A MIREX META-ANALYSIS OF HUBNESS IN AUDIO MUSIC SIMILARITY

Arthur Flexer, Dominik Schnitzer, Jan Schlüter
Austrian Research Institute for Artificial Intelligence (OFAI), Vienna, Austria
arthur.flexer|dominik.schnitzer|jan.schlueter@ofai.at

ABSTRACT

We use results from the 2011 MIREX "Audio Music Similarity and Retrieval" task for a meta analysis of the hub phenomenon. Hub songs appear similar to an undesirably high number of other songs due to a problem of measuring distances in high dimensional spaces. Comparing 17 algorithms we are able to confirm that different algorithms produce very different degrees of hubness. We also show that hub songs exhibit less perceptual similarity to the songs they are close to, according to an audio similarity function, than non-hub songs. Application of the recently introduced method of "mutual proximity" is able to decisively improve this situation.

1. INTRODUCTION

In a number of recent publications [21,27,28] the so-called "hubness" phenomenon has been described and explored as a general problem of machine learning in high dimensional data spaces. Hubs are data points which keep appearing unwontedly often in nearest neighbor lists of many other data points. This effect is particularly problematic in algorithms for similarity search, as the same "similar" objects are found over and over again. In Music Information Retrieval (MIR), the hub problem has been primarily studied in the context of music recommendation based on modeling of audio similarity. Songs which act as hubs are reported as being similar to very many other songs and hence keep a significant proportion of the audio collection from being recommended at all. This paper tries to answer the following questions concerning hubs in audio music similarity which so far have not been solved to a satisfactory degree: (i) Do different parameterizations and algorithms produce different hubs? (ii) Are hub songs perceptually meaningful?

This is done by conducting a meta-analysis of 17 algorithms and utilizing 8500 human gradings of the perceptual similarity of song pairs. A recently published method [29] ("mutual proximity") is applied to reduce the negative effects of hubness.

2. RELATED WORK

One of the central notions of Music Information Retrieval (MIR) is that of *music similarity*. Proper modeling of music similarity is at the heart of every application allowing automatic organization and processing of music data bases. A fundamental constituent to computation of music similarity is timbre similarity based on parameterization of audio using Mel Frequency Cepstrum Coefficients (MFCCs) plus Gaussian mixtures as statistical modeling [22]. It is precisely for this approach to music similarity where existence of hubs has been first documented and established in MIR by Aucouturier and Pachet in 2004 [3]. Hub songs were defined as songs which are, according to the audio similarity function, similar to very many other songs and therefore keep appearing unwontedly often in recommendation lists preventing other songs from being recommended at all. Such songs that do not appear in any recommendation list have been termed "orphans". The authors further stated that hub songs "objectively have nothing to do with the seed song" [3], i.e. they share no perceptual similarity with the songs they are recommended for according to the audio similarity function. Only anectodic evidence in the form of two examples is provided for this rather general statement. Since the data set for this study was quite small (350 songs from 37 artists), the authors remark that "a further study should be done with a larger database". Similar observations about false positives in music recommendation that are not perceptually meaningful have been made elsewhere [24] using an even smaller data set.

Following this initial report about the hub problem a number of results concerning hubness in the context of MIR have been established. Aucouturier and Pachet [4] showed that hubs are distributed along a scale-free distribution, i.e. non-hub songs are extremely common and large hubs are extremely rare. This is true for MFCCs modelled with different kinds of Gaussian mixtures as well as Hidden Markov Models, irrespective whether parametric Kullback-Leibler divergence or non-parametric histograms plus Euclidean distances are used for computation of similarity. But is also true that hubness is not the property of a song per se since non-parametric and parametric approaches produce very different hubs. The hub effect is not an artefact of using small data sets in computer experiments since it also exists in very large databases (> 250000 songs) and gets even worse with growing size of databases

