Investigating Different Term Weighting Functions for Browsing Artist-Related Web Pages by Means of Term Co-Occurrences

Markus Schedl and Peter Knees
{markus.schedl, peter.knees}@jku.at

Department of Computational Perception
Johannes Kepler University
Linz, Austria
http://www.cp.jku.at

Abstract. We present a user interface (UI) for browsing collections of web pages about music artists. Given such a collection, we use a term list to index its contents and to derive term co-occurrences. Based on these co-occurrences, we create a UI that employs a variant of the *Sunburst* visualization technique. The UI is embedded in *CoMIRVA*, our framework for music information retrieval and visualization.
We use two dictionaries of musically relevant terms and derive information about which terms occur on which web pages. Based on this information, subsets of the web page collection are created according to the terms that occur most frequently in the collection. The generated UI, which we call the *Co-Occurrence Browser (COB)*, thus allows for associating each artist with its most important (descriptive) terms and for browsing the respective web pages. To assess the usability of the *COB*, we carried out a small qualitative user study. Furthermore, different term weighting functions used to create the UI were tested and evaluated in a quantitative user study.

1 Introduction and Context

Automatically finding descriptive terms for a given music artist is an important question in music information retrieval (MIR). Such terms may describe, for example, the genre or style of the music performed by the artist under consideration and enable a wide variety of applications, e.g. enriching music players [14], recommending unknown artists based on the favorite artists of the user (recommender systems) [18], or enhancing user interfaces for exploring or browsing music collections [9, 12, 8, 10, 17].

One possibility for assigning musically relevant terms to a given artist is manual annotation by music experts, as it is usually employed by music information systems like the *All Music Guide* [1] or interfaces for music search like *musiclens* [2]. However, this is a very labor-intensive task and barely feasible for huge music collections. An alternative way, which we follow here, is to exploit

today's largest information source, the World Wide Web (WWW). Automatically deriving information about music artists from the WWW is advantageous since it does not only represent the views of a few experts, which are usually bound to one cultural context, but incorporates the opinions of a large number of different people, and thus embodies a kind of cultural knowledge.

In this paper, we present an application – which we henceforth call the *Co-Occurrence Browser (COB)* – that automatically indexes a set of web pages about music artists according to a dictionary of musically relevant terms and organizes these web pages by creating a number of subsets S_{T_i}, each of which is described by a set of r terms $T_i = \{t_{i1}, \ldots, t_{ir}\}$ from the dictionary. Each subset S_{T_i} thus represents those documents of the entire collection in which all terms of the set T_i occur. The subsets are then visualized using a variant of the well-established Sunburst technique [6, 16].

The purpose of the *COB* is twofold. First, it facilitates getting an overview of the web pages related to a music artist by structuring them according to co-occurring terms. Second, since the descriptive terms that most often occur on web pages related to a music artist X constitute an individual profile of X, the *COB* is also suited to reveal various meta-information about the artist, e.g. musical style, related artists, or instrumentation.

The *COB* has been implemented in the context of the *CoMIRVA* framework for music information retrieval and visualization. *CoMIRVA* is presented in [13] and can be downloaded from [3].

2 Web Retrieval and Co-Occurrence Analysis

Given a list of artist names, we first query *Google* with the scheme *"artist name"+music+review* to obtain the URLs of up to 200 web pages related to the artist. We then retrieve the content available at the extracted URLs via *wget* [4].

Subsequently, a term occurrence analysis step is performed. To this end, we use a dictionary containing musically relevant terms, which are searched in all web pages of each artist. We conducted experiments using two dictionaries T_s and T_l. T_s has been compiled by the authors and includes genre names taken from the *Yahoo! Directory* [5] and style names taken from the *All Music Guide*. T_l is basically the same dictionary as used in [11]. It was manually compiled by the authors of [11] by copying lists of genres, instruments, and other descriptive terms from various sources such as *Wikipedia*, *Yahoo! Directory*, and *All Music Guide*. We further added the names of all artists in the collection used for our experiments, which yields a total number of 544 and 1,506 terms for T_s and T_l, respectively.

The outcome of the term occurrence analysis is an inverted file index, i.e. a data structure that stores, for every term of the dictionary, pointers to the web pages that contain the term. From such an inverted file index of an artist X, we can easily extract subsets $S_{X,\{t_1,\ldots,t_r\}}$ of the web page collection of X which have in common the occurrence of all terms $t_1, \ldots, t_r$.

