The Quest for Ground Truth in Musical Artist Similarity

Dan Ellis, Brian Whitman, Adam Berenzweig, Steve Lawrence
Columbia, MIT, NEC

Outline

1. Similarity and Artists
2. Sources of Information
3. Evaluation
4. Applications and Conclusions

Lab
ROSA
Laboratory for the Recognition and
Organization of Speech and Audio

1. Similarity and music

- Want similarity for recommendation etc.

- Multiple possible information sources

  - Do they agree? Which is best?

- Exercise: define single 'best-fit' pairwise artist distance matrix:

	abba	ace of base	aerosmith	a-ha
abba	0	2.5	4.6	1.2
ace of base	2.5	0	4.4	1.4
aerosmith	4.6	4.4	0	4.1
a-ha	1.2	1.4	4.1	0

  - Training data for audio-based metrics?

Lab
ROSA
Laboratory for the Recognition and
Organization of Speech and Audio

Problems

- **Similarity judgment**
  - ... is individual
  - ... is deep
  - ... depends on the context/criteria
  - ... can be asymmetric

- **Many aspects of similarity**

- **But .. try it and see what happens**

Lab
ROSA
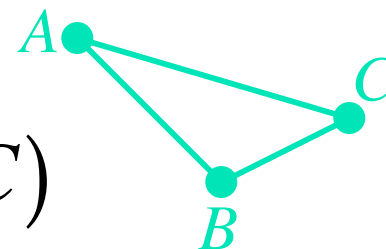Laboratory for the Recognition and
Organization of Speech and Audio

Distances

- Similarity (1..0) or distance (0..∞)?

  - How to interpret the numbers:
    $$Sim = \Pr(\text{artist A called artist B}) \ ?$$

  - Relate the two: $dist = \left(-\log\left[sim\right]\right)^{k}$

- Symmetry? "A is like B" ⟺ "B is like A"

- Triangle inequality:
  $$dist(A,C) \le dist(A,B) + dist(B,C)$$

- Distances → Geometric embedding
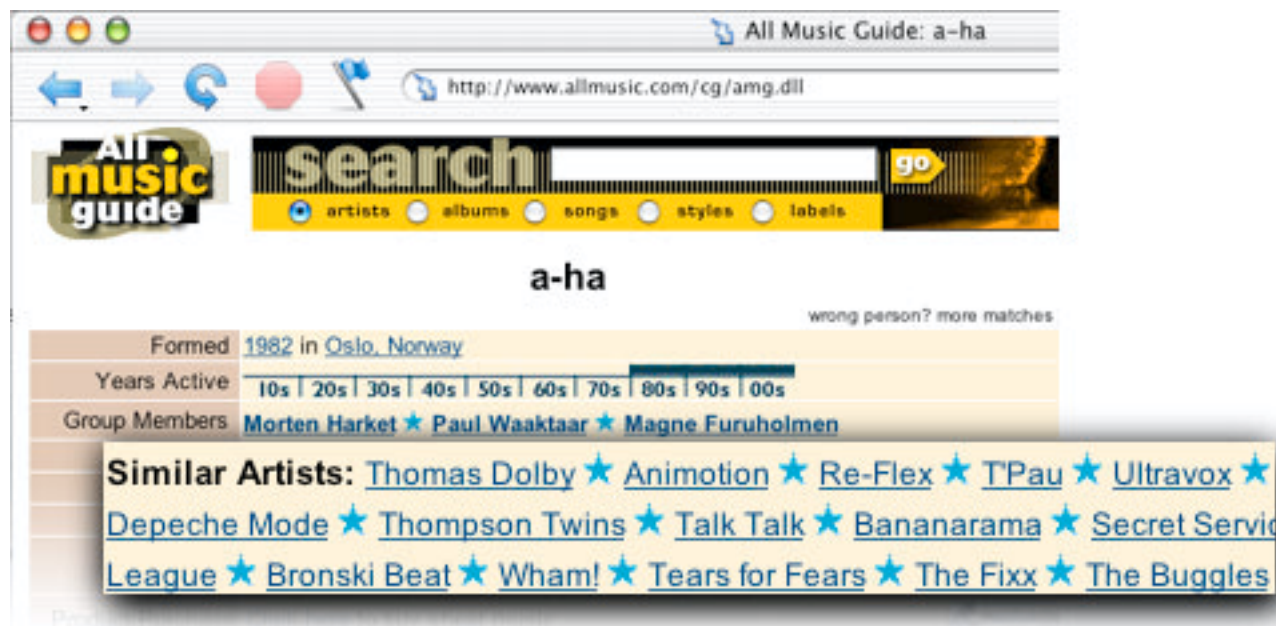
2. Data sources

