

# Context-based Music Similarity Estimation



**Peter Knees and Markus Schedl**  
music@jku.at | <http://www.cp.jku.at>

# CONTEXT-BASED MUSIC SIMILARITY ESTIMATION

**Peter Knees**  
**Markus Schedl**

**Department of Computational Perception**  
**Johannes Kepler University (JKU)**  
**Linz, Austria**

# Motivation for Context-based Features

## Learning Semantics of Audio Signals

Which semantics?

The “semantics” are determined by the outside world, i.e., the **context**

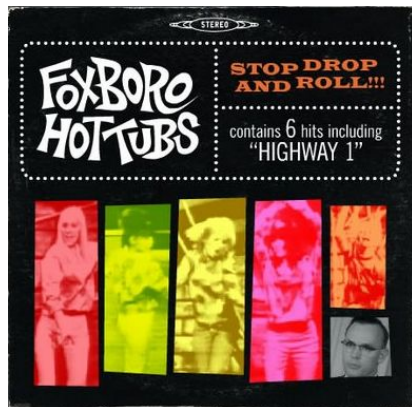
“Similarity” as perceived by humans is not exclusively determined by the audio

Influenced by, e.g., marketing strategies, political views of artists, language of songs, decade of activity, listening context, mood, peers, etc., etc., etc.

Analysis of audio alone probably won't do

# Content-based Similarity

Audio similarity may find that these two sound similar:



Foxboro Hot Tubs  
“Ruby Room”



The Stagers  
“Little Boy Blue”



But, for example, it won't tell you that...

- “Foxboro Hot Tubs” are better known as “Green Day”
- “The Stagers” are a band from Graz

# Context-based Similarity Trivia – Example 1

What do these songs have in common?



**NOFX**  
“Idiot Son of an Asshole”



**Eminem**  
“Mosh”



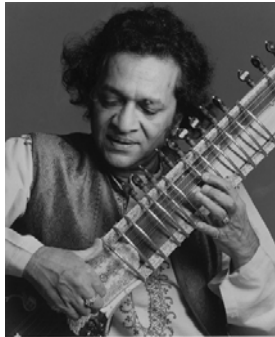
**Answer:**

**Both are Anti-Bush protest songs.**

## Context-based Similarity Trivia – Example 2

**What do these artists have in common?**

(Example borrowed from Lamere & Celma's Music Recommendation Tutorial)



**Ravi Shankar**



**Norah Jones**



**Answer:**

**Half of their DNA. Norah Jones is Ravi Shankar's daughter.**

## Context-based Similarity Trivia – Example 3

What do these songs have in common?



Antonio Carlos Jobim  
“Insensatez”

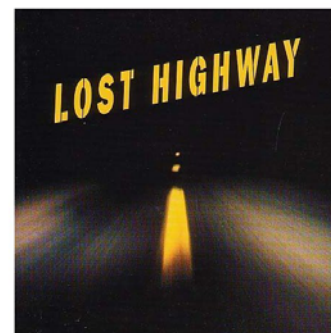


Rammstein  
“Rammstein”



**Answer:**

Both were featured on the Soundtrack of David Lynch’s movie “Lost Highway”



# Content vs. Context

## Advantages of Content Analysis

- Features can be extracted from any audio file
- No other data or community necessary
- No cultural biases (i.e., no popularity bias, no subjective ratings etc.)

## Advantages of Context Analysis

- Capture aspects beyond pure audio signal
- No audio file necessary
- Usually, user-based features are closer to what users want

## Challenge for both Content and Context Analysis

- Extraction of relevant features from (noisy) signal

# What's In This Talk

Manifold applications what can be done with contextual data. In this talk we focus on approaches that explicitly target the **estimation of similarity** of musical entities

## Not in this talk...

- Collaborative Filtering, User modeling, Foafing
- Server Log Mining (difficult to obtain)
- Social Network Mining (only preliminary results)

## In this talk...

### Text-based Similarity

- Web-Terms
- Tags
- Lyrics

### Co-Occurrence-based Similarity

- Playlists
- Page Counts
- P2P Networks

# Text-based Approaches: Basic Concepts

Using **traditional Text-IR concepts** to deal with textual data related to music

## Bag-of-Words approach

Text chunked into words (or n-grams): text = unsorted accumulation of terms

## Part-of-Speech (POS) Tagging

Determines the linguistic category for each word in a text; e.g., used to extract all adjectives

## Term weighting

Assigns a score to each term for each document.

