methods described above to obtain an intra-genre-similarity ranking for a test collection of 224 artists in 14 quite general genres. The results can be found in Table 1 and will be discussed in the next section.

Evaluating the quality of the results is a very difficult task since the prototypicality of an artist for a genre cannot be defined easily and is also affected by subjective and cultural opinions. For the latter reason, we decided to use the World Wide Web again for evaluation. We tried to create a ranking of the degree of artist membership and popularity for all artists of a genre to estimate a prototypicality ranking. To this end, we used the query scheme ***"artist"+"genre"*+music+review** and retrieved the page count for each artist. We then ranked the artists of a genre according to these page counts and computed the Spearman's rank correlation, e.g. (Hogg et al., 2004), between this ranking and the one given by the bl/fl ratios. The Spearman's rank correlation coefficients for each genre can be found in Table 2.

Using the page counts obtained by the queries ***"artist"+"genre"*+music+review** for evaluation works well if a taxonomy of well-defined genres is given. However, applying the bl/fl approach to an artist set which is structured according to another taxonomy (e.g. mood, nationality of the artist, or personal attributes a user may utilize) would probably need another evaluation method since the used query scheme may not give a useful ranking.

3.3 Results and Discussion

Table 1 shows the artist prototypicality ranking for each genre of the test collection. Taking a closer look reveals that the top-ranked artists are usually known to be very famous and typical for the respective genre, whereas the artists at the lower end of the ranking seem to be less typical, at least in some cases. However, we have to admit that the used artist collection has originally been composed for a different purpose and therefore barely contains artists which are unknown to the interested music listener. Thus, none of the artists of the test collection is really untypical for the respective genre. Further evaluations on a test set comprising more than 950 artists with highly varying popularities are currently in preparation (cf. Section 5).

Interestingly, artist names which are also used in everyday speech are always top-ranked, for example, "Kiss" from the genre "Heavy Metal/Hard Rock", "Bush", "Hole", and "Nirvana" from "Alternative Rock/Indie", or "Prince" and "Madonna" from "Pop". The reason for this is that such common speech words occur very often on