- **Many potential sources of subjective 'ground-truth' data:**
  - Explicit descriptions ("*A* is similar to *B*, *C*..")
  - Behavioral i.e. used in similar contexts
  - Described in similar terms
- **Issues:**
  - Converting to numerical values
  - Extending to cover every pair of artists

Lab
ROSA
Laboratory for the Recognition and Organization of Speech and Audio

Explicit similarity statements

- *All Music Guide* similar artists lists

- Binary similarity (in/not in list)

Erdös Distance

- ## Distance from A to B is minimum # hops

roxette —— a-ha

abba

pet shop boys
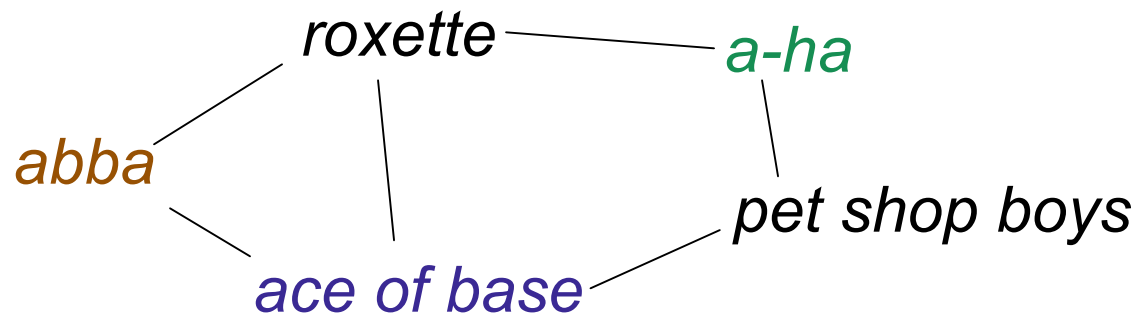
ace of base

	abba	ace of base	aerosmith	a-ha
abba	0	1	5	2
ace of base	1	0	4	2
aerosmith	5	4	0	4
a-ha	2	2	4	0

- ## also 'Resistive' Erdös

Lab
ROSA
Laboratory for the Recognition and
Organization of Speech and Audio

Collection co-occurrence

- **Assumption:**
  User *U* possess music by *A* and *B*
  $\Rightarrow$ *A* and *B* are similar

  - problematic for individuals, averages out

- *OpenNap* data:
  Use Napster protocol to record the collections of ~3000 users

  - 400,000 identified tracks

- $P(A|B) \times$ popularity mismatch discount

Similarity of associated text

- Find web pages mentioning particular artists (`"google abba music"`)

- *tf•idf* weighted terms for each artist

- Similarity of artists
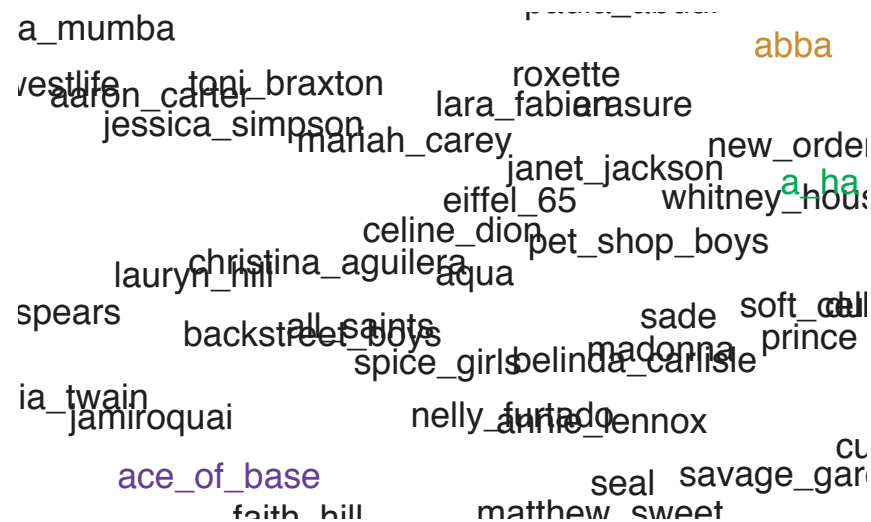 = weightings of common terms

