Automatically Extracting, Analyzing, and Visualizing Information on Music Artists from the Web

Workshop on Learning the Semantics of Audio Signals (LSAS)
21st June 2008, Paris, France

Markus Schedl, Peter Knees

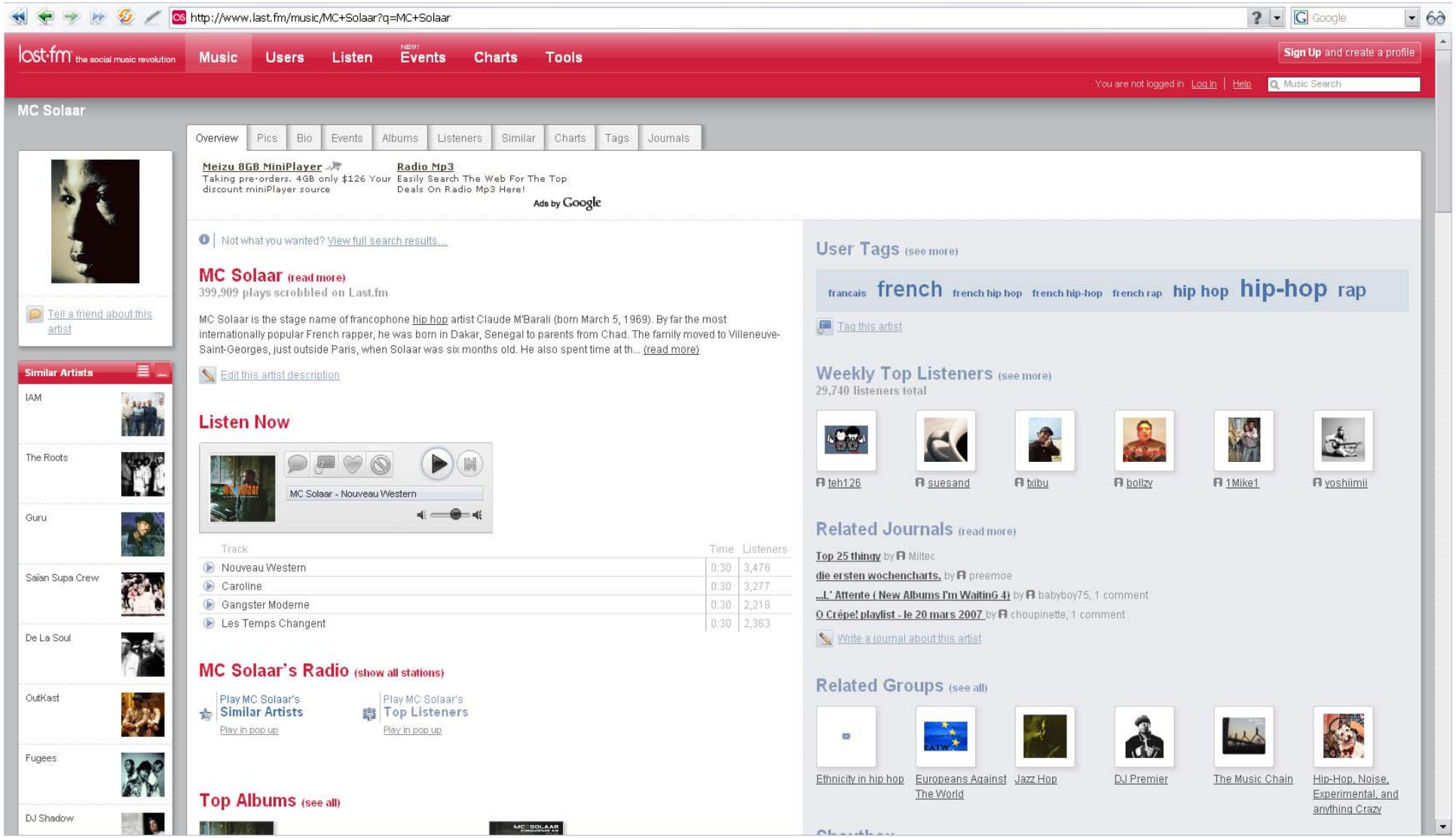
Department of Computational Perception
Johannes Kepler University Linz

markus.schedl@jku.at
http://www.cp.jku.at

Overview

- Introduction
- Motivation for an Automatically Generated Music Information System
- Data Processing Pipeline for Web Information Retrieval
- Information Extraction
  - Artist Similarity
  - Prototypicality of Artist for a Genre
  - Album Cover Artwork
  - Band Members and Instrumentation
  - Descriptive Terms (Tagging & Visualization via Co-Occurrence Browser)
- Future Work

Example of a Music Information System: *last.fm*



Example of a Music Information System: *allmusic*

The Big Picture

- Creating an automatically generated/populated music information system (AGMIS)

- How ?