Country		*Folk*		*Jazz*		*Blues*	
artist ranking	**bl/fl**	**artist ranking**	**bl/fl**	**artist ranking**	**bl/fl**	**artist ranking**	**bl/fl**
Johnny Cash	15:0	Bob Dylan	15:0	Miles Davis	15:0	BB King	15:0
Willie Nelson	14:1	Donovan	14:1	Duke Ellington	14:1	Taj Mahal	14:1
Tim McGraw	12:3	Leonard Cohen	13:2	Louis Armstrong	13:2	Muddy Waters	13:2
Dixie Chicks	12:3	Joni Mitchell	12:3	Count Basie	12:3	Etta James	11:4
Hank Williams	10:5	Cat Stevens	11:4	John Coltrane	10:5	Howlin' Wolf	10:5
Dolly Parton	10:5	John Denver	10:5	Ella Fitzgerald	10:5	John Mayall	10:5
Faith Hill	9:6	Joan Baez	9:6	Billie Holiday	9:6	John Lee Hooker	10:5
Kenny Chesney	8:7	Tracy Chapman	8:7	Nat King Cole	7:8	Albert King	8:7
Kenny Rogers	7:8	Pete Seeger	7:8	Herbie Hancock	6:9	T-Bone Walker	7:8
Garth Brooks	7:8	Don McLean	6:9	Thelonious Monk	5:10	Willie Dixon	6:9
Kris Kristofferson	6:9	Townes van Zandt	5:10	Charlie Parker	5:10	Lightnin' Hopkins	5:10
Roger Miller	4:11	Suzanne Vega	4:11	Nina Simone	5:10	Mississippi John Hurt	4:11
Jim Reeves	2:13	Crosby Stills & Nash	3:12	Glenn Miller	5:10	Blind Lemon Jefferson	4:11
Brooks and Dunn	2:13	Tim Buckley	2:13	Django Reinhardt	2:13	Otis Rush	2:13
Hank Snow	2:13	Steeleye Span	1:14	Dave Brubeck	2:13	Big Bill Broonzy	1:14
Lee Hazlewood	0:15	Woodie Guthrie	0:15	Cannonball Adderley	0:15	Blind Willie McTell	0:15
RnB/Soul		*Heavy Metal/Hard Rock*		*Alternative Rock/Indie*		*Punk*	
artist ranking	**bl/fl**	**artist ranking**	**bl/fl**	**artist ranking**	**bl/fl**	**artist ranking**	**bl/fl**
Alicia Keys	15:0	Kiss	15:0	Bush	15:0	Green Day	15:0
James Brown	14:1	Metallica	14:1	Hole	14:1	Ramones	14:1
Marvin Gaye	13:2	Slayer	13:2	Nirvana	13:2	Blink 182	13:2
Jill Scott	11:4	AC/DC	12:3	Beck	12:3	The Clash	12:3
The Temptations	11:4	Iron Maiden	11:4	Radiohead	11:4	Sum 41	11:4
Aretha Franklin	11:4	Anthrax	10:5	Sonic Youth	10:5	Sex Pistols	10:5
Al Green	9:6	Black Sabbath	9:6	Pearl Jam	9:6	Rancid	8:7
The Supremes	8:7	Def Leppard	8:7	Weezer	8:7	NoFX	8:7
Erykah Badu	7:8	Deep Purple	7:8	Smashing Pumpkins	7:8	Bad Religion	7:8
Otis Redding	6:9	Megadeth	6:9	Depeche Mode	6:9	Pennywise	6:9
Isaac Hayes	5:10	Pantera	5:10	Foo Fighters	5:10	Dead Kennedys	5:10
Sam Cooke	4:11	Alice Cooper	4:11	The Smiths	3:12	Buzzcocks	4:11
India Arie	3:12	Judas Priest	3:12	Alice in Chains	3:12	Patti Smith	4:11
Fats Domino	2:13	Sepultura	2:13	Belle and Sebastian	3:12	The Misfits	2:13
Solomon Burke	1:14	Skid Row	1:14	Jane's Addiction	1:14	Sid Vicious	1:14
The Drifters	0:15	Queensryche	0:15	Echo and the Bunnymen	0:15	Screeching Weasel	0:15
Rap/Hip-Hop		*Electronica*		*Reggae*		*Rock 'n' Roll*	
artist ranking	**bl/fl**	**artist ranking**	**bl/fl**	**artist ranking**	**bl/fl**	**artist ranking**	**bl/fl**
Eminem	15:0	Moby	15:0	Bob Marley	15:0	The Who	15:0
Jay-Z	14:1	Underworld	13:2	Peter Tosh	14:1	The Animals	14:1
Snoop Dogg	13:2	Prodigy	13:2	Inner Circle	13:2	Elvis Presley	13:2
50 Cent	12:3	Chemical Brothers	12:3	Shaggy	12:3	The Faces	12:3
DMX	11:4	Fatboy Slim	11:4	Sean Paul	11:4	The Rolling Stones	11:4
2Pac	10:5	Kraftwerk	10:5	UB40	10:5	Buddy Holly	9:6
Dr. Dre	9:6	Massive Attack	9:6	Jimmy Cliff	9:6	Chuck Berry	9:6
Ice Cube	8:7	Aphex Twin	8:7	Ziggy Marley	8:7	The Kinks	9:6
Public Enemy	7:8	Paul Oakenfold	6:9	Desmond Dekker	7:8	Jerry Lee Lewis	7:8
LL Cool J	6:9	Basement Jaxx	6:9	Bounty Killer	6:9	Little Richard	6:9
Cypress Hill	5:10	Daft Punk	6:9	Black Uhuru	5:10	Bo Diddley	5:10
Busta Rhymes	4:11	Mouse on Mars	4:11	Capleton	4:11	The Yardbirds	4:11
Run DMC	3:12	Moloko	3:12	Shabba Ranks	3:12	Carl Perkins	2:13
Missy Elliot	2:13	Carl Cox	3:12	Maxi Priest	2:13	Chubby Checker	2:13
Mystikal	1:14	Armand van Helden	1:14	Alpha Blondy	1:14	Gene Vincent	1:14
Grandmaster Flash	0:15	Jimi Tenor	0:15	Eddy Grant	0:15	Bill Haley	1:14
Pop		*Classical*					
artist ranking	**bl/fl**	**artist ranking**	**bl/fl**				
Prince	15:0	Johann Sebastian Bach	14:1				
Madonna	14:1	Tchaikovsky	14:1				
Britney Spears	13:2	Ludwig van Beethoven	13:2				
Michael Jackson	12:3	Wolfgang Amadeus Mozart	12:3				
Avril Lavigne	10:5	Richard Wagner	11:4				
Janet Jackson	10:5	Johannes Brahms	10:5				
Jennifer Lopez	9:6	Franz Schubert	9:6				
Christina Aguilera	8:7	Giuseppe Verdi	8:7				
Robbie Williams	7:8	Antonio Vivaldi	7:8				
ABBA	6:9	Gustav Mahler	6:9				
Justin Timberlake	4:11	Joseph Haydn	5:10				
N'Sync	4:11	Herbert von Karajan	4:11				
Shakira	3:12	Yehudi Menuhin	3:12				
Spice Girls	3:12	Antonin Dvorak	3:12				
O-Town	1:14	Frederic Chopin	1:14				
Nelly Furtado	1:14	Georg Friedrich Haendel	0:15				