[14]. Not all parameterizations of audio are equally prone to hubness. Fluctuation patterns (FP) [16, 23] have been shown to produce almost no hubs and a combination of MFCCs and FPs is able to reduce hubness while maintaining an overall high quality of audio similarity [12, 14]. It has also been noted that audio recorded from urban soundscapes, different from polyphonic music, does not produce hubs [2] since its spectral content seems to be more homogeneous and therefore probably easier to model. The same has been observed for monophonic sounds from individual instruments [17]. Direct interference with the Gaussian models during or after learning has also been explored (e.g. homogenization of model variances) although with mixed results. Whereas some authors report an increase in hubness [4], others observed the opposite [18]. Using a Hierarchical Dirichlet Process instead of Gaussians for modeling MFCCs seems to avoid the hub problem altogether [20]. The existence of the hub problem has also been reported for music recommendation based on collaborative filtering instead of on audio content analysis [8]. Similar effects exist for image [10, 19] and text retrieval [28] making this phenomenon a general problem in multimedia retrieval and recommendation.

Berenzweig [7] was probably the first to suspect a connection between the hub problem and the high dimensionality of the feature space. The hub problem was seen as a direct result of the curse of dimensionality [5], a term which refers to a number of challenges due to the high dimensionality of data spaces. Radovanović et al [27, 28] were able to provide more insight by linking the hub problem to the property of *concentration* [15] which occurs as a natural consequence of high dimensionality. Concentration is the surprising characteristic of all points in a high dimensional space to be at almost the same distance to all other points in that space. It is usually measured as a ratio between spread and magnitude, e.g. the ratio between the standard deviation of all distances to an arbitrary reference point and the mean of these distances. If the standard deviation stays more or less constant with growing dimensionality while the mean keeps growing the ratio converges to zero with dimensionality going to infinity. In such a case it is said that the distances concentrate. This has been studied for Euclidean spaces and other l^p norms [1, 15]. Radovanović et al [28] presented the argument that in the finite case, some points are expected to be closer to the center than other points and are at the same time closer, on average, to all other points. Such points closer to the center have a high probability of being hubs, i.e. of appearing in nearest neighbor lists of many other points. Points which are further away from the center have a high probability of being "orphans", i.e. points that never appear in any nearest neighbor list. This concentration of distances has also been reported for audio data [21].

Already in the context of concentration of distances it has been noted that the degree of concentration depends on the intrinsic rather than embedding dimension of the feature space [15]. Whereas the embedding dimension is the actual number of dimensions of a feature space the intrinsic dimension is the, often much smaller, number of dimensions necessary to represent a feature space without loss of information. It has also been demonstrated that hubness depends on the intrinsic rather than embedding dimensionality [28].

A direct consequence of the presence of hubs is that a large number of nearest neighbor relations in the distance space are asymmetric, i.e., a hub y is the nearest neighbor of x, but the nearest neighbor of the hub y is another point a ($a \neq x$). This is because hubs are nearest neighbors to very many data points but only k data points can be nearest neighbors to a hub since the size of a nearest neighbor list is fixed. This behavior is especially problematic if x and y belong to the same class but a does not, violating the pairwise stability of clusters [6]. In a recent publication [29] a general unsupervised method to attenuate the negative effects of hubness by repairing asymmetric nearest neighbor relations has been presented. It transforms arbitrary distance matrices to matrices of so-called probabilistic mutual proximity (MP). On a range of audio data sets it has been demonstrated that it is indeed able to decrease hubness while improving audio similarity as measured with genre classification accuracy. Since we will use this method to improve results for the MIREX data set it will be described in more detail in section 5.3. Please note that MP can be seen as a refinement of the so-called "P-norm" [26], which has been applied to the hub problem by other authors too [9].

3. DATA AND ALGORITHMS

For our meta-analysis of hubness we use the data from the recent 2011 "Audio Music Similarity and Retrieval" task [1] within the annual MIREX [11] evaluation campaign for MIR algorithms. Each of 18 competing algorithms was given 7000 songs (30 second audio clips). The data consists of 10 almost equally sized genre classes: 700 songs from BAROQUE, COUNTRY, EDANCE, JAZZ, METAL, RAPHIPHOP, ROCKROLL, ROMANTIC, 699 from BLUES, 701 from CLASSICAL. Every algorithm was given these 7000 song excerpts and returned either a full 7000×7000 distance matrix (algorithms CTCP1, CTCP2, CTCP3, DM2, DM3, ML1, ML2, ML3, SSKS3, SSPK2, STBD1, STBD2, STBD3) or a matrix of size 7000×100 (GKC1, HKHLL, PS1, YL1) containing the first 100 nearest neighbors to each song. The resulting distance matrix for algorithm ZYC2 is faulty containing the same distance of the same pair of songs over and over again and is therefore excluded from our analysis [2] . Please note that some of the systems are very closely related, sometimes using just different parameters for the same algorithm (e.g. CTCP1-3, DM2-3, ML1-3, STBD1-3).

