PAGE-LEVEL TEMPLATE DETECTION

Ravi Kumar

Yahoo! Research

Sunnyvale, CA

ravikumar@yahoo-inc.com

joint work with ...

Deepayan Chakrabarti, Yahoo! Research

Kunal Punera, UT Austin

- *Appeared in WWW 2007*

Talk outline

- Motivation
  - Potential applications
  - Related work
- Approach
  - Page-level template detection
  - Regularized isotonic regression
- Some experimental results

Templates: www.findbestcasinos.com

Templates: www.findbestcasinos.com

Advertisements

Look and feel

Links for navigation

Copyright message

Templates make surfing easy

- *One-click navigation within a site*
- *Place important links on multiple pages*
- *Common look and feel*
- *Surfers have been conditioned to look for sidebars and topbars*
- *Provide page update status*
- *Accommodate fine print*

But, good to detect them

- *Web ranking*
  - *Do not match query to text in templates*
- *Duplicate detection*
  - *Do not shingle text inside templates*
- *Summarization*
  - *Do not use text within templates for summary*
- *Indexing*
  - *Save space by indexing common part once*
- **...**

Two lines of attack

- *Site-level template detection*
- *Page-level template detection*

Site-level template detection

- *Templates = page-fragments that recur across several pages of a website.*
  - *eg, copyright, navigation links*
- *Page-fragment can be*
  - *HTML code (tags), visible text, DOM nodes (structure + text)*
- *Simple two-pass algorithm*
  - *Hash page-fragments and count occurrences*
  - *Mark templates in second pass*