3 Sunburst User Interface

Based on the inverted file index of an artist, we create a user interface that employs a variant of the Sunburst [6, 16], i.e. a radial, space-filling visualization technique for hierarchical data. In most publications related to the Sunburst, its usual application scenario is browsing the hierarchical tree structure of a file system. Since the number of directories and files in a file system is limited, in this application scenario, no further attention has to be paid to restricting the size of the Sunburst. In contrast, for the *COB*, we had to elaborate methods to limit the size of the visualization, as explained below, since this size is in principle only restricted by the number of possible combinations of all terms in the dictionary.

Starting with the entire set of web pages $S_{X,\{\}}$ of an artist X, a user-definable maximum number N of terms with highest value according to some term weighting function (e.g. document frequencies) is selected to create N subsets $S_{X,\{t_1\}}, \ldots, S_{X,\{t_N\}}$ of the collection. These subsets are visualized as filled arcs $A_{X,\{t_1\}}, \ldots, A_{X,\{t_N\}}$ around a centered circle (the root node) $A_{X,\{\}}$ that represents the entire set of web pages retrieved for artist X. The angular extent of each arc is proportional to the weight of the associated term t_i, i.e. to the number of documents containing t_i when using document frequencies for term weighting. To avoid very small arcs that are barely perceivable, arcs whose angular extent is smaller than a fixed threshold E are omitted. Furthermore, each arc is filled with the color given by the colormap selected in *CoMIRVA*'s user interface.

The term selection with respect to term weights and the corresponding visualization steps are recursively performed for all arcs, with a user-definable maximum recursion depth R. This eventually yields a complete Sunburst like the one shown in Figure 1, where each arc at a specific recursion depth r represents a set of web pages $S_{X,\{t_1,\ldots,t_r\}}$ in which all terms $t_1, \ldots, t_r$ co-occur.

Internally, the *COB* stores the Sunburst as a tree, where each arc is represented by a node. A node $A_{X,\{t_1,\ldots,t_r\}}$ at depth r in the tree thus represents the set of web pages that contain the term t_r and all terms $t_1, \ldots, t_{r-1}$ associated with the nodes on the shortest path from $A_{X,\{t_1,\ldots,t_r\}}$ to the root node.

Addressing the fact that in the application scenario of the *COB* the size of the Sunburst is in principle only restricted by the number of possible combinations of all terms in the dictionary, the user can define some stop criteria for complexity limitation: maximum sub nodes per node (by default, $N = 20$), maximum recursion depth (by default, $R = 8$), minimum angular extent of an arc (by default, $E = 1.0$ degree).

As the *COB* is intended to be used for browsing web page collections, user interaction is essential. It is provided in two ways. First, clicking with the left mouse button on an arbitrary arc generates a new Sunburst visualization with this arc as root node, i.e. only the web pages that are represented by the selected arc are used to create the new visualization. Second, a right mouse click on any arc displays a pop-up menu with the URLs of the web pages represented by

the respective arc. The user can then view a web page by selecting it from the pop-up menu.

4 Browsing Artist-Related Web Pages

To demonstrate the *COB*, we compiled a test collection of 112 well-known music artists for which we retrieved a total of 21,594 web pages. For each artist, two inverted indices were created, one using the dictionary T_s, the other using T_l (cf. Section 2). Figure 1 shows the Sunburst generated for the music artist *Britney Spears* using the dictionary T_l. The values in parentheses indicate the document frequency of the term of the respective arc. This sample visualization reveals which combinations of terms most frequently occur in web pages about *Britney Spears*. For example, the term *pop* occurs on 75 of the 124 web pages retrieved for *Britney Spears*. If the user wants to know, for example, in which web pages the terms *Britney Spears (BS)*, *song (s)*, and *vocal (v)* are mentioned together, s/he can easily display a list of the corresponding URLs by clicking on the arc $A_{S_{BS,\{BS,\ s,\ v\}}}$ as shown in Figure 1. A further click on one of the URLs opens the respective web page in the user's preferred web browser.

Figures 2 and 3 show the influence of different constraints for the size of the Sunburst. For the visualization depicted in Figure 2, we used the dictionary T_l. It can easily be seen that the terms that most often occur on web pages related to the music artist *Iron Maiden* are *metal*, *band*, *song*, *guitar*, *world*, *heavy metal*, and *hard*, which seems reasonable. The Sunburst shown in Figure 2 uses the default values for the complexity constraints (cf. Section 3). In contrast, for the visualization depicted in Figure 3, N was reduced to 4 and E to 0.5 degrees. Using these modified constraints, the generated Sunburst contains more arcs at deeper hierarchy levels and therefore provides more detailed information on term co-occurrences. However, this comes at the cost of lucidity. Figure 3 further shows that the terms *Metal*, *Rock*, *World*, and *Epic* most often occur on web pages retrieved for *Iron Maiden* when using the dictionary T_s for indexing.