From the 7000 songs, "100 songs were randomly selected from the 10 genre groups (10 per genre) as queries

[1] The 2011 results and details can be found at: `http://www.music-ir.org/mirex/wiki/2011:Audio_Music_Similarity_and_Retrieval_Results`

[2] Please note that ZYC2 did participate in the MIREX task and even scored in the mid-field of results. This seems to be due to nonrandom genre order during evaluation and not to its real performance.

and the first 5 most highly ranked songs out of the 7000 were extracted for each query (after filtering out the query itself, returned results from the same artist were also omitted). Then, for each query, the returned results (candidates) from all participants were grouped and were evaluated by human graders" [1]. For each individual query/candidate pair, a single human grader provided both a FINE score (from 0 (failure) to 100 (perfection)) and a BROAD score (not similar NS, somewhat similar SS, very similar VS) indicating how similar the songs are in their opinion. Since we use FINE scores only, this altogether gives $17 \times 100 \times 5 = 8500$ human gradings for our analysis.

4. EVALUATION

The following measures are used to evaluate the hubness phenomenon in section 5. Abbreviations correspond to labels in the result table 1.

k-occurrence statistics (H25/cov, H50/cov, maxH): As a measure of the hubness of a given song we use the so-called k-occurrence N_k [4], i.e. the number of times the song occurs in the first k nearest neighbors of all the other songs in the data base. Please note that the mean k-occurrence across all songs in a data base is equal to k. Any k-occurrence significantly bigger than k therefore indicates existence of a hub. Since human graders evaluated the five most similar songs we used $k = 5$. We compute the absolute number of the maximum k-occurrence $maxH$ (i.e. the biggest hub) and the number of songs for which the k-occurrence is bigger than 25 or 50 (i.e. the number of small hubs $H25$ and large hubs $H50$). Additionally we give the number of these hubs that appear as candidate songs in the human grading evaluation, e.g. "$H25/cov = 6/3$" means that out of six hub songs with k-occurrence bigger than 25 three songs have been evaluated (covered) by human graders.

Hubness (hub): We compute the hubness for each algorithm's distance matrix according to Radovanović et al. [28]. Hubness is defined as the skewness of the distribution of k-occurrences N_k of a whole data set. Positive skewness indicates high hubness, negative values low hubness.

Reachability (reach): This is the percentage of songs from the whole data base that are part of at least one of the nearest neighbor lists. If a song is not part of any of the recommendation lists of size $k = 5$ it is an orphan song which will never be recommended as a candidate song.

Number of hub gradings (#H): For every algorithm, 500 human gradings exist for further analysis. The number of hub gradings is the number of gradings where the candidate song in a query/candidate pair is a hub song. This is given using k-occurrences of 25 (H25) and 50 (H50) to distinguish between hubs and non-hubs.

Fine Score (fineH, fineNH, fine): To evaluate the perceptual quality of hubs and non-hubs we compute average fine scores. For every song in the whole data base we check whether it was the candidate in any of the query/candidate pairs in the human evaluation experiment. We average all respective fine scores for hub songs (fineH) and non-hub songs (fineNH) separately. This is done using k-occurrences of 25 (H25) and 50 (H50) to distinguish between hubs and non-hubs. We also include the average across all fine scores (fine) irrespective of whether candidate songs are hubs or not.