Site-level template detection

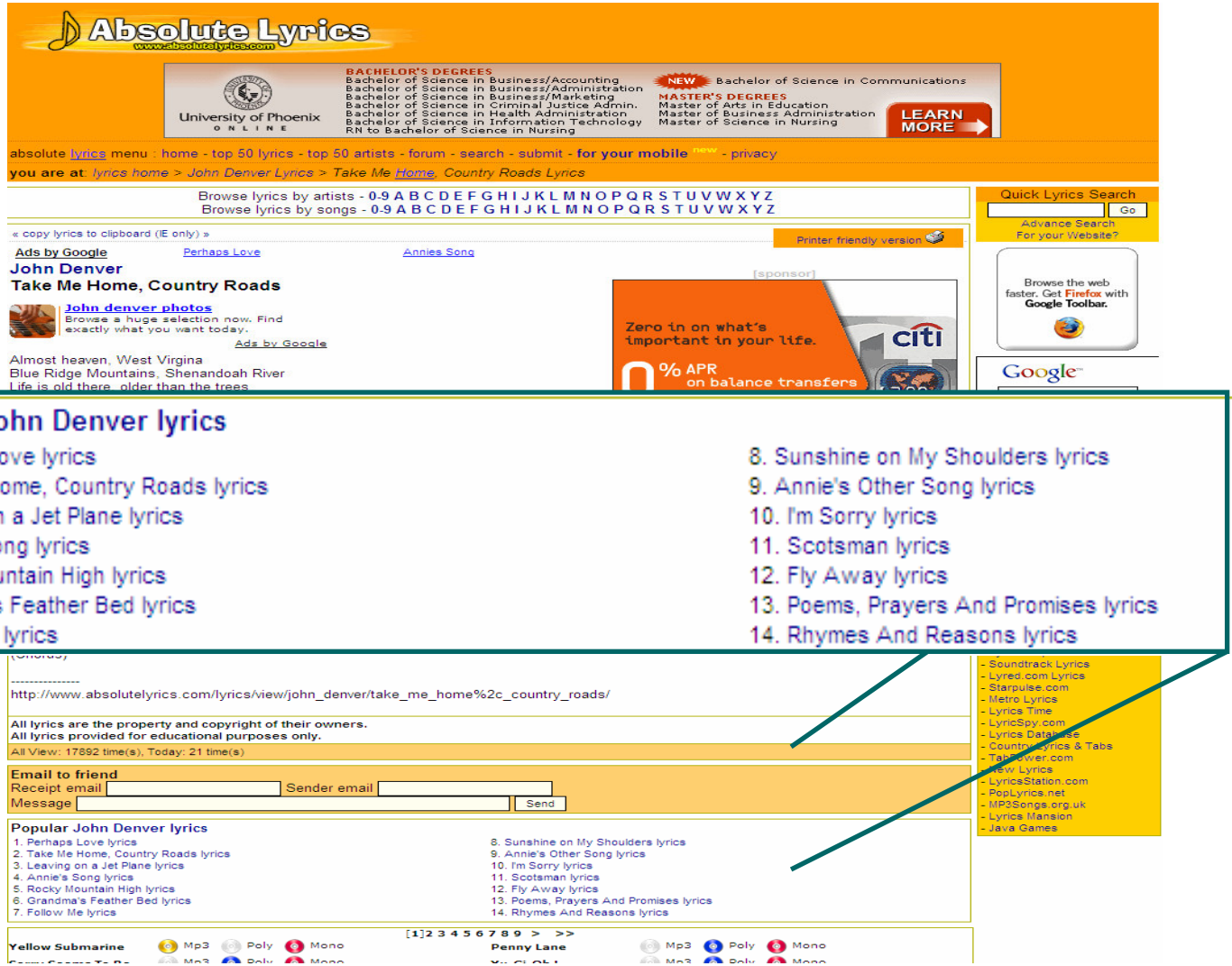
Advantages

- *No labeled training data needed*
- *Very high precision*

Issues

- *Inefficient when pages are not processed in site order*
  - *Eg, in a web crawler pipeline*
  - *Need to maintain hashes and counts for all sites*
  - *Marking site-level templates for new websites*
- *Not all templates are site-level in nature*
  - *Low recall*

A non-site level template



Features of site-level detection

Advantages:

- No labeled training data needed
- Very high precision

Issues:

- Inefficient when pages are not processed in site order
  - Eg: in a web crawl pipeline
  - Need to maintain hashes and count
  - Marking site-level templates for new websites
- Not all templates are site-level in nature
  - Low recall

Only use page-level information

Learn a general model for templates

Page-level model-based detection

- *Problem: Find templates*
  - *Using only information local to a webpage*
  - *Detect all templates: not just site-level*
  - *No manually labeled training data*
- *Our approach*
  - *Obtain training data via site-level approach*
  - *Learn a classification model for "templateness"*
    - *For each internal DOM node*
  - *Enforce a global monotonicity property of "templateness"*

Automatically labeling data

- Use site-level approach
  - 3,000 website (200 webpages per site)
  - Obtained ~1M labeled DOM nodes
- Labeled data has a bias
  - Some template DOM nodes labeled as non-templates
  - False negatives are noise
- Extract general structural and content cues from the DOM nodes
  - Generalize over the site-level training data

Cues used implicitly by humans

BGColor

Aspect Ratio

Placement on screen

Average Sentence Size

Fraction of text outside anchors

Link Density

"Templateness" classifier

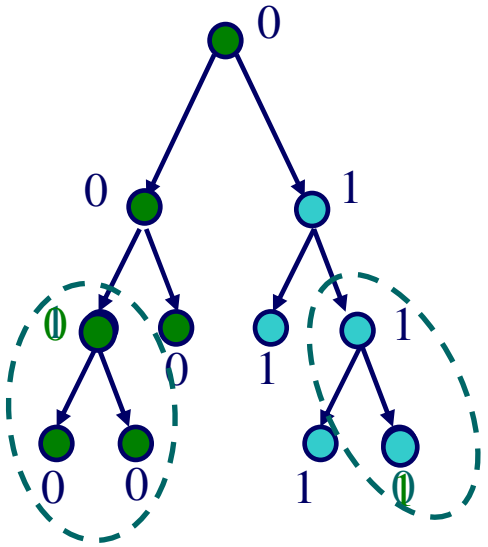
- Extract features of DOM nodes from cues
- Learn weights for these features
  - 2-class problem
  - Logistic regression classifier
  - Simple classifier, avoids fitting noise in the data
  - "Templateness" = probability of belonging to template class
  - Separate classifiers learned for nodes of different sizes
- Each node in the DOM tree is classified
  - Past work classify segments of web pages
  - Segmentation might mix template and non-template content

Monotonicity of "templateness"

A DOM node is a template
if and only if
all its children are templates

High-level idea

A node in the DOM tree is a template if and only if all its children are templates



- *Each node is classified in isolation*
  - *Classifier scores needn't be monotonic*
  - *Classifier might misclassify nodes*
- *Post-processing "templateness" scores*
  - *Enforces monotonicity*
  - *Corrects misclassifications by smoothing*
- *"Templateness" scores are real numbers x(i)*

Regularized isotonic regression

Given raw scores x(1),…,x(n) for nodes in a tree, find smoothed scores y(1),…,y(n) such that i = parent(j) implies y(i) ≤ y(j) and minimize

$$\sum \alpha_i \, | \, x(i) - y(i) \, | + \beta_i \, | \, y(i) - \max_{i = parent(j)} y(j) \, |$$