- *Klepmit* data

Lab
ROSA
Laboratory for the Recognition and
Organization of Speech and Audio

Multidimensional Scaling

- Map distances to spatial arrangement minimizing MSE of Euclidean distances

- Regularizes distance matrices:

a_mumba

abba

westlife toni_braxton lara_fabian erasure
aaron_carter
jessica_simpson mariah_carey new_order
janet_jackson a_ha
eiffel_65 whitney_hous
celine_dion pet_shop_boys
christina_aguilera aqua
lauryn_hill
spears all_saints sade soft_cell
backstreet_boys prince
spice_girls belinda_carlisle madonna
ia_twain furtado
jamiroquai nelly_furtado annie_lennox
ace_of_base seal savage_gar
faith_hill matthew_sweet

- Issues: # dimensions, distance warping

3. Evaluation

- How to compare distance metrics?

- Need *independent* test data

→ Collect new evaluation data set

  - Sparse sampling of artist judgments to validate the complete set

Lab
ROSA
Laboratory for the Recognition and
Organization of Speech and Audio

Information elicitation

- ~~"What is similarity of A and B?"~~
  - no consistency, arbitrary scale

- **Target** + Choices paradigm:

Which artist is most similar to:
Janet Jackson?

1. R. Kelly
2. Paula Abdul
3. Aaliyah
4. Milli Vanilli
5. En Vogue
6. Kansas
7. Garbage
8. Pink
9. Christina Aguilera

- Chosen is *more like* **target** than (known) others

Lab
ROSA
Laboratory for the Recognition and Organization of Speech and Audio

`musicseer.com`

- Web site to collect artist-similarity judgments from anonymous volunteers

- Two procedures:

  - `poperdos` game:
    Get from artist *A* to *B* in fewest # steps

  - Direct survey:
    Choose most similar artist from a list

Lab
ROSA
Laboratory for the Recognition and
Organization of Speech and Audio

poperdos game

- Competitive motivation of informants?
- Choices from AllMusic lists + OpenNap
  $\rightarrow$ roughly equidistant choices
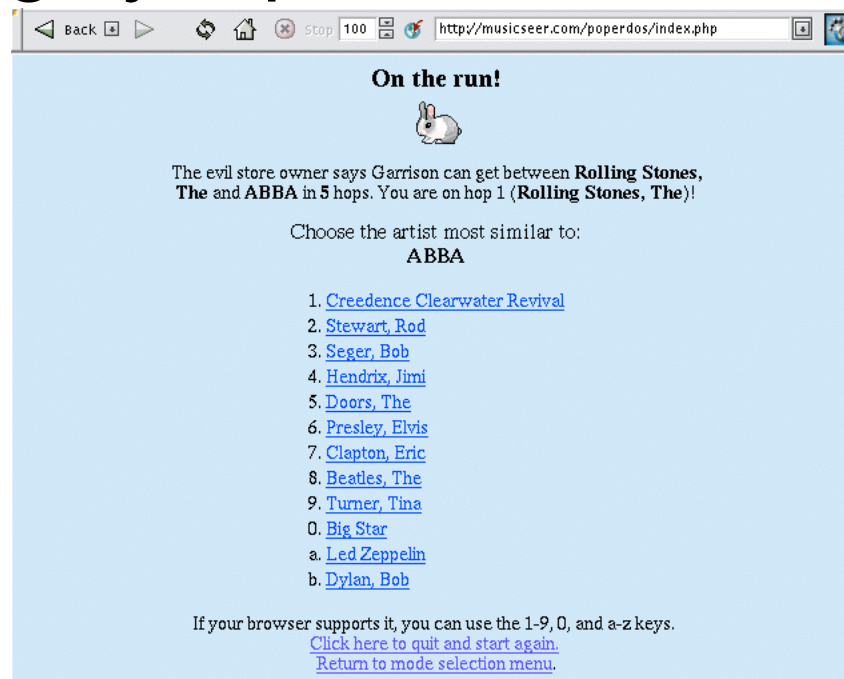
Survey