  → Web Content Mining (Text, Image, Audio, Video)

- Using techniques from Information Retrieval (IR) and Natural Language Processing (NLP)

Motivation for AGMIS

- No need for labor-intensive maintainance of the system (no music experts, nor large community needed)
- Not vulnerable to editors‘ cultural bias (allmusic), nor to vandalism (last.fm)
- Automatical incorporation of new information as soon as they become available on the Web

What AGMIS Will Look Like

AGMIS
The Search Engine for Music Artists

Artist or band name: search

Hammerfall

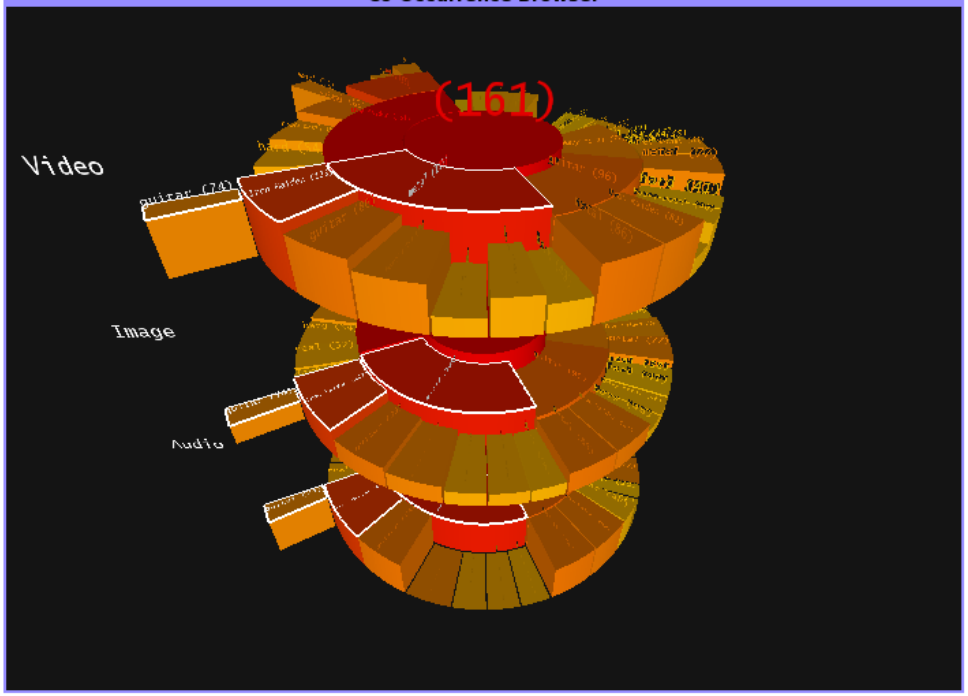
Genre:	**Rock**
Prototypicality Rank:	**n/a of 268878**

Similar Artists	
96	Dragonforce
84	Edguy
80	Blind Guardian
77	Helloween
73	Primal Fear

Descriptive Terms	
100	scandinavian metal
99	metal
99	true metal
98	power metal
97	swedish
84	melodic metal

Band Members	
Anders Johansson	drums, vocals
Fredrik Larsson	bass
Joacim Cans	vocals
Oscar Dronjak	guitar
Stefan Elmgren	guitar

