A Deeper Look into Web-based Classification of Music Artists

Peter Knees, Markus Schedl, Tim Pohle

Department of Computational Perception
Johannes Kepler University Linz, Austria

Department of
Computational
Perception

JOHANNES KEPLER
UNIVERSITY LINZ

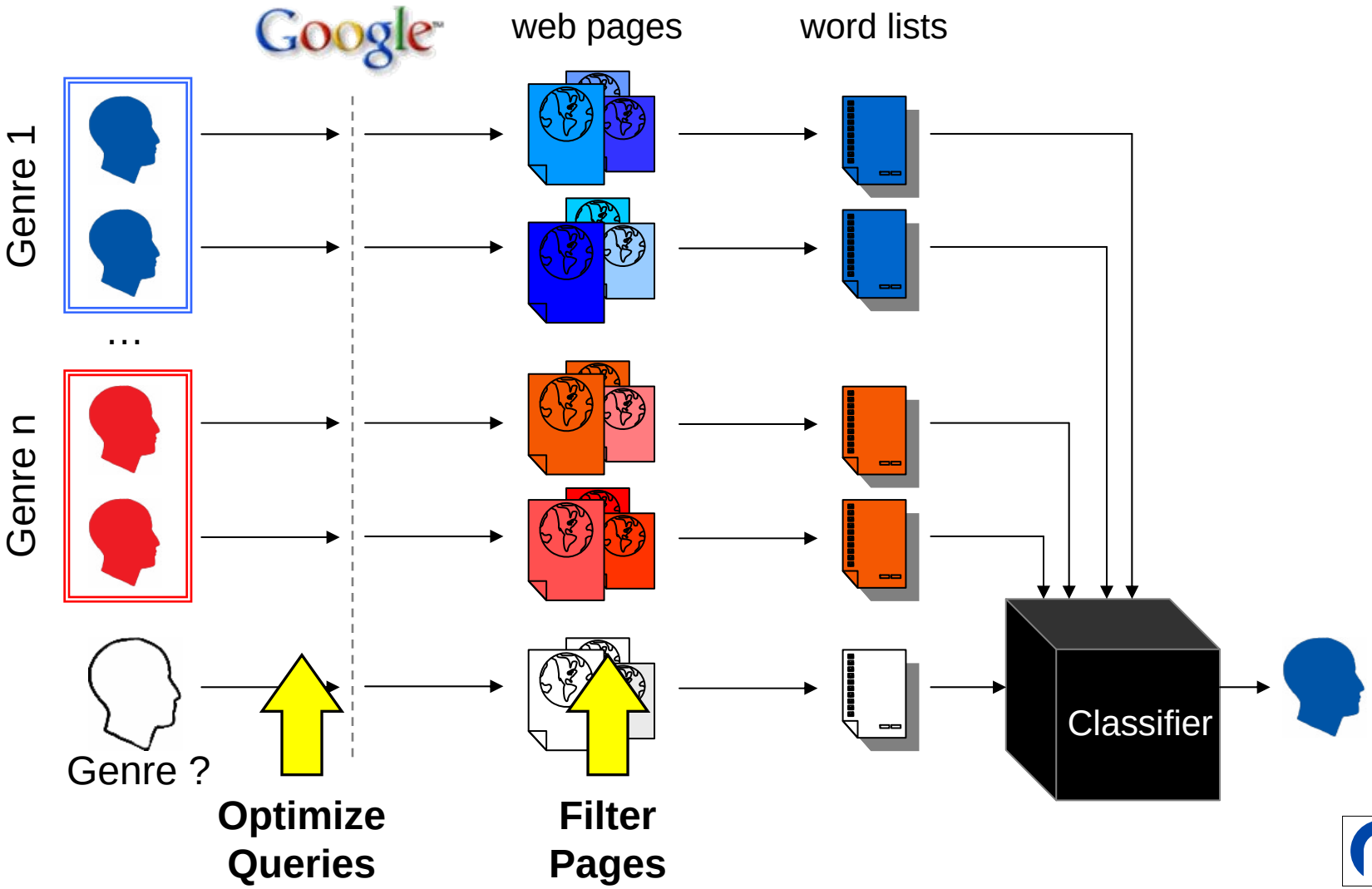
Overview

- Artist Classification with Web-based Data
- “Improvements”
  - Optimizing Queries
  - Page Filtering
  - Investigation of Results
- Simplified Approach
- Conclusions for Future Work