Table 1: Artist ranking according to prototypicality for each genre. Furthermore, the *backlink/forward link (bl/fl) ratio* is shown for every artist.

genre	rank correlation
Country	0.96
Folk	0.89
Jazz	0.92
Blues	0.96
RnB/Soul	0.67
Heavy Metal/Hard Rock	0.57
Alternative Rock/Indie	0.57
Punk	0.96
Rap/Hip-Hop	0.55
Electronica	0.76
Reggae	0.81
Rock 'n' Roll	0.76
Pop	0.95
Classical	0.69
mean	**0.79**

Table 2: Spearman's Rank Correlations between ranking of artist names according to *backlink/forward link prototypicality* and *artist-genre-page counts* for each genre.

artists' web pages and therefore produce a lot of backlinks for the respective artist with the same name. However, they usually do not refer to the artist, but simply denote the common speech word. To give an example, finding many co-occurrences of "Bush" and "Michael Jackson" does not necessarily mean that these artists create similar music. It could also mean that Michael Jackson had a meeting with the current president of the US or that Mr. Jackson likes bushes on his "Neverland"-ranch.

Such misleading co-occurrences are a shortcoming of web-based information retrieval methods and could also distort the prototypicality ranking. For example, the authors would not attest "Bush" a higher prototypicality than "Nirvana" for the genre "Alternative Rock/Indie". However, we could turn the tables and use our prototype detection approach to find artist names that equal common speech words by investigating the terms with outstandingly high bl/fl ratios. To this end, the bl/fl ratios should be calculated on the complete artist set rather than for each genre separately since common speech words appear on artists' web pages independently of their genre. Indeed, performing these computations on our test collection reveals that the artists which show by far the highest bl/fl ratios are "Bush" (223:0), "Prince" (222:1), "Kiss" (221:2), "Madonna" (220:3), and "Nirvana" (218:5).

As for the results of the evaluation, Table 2 shows the Spearman's rank correlation coefficients for each of the 14 genres of the test collection. The prototypicality ranking given by the bl/fl ratios and the evaluation ranking obtained by the artist-genre-page counts show strong or even very strong positive correlations. Especially for the genres "Country", "Blues", "Punk", and "Pop", the prototypicality ranking nearly perfectly correlates with the evaluation ranking. In contrast, the results for the genres "Heavy Metal/Hard Rock", "Alternative Rock/Indie", and "Rap/Hip-Hop" are situated at the other end of the performance scale with correlation coefficients of about 0.55 which is nevertheless an indication for strongly correlating rankings.

4 VISUALIZATION

In order to visualize the prototypical artists that were identified (together with similar artists), we developed a novel method which we call the *"Continuous Similarity Ring (CSR)"*. A sample screenshot taken from the music information retrieval and visualization framework "CoMIRVA"[6] which we are developing at our department can be found in Figure 1.

The basic idea is to display the prototypes – one for each genre – in the form of a circle. Since similar or related prototypes and the genres they represent should be placed close to each other, we formulate a *Traveling Salesman Problem (TSP)*, e.g. (Lawler et al., 1985; Skiena, 1997), and apply a simple heuristic algorithm. To this end, we use a symmetrized version P_s of the similarity matrix P (cf. Formula 1) which we obtain by calculating the arithmetical mean of p_{ij} and p_{ji} for every pair of artists i and j. Subsequently, we convert P_s into a distance matrix. The TSP-algorithm then tries to find the shortest path between all prototypes. Thus, it gives a tour that passes all prototypes and minimizes the overall distance. The resulting tour defines the arrangement of the artists within the circle of prototypes.