The algorithm

- *Lemma: For L_1 distance, each y_i must equal some x_j*
- *The optimal solution found using a dynamic program*
  - *Complexity: $O(n^2 \log n)$, n = tree size*
  - *Equals complexity of algorithms for non-regularized isotonic regression ($\beta = 0$)*
  - *On a Pentium 4, 3GHz, 512MB running FreeBSD: around 60ms for cnn.com (n=292)*

The complete picture

- *Obtain training data via site-level approach*
- *Classification model for "templateness"*
  - ✓ *Designed features for DOM nodes*
  - ✓ *Each DOM node labeled by a logistic regression classifier*
- *Enforce a global monotonicity property of "templateness"*
  - ✓ *Formulate it as regularized isotonic regression over trees*
  - ✓ *Optimal solution via dynamic program*

Accuracy: f-measure

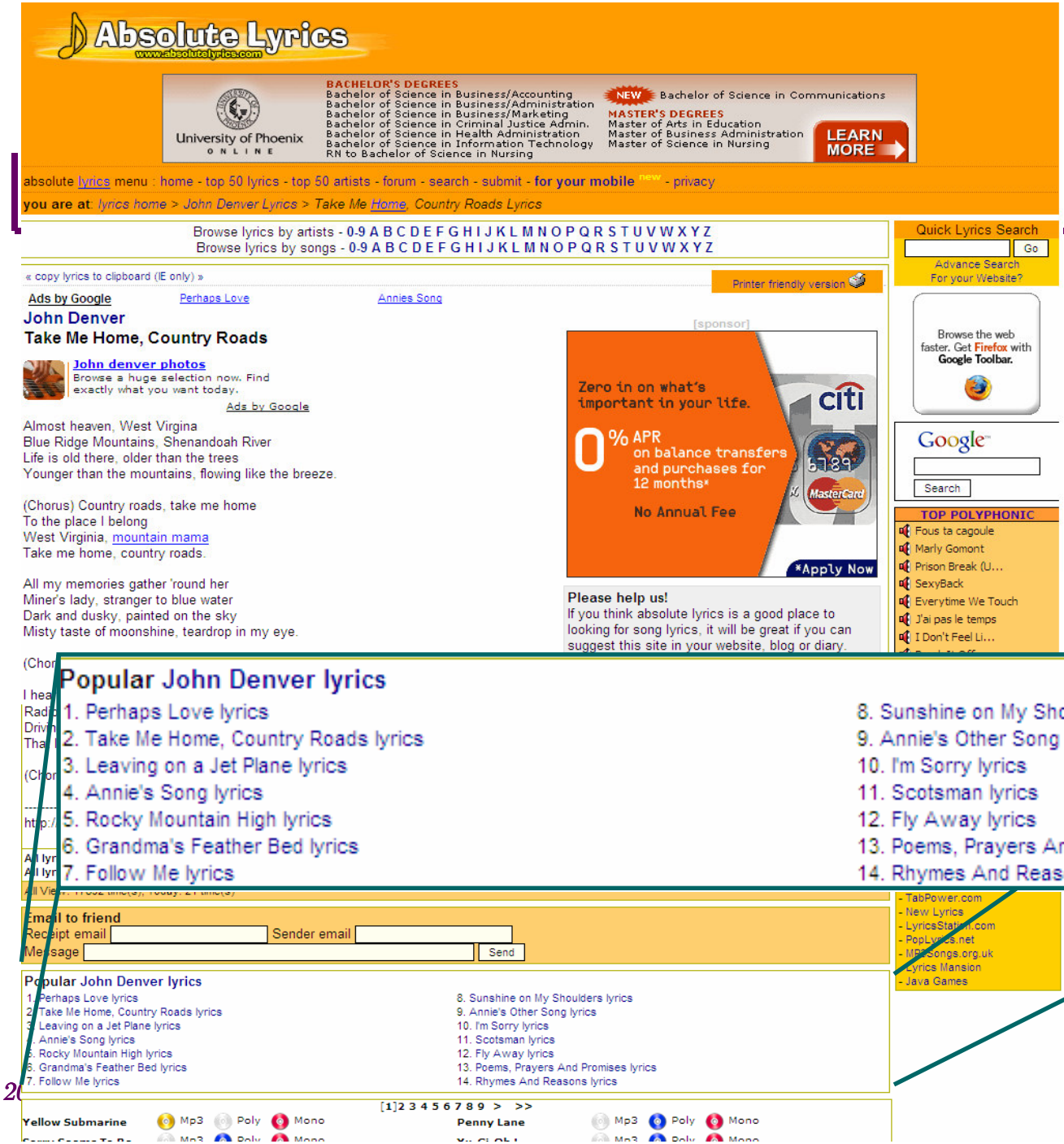
	Data Set	Basic	Smooth
	Text	0.56	0.60
Common	AT	0.65	0.71
	Links	0.69	0.73
	Text	0.63	0.66
Random	AT	0.71	0.93
	Links	0.75	0.77

