JOHANNES KEPLER UNIVERSITÄT LINZ | JKU

TNF

Technisch-Naturwissenschaftliche Fakultät

Indexing Content-Based Music Similarity Models for Fast Retrieval in Massive Databases

RIGOROSUM

Dipl.-Ing. Dominik Schnitzer

Dr. Simon Dixon (Prüfer)
Centre for Digital Music
School of Electronic Engineering and Computer Science
Queen Mary University of London

Univ.-Prof. Gerhard Widmer (Betreuung)
Univ.-Prof. Hanspeter Mössenböck (Vorsitz)

Johannes Kepler Universität Linz

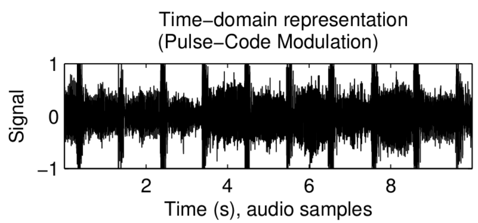
Overview

JOHANNES KEPLER UNIVERSITÄT LINZ | JKU

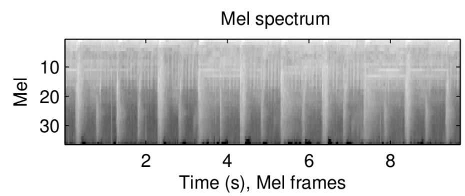
- **Aim**
  - scale existing content-based music similarity measures to the millions
  - high quality recommendation engine capable of searching millions of songs in a fraction of a second.
- **Introduction**
  - content-based music similarity
- **Problems**
- **Solutions, Contributions**
- **Prototype**
  - Wolperdinger: a fast, large-scale music similarity engine with 2.3 million songs

Similarity Measures

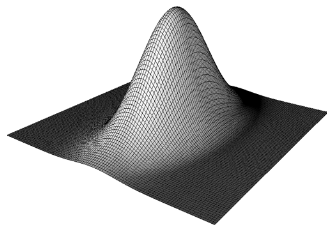
- feature extraction



audio signal



psychoacoustic modeling



similarity model

- similarity function
  - computes a similarity value between two similarity models
  - "two similar sounding music pieces should have a high similarity value"
- relevance in industry
  - increasing, with increasing number of music available
  - with it: the need for scalable solutions

Problems

Focus: **Mandel–Ellis (ME)**

Thesis: Elias-Pampalk (EP), Pohle-Schnitzer (PS)

- **similarity model**: single multivariate Gaussian, $X \sim \mathcal{N}(\hat{\mu}, \hat{\Sigma})$, computed from the Mel Frequency Cepstrum Coefficients (MFCCs)
- **similarity function**: symmetrized Kullback-Leibler (SKL) divergence

Problems

(1) missing indexing methods to scale search: non-vectorial, non-metric, high dimensional, expensive

(2) the "hub problem" in high dimensions & large datasets

Scaling Search

- preparations: retrieve a large test collection
  - 2.3 million MP3 snippets crawled from a web store
  - linear scan to find the nearest neighbors (**NN_{true}**) of a song: ~20s
- idea: design a filter-and-refine approach to find the **k** nearest neighbors (NN)
  - cheaply filter for possible nearest neighbors (**NN_{found}, $|NN_{found}| > k$**)
  - refine the ranking of possible nearest neighbors (original models)
  - find a fast configuration that identifies as many true NNs as possible, i.e., has high NN recall: $recall = \frac{|NN_{found} \cap NN_{true}|}{|NN_{true}|}$

Filter in a vector space created by FastMap

- multidimensional scaling, independent of features, only distances
- requires metric distances
- vector mapping

