Artist Classification with Web-based Data

Peter Knees, Elias Pampalk, Gerhard Widmer

Austrian Research Institute for
Artificial Intelligence (OeFAI)
Vienna, Austria

ISMIR 2004, October 11-14. Barcelona, Spain

Outline

1. Aims
2. Related Work
3. System Architecture
4. Experimental Results
5. Conclusions
6. Future Work

Aims

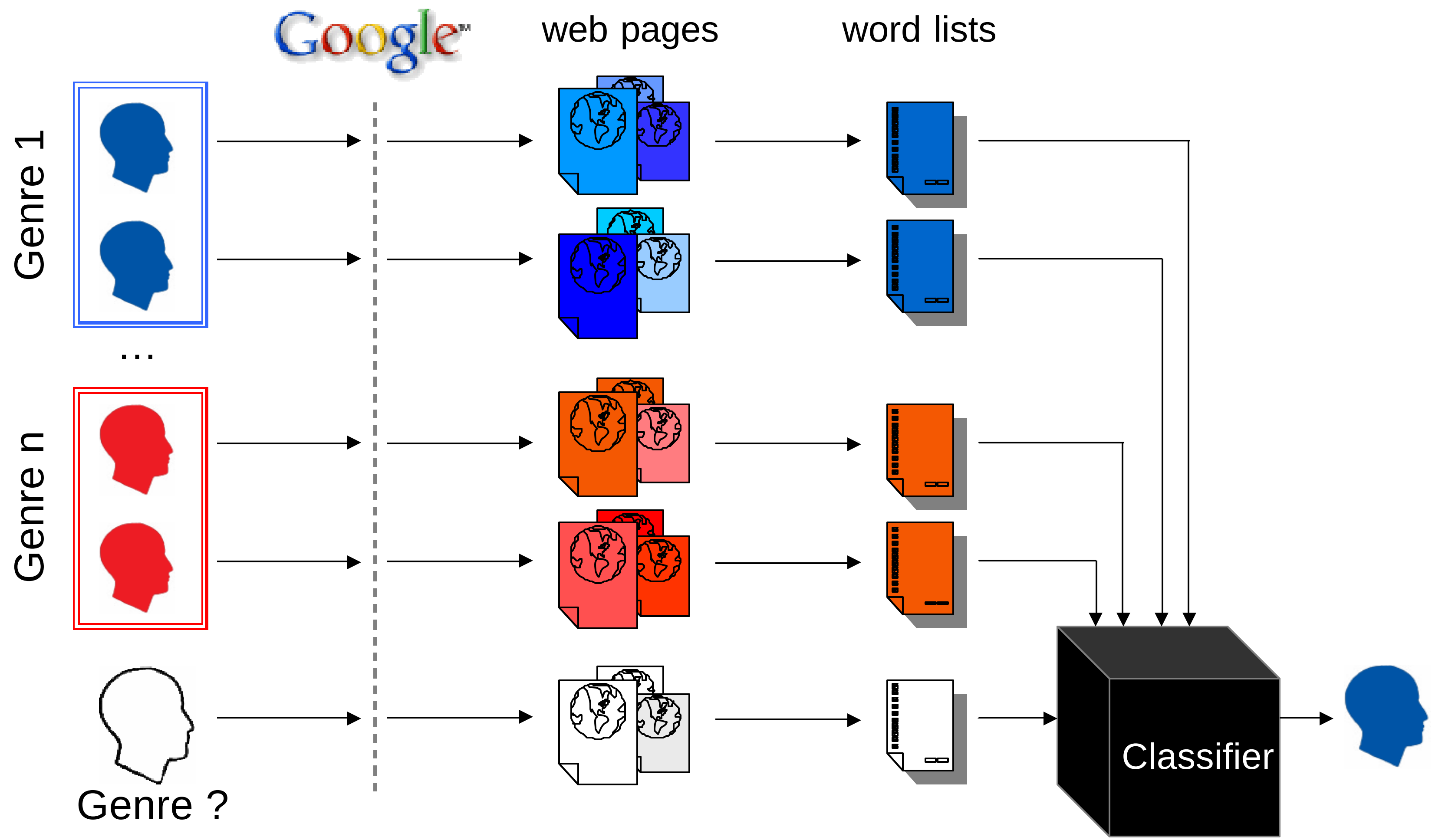
- Classification of artists into genres
- Using text features from the web