5 Qualitative Evaluation

We conducted a small qualitative user study to assess the usability of the *COB*. To this end, we asked some computer science students to choose a music artist they are familiar with (out of the 112 contained in our test collection). After having introduced the *COB* and let them play around with it (using both term lists T_s and T_l), we asked the participants the following questions:

- How would you rate the terms displayed with respect to their descriptiveness and their usefulness for browsing the artist's web pages?
- Did you discover any formerly unknown artists?
- Do you have any suggestions for improving the *COB*?

We received detailed responses for the artists *Eminem*, *Sonic Youth*, *Aphex Twin*, *Britney Spears*, *Fatboy Slim*, and *The Kinks*. Due to space limitations, in the following, we can only briefly summarize the most important results.

Almost independent of the artist, the vast majority of the top-ranked terms was rated descriptive or even very descriptive when using the dictionary T_l. Using T_s, however, for three of the six artists the results were unsatisfactory. Especially the top-ranked occurrence of general terms like *band*, *song*, *world*, or *personal* was disliked.

Displaying not only descriptive terms, but also related artists was appreciated by all participants. Two of them, however, requested some kind of highlighting of the artist names to make them clearly distinguishable from the descriptive terms. Since all artists of the collection are well-known, no formerly unknown artist was discovered. At least, all participants stated that the co-occurring artists are similar to the chosen one.

As for comments and suggestions, in general, using a Sunburst-based UI for the purpose of browsing collections of web pages was seen "a very interesting and appealing" application. However, the response times of the UI should definitely be improved. Two participants suggested making the term list extendable by user-defined terms, e.g. *love affair* for *Britney Spears* was mentioned. One user suggested showing a preview of each web page when performing a right mouse click on an arc instead of only displaying the URLs. Another participant requested a *back*-button to return to a higher level after having restricted the shown Sunburst to a subset of web pages.

6 Quantitative Evaluation

We experimented with three different term weighting functions (document frequency, term frequency, TF×IDF) for term selection in the Sunburst creation step, cf. Section 3. Given a set of web pages S of an artist, the *document frequency* DF_t of a term t is defined as the absolute number of documents on which t appears at least once. The *term frequency* TF_t of a term t is defined as the sum of all occurrences of t in S. The *term frequency inverse document frequency* measure $TF \times IDF_t$ of t is calculated as $TF_t \times \ln \frac{|S|}{DF_t}$.
To assess the influence of the term weighting function on the quality of the hierarchical clustering, the hierarchical layout, and thus on the visualization of the COB, we conducted a quantitative user study.

6.1 Setup

For the user study, we chose a collection of 112 well-known artists (14 genres, 8 artists each). Table 1 depicts a list of all artist names. The dictionary used for indexing contains 1,506 musically relevant terms, cf. Section 2. To create the evaluation data, for each artist, we calculated on the complete set of his/her retrieved and indexed web pages, the 10 most important terms using each of the three term weighting functions. To avoid biassing of the results, we combined, for

each artist, the 10 terms obtained by applying each weighting function. Hence, every participant was presented a list of 112 artist names and, for each of these, a set of associated terms (as a mixture of the terms obtained by the three weighting functions). Since the authors had no a priori knowledge of which artists were known by which participant, the participants were told to evaluate only those artists they are familiar with. Their task was then to rate the associated terms with respect to their appropriateness for describing the artist or his/her music. To this end, they had to associate every term to one of the three classes *+ (good description)*, *– (bad description)*, and $\sim$ *(indifferent or not wrong, but not a description specific for the artist)*.
Due to time constraints, we had to limit the number of participants in the user study to five. Three of them are computer science students, the other two researchers in computer science. All of them are male and all stated to listen to music often.