  (a) select a vector space dimension (**k**) to map the models to

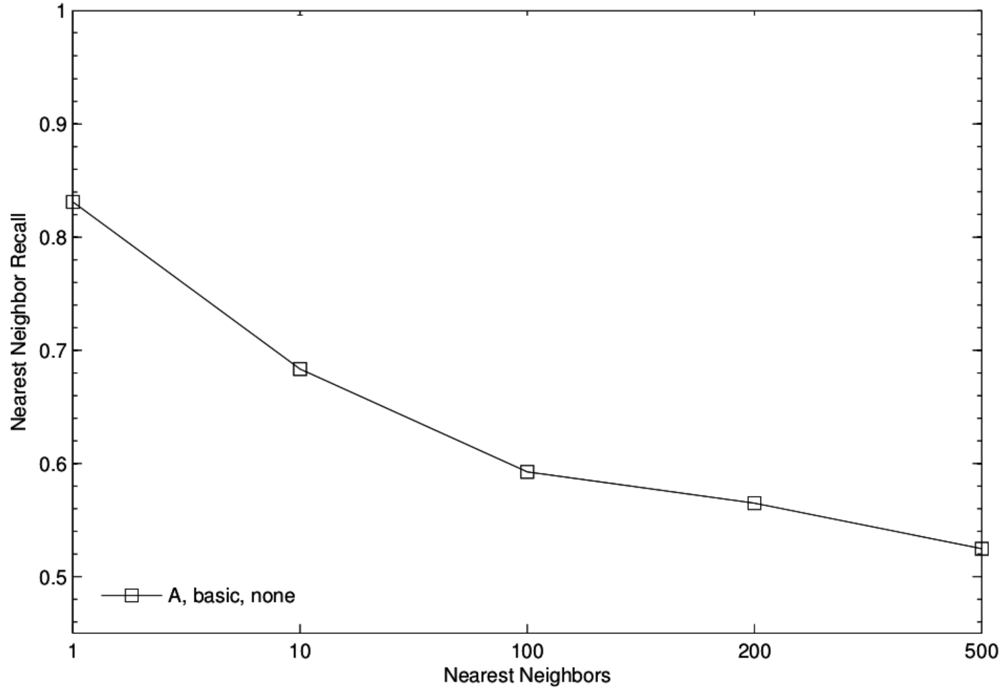
  (b) find two pivot models ($\mathbf{x}_1$, $\mathbf{x}_2$) for each dimension

  (c) to map models into the vector space, compute mapping for each dimension **k** according to FastMap

- everything is in place to use F&R method: filter in vector space, refine using SKL

First Attempt: Original FastMap in a filter&refine Search

- SKL, ME, N = 100 000, k = 40, filter size = 10 000 (10% N)



standard FastMap

Two modifications to the original FastMap algorithm

- **new pivot object selection heuristic: select pivots at median distance**
  - select random object R, compute distances from R to all other objects
  - select pivot $\mathbf{x_1}$ at median distance to R, all distances distance from $\mathbf{x_1}$
  - select pivot $\mathbf{x_2}$ at median distance to pivot $\mathbf{x_1}$

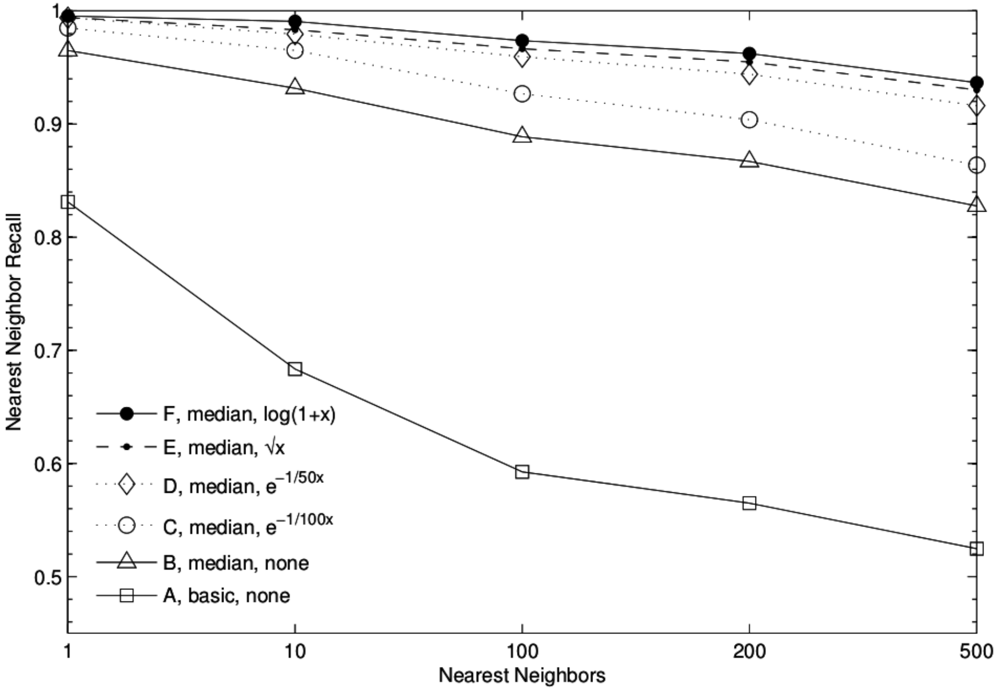
- **divergence rescaling**
  - symmetrized KL (SKL): symmetry and identity
  - triangle inequality: tested in a random collection of 100 000 models
  - make more metric: rescaling with exp(), log() sqrt()