Very frequently a variant of the **TF-IDF scheme**:

*tf*...term freq., # term occurrences in doc,

*df*...doc freq., # docs containing term (*idf*...inverse *df*)

$$w_{t,d} = tf_{t,d} \cdot \log \frac{N}{df_t}$$

## Vector Space Model

Each term represents a dimension, value = weight

Each doc is represented as a vector which dimensionality equals the number of distinct terms

## Text-based Approaches: Basic Concepts (2)

### Latent Semantic Analysis (LSA)

Transforming term-document space into (lower dimensional) concept-document space.  
Automatic detection of latent topics. Uses Single Value Decomposition.

### Similarity/Distance Calculation

- **Euclidean Distance**

$$d(a, b) = \sqrt{\sum (a_i - b_i)^2}$$

- **Cosine Similarity**

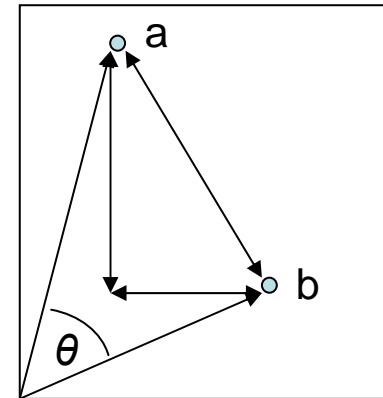
$$\text{sim}(a, b) = \cos \theta = \frac{\sum a_i b_i}{\sqrt{\sum a_i^2} \sqrt{\sum b_i^2}}$$

- **L<sub>1</sub> (Manhattan Distance)**

$$d(a, b) = \sum |a_i - b_i|$$

- **Overlap Score**

sum up terms  $t_i$  for which  $a_i \neq 0$  &&  $b_i \neq 0$



# Text-based Approaches: Web-Terms

## Basic Idea

- Analyze arbitrary texts about music (artists) from Web
- Create characteristic term profiles for each piece

## The “Let’s Use Google” Approaches

[Whitman & Lawrence, 2002], [Baumann & Hummel, 2003],  
[Knees et al., 2004]

- Take name of artist, add some constraints (e.g., *+music +review*), send it to Google
- Retrieves a list of relevant Web pages
- Optionally, apply filtering of noisy pages
- Throw retrieved texts together and calculate features (TF-IDF)
- Works best on artist level
- No restriction to specific sites
- Unstructured text data
- Number of terms very high (dimensionality!)



Web [Show options...](#) Results 1 - 10 of about 77,100 for "burt bacharach"

[Burt Bacharach - news, pictures, reviews, tickets, biography ...](#)  
Get the Latest on **Burt Bacharach** - news, pictures, album reviews, tickets, ... Great Jewish Music: **Burt Bacharach** (disc 1) - 20/05/1997 (Tzadik/US) ...  
[www.nme.com/artists/burt-bacharach](#) - [Cached](#) - [Similar](#)

[Living Together - Burt Bacharach - Music Reviews](#)  
**Burt Bacharach** MP3 Downloads - MP3.com offers legal **Burt Bacharach** music downloads as well as all of your favorite **Burt Bacharach** music videos.  
[www.mp3.com/albums/53519/reviews.html](#) - [Cached](#) - [Similar](#)

[Burt Bacharach: At This Time | Music Review | Slant Magazine](#)  
With the glut of pop records trying desperately to be anti-war these days, it was with considerable dubiousness that I cracked open **Burt Bacharach's** newest ...  
[www.slantmagazine.com/music/music\\_review.asp?ID=683](#) - [Cached](#) - [Similar](#)

[BBC - Music - Review of Burt Bacharach - Classics](#)  
BBC Music Review of Classics by **Burt Bacharach**. ... There's always something there to remind us of **Burt Bacharach**. His songs are endlessly reinterpreted and ...  
[www.bbc.co.uk/music/reviews/6qhj](#) - [Cached](#)