Co-Occurrence Browser



Discography

Crimson Thunder

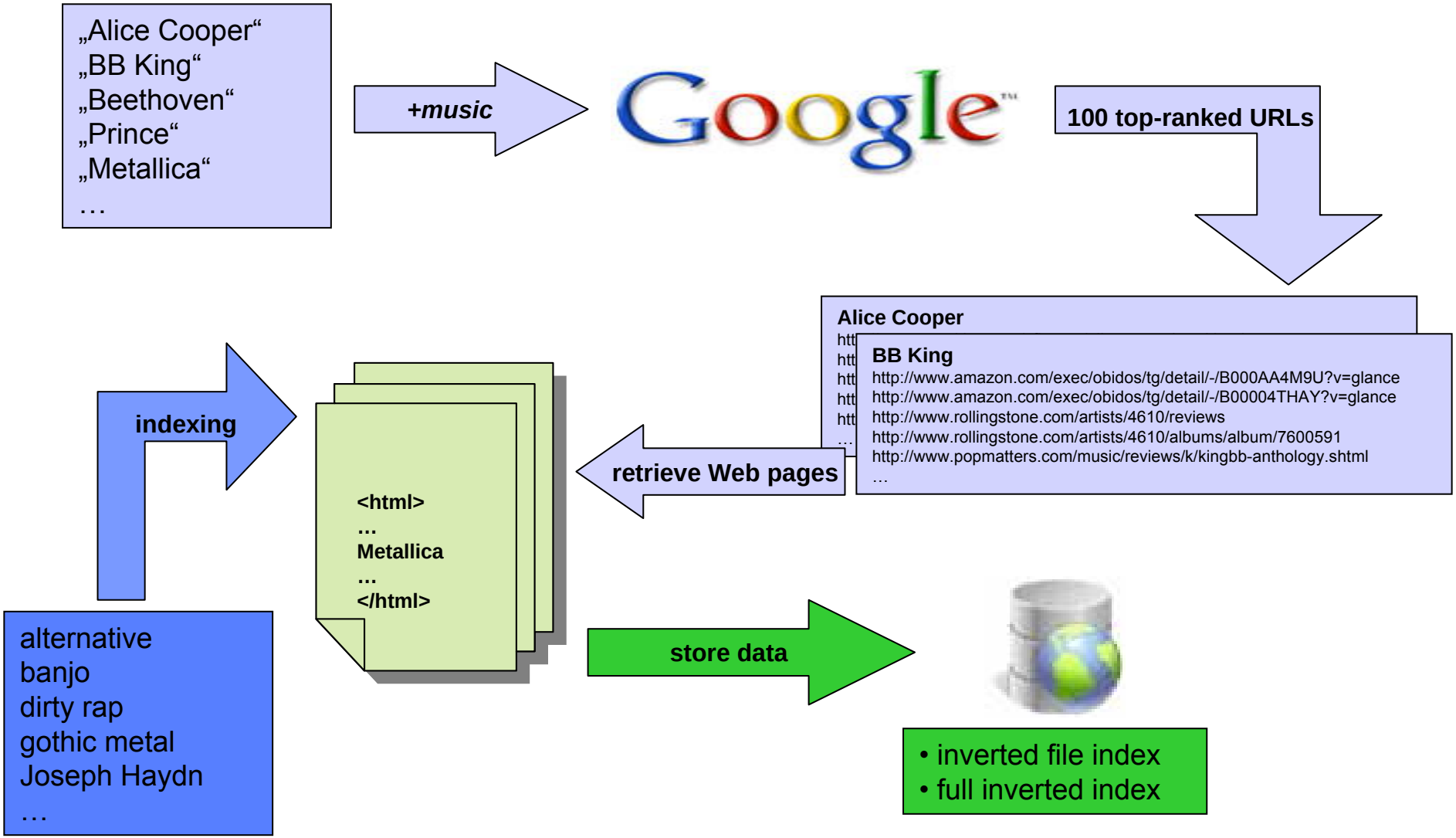
Legacy of Kings

Renegade

Parts of AGMIS

- Similar and prototypical artist detection
- Album cover retrieval
- Band member and instrumentation detection
- Automatic attribution/tagging of artists
- UI to browse artist-related Web pages (Co-Occurrence Browser)

Data Processing Pipeline



Similar and Prototypical Artist Detection

- Data source: inverted file index
- Calcuate document frequency (DF) of artist name *v*
  on Web pages retrieved for artist *u*
- Estimate conditional probability for artist *v* to be found on an arbitrary Web page of *u* (relative frequency DF_{uv} / DF_{uu})
  → asymmetric conditional probabilities
- Compute arithmetric mean to derive a symmetric **artist similarity measure**
- Use asymmetric probabilities to estimate **prototypicality of an artist for a genre** (idea: within a genre, Web pages about less prototypical artists tend to mention more prototypical artists more frequently than vice versa)

