

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

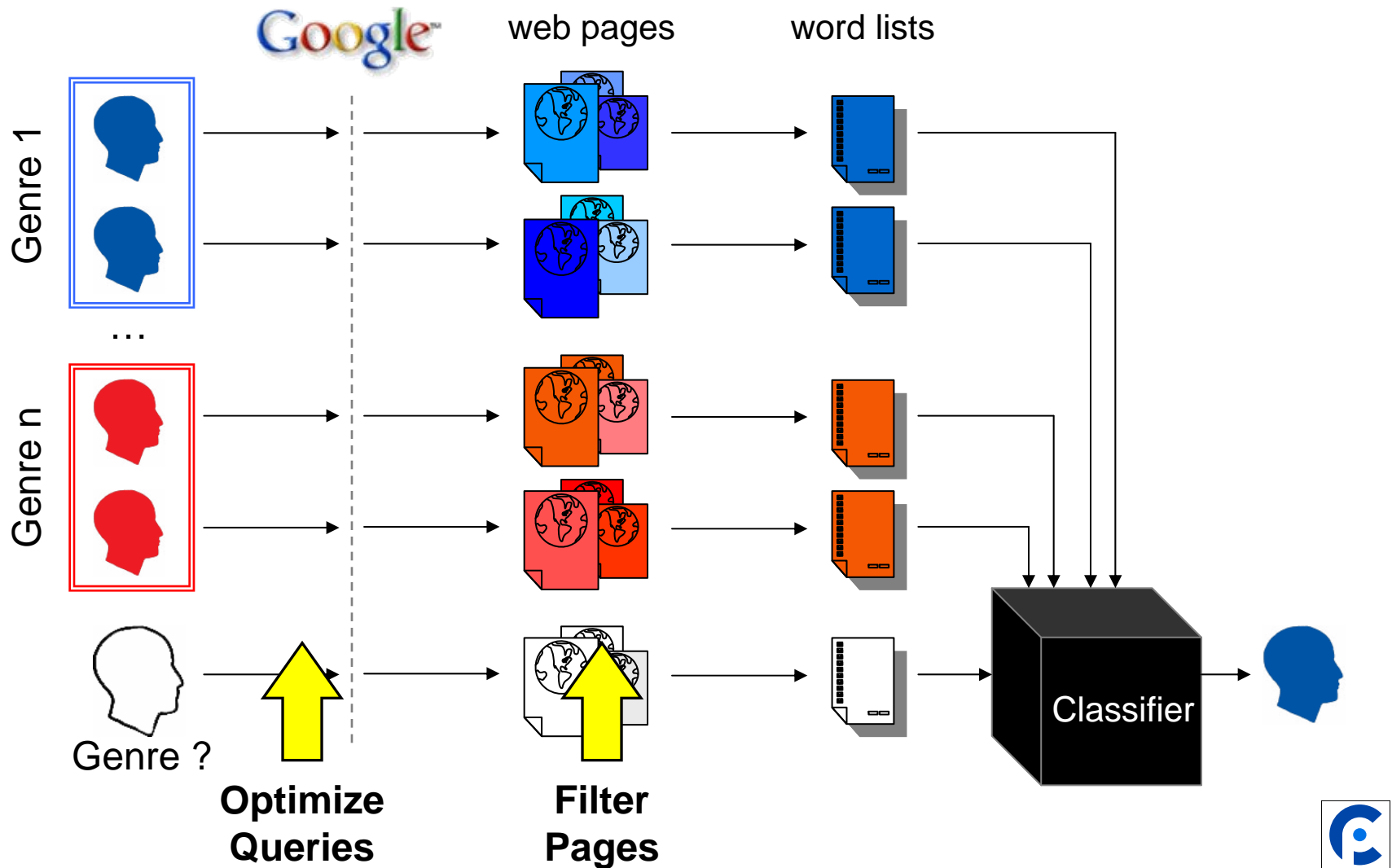


Introduction

- 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 |
|-------------------------|----------------|
| <i>+like -mp3</i> | 0.278 |
| <i>+like</i> | 0.277 |
| <i>+like -videos</i> | 0.272 |
| <i>+work -prices</i> | 0.271 |
| <i>+work -mp3</i> | 0.269 |
| <i>+work -services</i> | 0.268 |
| <i>+like -download</i> | 0.267 |
| <i>+like -tickets</i> | 0.265 |
| <i>+like -cart</i> | 0.262 |
| <i>+like -login</i> | 0.261 |
| <i>+like +time</i> | 0.261 |
| <i>+work</i> | 0.260 |
| <i>+like -prices</i> | 0.260 |
| <i>+work -format</i> | 0.259 |
| <i>+work -health</i> | 0.258 |
| <i>+like +people</i> | 0.257 |



Optimizing Queries - Results

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

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

| | | <i>c224a</i> | | | | <i>c103a</i> | | | | <i>uspop2002</i> | | | |
|---|------------|-------------------|-------|-----------------|-------|-------------------|-------|-----------------|-------|-------------------|-------|-----------------|-------|
| | | <i>unfiltered</i> | | <i>filtered</i> | | <i>unfiltered</i> | | <i>filtered</i> | | <i>unfiltered</i> | | <i>filtered</i> | |
| | <i>pg.</i> | SVM | NN | SVM | NN | SVM | NN | SVM | NN | SVM | NN | SVM | NN |
| <i>+music</i> | 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 |
| <i>+music</i> <i>+review</i> | 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 |
| | 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 |
| <i>+music</i> <i>+genre</i> <i>+style</i> | 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 |
| | 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 |
| | 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 |
| <i>+music</i> <i>+like</i> | 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 |
| | 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 | incentire | 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