6.2 Results and Discussion

We received a total of 172 assessments for sets of terms assigned to a specific artist. 92 out of the 112 artists were covered. To analyze the results, we calculated, for each artist and weighting function, the sum of all points obtained by the assessments. As for the mapping of classes to points, each term in class + contributes 1 point, each term in class – gives -1 point, and each term in class $\sim$ yields 0 points.
Summing up all points for every artist and term weighting function gives the results shown in columns 3, 4, and 5 of Tables 2 and 3. Furthermore, these tables depict, for every artist assessed at least once, the number of assessments, i.e. the number of participants which assessed the artist (column 2). Since the performance of the term weighting functions are hardly comparable between different artists using the summed up points, columns 6, 7, and 8 illustrate the averaged scores, which are obtained by dividing the summed up points by the number of assessments.
These averaged points reveal that the quality of the terms vary strongly between different artists. Nevertheless, we can state that, for most of the artists, the number of descriptive terms exceeds the number of the non-descriptive ones. Combining the averaged points of all artists separately for each term weighting function to obtain a performance measure for the weighting functions, we calculated the means of columns 6, 7, and 8. These were 2.22, 2.43, and 1.53 for TF, DF, and TF×IDF, respectively. Due to the performed mapping from classes to points, these values can be regarded as the average excess of the number of good terms over the number of bad terms. Hence, overall, we assume that the document frequency measure performed best, the term frequency second best, and the TF×IDF worst.
To test for the significance of these results, we performed Friedman's non-parametric two-way analysis of variance (cf. [7, 15]). This test is similar to the two-way ANOVA, but does not assume a normal distribution of the data. The test yielded a p value of 0.000024. Therefore, we can state that the variance

differences in the results are significant with a very high probability. Moreover, pairwise comparisons between the results given by the three term weighting functions showed that TF×IDF performed significantly worse than both TF and DF, whereas no significant difference could be made out between the results obtained using DF and those obtained using TF.
The laborious task of combining and analyzing the different assessments of the participants in the user study further allowed us to take a qualitative look at the terms. Overall, the majority of the terms was judged descriptive. However, we discovered some interesting flaws. First, the term "musical" occurred on quite a lot of web pages and was therefore often contained in the set of the top-ranked terms. None of the participants judged this term as descriptive for none of the artists. A similar observation could be made for the term "real". In this case, however, one participant stated that this is a term commonly used in the context of Hip-Hop music and therefore can be regarded as being descriptive to some extent for artists of this particular music style. Furthermore, the term "christmas" was associated occasionally to some artists. These associations seem quite random since none of the artists is known for his/her performance of Christmas carols. Another reason for erroneously assigning a term to an artist are terms which are part of artist, album, or song names, but are not suited well to describe the respective artist. Examples for this problem category are "infinite" for the artist "Smashing Pumpkins" and "human" as well as "punk" for the artist "Daft Punk".

7 Conclusions and Future Work

We presented the *Co-Occurrence Browser (COB)*, a user interface for browsing collections of web pages related to music artists via co-occurring terms. The *COB* employs a variant of the Sunburst visualization technique, which we had to adapt to handle the data provided by the applied co-occurrence analysis. We further conducted a small qualitative user study which showed that the *COB* is able to provide interesting views on a set of artist-related web pages and to reveal various descriptive, artist-related properties.

As for future work, we aim at extending the current implementation to index multimedia content found on the retrieved web pages and incorporate this content in the visualization. Furthermore, we are elaborating a three-dimensional version of the user interface.

Moreover, we reported on the results of a user study that was carried out to investigate the performance of different term weighting functions used in the visualization of the COB to determine the sizes of the individual Sunburst arcs. We found that using TF×IDF yielded significantly worse results than the simple TF and DF measures with respect to the appropriateness to describe the music artists. In contrast, comparing the measures TF and DF, no significant difference in the performance was detected.

Moreover, the conducted user study showed that very general terms like *band*, *song*, or *world* occur on many web pages and are thus rated highly relevant by the

simple document frequency measure which is applied to determine the arc sizes. We tried to address this shortcoming by using TF×IDF instead of the simple DF as relevance measure for determining the arc sizes. However, it turned out that using TF×IDF does not significantly reduce the size of those arcs which are associated with very general terms. Hence, we will experiment with techniques for down-ranking terms with exorbitant high popularity.

8 Acknowledgments

This research is supported by the Austrian Fonds zur Förderung der Wissenschaftlichen Forschung (FWF) under project number L511-N15.