Accuracy (acc): To evaluate the quality of audio similarity for the whole data base, and not just for the songs which have been evaluated by human graders, we computed the genre classification performance. Since songs within a certain genre will also sound more similar than songs from different genres, high genre classification results indicate good audio similarity measures. It has also been demonstrated that algorithms achieving high genre classification results are able to produce results that correlate higher with human music similarity judgements [25, p. 26]. We compute the $k = 5$-nearest neighbor classification accuracy, i.e. the percentage of the five nearest songs that have the same label as the query song. Songs from the same artist as a query song are omitted from the nearest neighbor list by using an artist filter [13].

5. RESULTS

All results discussed in the following section can be found in table 1 listing all evaluation measures (from column hub to $fineNH$, see section 4) for all algorithms (from CTCP1 to YL1, see section 3) plus improved results using "mutual proximity" (rows mp, see section 5.3).

5.1 Hubness across algorithms

As already discussed in section 2 hubness is not a property of an individual song but is connected to what features are computed from the audio, what models are being learned from the features and how these processing steps are affected by the concentration of distances in high dimensional spaces. The MIREX audio similarity results provide the opportunity to compare 17 different algorithms with respect to their hubness. Looking at the results in table 1 it is apparent that the different algorithms produce very different degrees of hubness. The hubness values (column hub) range from 0.96 (HKHLL1) to 3.98 (STBD3) indicating that all distributions of k-occurrences are skewed to the right, i.e. are prone to hubness. Looking at the numbers of small ($H25$) and large ($H50$) hubs it is also clear that the different algorithms produce very different numbers of hubs. The number of small hubs range from 0 (SSKS3) to 256 (STBD3), those of large hubs from 0 (CTCP1-3, GKC1, HKHLL1, ML1, ML3, SSKS3, SSPK2) to 60 (STBD3). The largest k-occurrence $maxH$ ranges from 21 (HKHLL1) to 122 (STBD3). The average correlation of k-occurrences of all songs across all pairs of algorithms is only 0.14 showing that different algorithms produce very different hubness for a song. It is interesting to note that closely related algorithms show much higher correlations (e.g. an average of 0.76 for CTCP1, CTCP2 and CTCP3). The reachability $reach$ ranges from small 65.5% (STBD3) up to 95.9% (SSKS3).

								H25			H50		
algo	hub	acc	H25/cov	H50/cov	maxH	reach	fine	#H	fineH	fineNH	#H	fineH	fineNH
CTCP1	1.28	59.42	6/3	0/0	31	93.8	57.3	3	60.0	57.3	0	-	57.3
mp	1.32	58.75	9	0	35	92.5							
CTCP2	1.08	59.66	1/1	0/0	26	94.9	58.6	2	41.0	58.7	0	-	58.6
mp	1.02	59.20	0	0	25	94.4							
CTCP3	1.41	60.07	12/3	0/0	35	92.3	56.2	3	44.3	56.3	0	-	56.2
mp	1.41	59.38	10	0	36	90.3							
PS1	2.43	59.52	32/15	2/2	81	88.0	57.7	19	55.9	57.8	2	76.5	57.6
mp	1.02	54.80	1	0	31	99.1							
SSKS3	0.93	60.12	0/0	0/0	25	95.9	58.1	0	-	58.1	0	-	58.1
mp	1.13	59.61	3	0	31	93.8							
SSPK2	1.19	59.67	5/3	0/0	33	94.5	58.6	3	73.0	58.6	0	-	58.6
mp	1.33	58.95	5	0	38	93.7							
DM2	2.80	50.62	68/22	4/3	90	84.2	50.5	29	48.2	50.6	5	36.2	50.6
mp	2.15	51.19	40	1	58	88.9							
DM3	2.83	50.69	76/32	7/3	85	84.4	50.3	43	47.2	50.6	4	42.0	50.4
mp	2.15	51.03	48	1	58	88.9							
GKC1	1.31	25.83	11/3	0/0	34	90.8	31.8	3	30.0	31.9	0	-	31.8
mp	0.16	26.52	0	0	15	97.8							
HKHLL1	0.96	38.40	0/0	0/0	23	93.3	42.2	0	-	42.2	0	-	42.2
mp	0.48	38.76	0	0	24	98.3							
ML1	2.19	45.95	61/22	0/0	47	85.9	47.8	26	46.5	47.9	0	-	47.8
mp	1.06	48.22	2	0	30	93.8							
ML2	2.17	44.22	60/19	1/0	54	87.4	47.3	25	36.1	47.9	0	-	47.3
mp	1.08	45.74	2	1	30	94.4							
ML3	1.49	45.17	12/3	0/0	38	90.8	47.8	3	47.7	47.8	0	-	47.8
mp	0.94	44.74	1	0	26	95.2							
STBD1	3.46	26.62	161/71	17/10	90	81.0	33.9	86	29.5	34.8	16	22.9	34.3
mp	1.40	28.81	5	0	45	93.7							
STBD2	2.88	25.91	135/60	9/5	70	82.0	30.6	72	28.0	31.0	6	21.3	30.7
mp	1.24	27.37	3	0	34	95.5							
STBD3	3.98	25.38	256/113	60/35	122	65.5	30.4	149	29.3	30.9	54	23.0	31.3
mp	2.18	26.88	64	2	57	86.1							
YL1	2.16	41.14	65/21	1/0	54	84.4	42.4	27	33.4	42.9	0	-	42.4
mp	1.82	41.01	2	1	65	95.6							