Divergence	% triangle inequality
SKL	90.08%
$1 - e^{\lambda SKL}, \lambda = -\frac{1}{100}$	94.27%, used in [Pam06]
$1 - e^{\lambda SKL}, \lambda = -\frac{1}{50}$	97.54%, used in [Pam06]
$\sqrt{SKL}$	99.32%, used in [SFW09]
$\log(1 + SKL)$	99.99%

JOHANNES KEPLER UNIVERSITÄT LINZ | JKU

Effect of Modifications

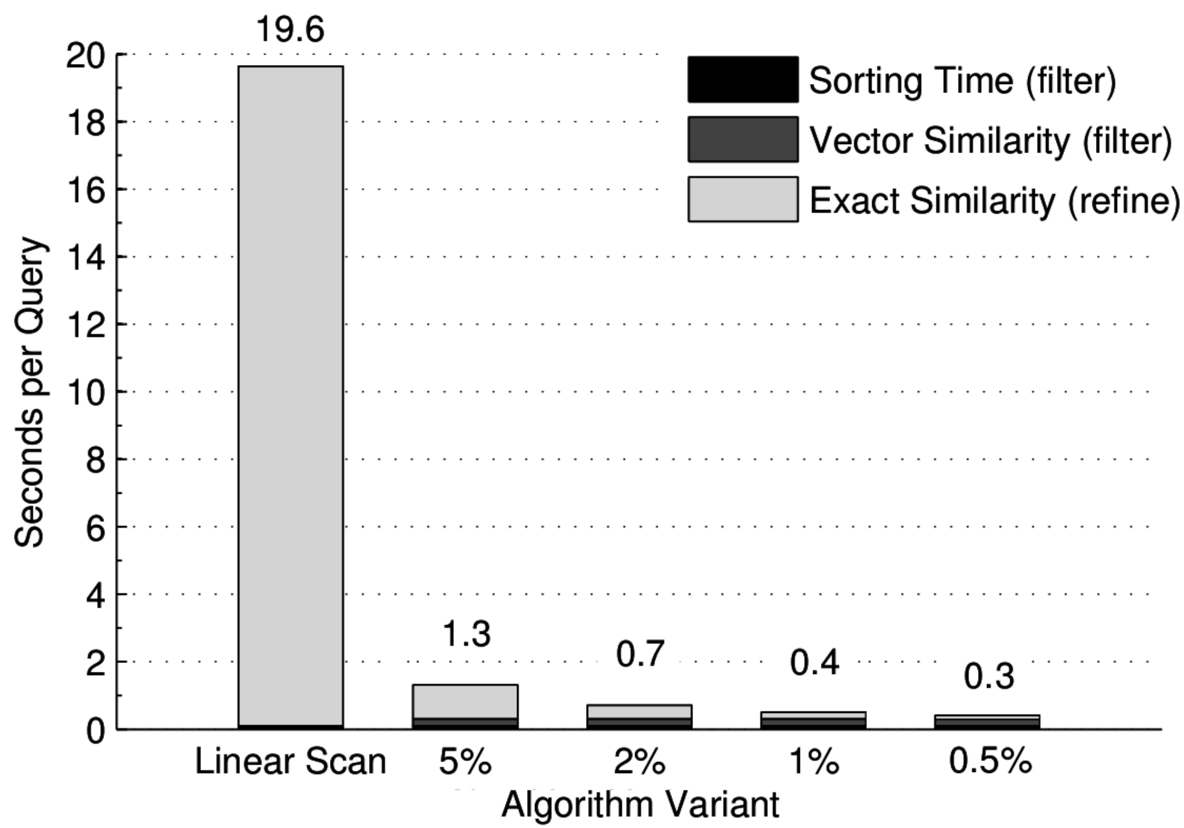
- SKL, ME, N = 100 000, k = 40, filter size: 10.000 (10% N)



standard FastMap

Scaling Search

2.3 million songs collection, PS similarity measure



effective query speed: 28x speedup

0.7 sec/query, 90%/85% 1/10 NN recall

Scaling Search

Contributions

- filter and refine search algorithm for music similarity with high recall, large speedups compared to a linear scan
- speedup with 2.3M : 28x with 90%/85% 1/10-NN recall (0.68s/query)