Similar and Prototypical Artist Detection: Evaluation

- Artist similarity:

  On collection of 224 well known artists from 16 general genres (Rock, Classical, Blues, …):

  classification accuracy (k-NN, leave-one-out CV) of about 85%

  On collection of 103 artists grouped in 22 quite specific genres (Bossa Nova, Death Metal, Jazz Guitar, German Hip-Hop, …):

  classification accuracy (k-NN, leave-one-out CV) up to 70%

- Artist prototypicality:

  On collection of 1,995 artists from 9 genres:

  overall agreement with importance ranking by AMG: 60-65%

Album Cover Retrieval

- Data source: full inverted index (word level + HTML tags)
- Image Pre-Filtering (quadratic, scanned CDs)
- Different approaches for image selection:
  - char/tag distance of artist and album names to <img> tag, select image with lowest distance
  - calculate an average histogram, select image which is nearest to it
  - use the first image returned by Google‘s image search (baseline)

Album Cover Retrieval: Evaluation

- Test set: 3,311 album names
- Best results using pre-filtering (quadratic constraint and scanned compact disc filter):

approach	*correct*
Google‘s image search (baseline)	56.7%
Avg. Histogram	10.0%
Tag distance	**58.9%**
Char distance	57.9%

Band Members and Instrumentation

- Data source: full inverted index
- Named Entity Detection to find candidate members (N-grams of capitalized words, filtering of common speech words)
- Rule-based Linguistic Analysis
    1. M plays the I
    2. M who plays the I
    3. R M
    4. M is the R
    5. M, the R
    6. M I
    7. M R

    M: member, I: instrument,
    R: role (singer, guitarist, bassist, drummer, keyboardist)

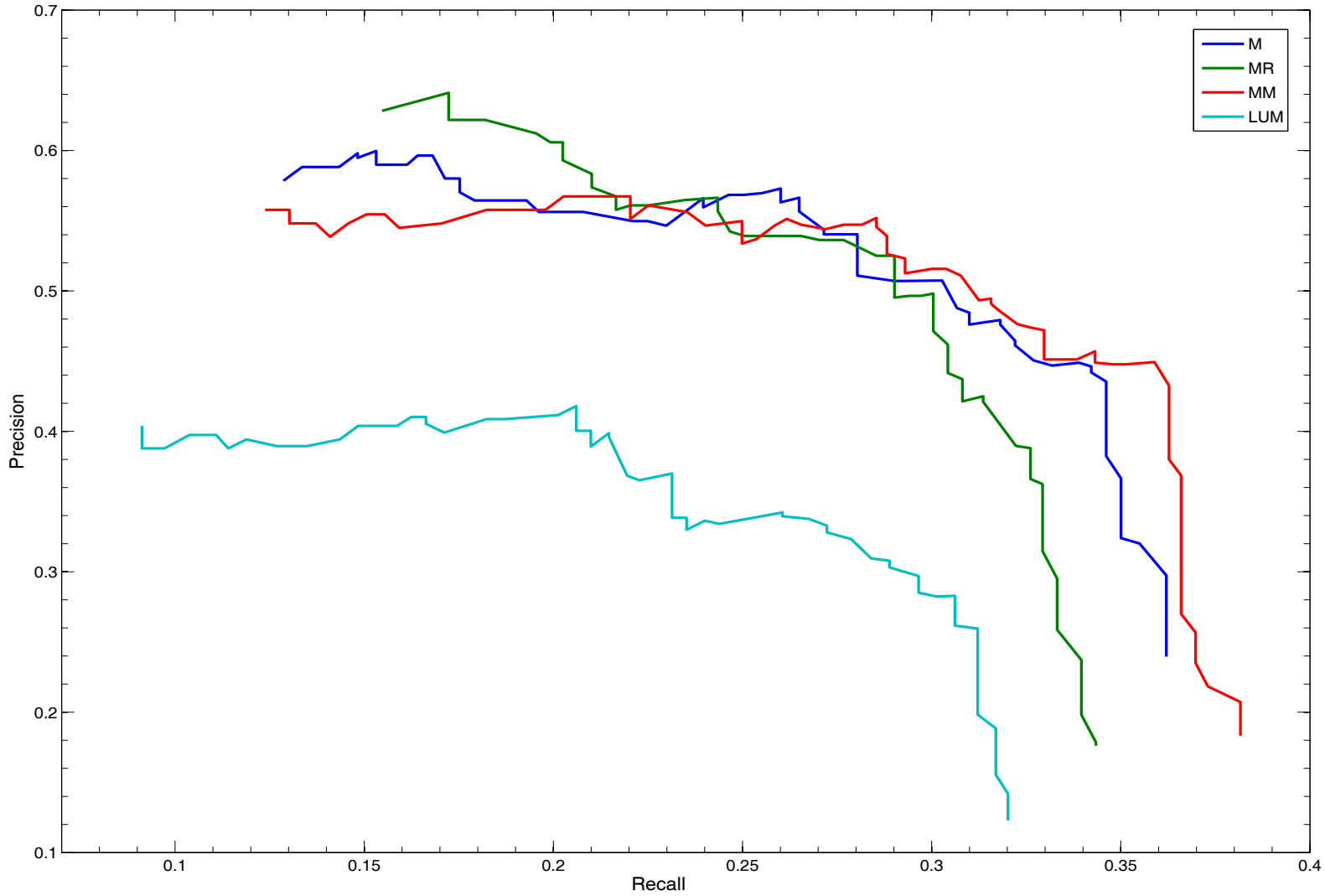
Band Members and Instrumentation (2)