Introduction

- Idea: Classify music artists into genres based on related Web pages
- Obtain related Web pages via search engine
  - Then: *Text Categorization* task
  - *tf* x *idf* weighted term vectors describe artists
  - χ^2-test for dimensionality reduction
- No audio signal involved (no semantics either…)

Artist Classification with Web-based Data (ISMIR 2004)



Evaluation

- On 3 different genre taxonomies
  - *c224a*: from ISMIR'04 paper (**224 artists, 14 genres, baseline 7.4%)**
  - *uspop2002*: Berenzweig et al., CMJ 28(2) 2004 (**400a, 10g, bl 73.3%**)
  - *c103a*: Pampalk et al., ISMIR'05 (**103a, 22g, bl 5.8%**)
- n-fold Cross Validation
- SVM and Nearest Neighbor Classification

Optimizing Queries

- “Let Google do the filtering”
- Saves bandwidth and time
- Find terms that indicate relevant pages analytically
- To this end: Create a ground truth set of Web pages labelled either ”informative” or “uninformative”

Optimizing Queries (2)

- Starting with 700 random pages retrieved via "*artist name*"+music (35 new artists á 20pg)
- Labelling done by 3 experts: full agreement on 538 pages (198 informative, 340 not)
- χ^2-test to identify most discriminative terms
- also done for binary combinations of terms

+term1 +term2, +term1 –term2, -term1 +term2, -term1 -term2

Optimizing Queries (3)

terms/term combinations	χ^2 value
+like -mp3	0.278
+like	0.277
+like -videos	0.272
+work -prices	0.271
+work -mp3	0.269
+work -services	0.268
+like -download	0.267
+like -tickets	0.265
+like -cart	0.262
+like -login	0.261
+like +time	0.261
+work	0.260
+like -prices	0.260
+work -format	0.259
+work -health	0.258
+like +people	0.257

Optimizing Queries - Results

- Classification Accuracy (avg. over 50-fold CV)

	c224a		*c103a*		*uspop2002*	
	SVM	NN	SVM	NN	SVM	NN
+music	95.69	93.90	65.00	73.00	89.75	87.25
+music +review	92.69	83.40	60.00	70.00	86.50	85.25
+music +genre +style	90.90	89.10	58.00	63.66	87.25	85.75
+music +biography	91.19	84.70	58.33	68.66	89.00	80.75
+music +like -mp3	92.70	87.80	57.66	72.00	88.50	86.00
+music +like -videos	92.70	86.30	60.00	73.00	88.75	87.50
+music +like	94.90	91.99	59.66	72.66	89.25	85.00
+music +work -prices	89.99	83.20	52.33	59.00	86.50	84.25
+music +work -mp3	89.09	81.00	58.33	62.00	86.50	82.75
+music +work -services	89.49	83.70	56.66	57.00	87.50	83.75

Page Filtering

- Remove “uninformative” pages from retrieved set (worked for Baumann et al, WEDELMUSIC’03)
- Use ground truth set to train classifier

  Features: *tf* x *idf* weigths

  + HTML structure info (tag frequencies)
- Used RIPPER rule learner
  (estimated prediction acc.: 83%)

Page Filtering (2)

- Obtained rule set

---

if just >= 0.055528 and two >= 0.051821 **then**
  ⇒ informative
else if <p> >= 0.03515 and <i> >= 0.042748 **then**
  ⇒ informative
else if <p> >= 0.04258 and life >= 0.050582 **then**
  ⇒ informative
else if work >= 0.075633 **then**
  ⇒ informative
else if album >= 0.083651 and review >= 0.111605 and privacy <= 0.071766 **then**
  ⇒ informative
else
  ⇒ not informative
end if