- Explore
  - number of examples necessary to define genres
  - application as similarity measure
  - validity of descriptors over time

Related Work

- Artist Similarity using Metadata
  - Whitman & Lawrence (2002)
  - Baumann & Hummel (2003)
  - Cano, Koppenberger & Celma (2004)
- “Style” Classification (Audio and Metadata combined)
  - Whitman & Smaragdis (2002)
- General
  - Pachet & Cazaly (2000)
  - Pachet, Westermann & Laigre (2001)
  - Ellis, Whitman, Berenzweig & Lawrence (2002)

System Overview



Data Source

- Retrieval of web pages containing the artist’s name
- Top 50 sites from Yahoo! or Google (in avg. 40 available)
- Query string:
  - “*artist name*” +music +genre +style
  - “*artist name*” +music +review

    (cf. Whitman & Lawrence, 2002)

Method (1)

- Remove all HTML tags
- Ignore stop words
- Count term (*tf*) and document frequency (*df*)
- Calculate *tf* x *idf* $$tfidf_{ta} = tf_{ta} \log_2 \frac{N}{df_t}$$
  - Rewards frequent terms (high tf)
  - Penalizes terms that occur in many documents (high df — low inverse df)

Method (2)

- Term Selection
  - For each genre find most discriminative terms
  - Use ?^{2} test for ranking
  - Best results with top 100 terms per genre
- Cosine normalization
- Classification with Support Vector Machines

Example of ?^{2} ranked Term List

Top 40 for Heavy Metal

defined through Black Sabbath, Pantera, Metallica, Def Leppard

100	sabbath	40	dimebag	28	hammett	24	halford
97	pantera	40	anselmo	27	bloody	24	dio
89	metallica	40	pyromania	27	thrash	23	reinventing
72	leppard	40	paranoid	27	phil	23	lange
58	metal	39	osbourne	26	lep	23	newsted
56	hetfield	37	def	26	heavy	21	leppards
55	hysteria	34	euphoria	26	ulrich	21	adrenalize
53	ozzy	32	geezer	26	vulgar	21	mutt
52	iommi	29	vinnie	25	megadeth	20	kirk
42	puppets	28	collen	25	pigs	20	riffs

(normalized: max. = 100)

Experiments

- Using 14 genres:
  Country, Folk, Jazz, Blues, R&B/Soul, Heavy Metal, Alternative/Indie, Punk, Rap/HipHop, Electronic, Reggae, Classical, Rock'n'Roll, Pop
- Manually gather 16 representatives per genre
- 50 hold out experiments
  - 2, 4, and 8 artists to learn a genre (rest as test set)
  - For two search engines and both query strings

Results (1)

query string	*+music +genre +style*			*+music +review*		
# artists for training	*2*	*4*	*8*	*2*	*4*	*8*
Google	70%	80%	86%	71%	81%	**87%**
Yahoo!	61%	72%	79%	65%	78%	**87%**

Mean accuracy from 50 hold out experiments
(using 100 top ranked terms per genre)

Results (2)

Real Class \ Classification Results	Country	Folk	Jazz	Blues	R&B/S	HM	A/I	Punk	R/HH	Elctr	Reggae	Class	R&R	Pop
Country	91 ±7	2 ±6					2 ±4			1 ±3			2 ±4	
Folk	11 ±8	55 ±22		1 ±3	5 ±8	2 ±5	9 ±7	1 ±3		1 ±3			15 ±16	1 ±3
Jazz		1 ±2	93 ±4		4 ±4								1 ±2	1 ±3
Blues				88 ±14	4 ±6								8 ±14	
R&B/Soul				1 ±3	75 ±12				4 ±7				14 ±9	6 ±8
Heavy Metal		2 ±4			1 ±2	75 ±16	9 ±11	6 ±11					2 ±4	6 ±8
Alt/Indie		1 ±3			1 ±3	7 ±10	57 ±23	13 ±14		12 ±9			1 ±4	8 ±10
Punk		1 ±2			1 ±2	8 ±11	11 ±12	69 ±14						10 ±7
Rap/HipHop					1 ±3				91 ±9	4 ±4				3 ±6
Electro							5 ±7	1 ±2		94 ±8				
Reggae					3 ±6		1 ±4		4 ±5	2 ±4	86 ±9		1 ±2	2 ±4
Classical												100 ±0		
Rock n' Roll	4 ±4	3 ±6		2 ±5	2 ±6	5 ±4	8 ±10	2 ±5		1 ±3			74 ±15	1 ±3
Pop		1 ±2			2 ±5	1 ±2	5 ±6	1 ±3	1 ±3	2 ±4			1 ±3	87 ±8