[The Burt Bacharach Album: Broadway Sings the Best o...: Album ...](#)  
Album Review: The **Burt Bacharach** Album: Broadway Sings the Best of **Burt** ... Who is the singer most closely associated with the music of **Burt Bacharach**? ...  
[www.answers.com/.../the-burt-bacharach-album-broadway-sings-the-best-of-burt-bacharach](#) - [Cached](#)

[Various Artists: Great Jewish Music: Burt Bacharach : Music ...](#)  
Home : artists : Sean Lennon : reviews : Great Jewish Music: **Burt Bacharach**. RSS  
Subscribe to Rolling Stone Sean Lennon feed ...  
[www.rollingstone.com/artists/.../great\\_jewish\\_music\\_burt\\_bacharach](#) - [Cached](#)

[Amazon.com: Painted from Memory: Elvis Costello & Burt Bacharach ...](#)  
Elvis teaming up with **Burt Bacharach** to make beautiful music together did not surprise ...  
Share your thoughts with other customers: Create your own review ...  
[www.amazon.com](#) > Music > Alternative Rock > General - [Cached](#) - [Similar](#)

[Great Jewish Music: Burt Bacharach | Music | EW.com](#)  
EW Home // Music // Great Jewish Music: **Burt Bacharach**. Music Review ... Details Lead Performances: **Burt Bacharach** and John Zorn; Genre: Experimental ...  
[www.ew.com/ew/article/0,,288020,00.html](#) - [Cached](#) - [Similar](#)

## Text-based Approaches: Web-Terms (2)

### The “Let’s Use Google and a Dictionary” Approaches

Only specific terms are considered in order to lower dimensionality and exclude noisy features

- manually compiled list of (musically relevant) terms [Pampalk et al., 2005]
- automatically generated using tags from Last.fm [Pohle et al., 2007]

Not necessarily better, especially important if features are presented to user

### Retrieving Texts from Specific Sources

- Mining texts from review pages such as *epinions.com* [Hu et al., 2005]
- Mining texts from mp3-Blogs (RSS feeds) [Celma et al., 2006]

Structure of data known, facilitates extraction of relevant text, limited to tracks/album/artists included in the service

# Text-based Approaches: Tags

## Basic Idea

- Use user/community generated tags for music as textual input

## Steps

- Retrieve tags for artist or track from Last.fm
  - Cleaning of noisy and redundant tags
    - manually
    - automatically [Geleijnse et al., 2007]
  - List of collected terms is treated as text document and TF-IDF'd [Levy & Sandler, 2007]
  - Optionally, LSA to reduce dimensionality
  - Comparison of vectors via cosine similarity (or overlap score)
- 
- Data available in standardized fashion
  - Dedicated terms for music
  - Lower dimensionality (13,500 tags vs. >200,000 Web terms [Levy & Sandler, 2007])
  - Depends on community



Burt Bacharach

### Tags

1960s 60s acoustic american **bacharach** baroque baroque pop  
boltonesque brill building pop **burt bacharach** chill classic **composer** disco driving  
easy **easy listening** everything favorite artists favorites  
film music film score fusion genius god great innovators guitar hal david inspirerande  
instrumental jazz **lounge** male male vocalists master melancholy music to  
warm the heart and hands my ancients my tag **oldies** outstanding **pop** relax  
rock score sexy singer-songwriter smooth **songwriter** sophistopop soul  
**soundtrack** space age pop swing symphonic pop us usa virtuoso vocal 2005

Tag

# Text-based Approaches: Lyrics

## Basic Idea

- Analyze the lyrics for a song (lyrics are usually easily available)

## Topic Features

[Logan et al., 2004]

- Typical topics for lyrics are distilled from a large corpus using (P)LSA (“Hate”, “Love”, “Blue”, Gangsta, Spanish)
- Lyrics are transformed to topic-based vectors, similarity is calculated via  $L_1$  distance
- Alternative approaches use TF-IDF with optional LSA and Stemming for Mood Categorization [Laurier et al., 2008], [Hu et al., 2009]

## Rhyme Features [Mayer et al., 2008]

- Phonetic transcription is searched for patterns of rhyming lines (AA, ABAB, AABB)
- Frequency of patterns + statistics like *words per minute*, *punctuation freq.* etc.