Page Filtering - Results

- Classification Accuracy (avg. over 10-fold CV)

		c224a				*c103a*				*uspop2002*			
		unfiltered		*filtered*		*unfiltered*		*filtered*		*unfiltered*		*filtered*	
	pg.	SVM	NN	SVM	NN	SVM	NN	SVM	NN	SVM	NN	SVM	NN
+music	10	92.07	91.12	92.49	86.60	66.81	72.63	64.99	72.00	86.74	84.75	87.50	83.75
	25	92.01	86.62	95.61	91.14	68.72	71.63	61.00	66.90	87.25	85.74	88.24	83.75
	50	95.17	91.95	93.41	91.12	64.81	72.81	60.09	65.81	89.00	86.25	87.50	84.75
+music	10	92.92	85.23	92.03	83.91	62.09	66.90	57.18	63.18	86.50	84.49	88.00	84.75
+review	25	93.79	87.47	91.58	85.71	60.18	66.90	54.18	72.81	88.75	86.50	86.99	85.25
	50	92.92	83.85	92.01	81.24	56.27	64.18	52.45	68.09	85.50	87.00	85.00	84.50
+music	10	86.20	80.79	83.03	79.94	52.18	54.45	55.18	55.45	84.25	79.49	81.99	80.49
+genre	25	90.25	83.10	90.23	84.44	56.18	62.90	54.18	60.09	86.25	82.49	87.00	83.00
+style	50	92.47	88.02	89.90	85.29	56.18	63.81	51.36	65.72	86.75	85.50	86.75	86.00
+music	10	93.33	87.01	92.01	88.37	68.81	74.72	61.27	66.99	87.50	85.74	88.00	84.49
+like	25	93.81	92.47	92.92	90.67	67.90	72.72	58.18	70.81	88.24	85.00	87.50	85.00
	50	93.37	90.27	92.94	86.60	60.00	72.63	52.27	66.72	88.00	85.00	86.75	83.75

Discussion

- Neither Query Optimization nor Page Filtering consistently improved classification accuracy
- Problem seems to be the “ground truth page set”
- Users’ “informativeness” judgments not useful for genre classification
- What is useful for genre classification?

100 Most Relevant Terms for “Country”

mcgraw	traditionalist	curb	ropin	newman
cma	dunn	gallimore	buffett	helplessly
nashville	gentry	faith	mcentire	viacom
garth	chely	tulsa	hill	dolly
brooks	tnn	alan	raye	messina
chesney	acm	maines	trisha	collin
leann	chicks	tritt	parton	robison
rimes	honky	martina	loveless	seidel
shania	tonk	aboutcountry	outlaw	tennessee
toby	montgomery	funstuff	clint	daryle
cyrus	keith	gac	getcha	cmas
twain	honkytonk	chattahoochee	oklahoma	patsy
country	andrews	shockn	cowboylyrics	tippin
fireflies	tim	haggard	tonks	flatts
cmt	mcbride	lila	hgtv	tn
strait	shave	martie	hgtvpro	jacked
deana	opry	merle	yearwood	nascar
dixie	wade	mccann	spaces	paisley
achy	reba	entertainer	stroud	diffie
breaky	kenny	cowboy	hayes	tillis

artist name (58)

location/institution (21)

instrument, role (1)

album/track title (11)

genre, style (8)

adjectives (0)