Publications

- Journal
  - A Fast Audio Similarity Retrieval Method for Millions of Music Tracks, Schnitzer D., Flexer A., Widmer G., Multimedia Tools and Applications, Springer, 2010.
- Conference Publication (oral)
  - A Filter-and-Refine Indexing Method for Fast Similarity Search in Millions of Music Tracks, Schnitzer D., Flexer A., Widmer G., Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR'09), Kobe, Japan, 2009.
- US/EU Patent (pending):
  - A method and a system for identifying similar audio tracks, Schnitzer D., EP Application 10167208.7, US Application 2011/0004642 A1

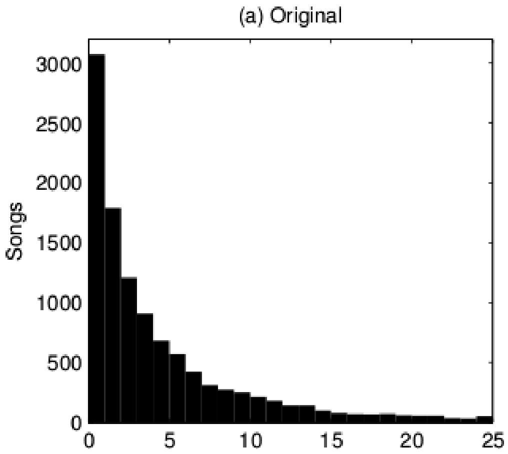
The Hub/Orphans Problem

- music similarity measures very prone to hubs/orphans

- hubs: data points which keep appearing unwontedly often as nearest neighbors of a large number of other data points
- orphans (anti-hubs): they never occur as nearest neighbor
  - occurs in high dimensional spaces, due to the property of distance concentration
  - newly found facet of the "curse of dimensionality"

Problem in a Music Recommendation System

FM4 Soundpark player (11 000pcs)
by M. Gasser & A. Flexer

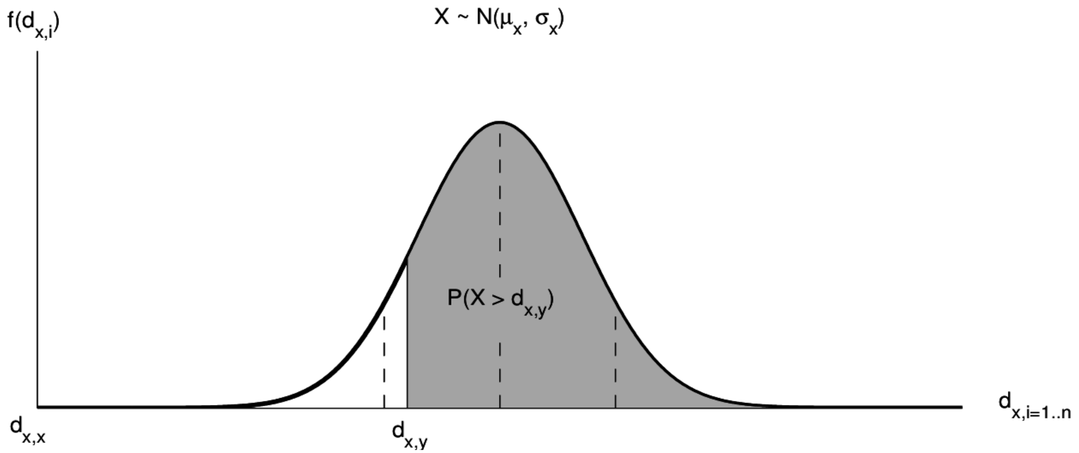


k=5 occurrence

- **hubs lead to asymmetric nearest neighbor relations**
  - "hub is nearest neighbor of N objects, but only one object can be the nearest neighbor of hub"