- Calculate number of rule appliance
  → (member, instrument, rule, DF)
- Combine information over all rules
  → (member, instrument, ∑DF)
- Discard uncertain information, i.e., (member, instrument)-pairs with ∑DF value below a threshold t_{DF}
- Predict remaining (member, instrument)-pairs
  → m:n assignment between member and instrument

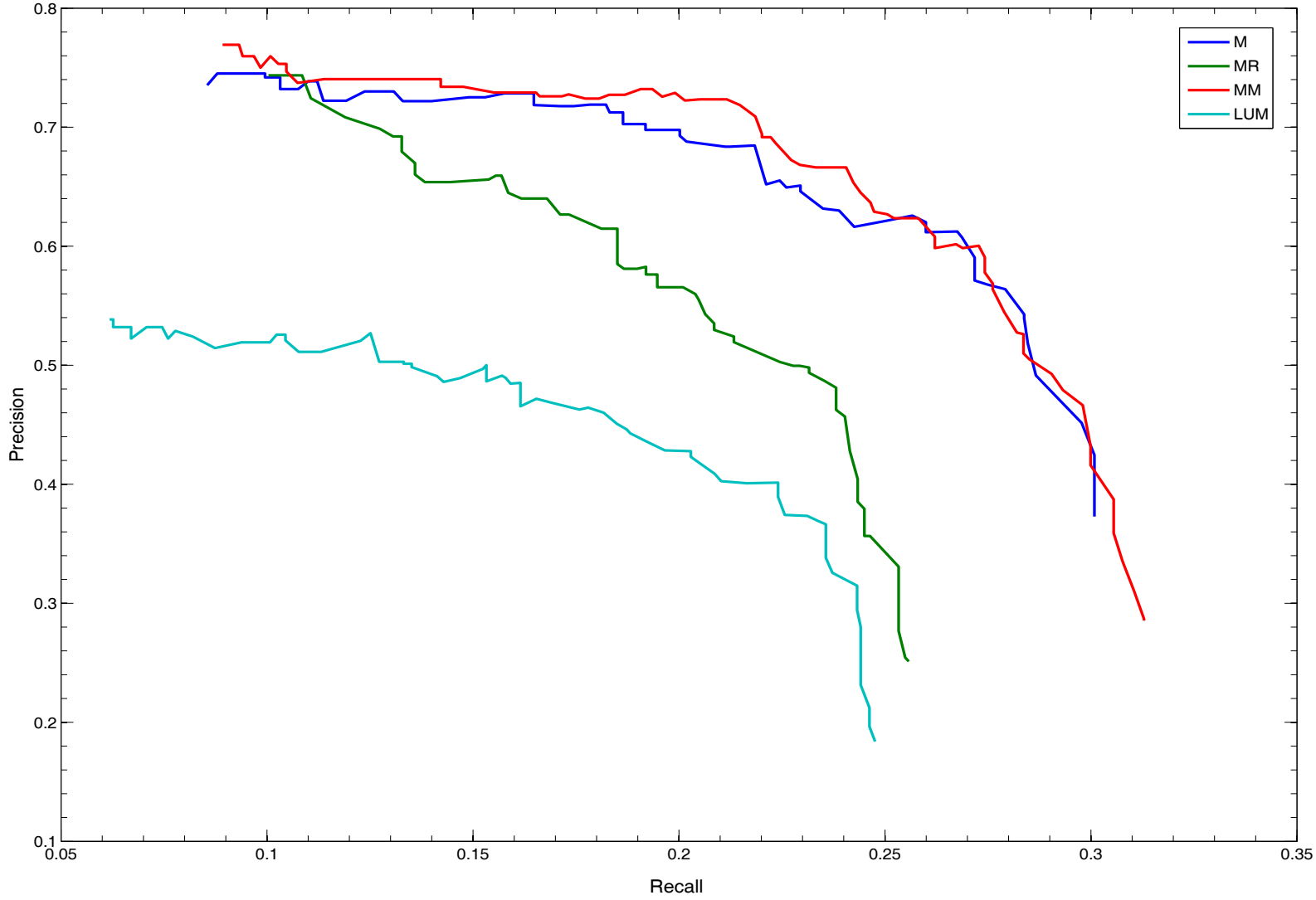
Band Members and Instrumentation: Evaluation