Since we also want to show which artists are similar to which prototypes, we again use the symmetrized similarity matrix P_s and select, for each prototype r, a fixed number k of artists with minimal distance to r. These k neighbors are chosen from the complete artist set regardless of their genre assignment, which enables the user to easily detect artists that are inspired by musicians of different genres. Hence, unlike the prototype detection approach, the CSR-visualization technique does not rely on genre information, provided that a list of prototypes is available.

Given the set of nearest neighbors N_r for each prototype r, we investigate which artists are neighbor of only one prototype (inserted into artist set O), and which neighbor more than one (inserted into artist set I). The goal is to point out artists which cannot be classified exactly into one genre and thus neighbor several prototypes. For visualizing, we use the region outside of the circle of prototypes to display the artists contained in O since they need to be connected only to their single prototype. Artists of the set I are mapped to the area inside of the circle of prototypes and are connected to all prototypes r containing them in N_r. For these artists, special handling is necessary since we want to preserve the original distances between the artists and their prototypes as given by P_s. Furthermore, the length of the edges connecting prototypes and neighbors should be minimized in order to avoid overloading of the visualization. Thus, we use a heuristic cost-minimizing algorithm to position the artists of set I. The costs c_n for an artist $n \in I$ are calculated as shown in Formula 5, where P_n is the set of prototypes that are connected to artist n, $origDist(r, n)$ is the original distance between prototype r and neighbor n according to the similarity matrix, $origDistSum$ is the sum of the original distances between n and all elements of P_n, $screenDist(r, n)$ is the distance on the screen between

[6] *http://www.cp.jku.at/comirva*

the vertex representing prototype r and the vertex representing neighbor n, and $screenDistSum$ is the sum of the screen distances between the vertex representing n and all vertices that represent an element of P_n.

$$c_n = \sum_{r \in P_n} \left(\frac{origDist(r,n)}{origDistSum} - \frac{screenDist(r,n)}{screenDistSum} \right) \quad (5)$$

The algorithm for positioning the vertex of a neighbor $n \in I$ comprises three steps which are performed iteratively (5000 iterations seemed to be a good choice for the used artist set).

1. The vertex of the current neighbor $n \in I$ is initially positioned in the center of the screen.
2. This position is then randomly modified by a small amount (we restricted the movement to a maximum of 10 pixels in each direction).
3. The costs for this new position are calculated and the vertex is moved to the new position if an improvement in costs and in the $screenDistSum$ (for minimizing the length of the edges) can be achieved.

Figure 1 shows a screenshot of a CSR-visualization which is based on the prototypes of the used artist collection. The three nearest neighbors of each prototype are depicted and edges connecting these neighbors with the respective prototypes are drawn. Varying thickness and color of the edges reveal information about the similarity values of the artists they connect. Thick and dark edges connect very similar artists, whereas thinner and lighter edges connect artists with lower similarity values. Regarding Figure 1, it can be seen that the only prototype whose neighbors are not connected to any other prototype is "Johann Sebastian Bach". Thus, we can state that classical artists are very well distinguishable from artists of other genres. We also see that "Nirvana" is one of the three nearest neighbors of "Green Day", "Kiss", and "Johnny Cash" which does make sense to some extent since "Alternative Rock/Indie" is related to "Punk" and also "Heavy Metal/Hard Rock" is not that far away. Unfortunately, artists whose name equal common speech words dominate the region inside of the circle of prototypes. However, this problem arises only for small values of k (number of displayed neighbors for each prototype). Using $k = 5$, for example, reveals more interesting relations. Due to limited space in and resolution of this paper we unfortunately cannot depict such a detailed CSR-visualization. The interested reader is invited to experiment with the CoMIRVA-framework and create his/her own CSR.

A possible application scenario for the CSR-visualization technique could be its usage in online music stores. Prototypical artists according to a set of genres, or any other useful taxonomy (e.g. mood), can be seen as reference points for the user since they are usually well-known. Starting at these prototypes, the user could utilize the CSR-visualization to explore similar but less known artists. Moreover, focusing an artist which has been selected arbitrarily by the user, the influence of different prototypical artists and their genres (or other attributes according to the used taxonomy) on the artist under consideration could easily be made out when using a CSR-visualization.