## Other Features [Mahedero et al., 2005]

- Language, Structure

## Text-based Approaches: Summary

	Web-Terms	Tags	Lyrics
<b>Source</b>	arbitrary Web pages	Web service	lyrics portal
<b>Community-based</b>	depends	yes	no
<b>Level</b>	artists	artists (tracks)	tracks (artists)
<b>Feature Dimensionality</b>	very high	moderate	possibly high
<b>Specific Bias</b>	low	community	none
<b>Potential Noise</b>	high	moderate	low

# Approaches based on Co-Occurrences: Playlists

[Pachet et al., Proc. WEDELMUSIC, 2001]

analysis of co-occurrences of artists and songs on

- radio station playlists (French radio station *Fip*)
- compilation CD databases (CDDB)

"co-occurrence" of an entity with itself is its number of occurrence

normalizing in order to account for different frequency/popularity of entities

similarity of 2 entities  $A_i$  and  $A_j$ :

$$sim_{pl\_cooc}(A_i, A_j) = \frac{1}{2} \cdot \left[ \frac{cooc(A_i, A_j)}{cooc(A_i, A_i)} + \frac{cooc(A_j, A_i)}{cooc(A_j, A_j)} \right]$$

shortcoming: cannot capture indirect links

# Approaches based on Co-Occurrences: Playlists

[Pachet et al., Proc. WEDELMUSIC, 2001]

correlation on vector representation to capture indirect links:

$$\mathit{sim}_{pl\_corr}(A_i, A_j) = \frac{\mathit{Cov}(A_i, A_j)}{\sqrt{\mathit{Cov}(A_i, A_i) \cdot \mathit{Cov}(A_j, A_j)}}$$

*insights:*

- co-occurrence performed better than correlation  
(too much irrelevant info in feature vectors?)
- rather small test collection

# ART OF THE MIX

[Cano

HOME FIND A MIX COMMUNITY MY ART OF THE MIX SUBMIT A MIX

## FEATURED MIXES



**THE LETTER FROM ENGLAND (RADIO ...**  
by Kostas

Playlist | Indie

My first broadcast for Radio Bubble internet radio. A mix of mostly lesser known indie pop, folk and punk. The link leads to a page where you can listen and legally download the mix (Radio Bubble pays ...

11/12/2009 1:31:00 PM



**I HAD A STRANGE DREAM LAST ...**  
by mallorys\_beehive

Cassette | Theme

90 minutes. April 2, 2007. It was spring break my junior year and I didn't go anywhere. I just stayed home and listened to some new records I had gotten and obsessed the whole time about this new guy ...

11/13/2009 11:41:00 AM



**WHAT'S IN A NAME?**  
by BizarroAnnie

Playlist | Theme

Playlist I made a while ago. Sorry it is in alphabetical order.

11/14/2009 10:50:00 AM

[VIEW ALL FEATURED MIXES >](#)

## RECENTLY POSTED MIXES



**ISLAND PARTY MIX**  
by Trish Mullins

Playlist | Reggae

Take a virtual Trip to de islands mon then when ya saved up all ya quarters take a real one Mon! No Problem mon!



**PLAYER'S ANTHEMS: THE BEST OF ...**  
by EzraPound

CD | Single Artist

I formulated this compilation in response to the godawful "Greatest Hits" CD currently available from The



**THE ANIMAL REVOLUTION**  
by Sam Rosehip

Cassette | Theme

This mix i made for a mix tape group. The theme for the month was "animals", so I decided to make up a

## ART OF THE MIX



Welcome to the website dedicated to making mixed tapes and cds. **SEARCH THE ARCHIVES** of over 100,000 mixes or check out **RECENT SUBMISSIONS**. Submit a **MIXED TAPE** or **PLAYLIST** yourself. Check out the **EXHIBITS**, **FORUMS** and **BLOG**. For more information about the site, review the **FREQUENTLY ASKED QUESTIONS**.