Table 1. All results (please see section 5 for details) for all evaluation measures (from column hub to $fineNH$, see section 4) for all algorithms (from CTCP1 to YL1, see section 3) plus improved results using "mutual proximity" (rows mp, see section 5.3). Algorithms CTCP1 to SSPK2 (top six rows) already use "P-norm" and therefore do not show improvements due to mp, see section 5.3.

To sum up, different algorithms indeed produce very different degrees of hubness.

5.2 Hubness and perceptual quality

The next question we like to clarify is whether hub songs exhibit less perceptual similarity to the songs they are close to (according to an audio similarity function) than non-hub songs.

The correlation between hubness and average fine score of all algorithms (columns hub and $fine$ in table 1) is -0.56. This indicates that algorithms generating large hubness show low fine scores, i.e. overall bad perceptual similarity. Notable exceptions are maybe GKC1 with low hubness and low fine scores ($hub = 1.31$, $fine = 31.8$) and PS1 with rather high hubness and high fine scores ($hub = 2.43$, $fine = 57.7$).

Analyzing the differences in fine scores between hubs and non-hubs, we can see that the average fine scores for small hubs ($fineH$,$H25$) are almost always smaller than those for non-hubs ($fineNH$,$H25$). The only exceptions are algorithms CTCP1 and SSPK2. The average difference in fine scores is 3.65. This average is taken across all algorithms where grading information for hubs is available ($\#H > 0$). The average fine scores for large hubs ($fineH$,$H50$) are again almost always smaller than those for non-hubs ($fineNH$,$H50$), with algorithm PS1 being the only exception. The average difference in fine scores is 5.49. Again this average is taken across all algorithms where grading information for hubs is available ($\#H > 0$). Please note that e.g. for PS1, only two human gradings of large hub songs do exist.

To sum up, both small and large hubs seem to be less perceptually meaningful than non-hub songs but the average difference in human gradings is only 3.65 to 5.49 on

a scale from 0 to 100. Audio similarity computed with algorithms showing high hubness seems to be less perceptually meaningful than that of algorithms with low hubness in general.

5.3 Reducing hubness

To reduce the negative effects of hubness we apply "mutual proximity" (MP) [29] to the distance matrices from all algorithms. MP takes a distance matrix and (i) transforms distances between points x and y into probabilities that y is closest neighbor to x given the distribution of all distances to x in the data base, (ii) combines these (asymmetric) probabilistic distances from x to y and y to x via the product rule. The first step of transformation to probabilities re-scales and normalizes the distances like a z-transform. The second step combines the probabilities to a mutual measure thereby repairing sometimes contradicting, asymmetric nearest neighbor information which seems to cause hubness in similarity measures. Please note that MP requires knowledge of the full distance matrix since it needs to compute means and variances across full rows and columns during the re-scaling step. For algorithms GKC1, HKHLL, PS1 and YL1 we only have distances to the first 100 nearest neighbors of each song. In this case we use this limited set of distances instead of full rows and columns for computation of MP.