Alleviating Hubness: Mutual Proximity (MP)

- Mutual Proximity: symmetrize asymmetric nearest neighbors relations
  - for each object, compute the mean and standard deviation of its distances
  - for each object, transform all distances to nearest neighbor probabilities:

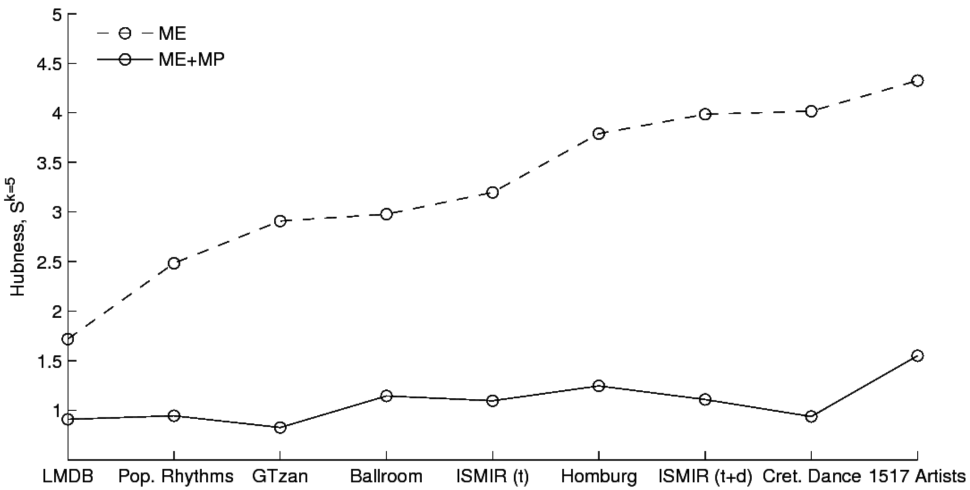


  - Combine the "opposite" probabilities into a new mutual measure, Mutual Proximity: X is a nearest neighbor of Y, and Y is a nearest neighbor of X

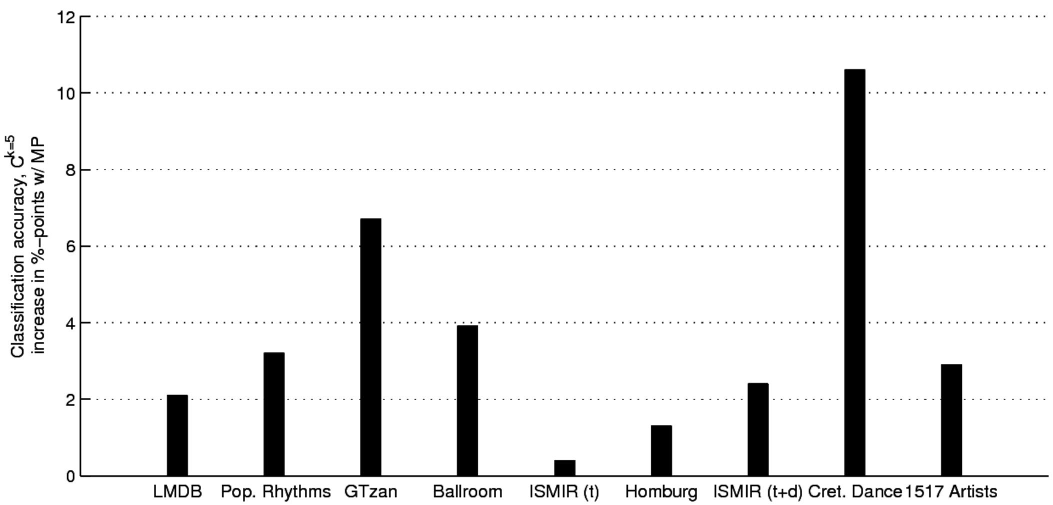
$$MP(d_{x,y}) = P(X > d_{x,y} \cap Y > d_{x,y}) = P(X > d_{x,y}) \cdot P(Y > d_{x,y})$$

JOHANNES KEPLER UNIVERSITÄT LINZ | JKU

Positive effects of Mutual Proximity



(1) decrease of hubs (and orphans)



(2) increase of k-nearest neighbor classification accuracies

- increase in class coherence, measured e.g. by the Goodman-Kruskal index
- effects also shown with 30 general machine learning databases (e.g., UCI)