## RECENT FEEDBACK

*that was the title of a Wings album! Paul McCartney wont be happy if he finds out you took the title!*

feedback by **TRISH MULLINS**  
11/29/2009 5:43:00 PM

ences



Department of  
Computational  
Perception

**JKU**  
JOHANNES KEPLER  
UNIVERSITY LINZ

Peter Knees and Markus Schedl

3<sup>rd</sup> Workshop on Learning the Semantics of Audio Signals, Graz, Austria, December 2009

# Approaches based on Co-Occurrences: Playlists

[Baccigalupo et al., Proc. ISMIR, 2008]

> 1 mio. playlists from "MusicStrands"

subset of 4,000 most popular artists

distance between the occurrence of two artists in a playlist taken into account

→  $d_h(A_i, A_j)$ : co-occurrence count of song by  $A_i$  and song by  $A_j$  at distance of  $h$

$$dist_{pl\_d}(A_i, A_j) = \sum_{h=0}^2 \beta_h \cdot [d_h(A_i, A_j) + d_h(A_j, A_i)]$$

$$\beta_0 = 1, \beta_1 = 0.8, \beta_2 = 0.64$$

# Approaches based on Co-Occurrences: Playlists

[Baccigalupo et al., Proc. ISMIR, 2008]

normalization w.r.t. popularity:

$$dist_{|pl\_d|}(A_i, A_j) = \frac{dist_{pl\_d}(A_i, A_j) - \widehat{dist_{pl\_d}}(A_i)}{\left| \max \left( dist_{pl\_d}(A_i, A_j) - \widehat{dist_{pl\_d}}(A_i) \right) \right|}$$

no evaluation relevant to similarity measurement

# Approaches based on Web Co-Occurrences: General Remarks

- Web as very rich source of context information
- context defined similarly to "text-based approaches" as textual surrounding of a music entity on a Web page (artist name)
- BUT: how to find the musically relevant Web pages?
  - build own crawler and indexer (focused crawling)
    - + high quality, relevant pages
    - slow
  - rely on results of search engines
    - noisy
    - "black box"
    - restricted
    - + easy
    - + can be very fast

### [George W. Bush](#) - Wikipedia, the free encyclopedia

**Bush** is the eldest son of George H. W. **Bush** (the 41st President) and Barbara **Bush**, making him one of only two American presidents to be the son of a ...

[Childhood to mid-life](#) - [Marriage and family](#) - [Early career](#)

[en.wikipedia.org/wiki/George\\_W.\\_Bush](http://en.wikipedia.org/wiki/George_W._Bush) - [Cached](#) - [Similar](#)

### [Bush \(band\)](#) - Wikipedia, the free encyclopedia

**Bush** were a British alternative rock band formed in London in 1992 by singer/guitarist Gavin Rossdale and guitarist Nigel Pulsford. ...

[History](#) - [Discography](#) - [References](#) - [External links](#)

[en.wikipedia.org/wiki/Bush\\_\(band\)](http://en.wikipedia.org/wiki/Bush_(band)) - [Cached](#) - [Similar](#)

[+ Show more results from en.wikipedia.org](#)

### [Gavin Rossdale Fans: gavinrossdalefans.com | bush-music.com](#)

Official site. Offers news, biographies, lyrics, articles, and images.

[www.bush-music.com/](http://www.bush-music.com/) - [Cached](#)

### [Image results for bush](#) - [Report images](#)



### [News results for bush](#)



[CBC.ca](#)

[Bush 'raised Iraq issue after 9/11'](#) - 1 hour ago

George **Bush** raised the issue of Iraq with Tony Blair just three days after the 9/11 attacks, Mr Blair's former foreign policy adviser has said. ...

[The Press Association](#) - [836 related articles »](#)

[Blaming Bush for the Budget Deficits](#) - [Atlantic Online \(blog\)](#) - [62 related articles »](#)

[Perino: No Terrorist Attacks In America Under Bush](#) -

[Huffington Post \(blog\)](#) - [18 related articles »](#)

### News results for **britney spears**



[NEWS.com.au](#)

[Britney Spears sends lover Jason Trawick home after rejected ...](#) - 14 hours ago  
**BRITNEY Spears** ended her near month-long Australian tour by hitting a nightclub with her entourage for the first time since arriving - but boyfriend Jason ...

[Herald Sun](#) - [119 related articles »](#)

[Britney Spears Channeled 'Classy' Madonna In 'Radar' Video](#) -

[MTV.com](#) - [117 related articles »](#)

[Hollywood's bad girls clean up as public tires of 'the flippant ...](#) -  
[guardian.co.uk](#) - [2 related articles »](#)

### [Britney Spears](#) - official web site and blog

29 Nov 2009 ... Access **Britney Spears** photos, galleries and videos. Get the latest news direct from **Britney** on her official blog.