Before discussing the improvements due to MP it has to be said that six of the competing algorithms (CTCP1-3, PS1, SSKS3, SSPK2) already use a transformation of the distance matrix similar to MP. Usage of the so-called "P-norm" [26], which can be seen as a predeccesor to MP, together with application of MP does not seem to improve hubness. As a matter of fact, it sometimes even worsens the hubness situation. On the other hand, the six algorithms employing the "P-norm" already are the six best ranking systems according to their average fine score. Therefore we now discuss only those algorithms that do not use the "P-norm" already (DM2-3, GKC1, HKHLL1, ML1-3, STBD1-3, YL1).

Comparing hubness indices (column hub in table 1) between original algorithms and their improved MP version (rows mp) it can be seen that all values improve. The average decrease in hubness (hub) is 45.5%. The number of small hubs $H25$ also always decrease with an average of 83% less hubs. The average decrease in the number of large hubs $H50$ is 79.6%. The average decrease of the largest hub $maxH$ is 33.2%. Only HKHLL1 and YL1 show a slightly larger $maxH$, with all other indices also diminishing. The reachability $reach$ for all algorithms is enhanced, on average by 8.95 percentage points. This means that audio simialrity re-scaled with MP produces less orphan songs and includes a larger part of the data base in the nearest neighbor lists. All these improvements seem to be accompanied with unchanged quality in audio similarity. At least all genre classification results (column acc) remain more or less constant, with some insignificant increases and decreases.

To sum up, mutual proximity (MP) is able to decisively improve the hubness situation while not changing the overall performance in audio similarity.

6. DISCUSSION AND CONCLUSION

In this paper we were able to explore two important questions concerning hubness in audio music similarity which so far have not been answered satisfactory. We have corroborated earlier results indicating that different features computed from the audio in combination with different models being learned produce very different degrees of hubness. This was done by comparing a yet unprecedented number of different approaches (17 algorithms from the 2011 MIREX "Audio Music Similarity and Retrieval" task). We were also able to show that hub songs, when being recommended as being very similar, are judged to be less perceptually meaningful than non-hub songs by human evaluators. This was done by conducting the first systematic and extensive study on the perceptual quality of hub songs based on human evaluations again using MIREX data. Last but not least we were able to show that it is possible to reduce the many negative effects of hubness by applying "mutual proximity" to re-scale audio similarity distances.

7. ACKNOWLEDGEMENTS

Many thanks are due to the spiffy MIREX people (especially Stephen Downie and Andreas Ehmann) for conducting the audio music similarity and retrieval evaluation in the first place and for making the data needed for our research available. This research is supported by the Austrian Science Fund (FWF, grant P24095) and the Vienna Science and Technology Fund (WWTF, project "Audiominer" MA09-024).

8. REFERENCES

[1] Aggarwal C.C., Hinneburg A., Keim D.A.: On the Surprising Behavior of Distance Metrics in High Dimensional Spaces, Proceedings of the 8th International Conference on Database Theory, Springer-Verlag London, UK, pages 420-434, 2001.

[2] Aucouturier J.-J., Defreville B., Pachet F.: The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music, Journal of the Acoustical Society of America, 122 (2), 881-891, 2007.

[3] Aucouturier, J.-J., Pachet F.: Improving Timbre Similarity: How high is the sky?, Journal of Negative Results in Speech and Audio Sciences, 1(1), 2004.

[4] Aucouturier J.-J., Pachet F.: A scale-free distribution of false positives for a large class of audio similarity measures, Pattern Recognition, Vol. 41(1), pp. 272-284, 2007.

[5] Bellman R.E.: Adaptive Control Processes: A Guided Tour, Princeton University Press, 1961.

[6] Bennett K.P., Fayyad U., Geiger D.: Density-based indexing for approximate nearest-neighbor queries, Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '99, pages 233-243, New York, NY, USA, 1999.

[7] Berenzweig A.: Anchors and hubs in audio-based music similarity, PhD thesis, Columbia University, New York, USA, 2007.

[8] Celma, O.: Music Recommendation and Discovery in the Long Tail, PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2008.