[Advanced Search](#)
[Preferences](#)

Web

Results 1 - 10 of about 68,100 for "[britney spears](#)" "[similar art](#)

[Britney Spears Similar Artist on Yahoo! Music](#)

Find artists similar to **Britney Spears** on Yahoo! Music. Also check out **Britney Spears** downloads, biography, discography, videos, news, photos, reviews, ...

[music.yahoo.com/ar-289282-similar--Britney-Spears](#) - 80k - [Cached](#) - [Similar pages](#)

[Britney Spears on Yahoo! Music](#)

More than any other single artist, **Britney Spears** was the driving force behind the return of teen pop in the late ... more **Britney Spears similar artists** > ...

[music.yahoo.com/ar-289282---Britney-Spears](#) - 74k - [Cached](#) - [Similar pages](#)

[Similar to Britney Spears – Listen free at Last.fm](#)

Learn more about **Britney Spears** at Last.fm, the world's largest social music platform. ... **Similar Artists** · [Jennifer Lopez](#) · **Britney Spears** Feat. ...

[www.last.fm/music/Britney+Spears/+similar](#) - 96k - [Cached](#) - [Similar pages](#)

[Britney Spears on MSN Music](#)

Britney Spears: Related Artists. Name. *NSYNC · *NSYNC · Songs | Albums | Bio | **Similar Artists** | Credits · [Backstreet Boys](#) · [Backstreet Boys](#) ...

[music.msn.com/music/artist-related-artists/britney-spears/](#) - 33k - 15 hours ago - [Cached](#) - [Similar pages](#)

Simplified Approach - Results

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

| | | <i>c224a</i> | | <i>c103a</i> | | <i>uspop2002</i> | |
|----------------------------|----------------|--------------|-------|--------------|-------|------------------|-------|
| | <i>results</i> | SVM | NN | SVM | NN | SVM | NN |
| <i>+music</i> | 50 | 95.69 | 93.90 | 65.00 | 73.00 | 89.75 | 87.25 |
| <i>+ "similar artists"</i> | 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 |
| <i>+ "related artists"</i> | 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