[Photos](#) - [Tour](#) - [Videos](#) - [Got Naked \(They Had A Plan\)](#)

[www.britneyspears.com/](#) - [Cached](#) - [Similar](#)

### [Concert Photos: Brisbane Night 3 - BRITNEY SPEARS](#)

29 Nov 2009 ... Here's your second-to-last gallery of Circus concert photos! Flip through to see the images from **Britney's** third night in Brisbane.

[www.britneyspears.com/.../concert-photos-brisbane-night-3.php](#) - 16 hours ago

[+ Show more results from www.britneyspears.com](#)

### [Britney Spears](#) - Wikipedia, the free encyclopedia

**Britney Jean Spears** (born December 2, 1981) is an American singer and entertainer. Born in Mississippi and raised in Louisiana, **Spears** first appeared on ...

[Discography](#) - [Videography](#) - [Products](#) - [Filmography](#)

[en.wikipedia.org/wiki/Britney\\_Spears](#) - [Cached](#) - [Similar](#)

### Image results for **britney spears** - [Report images](#)



# Web Co-Occurrences / Page Counts: Simple Approach

[Schedl et al., Proc. CBMI, 2005], similar to [Zadel and Fujinaga, Proc. ISMIR, 2004]

"Alice Cooper"  
"Alice Cooper" + "BB King"  
"Alice Cooper" + "Beethoven"  
...  
"Alice Cooper" + "ZZ Top"  
"BB King"  
...  
"ZZ Top"

+music+review



page counts

100	3	5	4
0	91	27	2
13	8	96	19
0	1	12	84

(co-occurrence) page counts

# Web Co-Occurrences / Page Counts: Simple Approach

[Zadel and Fujinaga, Proc. ISMIR, 2004]

Web services for artist recommendation

retrieve possibly related artists to seed artist  
from "Amazon Listmania!"

Google Web API to obtain a (sparse)  
co-occurrence matrix

similarity normalized for popularity via *min()*

$$sim_{pc\_min}(A_i, A_j) = \frac{pc(A_i, A_j)}{\min(pc(A_i), pc(A_j))}$$

only empirical evaluation

100	3	5	4
0	91	27	2
13	8	96	19
0	1	12	84

(co-occurrence) page counts

# Web Co-Occurrences / Page Counts: Simple Approach

[Schedl et al., Proc. CBMI, 2005]

aim: artist similarity estimation

model of conditional probability

normalization and symmetrization via  
cross-artist-probability

100	3	5	4
0	91	27	2
13	8	96	19
0	1	12	84

(co-occurrence) page counts

$$sim_{pc\_cp}(A_i, A_j) = \frac{1}{2} \cdot \left( \frac{pc(A_i, A_j)}{pc(A_i)} + \frac{pc(A_i, A_j)}{pc(A_j)} \right)$$

complete similarity matrix

also combination of different  
query schemes ("googling-approaches")

evaluation via genre classification task (224 artists, 14 genres): 85% acc.

## Web Co-Occurrences / Page Counts: Fetching Pages

[Schedl et al., Proc. CBMI, 2005]

*problem:* quadratic complexity in #artists

*solution:* retrieve content of top-ranked pages and index them with artist names