Confusion Matrix from 50 hold out experiments

upper value: mean
lower value: std

Google,
+music +review,
top 100 terms per genre,
used 4 artists each to learn a genre

Results (3)

Similarity (SOM-Visualization with 224 artists)

REGGAE (14)	country (1) rnbsoul (1)	COUNTRY (1) folk (2) rocknroll (1)	ROCKNROLL (8) folk (2) blues (1) rnbsoul (1)	BLUES (14) folk (1) rnbsoul (1) rocknroll (1)	CLASSIC (16)
altindie (3) rocknroll (3) folk (2) punk (2) electro (2) pop (1)	FOLK (5)	rnbsoul (4) folk (4)	rnbsoul (3) jazz (1) pop (1)	blues (1)	rocknroll (1)
altindie (5) punk (4) rocknroll (2)	altindie (1) electro (1) pop (1)	POP (5)	RNBSOUL (5) pop (2)		JAZZ (15)
HEAVY (15) PUNK (9) ALTINDIE (6)	electro (2) altindie (1) punk (1)	ELECTRO (10) pop (1)	RAPHIPHOP (13) pop (1)	raphiphop (2) reggae (1) pop (1)	pop (3) folk (2) rnbsoul (1) heavy (1) raphiphop (1) electro (1) reggae (1)

Results (3)

Similarity (SOM-Visualization with 224 artists)

REGGAE (14)	country (1) rnbsoul (1)	COUNTRY (1) folk (2) rocknroll (1)	ROCKNROLL (8) folk (2) blues (1) rnbsoul (1)	BLUES (14) folk (1) rnbsoul (1) rocknroll (1)	CLASSIC (16)
altindie (3) rocknroll (3) folk (2) punk (2) electro (2) pop (1)	FOLK (5)	rnbsoul (4) folk (4)	rnbsoul (3) jazz (1) pop (1)	blues (1)	rocknroll (1)
altindie (5) punk (4) rocknroll (2)	altindie (1) electro (1) pop (1)	POP (5)			
HEAVY (15) PUNK (9) ALTINDIE (6)	electro (2) altindie (1) punk (1)	ELECTRO (10) pop (1)	R pop (1)	reggae (1) pop (1)	heavy (1) raphiphop (1) electro (1) reggae (1)

k-NN … 77%

k=7, 8 artists defining each genre,
Google, +review, no ?² cut-off

Results (4)

- Stability of descriptors over time
  - Repeatedly sent queries for 12 artists
  - For a period of 4 months (56 times)
- SOM-Visualization:

AK … Alicia Keys
DP … Daft Punk
MJ … M. Jackson
Moz … Mozart
RW … Robbie Williams
SO … Stacie Orrico
Stro … The Strokes
Sub … Sublime
YND … Youssour N'Dour

Sub 56			RW 56		SO 54		YND 56
					SO 2		
Moz 56							AK 21
			MJ 56				AK 35
MM 56	Em 24	Em 32		DP 56		Pulp 56	Stro 56

Marshall Mathers = Eminem !

Conclusions

- 87% accuracy for 14 genres
  (using only web-based data)
- Features can be used for similarity
- Seems robust to fluctuations over time

But …

- “popularity” of artist (number of webpages)
- uniqueness of artist name

Future Work

- Hierarchical genre taxonomies
- Improve filtering techniques
- Extend to song/album level
- Larger evaluation sets / longer time periods
- Combine with audio signal based approaches
- Combine with the Islands of Music approach to explore different views of music collections