- ## Sequence of single judgments
  - ### More uniform sampling of targets
  - ### Infer which bands are known to informant
  - ### Dummy entries 'catch' random guessing
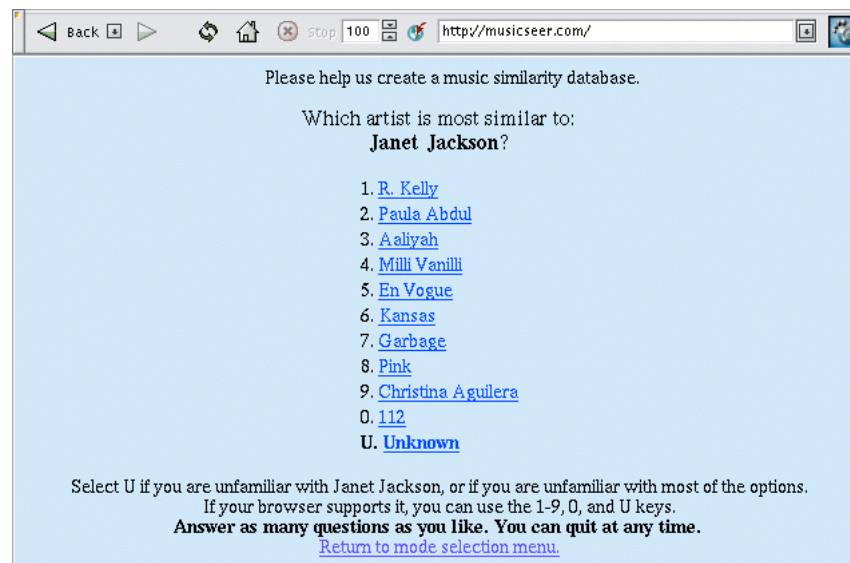
"10,000 random band names"

- **Generative model of band names**
  - trained on real band name 'grammars' + vocabulary of music-related web pages

- **Most popular decoy choices:**
  - *Neither Palindrome Taker*
  - *Archbishop Riot*
  - *Skylight Cuisine*
  - *Blond and Bipolar*

Lab ROSA
Laboratory for the Recognition and Organization of Speech and Audio

`musicseer` data

- ## 1,782 users
  - ### best user: 748 judgments, ten above 200
- ## 22,470 judgments
  - ### ~140k triplets `<target chosen unchosen>`
  - ### Spike after "10,000 names" blogging

Musicseer Trials/Day

- `http://musicseer.com/results/`

Lab
ROSA
Laboratory for the Recognition and
Organization of Speech and Audio
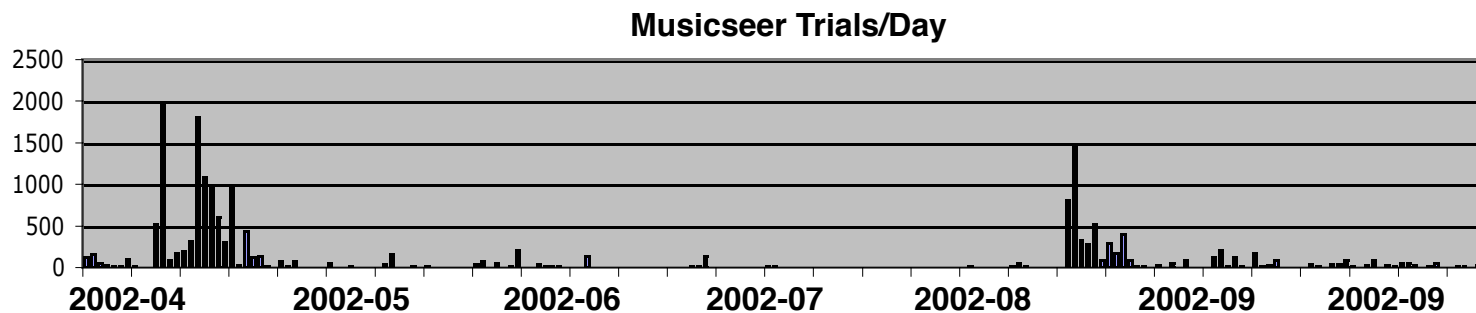
Scoring the similarity metrics

- Avg. ranking of user choice by metric

- Agreement percentages for...