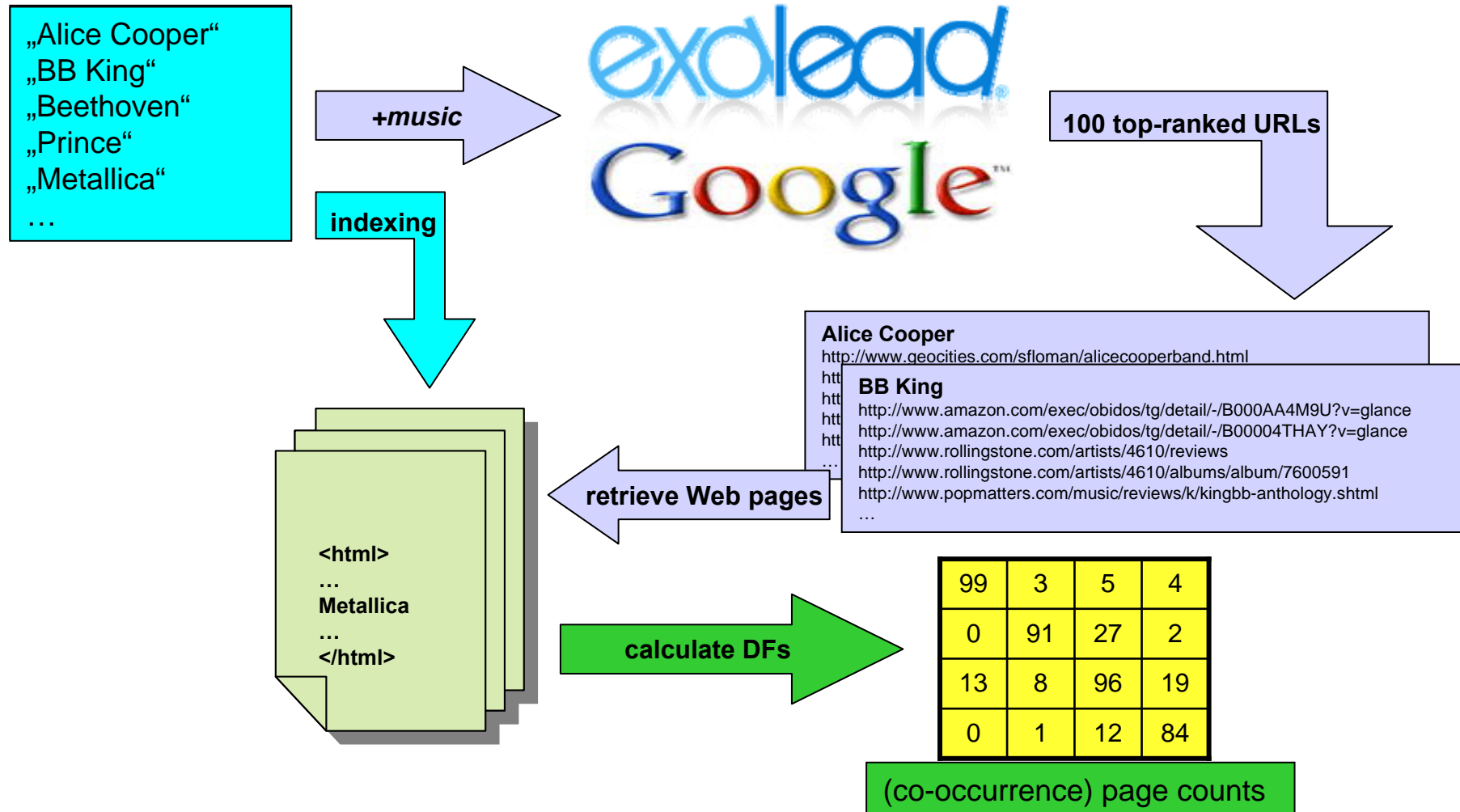
→ #queries linear with #artists

[Schedl, PhD thesis, 2008][Chapter 3]

100	3	5	4
0	91	27	2
13	8	96	19
0	1	12	84

(co-occurrence) page counts

# Web Co-Occurrences / Page Counts: Data Processing Pipeline



# Approaches based on Co-Occurrences: P2P Networks

make use of meta-data transmitted as files names or ID3 tags in peer-to-peer networks

information gathered from users' shared folders (no file downloads)

early work using "OpenNap":

[Whitman and Lawrence, Proc. ICMC, 2002]

[Ellis et al., Proc. ISMIR, 2002]

[Logan et al., Eval. MIR Systems Workshop @ SIGIR, 2003]

[Berenzweig et al., Proc. ISMIR, 2003]



# Approaches based on Co-Occurrences: P2P Networks

[Logan et al., 2003] and [Berenzweig et al., 2003]

400 most popular artists in mid 2002

175,000 user-to-artist relations from 3,200 shared collections

similarities via artist co-occurrences in collections (cond. prob.)

sparsity of the co-occurrence matrix

compared with "Art of the Mix" co-occurrences, AMG's similar artists, similarity data from a Web survey and content-based similarity (MFCCs)

evaluation via overlap between top N most similar artists for different sources

*insights:*

- "OpenNap" and "Art of the Mix" revealed a high overlap (0.6)

- Web survey and AMG showed a moderate overlap (0.4)