- Data: manually classified DOM nodes in webpages
- Results:
  - Page-level approach works very well
  - Smoothing improves classification accuracy

Application: Duplicate detection

	Total pairs	Page-level	Site-level	Full text
Dups	1711	1299 (76%)	730 (42.7%)	529 (30.9%)
Non-dups	2058	1885 (91.6%)	1712 (83.2%)	1781 (86.5%)

- Data: 2359 pages from 3 lyrics websites
  - 1711 duplicate pairs (same song, different websites)
  - 2058 non-duplicate pairs (different songs, same website)
- Errors occur when shingles hit template conten
- PageLevel detects more templates than SiteLevel


Absolute Lyrics
www.absolutelyrics.com
University of Phoenix ONLINE
BACHELOR'S DEGREES
Bachelor of Science in Business/Accounting
Bachelor of Science in Business/Administration
Bachelor of Science in Business/Marketing
Bachelor of Science in Criminal Justice Admin.
Bachelor of Science in Health Administration
Bachelor of Science in Information Technology
RN to Bachelor of Science in Nursing
NEW Bachelor of Science in Communications
MASTER'S DEGREES
Master of Arts in Education
Master of Business Administration
Master of Science in Nursing
LEARN MORE
absolute lyrics menu : home - top 50 lyrics - top 50 artists - forum - search - submit - for your mobile new - privacy
you are at: lyrics home > John Denver Lyrics > Take Me Home, Country Roads Lyrics
Browse lyrics by artists - 0-9 A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
Browse lyrics by songs - 0-9 A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
« copy lyrics to clipboard (IE only) »
Printer friendly version
Ads by Google
Perhaps Love
Annies Song
John Denver
Take Me Home, Country Roads
John denver photos
Browse a huge selection now. Find exactly what you want today.
Ads by Google
Almost heaven, West Virgina
Blue Ridge Mountains, Shenandoah River
Life is old there, older than the trees
Younger than the mountains, flowing like the breeze.
(Chorus) Country roads, take me home
To the place I belong
West Virginia, mountain mama
Take me home, country roads.
All my memories gather 'round her
Miner's lady, stranger to blue water
Dark and dusky, painted on the sky
Misty taste of moonshine, teardrop in my eye.
[sponsor]
Zero in on what's important in your life.
0% APR on balance transfers and purchases for 12 months*
No Annual Fee
citi
MasterCard
*Apply Now
Please help us!
If you think absolute lyrics is a good place to looking for song lyrics, it will be great if you can suggest this site in your website, blog or diary.
Quick Lyrics Search
Go
Advance Search
For your Website?
Browse the web faster. Get Firefox with Google Toolbar.
Google
Search
TOP POLYPHONIC
Fous ta cagoule
Marly Gomont
Prison Break (U...
SexyBack
Everytime We Touch
J'ai pas le temps
I Don't Feel Li...
Popular John Denver lyrics
1. Perhaps Love lyrics
2. Take Me Home, Country Roads lyrics
3. Leaving on a Jet Plane lyrics
4. Annie's Song lyrics
5. Rocky Mountain High lyrics
6. Grandma's Feather Bed lyrics
7. Follow Me lyrics
8. Sunshine on My Shoulders lyrics
9. Annie's Other Song lyrics
10. I'm Sorry lyrics
11. Scotsman lyrics
12. Fly Away lyrics
13. Poems, Prayers And Promises lyrics
14. Rhymes And Reasons lyrics
Email to friend
Receipt email
Sender email
Message
Send
- TabPower.com
- New Lyrics
- LyricsStation.com
- PopLyrics.net
- MP3Songs.org.uk
- Lyrics Mansion
- Java Games
[1] 2 3 4 5 6 7 8 9 > >>
Yellow Submarine
Penny Lane
Mp3
Poly
Mono

General paradigm

- *Identify a global property to be satisfied by a function on a structure*
- *Operate on local structures*
- *Smooth by enforcing the global property*

- *Eg, unimodality and class membership in a hierarchy*
- *Recently applied to some bio data*

Conclusions

- *Page-level model-based template detection*
- *Used no manually labeled training data*
- *"Templateness" monotonicity*
- *Regularized isotonic regression*
  - *might be of independent interest*
- *Page-level generalizes over the site-level data*

Thank you!

ravikumar@yahoo-inc.com