- 2 ground truth sets containing line-up of 51 bands
  - M_c 240 current band members
  - M_f 499 current and former members
- Measure precision and recall (set of predicted band members vs. set of band members given by ground truth)
- Upper limit for achievable recall: about 60%

Band Members and Instr.: Results M_c



Band Members and Instr.: Results M_f



Automatic Attribution/Tagging of Artists

- Data source: full inverted file index
- Different term weighting functions (TF, DF, TFxIDF) to rank terms from music dictionary occurring on corpus of artist‘s web pages
- User study to assess descriptivenes of highest ranked terms, using the three different weighting functions:
  - 112 well known artists from 14 genres
  - Web page indexing using dictionary of 1,506 musically relevant terms
  - 10 highest ranked terms of the 3 weighting functions merged
    → 1 term set for each artist
  - 5 participants, each told to rate terms for the artists they knew well (categorizing each term in three classes: +, -, ~)

Automatic Attribution/Tagging of Artists: Results

- 172 individual artist ratings returned
- 92 of 112 artists covered
- Overall excess of good terms (+) over bad terms (-)
  - TF: 2.22
  - DF: 2.42
  - TFxIDF: 1.53
- TF and DF performed significantly better than TFxIDF, no significant difference between TF and DF

Browsing Artist-Related Web Pages via COB

- Data source: full inverted file index

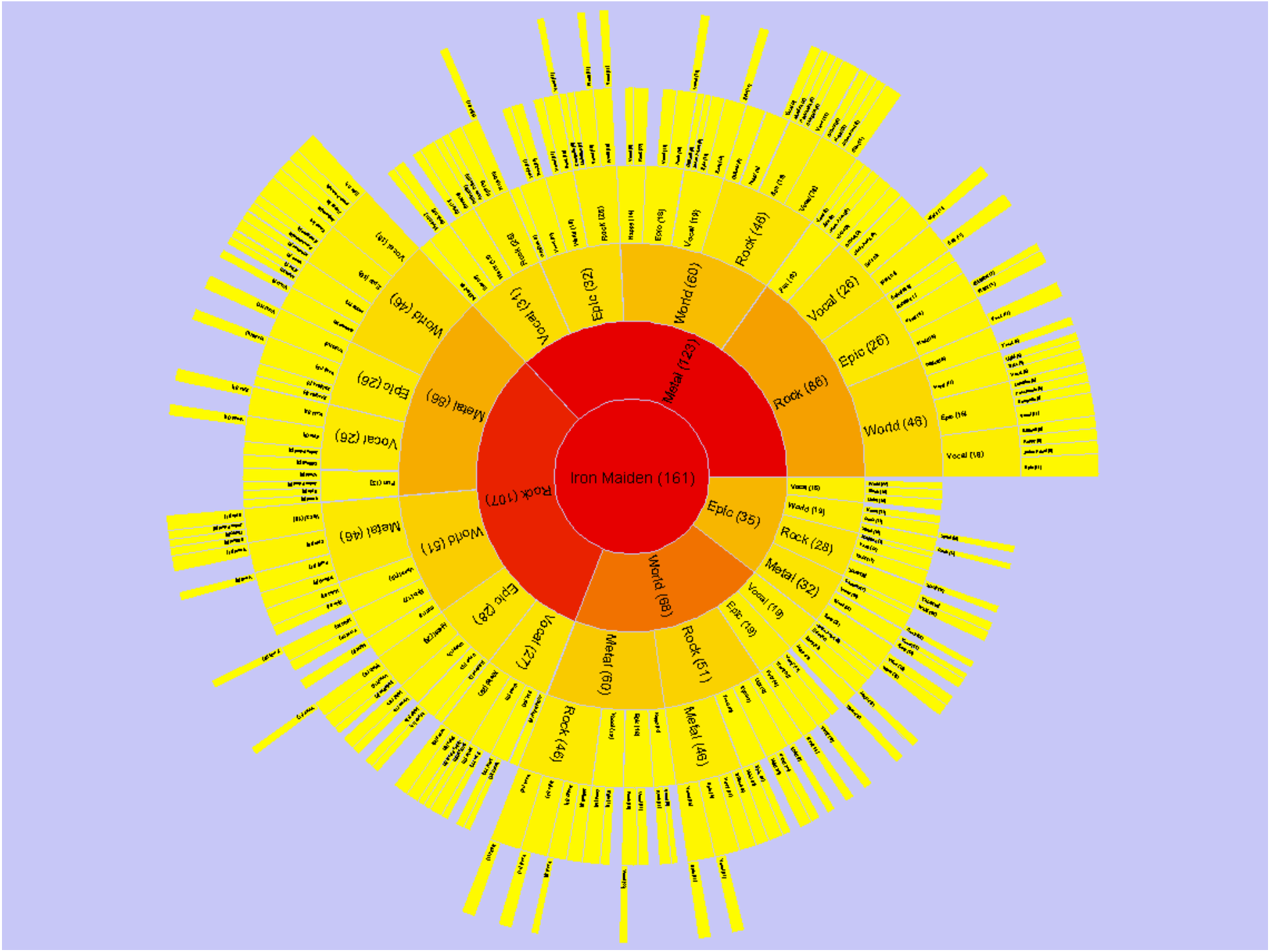
- Create co-occurrence tree and visualize it

  Algorithmic outline:

  1. Start at root node (set of all Web pages retrieved for artist)
  2. Select i most important terms t_i (according to term weighting func.)
  3. Create new sets of Web pages containing terms: $\{t_1\}$, …, $\{t_i\}$
  4. For each of these sets, goto 2. (until maximum depth reached)