References

1. http://www.allmusic.com.
2. http://www.musiclens.de.
3. http://www.cp.jku.at/CoMIRVA.
4. http://www.gnu.org/software/wget.
5. http://dir.yahoo.com/Entertainment/Music/Genres.
6. K. Andrews and H. Heidegger. Information Slices: Visualising and Exploring Large Hierarchies using Cascading, Semi-Circular Discs. In *Proceedings of IEEE Information Visualization 1998 (InfoVis'98)*, Research Triangle Park, NC, USA, October 1998.
7. M. Friedman. A Comparison of Alternative Tests of Significance for the Problem of m Rankings. *The Annals of Mathematical Statistics*, 11(1):86–92, March 1940.
8. M. Goto and T. Goto. Musicream: New Music Playback Interface for Streaming, Sticking, Sorting, and Recalling Musical Pieces. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR'05)*, London, UK, September 2005.
9. P. Knees, M. Schedl, T. Pohle, and G. Widmer. An Innovative Three-Dimensional User Interface for Exploring Music Collections Enriched with Meta-Information from the Web. In *Proceedings of the 14th ACM International Conference on Multimedia (ACM MM'06)*, Santa Barbara, CA, USA, October 2006.
10. F. Mörchen, A. Ultsch, M. Nöcker, and C. Stamm. Databionic Visualization of Music Collections According to Perceptual Distance. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR'05)*, London, UK, September 2005.
11. E. Pampalk, A. Flexer, and G. Widmer. Hierarchical Organization and Description of Music Collections at the Artist Level. In *Proceedings of the 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL'05)*, Vienna, Austria, September 2005.
12. E. Pampalk and M. Goto. MusicRainbow: A New User Interface to Discover Artists Using Audio-based Similarity and Web-based Labeling. In *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR'06)*, Victoria, Canada, October 2006.
13. M. Schedl, P. Knees, K. Seyerlehner, and T. Pohle. The CoMIRVA Toolkit for Visualizing Music-Related Data. In *Proceedings of the 9th Eurographics/IEEE VGTC Symposium on Visualization (EuroVis'07)*, Norrköping, Sweden, May 2007.

14. M. Schedl, T. Pohle, P. Knees, and G. Widmer. Assigning and Visualizing Music Genres by Web-based Co-Occurrence Analysis. In *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR'06)*, Victoria, Canada, October 2006.
15. D. J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman and Hall/CRC, Boca Raton, London, New York, Washington, D.C., 3rd edition, 2004.
16. J. Stasko and E. Zhang. Focus+Context Display and Navigation Techniques for Enhancing Radial, Space-Filling Hierarchy Visualizations. In *Proceedings of IEEE Information Visualization 2000 (InfoVis'00)*, Salt Lake City, UT, USA, October 2000.
17. F. Vignoli, R. van Gulik, and H. van de Wetering. Mapping Music in the Palm of Your Hand, Explore and Discover Your Collection. In *Proceedings of the 5th International Symposium on Music Information Retrieval (ISMIR'04)*, Barcelona, Spain, October 2004.
18. 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.

Country			
Johnny Cash	Willie Nelson	Dolly Parton	Hank Williams
Faith Hill	Dixie Chicks	Garth Brooks	Kenny Rogers
Folk			
Bob Dylan	Joni Mitchell	Leonard Cohen	Joan Baez
Townes van Zandt	Pete Seeger	Suzanne Vega	Tracy Chapman
Jazz			
Miles Davis	Dave Brubeck	Billie Holiday	Duke Ellington
Django Reinhardt	Glenn Miller	Ella Fitzgerald	Louis Armstrong
Blues			
John Lee Hooker	Muddy Waters	Taj Mahal	John Mayall
Big Bill Broonzy	BB King	Howlin' Wolf	Willie Dixon
RnB / Soul			
James Brown	Marvin Gaye	Otis Redding	Solomon Burke
Sam Cooke	Aretha Franklin	Al Green	The Temptations
Heavy Metal / Hard Rock			
Iron Maiden	Megadeth	Slayer	Sepultura
Black Sabbath	Anthrax	Alice Cooper	Deep Purple
Alternative Rock / Indie			
Nirvana	Beck	Smashing Pumpkins	Radiohead
Belle and Sebastian	Alice in Chains	Echo and the Bunnymen	Sonic Youth
Punk			
Patti Smith	Sex Pistols	Sid Vicious	Ramones
Bad Religion	The Clash	NoFX	Dead Kennedys
Rap / Hip-Hop			
Eminem	Dr. Dre	Public Enemy	Missy Elliot
Cypress Hill	50 Cent	Run DMC	Grandmaster Flash
Electronica			
Aphex Twin	Daft Punk	Kraftwerk	Chemical Brothers
Fatboy Slim	Basement Jaxx	Carl Cox	Moloko
Reggae			
Bob Marley	Jimmy Cliff	Peter Tosh	Ziggy Marley
Sean Paul	Alpha Blondie	Shaggy	Maxi Priest
Roch 'n' Roll			
The Rolling Stones	The Animals	The Faces	The Kinks
The Who	Elvis Presley	Chuck Berry	Little Richard
Pop			
Madonna	Britney Spears	N'Sync	Justin Timberlake
ABBA	Michael Jackson	Janet Jackson	Prince
Classical			
Wolfgang Amadeus Mozart	Ludwig van Beethoven	Johann Sebastian Bach	Joseph Haydn
Johannes Brahms	Frederic Chopin	Antonin Dvorak	Gustav Mahler