[9] Charbuillet C., Tardieu D., Peeters G.: GMM supervector for content based music similarity, Proceedings of the 14th Int. Conference on Digital Audio Effects (DAFx-11), Paris, France, 2011.

[10] Doddington G., Liggett W., Martin A., Przybocki M., Reynolds D.A.: SHEEP, GOATS, LAMBS and WOLVES: A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation, in Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP'98), Sydney, Australia, 1998.

[11] Downie J.S.: The Music Information Retrieval Evaluation eXchange (MIREX), D-Lib Magazine, Volume 12, Number 12, 2006.

[12] Flexer A., Gasser M., Schnitzer D.: Limitations of interactive music recommendation based on audio content, Proc. of the 5th Audio Mostly Conference, pp. 96-102, 2010.

[13] Flexer A., Schnitzer D.: Effects of Album and Artist Filters in Audio Similarity Computed for Very Large Music Databases, Computer Music Journal, Volume 34, Number 3, pp. 20-28, 2010.

[14] Flexer A., Schnitzer D., Gasser M., Pohle T.: Combining features reduces hubness in audio similarity, Proc. of the Eleventh International Society for Music Information Retrieval Conference (ISMIR 2010), 2010.

[15] Francois D., Wertz V., Verleysen M.: The concentration of fractional distances, IEEE Transactions on Knowledge and Data Engineering, Vol. 19, No. 7, pp. 873-886, 2007.

[16] Fruehwirt M., Rauber A.: Self-Organizing Maps for Content-Based Music Clustering, Proceedings of the Twelth Italian Workshop on Neural Nets, IIAS, 2001.

[17] Gasser M., Flexer A., Schnitzer D.: Hubs and Orphans - an Explorative Approach, Proceedings of the 7th Sound and Music Computing Conference (SMC'10), 2010.

[18] Godfrey M.T., Chordia P.: Hubs and Homogeneity: Improving Content-Based Music Modeling, Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR'08), 2008.

[19] Hicklin R.A., Watson C.I., Ulery B.: The Myth of the Goats: How Many People Have Fingerprints that are Hard to Match?, The National Institute of Standards and Technology (NIST), NIST Interagency/Internal Report (NISTIR) - 7271, 2005.

[20] Hoffman M., Blei D., Cook P.: Content-Based Musical Similarity Computation Using the Hierarchical Dirichlet Process, Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR'08), 2008.

[21] Karydis I., Radovanović M., Nanopoulos A., Ivanović M.: Looking through the "glass ceiling": A conceptual framework for the problems of spectral similarity, Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR'10), pages 267-272, 2010.

[22] Logan B.: Mel Frequency Cepstral Coefficients for Music Modeling, Proceedings of the International Symposium on Music Information Retrieval (ISMIR'00), 2000.

[23] Pampalk E.: Computational Models of Music Similarity and their Application to Music Information Retrieval, Vienna University of Technology, Austria, Doctoral Thesis, 2006.

[24] Pampalk E., Dixon S., Widmer G.: On the Evaluation of Perceptual Similarity Measures for Music, Proceedings of the 6th International Conference on Digital Audio Effects (DAFx-03), pp. 7-12, London, U.K., 2003.

[25] Pohle T.: Automatic Characterization of Music for Intuitive Retrieval, PhD Thesis, Johannes Kepler University, Linz, Austria, 2010.

[26] Pohle T., Schnitzer D., Schedl M., Knees P., Widmer G.: On rhythm and general music similarity, Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR'09), 2009.

[27] Radovanović M., Nanopoulos A., Ivanović M.: Nearest neighbors in high-dimensional data: The emergence and influence of hubs, Proceedings of the 26th International Conference on Machine Learning (ICML'09), ACM International Conference Proceeding Series, volume 382, pages 865-872, 2009.

[28] Radovanović M., Nanopoulos A., Ivanović M.: Hubs in space: Popular nearest neighbors in high-dimensional data, Journal of Machine Learning Research, 11:2487-2531, 2010.

[29] Schnitzer D., Flexer A., Schedl M., Widmer G.: Using Mutual Proximity to Improve Content-Based Audio Similarity, in Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR'11), Miami, Florida, USA, 2011.