Sunburst / InterRing Visualization

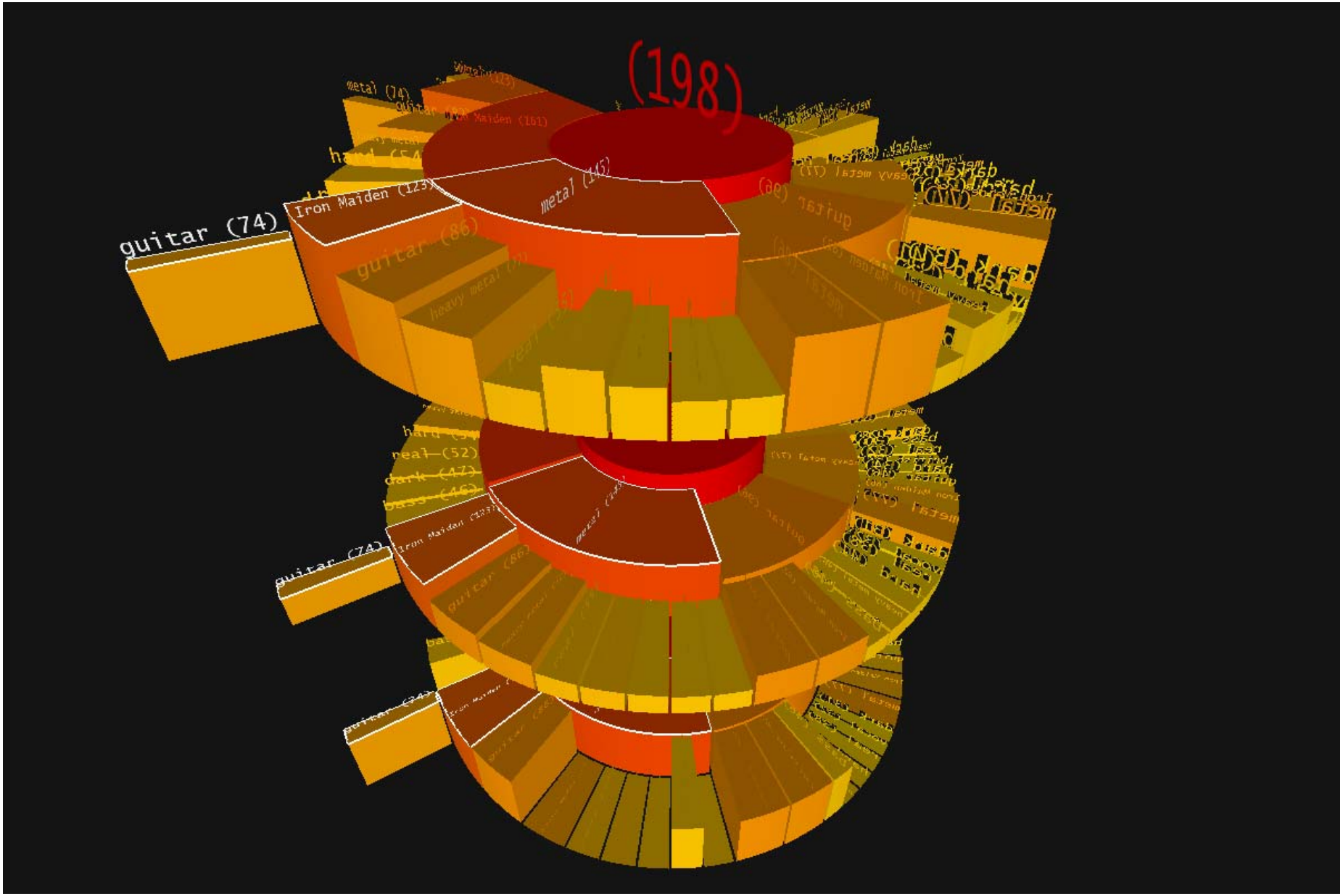
- Circular, space-filling visualization technique
- Center represents root node
- Deeper elements in hierarchy are drawn further away from center
- Children are drawn within angular borders of their parent


Iron Maiden (161)
Metal (123)
Rock (86)
World (46)
Rock (107)
Metal (86)
Vocal (26)
Epic (26)
Epic (35)
Rock (28)
Metal (32)
Vocal (19)
Epic (19)
World (66)
Rock (51)
Metal (60)
Rock (46)
Metal (46)
World (51)
Vocal (27)
Epic (28)
Metal (46)
Vocal (26)
Epic (26)
World (46)
Vocal (31)
Epic (32)
World (60)
Rock (48)
Vocal (26)
Epic (26)
World (46)

Co-Occurrence Browser

- Brings the Sunburst to 3D
- Additional data dimension can be encoded in height of each arc
- Stacking a number of such 3D-Sunbursts offers even more dimensions
- Arcs illustrate Web pages with certain term combinations (the most important ones according to some term weighting function)
- Amount of multimedia data found on the artist-related Web pages is visualized (three layers – for audio, image, and video files)
- User interaction by rotating, zooming, changing the view angle, displaying Web pages and multimedia content

Co-Occurrence Browser



Future Work

- Combine with audio signal-based MIR (incorporate real music)
- Automatically detect country of origin for artists
- Automatic biography generation
- Automatically detect new artists on the Web