Table 1. List of the 112 artist names used in the user study.

artist	assessments	TF	DF	TF×IDF	TF (avg)	DF (avg)	TF×IDF (avg)
50 Cent	3	17	16	19	5.67	5.33	6.33
ABBA	3	10	11	5	3.33	3.67	1.67
Al Green	1	-2	0	-4	-2.00	0.00	-4.00
Alice Cooper	3	8	5	1	2.67	1.67	0.33
Alice in Chains	2	10	12	7	5.00	6.00	3.50
Alpha Blondie	1	-10	-8	-8	-10.00	-8.00	-8.00
Anthrax	2	6	9	5	3.00	4.50	2.50
Antonin Dvorak	2	6	9	9	3.00	4.50	4.50
Aphex Twin	2	13	13	9	6.50	6.50	4.50
Aretha Franklin	3	9	8	9	3.00	2.67	3.00
Bad Religion	3	4	17	8	1.33	5.67	2.67
Basement Jaxx	1	7	8	7	7.00	8.00	7.00
BB King	3	-1	0	-1	-0.33	0.00	-0.33
Beck	3	-4	-6	0	-1.33	-2.00	0.00
Belle and Sebastian	2	-1	-3	-2	-0.50	-1.50	-1.00
Big Bill Broonzy	1	4	4	3	4.00	4.00	3.00
Billie Holiday	2	9	8	7	4.50	4.00	3.50
Black Sabbath	3	10	10	11	3.33	3.33	3.67
Bob Dylan	3	4	8	10	1.33	2.67	3.33
Bob Marley	3	-5	-3	1	-1.67	-1.00	0.33
Britney Spears	3	10	18	15	3.33	6.00	5.00
Carl Cox	1	8	7	8	8.00	7.00	8.00
Chemical Brothers	3	5	8	6	1.67	2.67	2.00
Chuck Berry	1	1	1	3	1.00	1.00	3.00
Cypress Hill	2	6	2	6	3.00	1.00	3.00
Daft Punk	2	6	9	3	3.00	4.50	1.50
Dave Brubeck	2	5	4	1	2.50	2.00	0.50
Dead Kennedys	1	5	6	4	5.00	6.00	4.00
Deep Purple	3	6	7	3	2.00	2.33	1.00
Dixie Chicks	1	6	5	6	6.00	5.00	6.00
Django Reinhardt	2	9	9	8	4.50	4.50	4.00
Dolly Parton	1	4	4	1	4.00	4.00	1.00
Dr. Dre	2	11	12	3	5.50	6.00	1.50
Duke Ellington	3	11	10	5	3.67	3.33	1.67
Elvis Presley	4	-3	-4	-5	-0.75	-1.00	-1.25
Eminem	4	22	15	15	5.50	3.75	3.75
Faith Hill	1	4	4	2	4.00	4.00	2.00
Fatboy Slim	2	5	6	1	2.50	3.00	0.50
Frederic Chopin	3	4	-1	0	1.33	-0.33	0.00
Garth Brooks	1	3	3	2	3.00	3.00	2.00
Glenn Miller	1	0	0	0	0.00	0.00	0.00
Grandmaster Flash	1	1	3	3	1.00	3.00	3.00
Hank Williams	1	4	3	2	4.00	3.00	2.00
Howlin' Wolf	1	1	1	-2	1.00	1.00	-2.00
Iron Maiden	3	10	11	11	3.33	3.67	3.67
James Brown	2	-1	1	-1	-0.50	0.50	-0.50

Table 2. Results of the user study. Only the 92 artists which were assessed at least once are depicted. The column labeled *assessments* shows the number of assessments made, i.e. the number of test persons which evaluated the respective artist. The next three columns reveal, for each of the weighting functions, the summed up ratings (in points) over all terms. The last three columns show the averaged ratings.