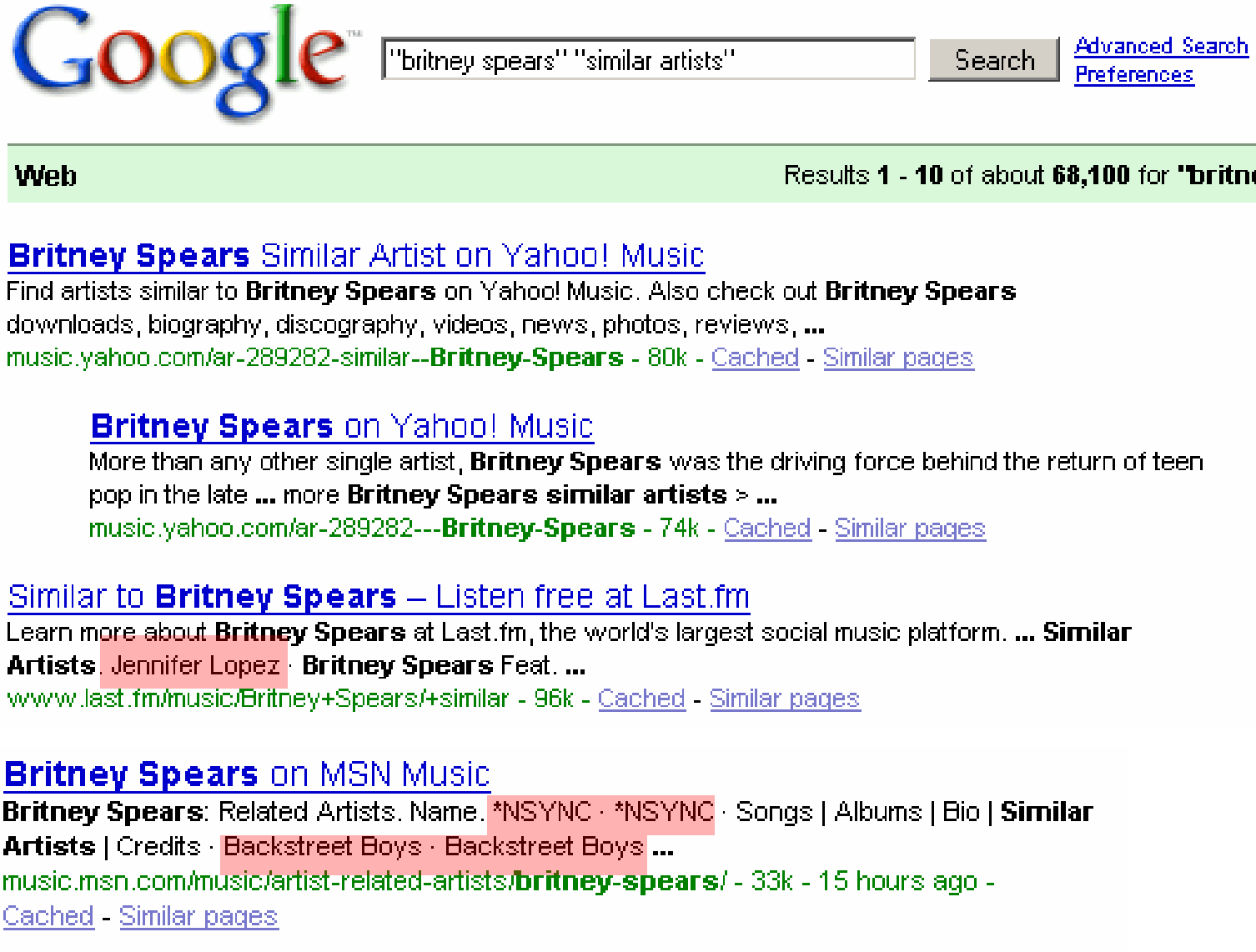
Simplified Approach

- Proper nouns (especially prototypical artist names) are very important for class.
- Modify queries

  “*artist name*” +“similar artists”

  “*artist name*” +“related artists”
- Parse directly Google result pages (results are contained in snippets)

Google Snippets



Simplified Approach - Results

- Classification Accuracy (avg. over 50-fold CV)

		c224a		*c103a*		*uspop2002*	
	results	SVM	NN	SVM	NN	SVM	NN
+music	50	95.69	93.90	65.00	73.00	89.75	87.25
+"similar artists"	10	89.69	79.69	56.66	61.66	85.25	77.00
	50	95.10	93.49	71.33	72.33	88.25	80.75
	100	95.69	93.49	68.33	72.33	87.75	78.25
+"related artists"	10	92.59	86.60	54.00	58.66	87.00	80.25
	50	94.69	91.59	65.66	71.66	96.32	86.66
	100	94.99	90.10	68.66	73.66	89.75	81.50

Conclusions

- No improvements through
  Query Optimization or Page Filtering
- Genre classification (with χ^2-test) heavily dependent on proper nouns; degrades to co-occurrence analysis
- Extensional Genre Definition
- Other Web-based MIR tasks more interesting