Mutual Proximity with Large Collections

- as presented: MP does not scale (full similarity matrix)
  - algorithm that uses a k-means clustering to estimate the parameters of the Gaussian distance distribution needed for MP
  - k-means clustering for multivariate Gaussians
  - usable for very large music collections

Contributions

- alleviation to the general problem of hubs: Mutual Proximity method
- method to estimate Mutual Proximity method parameters for large scale collections
- open-source Octave/Matlab toolbox to cluster multivariate Gaussians

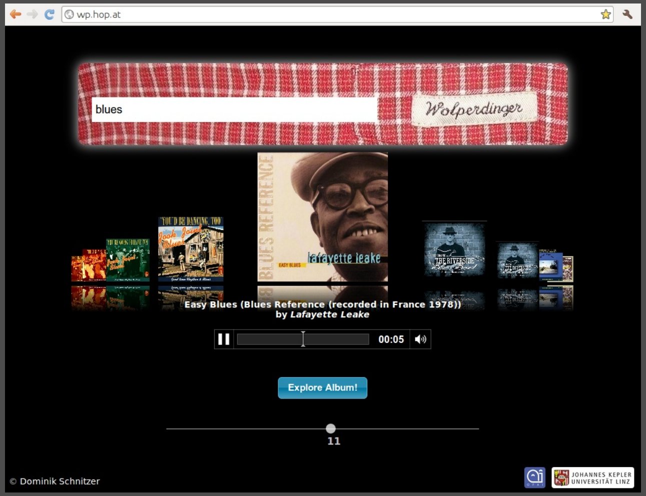
Publications

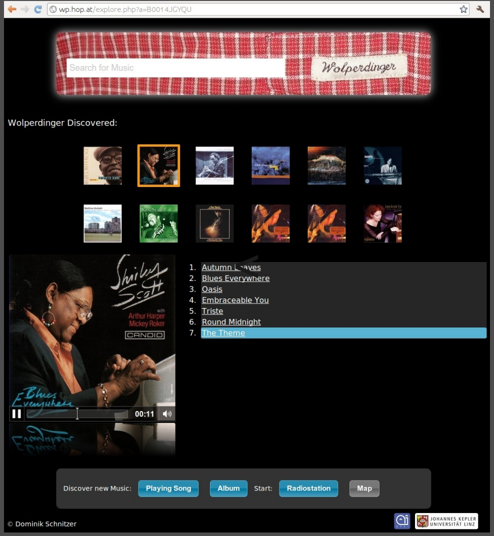
- Conference Publications
  - Using Mutual Proximity to Improve Content-Based Audio Similarity, Schnitzer D., Flexer A., Schedl M., Widmer G., Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR'11), Miami, FL, USA, 2011.
  - Islands of Gaussians: The Self Organizing Map and Gaussian Music Similarity Features, Schnitzer D., Flexer A., Widmer G., Gasser M., Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR'10), Utrecht, NL, 2010.
- FWF Project (Starts: Tomorrow, Feb. 1st 2012–14)
  - P24095: Die Vermeidung von Hubness in Music Information Retrieval

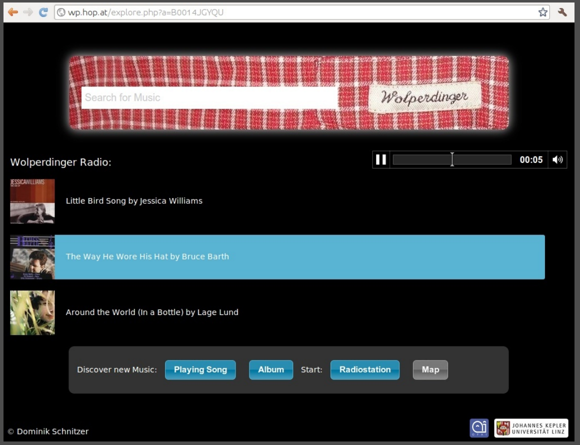
Wolperdinger: uses PS measure with MP and proposed F&R method

http://wp.ofai.at/

a music similarity engine working with 2.3 million songs







TNF

Technisch-Naturwissenschaftliche
Fakultät

Thank You!

Web:
http://www.schnitzer.at/dominik/phd-thesis/