- MFCCs had a low agreement with all other sources (0.1)

→ content-based features capture very different similarity aspects



# Approaches based on Co-Occurrences: P2P Networks

[Whitman and Lawrence, Proc. ICMC, 2002]



alleviate popularity bias in the similarity measure:

$$\text{sim}_{p2p\_wl}(A_i, A_j) = \frac{C(A_i, A_j)}{C(A_j)} \cdot \left( 1 - \frac{|C(A_i) - C(A_j)|}{C(A_k)} \right)$$

$C(A_i)$  number of users that share artist  $A_i$

$C(A_i, A_j)$  number of users that have both  $A_i$  and  $A_j$  in their shared collection

$A_k$  most popular artist in the corpus

# Approaches based on Co-Occurrences: P2P Networks

[Ellis et al., Proc. ISMIR, 2002]

400,000 user-to-song relations from "OpenNap"

about 3,000 artists

similarity defined as Erdős distance in artist-similarity-graph:

$d(A_i, A_j)$  equals number of intermediate artists on shortest path from  $A_i$  to  $A_j$

accounts for indirect links (per definition)

alternative formulation "Resistive Erdős":

assumes that 2 artists are more similar if they are connected via many paths

$$dist_{p2p\_res}(A_i, A_j) = \frac{1}{\sum_{p \in Paths(A_i, A_j)} \frac{1}{|p|}}$$

no improvement with "Resistive Erdős"  
due to popularity bias



# Approaches based on Co-Occurrences: P2P Networks

[Shavitt and Weinsberg, Proc. IEEE ISM: AdMIRe Workshop, 2009]

meta-data of shared files in "Gnutella" network  
retrieved in November 2007 (.mp3 and .wav)

1.2 million users; 530,000 songs

distance measure on the *song* level

accounts for popularity bias

$$dist_{p2p-pop}(S_i, S_j) = -\log_2 \left( \frac{uc(S_i, S_j)}{\sqrt{C_i \cdot C_j}} \right)$$

$uc(S_i, S_j)$  number of users that share songs  $S_i$  and  $S_j$

$C_i, C_j$  popularity of  $S_i$  and  $S_j$ , measured as total number of occurrences

evaluation in a recommendation setting (30% of songs of each user's collection  
used to predict remaining 70%): about 12% prec., 13% rec.

heavy inconsistencies in meta-data (ID3 tags)



# Co-Occurrence-based Approaches: Summary

	<b>Playlists</b>	<b>Web Co-Ocs</b>	<b>P2P nets</b>
<b>Source</b>	radio, compilation CDs, Web services	search engines, Web pages	shared folders
<b>Community-based</b>	depends on source	no	yes
<b>Level</b>	artists (tracks)	artists	artists (tracks)
<b>Specific Bias</b>	low	"wikipedia"-bias	community
<b>Potential Noise</b>	low	high	high

## Summary and Discussion

- estimating similarity based on the context of a music entity
- "context" can be defined in various ways
- "cultural knowledge", "community meta-data", "context-based features"
- context of music entity as complementary source of information
- combination of context- and content-based features
- still no common evaluation data set like in TREC
  - each publication uses its own data set
  - need to establish common "ground truth"

# Challenges / Shortcomings of Context-based Methods

- *data sparsity* (especially in "long tail")
- *popularity bias*: disproportionately more info is available for popular artists than for lesser known one
- *community/population bias*:
  - only participants of the community under consideration are taken into account (e.g., certain P2P network, last.fm, myspace, ...);
  - users of certain communities may not represent the average music listener, but rather share similar music tastes

## Future Challenges: What we believe

- establish a differentiated, multi-granular concept of music similarity that takes into account
  - cultural areas,
  - regional particularities,
  - individual views and tastes (→ personalization)

- transcend the current focus of research on Western music

- integrate multi-faceted music similarity measures in music applications that should combine both content- and context-based information

→ enable retrieval systems to deal with queries like  
*"give me rhythmically similar stuff to the most recent chart hits in Canada, but which was released in the 1970s (probably somewhere else)"*

- more (richer?) info sources to discover and make accessible

This is the end!

*thank you !*

*[peter.knees@jku.at](mailto:peter.knees@jku.at)*

*[markus.schedl@jku.at](mailto:markus.schedl@jku.at)*