Also (music) search engines could apply the prototypicality ranking technique (maybe together with the CSR-visualization approach) to support their users in discovering less known artists based on an entered or selected prototype. On the other hand, if the user entered a less prototypical artist, the system could provide a list of artists that may have influenced the entered one.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we showed how to use co-occurrences of artist names on web pages to calculate an asymmetric similarity matrix. Based on this similarity matrix, we estimate a prototypicality ranking for the artists using *backlink/forward link ratios*. We evaluated our approach on a test collection containing 224 artists of 14 genres. Using the page counts obtained by search queries that comprise artist name and genre, we calculated the rank correlation and showed that the prototypicality ranking given by the bl/fl ratios correlates well with the evaluation ranking given by the artist-genre-page counts. Furthermore, we presented a visualization approach called *"Continuous Similarity Ring (CSR)"* that makes use of the extracted prototypes of each genre.

A shortcoming of the used test collection is that most of its artists are quite popular and typical of their genre. Addressing this issue, it is planned to evaluate our approach on a larger artist set containing 953 artists from 15 genres. To create this artist set, we used the artist database of the "All Music Guide". We chose ten very general and five quite specific genres and selected the artists assigned the highest and the lowest tier in the genre-specific artist list of the "All Music Guide". This provides a mix of very well-known artists and artists which are not that popular. Unfortunately, calculating the co-occurrences of such a large artist set would include raising more than 450.000 queries. Since the Google Web-API[7] allows only 1.000 queries per day, using it is out of the question. Thus, we have to elaborate other approaches, which will be done in the near future. For example, we could use the Google Web-API to retrieve the URLs of the top-ranked web pages for each artist. The content of the web pages addressed by these URLs may then be extracted and scanned for the names of all other artists in the artist set. Storing the relative frequecies of artist names appearing on other artists top-ranked web pages would give us a co-occurrence matrix which could be used to estimate similarities again.

Another interesting issue would be the evaluation of the bl/fl prototypicalities on a ranking created by musical experts, ideally from different cultures. Since this would be hardly feasible, a web-based survey of music lovers from all over the world could be conducted instead.

ACKNOWLEDGEMENTS

This research is supported by the Austrian Fonds zur Förderung der Wissenschaftlichen Forschung (FWF) un-