artist	assessments	TF	DF	TF×IDF	TF (avg)	DF (avg)	TF×IDF (avg)
Janet Jackson	2	3	5	1	1.50	2.50	0.50
Jimmy Cliff	1	-1	-2	1	-1.00	-2.00	1.00
Joan Baez	1	7	7	5	7.00	7.00	5.00
Johann Sebastian Bach	1	4	4	4	4.00	4.00	4.00
Johannes Brahms	2	11	11	11	5.50	5.50	5.50
John Lee Hooker	1	0	0	2	0.00	0.00	2.00
John Mayall	1	-1	-1	-3	-1.00	-1.00	-3.00
Johnny Cash	2	11	11	7	5.50	5.50	3.50
Justin Timberlake	3	-2	0	-2	-0.67	0.00	-0.67
Kraftwerk	1	6	4	2	6.00	4.00	2.00
Little Richard	2	-3	-1	-3	-1.50	-0.50	-1.50
Louis Armstrong	2	-3	-4	-3	-1.50	-2.00	-1.50
Ludwig van Beethoven	1	5	6	1	5.00	6.00	1.00
Madonna	3	13	6	7	4.33	2.00	2.33
Marvin Gaye	1	3	4	0	3.00	4.00	0.00
Megadeth	1	0	3	-2	0.00	3.00	-2.00
Michael Jackson	2	-9	-9	-10	-4.50	-4.50	-5.00
Miles Davis	1	-2	-3	0	-2.00	-3.00	0.00
Missy Elliot	2	9	11	11	4.50	5.50	5.50
Moloko	2	11	9	7	5.50	4.50	3.50
Muddy Waters	1	0	-2	-2	0.00	-2.00	-2.00
N'Sync	4	5	6	4	1.25	1.50	1.00
Nirvana	1	1	0	3	1.00	0.00	3.00
NoFX	2	15	15	-6	7.50	7.50	-3.00
Patti Smith	1	1	4	4	1.00	4.00	4.00
Prince	2	-1	-1	1	-0.50	-0.50	0.50
Public Enemy	2	10	12	7	5.00	6.00	3.50
Radiohead	1	6	6	6	6.00	6.00	6.00
Ramones	1	3	6	-1	3.00	6.00	-1.00
Run DMC	3	9	1	1	3.00	0.33	0.33
Sepultura	2	11	5	4	5.50	2.50	2.00
Sex Pistols	2	6	8	4	3.00	4.00	2.00
Shaggy	2	3	-2	3	1.50	-1.00	1.50
Sid Vicious	1	-1	1	1	-1.00	1.00	1.00
Slayer	1	-2	0	-3	-2.00	0.00	-3.00
Smashing Pumpkins	2	-2	-2	-2	-1.00	-1.00	-1.00
Solomon Burke	1	2	2	3	2.00	2.00	3.00
Sonic Youth	1	4	7	5	4.00	7.00	5.00
Suzanne Vega	2	4	6	2	2.00	3.00	1.00
The Animals	1	-4	-4	-4	-4.00	-4.00	-4.00
The Clash	1	2	0	-2	2.00	0.00	-2.00
The Kinks	1	1	0	1	1.00	0.00	1.00
The Rolling Stones	4	-1	5	-3	-0.25	1.25	-0.75
Tracy Chapman	1	2	4	1	2.00	4.00	1.00
Wolfgang Amadeus Mozart	2	12	12	8	6.00	6.00	4.00
Ziggy Marley	1	1	1	4	1.00	1.00	4.00
Sum	*172*	*386*	*413*	*271*	*204.08*	*224.00*	*141.08*

Table 3. Continuation of Table 2.