  - judgments
  - triplets

Target: **Police**

Choices: U2
(ranked by Pretenders
metric) Sting
Queen
INXS
Phil Collins

Informant choice ranked **#2** by metric

⇒ **no agreement** for this *judgment* ... but 4 of 5 *triplets* do agree

Results

- Judgment-level agreement %:
  i.e. user choice ranks #1 under metric

Top rank agreement
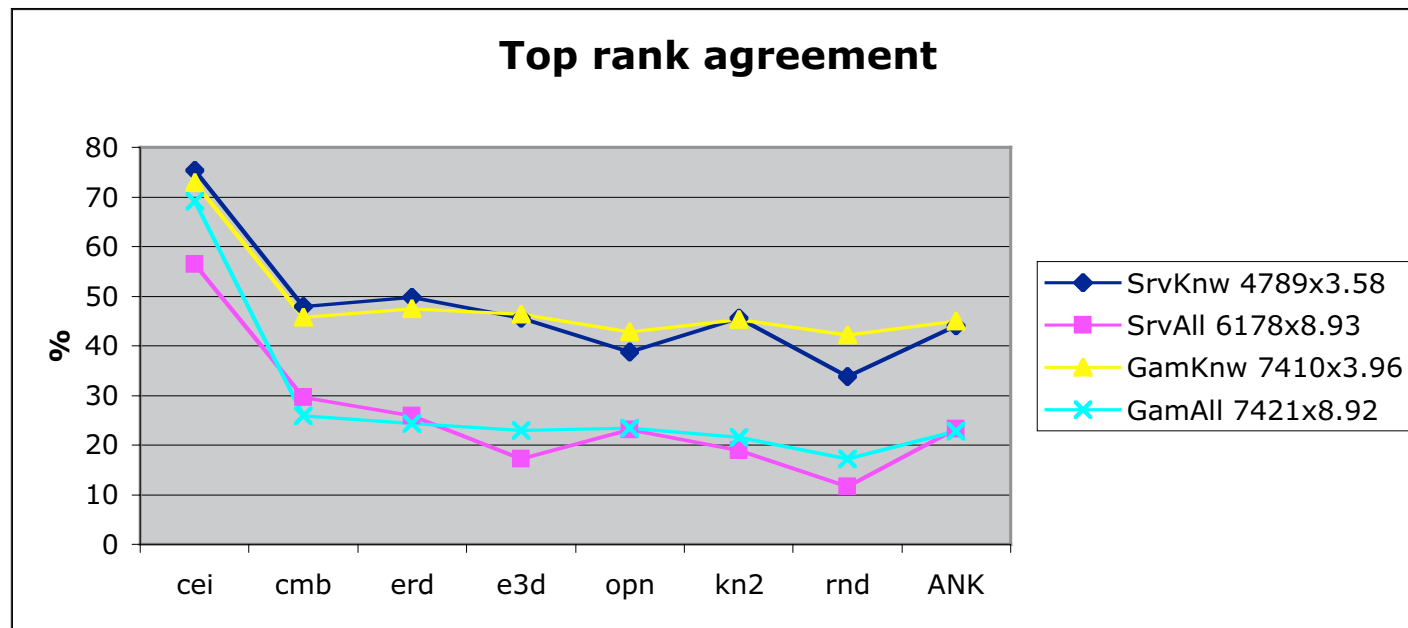
- "Ceiling" reflects inconsistency limit

4. Applications

- Original motivation:
  - Comparing ground-truth data sources
  - Targets for training music similarity classifier
  - What happens when you try?
- Actual outcome
  - `musicseer` data
  - Evaluation metrics

Lab
ROSA
Laboratory for the Recognition and
Organization of Speech and Audio

Using evaluation data

- ## Significant information in `musicseer` evaluation data

  - ### basis for a metric in itself?

- ## Use these evaluation procedures for acoustic-based similarity measures

- ## 'Inconsistency' may point to *aspects*...

Lab ROSA

Laboratory for the Recognition and Organization of Speech and Audio

Playola

- **Acoustic similarity browser**

Lab
ROSA
Laboratory for the Recognition and
Organization of Speech and Audio

Conclusions

- **Musical artist similarity - a tricky concept**
  - but *how* tricky?
  - best metric agreed with 50% of judgments

- **Evaluation procedure + data has broader applications**
  - Most valuable outcome?

- **Web-based user data collection is viable**

Lab
ROSA
Laboratory for the Recognition and
Organization of Speech and Audio