[7] *http://www.google.com/apis*

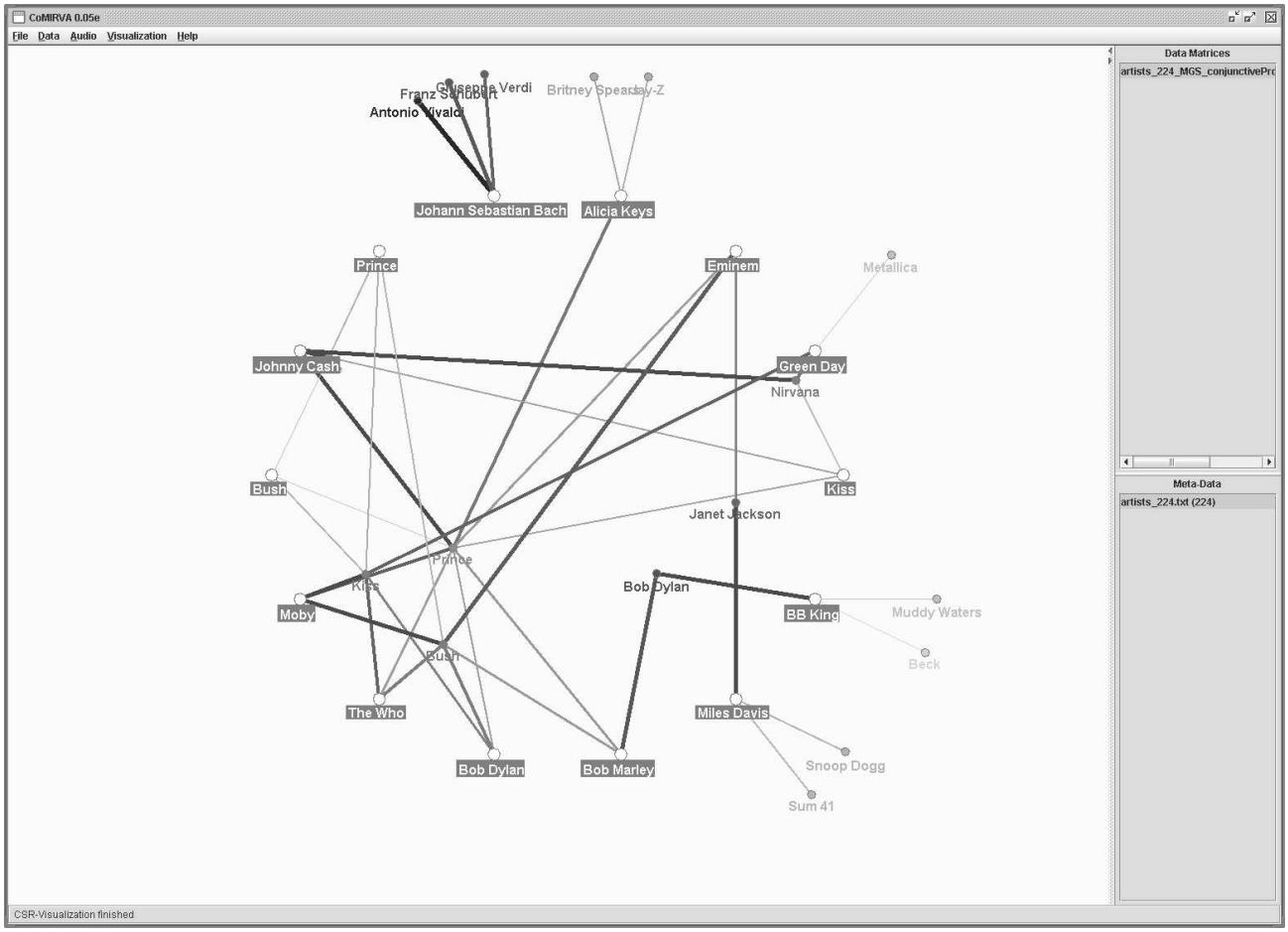

Figure 1: A CSR-visualization based on the prototypes of the test collection and the similarity matrix. Every prototype is connected to its three nearest neighbors.

der 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.

REFERENCES

J.-J. Aucouturier and F. Pachet. Scaling up music playlist generation. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*. Lausanne, Switzerland, August 2002.

P. Cano and M. Koppenberger. The Emergence of Complex Network Patterns in Music Artist Networks. In *Proceedings of the 5th International Symposium on Music Information Retrieval (ISMIR'04)*, Barcelona, Spain, October 2004.

D. P. W. Ellis, B. Withman, A. Berenzweig, and S. Lawrence. The Quest for Ground Truth in Musical Artist Similarity. In *Proceedings of the 3rd International Symposium on Music Information Retrieval (ISMIR'02)*, Paris, France, 2002.

R. V. Hogg, A. Craig, and J. W. McKean. *Introduction to Mathematical Statistics*. Prentice Hall, 6th edition, June 2004.

P. Knees, E. Pampalk, and G. Widmer. Artist Classification with Web-based Data. In *Proceedings of the 5th International Symposium on Music Information Retrieval (ISMIR'04)*, pages 517–524, Barcelona, Spain, October 2004.

E. L. Lawler, J. K. Lenstra, A. H. G. R. Kan, and D. B. Shmoys. *The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization*. Wiley Series in Discrete Mathematics and Optimization. John Wiley and Sons, September 1985.

B. Logan. Content-based Playlist Generation: Exploratory Experiments. In *Proceedings of the 3rd International Symposium on Music Information Retrieval (ISMIR'02)*, Paris, France, October 2002.

F. Pachet, G. Westerman, and D. Laigre. Musical Data Mining for Electronic Music Distribution. In *Proceedings of the 1st WedelMusic Conference*, 2001.

L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. In *Proceedings of the Annual Meeting of the American Society for Information Science (ASIS'98)*, pages 161–172, January 1998.

E. Pampalk, S. Dixon, and G. Widmer. Exploring Music Collections by Browsing Different Views. In *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR'03)*, Washington, D.C., USA, October 2003.

M. Schedl. An Explorative, Hierarchical User Interface to Structured Music Repositories. Master's thesis, Vienna University of Technology, Austria, December 2003.

M. Schedl, P. Knees, and G. Widmer. A Web-Based Approach to Assessing Artist Similarity using Co-Occurrences. In *Proceedings of the Fourth International Workshop on Content-Based Multimedia Indexing (CBMI'05)*, Riga, Latvia, June 2005.

S. S. Skiena. *The Algorithm Design Manual*. Springer, November 1997.

B. Whitman and S. Lawrence. Inferring Descriptions and Similarity for Music from Community Metadata. In *Proceedings of the 2002 International Computer Music Conference*, pages 591–598, Goeteborg, Sweden, September 2002.

M. Zadel and I. Fujinaga. Web Services for Music Information Retrieval. In *Proceedings of the 5th International Symposium on Music Information Retrieval (ISMIR'04)*, Barcelona, Spain, October 2004.