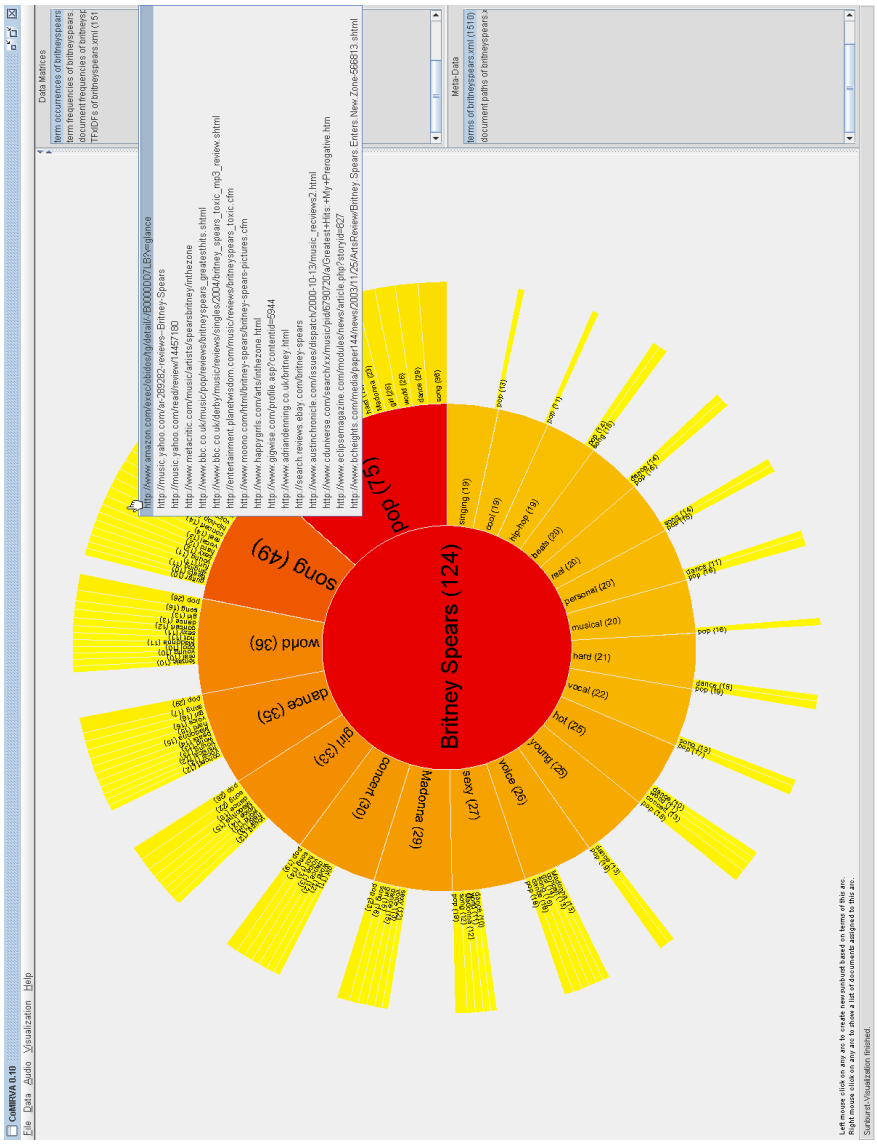

Fig. 1. A screenshot of *COB*'s user interface embedded in the *CoMIRVA* framework. The visualization is based on the web pages found for the music artist *Britney Spears* and on the dictionary T_l. In this example, the user has chosen to display a list of web pages mentioning the artist *Britney Spears* together with the terms *song* and *vocal*.

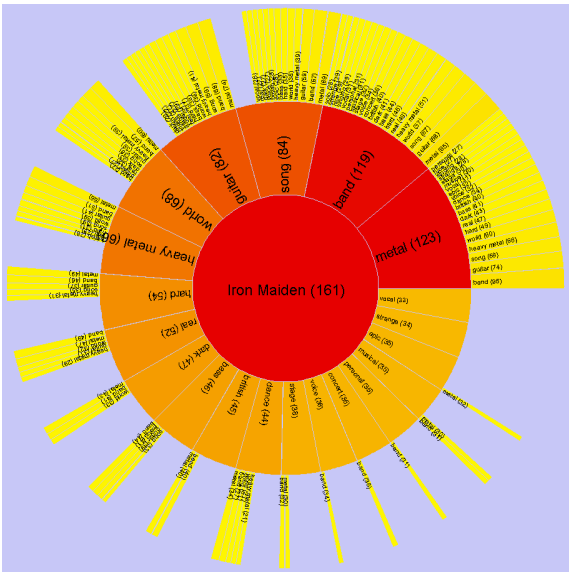

Fig. 2. A screenshot of a Sunburst generated from web pages about the music artist *Iron Maiden*. For this visualization, the dictionary T_l was used and the default values for complexity limitation were applied.

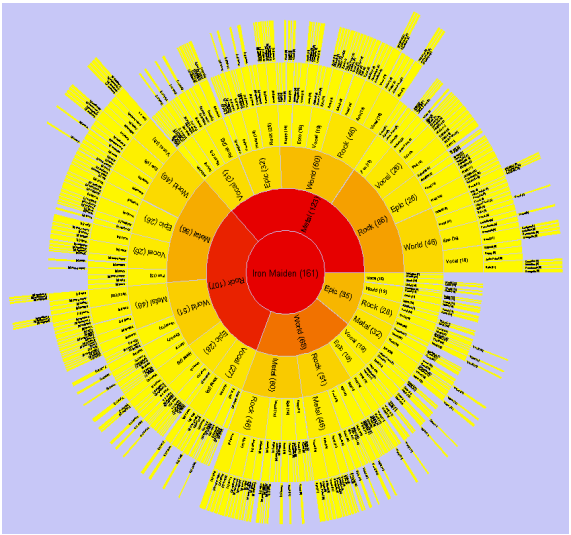

Fig. 3. A screenshot of a Sunburst generated from web pages about the music artist *Iron Maiden*. For this visualization, the dictionary T_s was used, the maximum number of sub nodes per node was set to 4, and the minimum angular extent of an arc was set to 0.5 degree.