



PAGE-LEVEL TEMPLATES 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

BEST CASINOS [Articles](#) - [Tell a Friend](#) - [Bookmark](#)

 **UP TO \$110** **CLICK HERE!**

[Home](#) | [Best Casinos](#) | [Slot Machines](#) | [Video Poker](#) | [Table Games](#)

Casino Reviews

Reviews of the Top Online Casinos

Specials!
Get a 100% match up to \$75 at Spin Palace!
Get a 100% free bonus up to \$200!



Monaco Gold Casino ★★★★★
Monaco Gold Casino is designed to pass on the majestic and prestigious experience of a land based casino. Combining the magnificent style of Monte Carlo and the art of gaming, Monaco Gold Casino guarantees an unforgettable online gaming experience and invites you to sample The Royal Side of Gaming! [Online Review](#)
[Visit Monaco Gold](#)



All Poker Casino ★★
All Poker Casino offers exceptional Video Poker games, each with leading-edge graphics, sound and play features. Plus, you can enter to win great prizes in our weekly tournaments, contests and promotions! [Online Review](#)
[Visit All Poker](#)

[Newsletter](#) - [Webmasters](#) - [Partners](#) - [Disclaimer](#) - [Privacy Policy](#) - [Contact Us](#)

Baccarat	Download Baccarat	Online Casino Best
Best Online Casinos	Online Machines Best Slot	Black Jack
Game Black Jack Casino	Black Jack Casino Games	Jack Download Black
Black Jack Game	Blackjack	Game Blackjack Casino

[Find Credit Cards](#) | [How to Lookup People](#) | [Free Credit History Check](#) | [Insurance Broker](#) | [Play Blackjack Online](#)

Copyright © 2005 FindBestCasinos.com
Your guide to the **Best Casinos**

Templates: www.findbestcasinos.com

BEST CASINOS [Articles](#) - [Tell a Friend](#) - [Bookmark](#)



UP TO \$110

CLICK HERE!

[Home](#) | [Best Casinos](#) | [Slot Machines](#) | [Video Poker](#) | [Table Games](#)

Specials!
Get a 100% match
up to \$75 at Spin
Palace!

Get a 100% free
bonus up to \$200!

*Look and
feel*

Advertisements

*Links for
navigation*

[Newsletter](#) - [Webmasters](#) - [Partners](#) - [Disclaimer](#) - [Privacy Policy](#) - [Contact Us](#)

Baccarat	Download Baccarat	Online Casino Best
Best Online Casinos	Online Machines Best Slot	Black Jack
Game Black Jack Casino	Black Jack Casino Games	Jack Download Black
Black Jack Game	Blackjack	Game Blackjack Casino

[Find Credit Cards](#) | [How to Lookup People](#) | [Free Credit History Check](#) | [Insurance Broker](#) | [Play Blackjack Online](#)

*Copyright
message*

June 19, 2007

Copyright © 2005 FindBestCasinos.com
Your guide to the **Best Casinos**



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

Absolute Lyrics
www.absolute-lyrics.com

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 - 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

John Denver
Take Me Home, Country Roads

John Denver photos
 Browse a huge selection now. Find exactly what you want today.

Almost heaven, West Virginia
 Blue Ridge Mountains, Shenandoah River
 Life is old there, older than the trees

Zero in on what's important in your life. **citi**
 % APR on balance transfers

Quick Lyrics Search
 Advance Search For your Website?
 Browse the web faster. Get Firefox with Google Toolbar.
 Google

Popular John Denver lyrics

- | | |
|---------------------------------------|--|
| 1. Perhaps Love lyrics | 8. Sunshine on My Shoulders lyrics |
| 2. Take Me Home, Country Roads lyrics | 9. Annie's Other Song lyrics |
| 3. Leaving on a Jet Plane lyrics | 10. I'm Sorry lyrics |
| 4. Annie's Song lyrics | 11. Scotsman lyrics |
| 5. Rocky Mountain High lyrics | 12. Fly Away lyrics |
| 6. Grandma's Feather Bed lyrics | 13. Poems, Prayers And Promises lyrics |
| 7. Follow Me lyrics | 14. Rhymes And Reasons lyrics |

http://www.absolute-lyrics.com/lyrics/view/john_denver/take_me_home%2c_country_roads/

All lyrics are the property and copyright of their owners.
 All lyrics provided for educational purposes only.

All View: 17892 time(s), Today: 21 time(s)

Email to friend
 Receipt email Sender email
 Message

Popular John Denver lyrics

1. Perhaps Love lyrics	8. Sunshine on My Shoulders lyrics
2. Take Me Home, Country Roads lyrics	9. Annie's Other Song lyrics
3. Leaving on a Jet Plane lyrics	10. I'm Sorry lyrics
4. Annie's Song lyrics	11. Scotsman lyrics
5. Rocky Mountain High lyrics	12. Fly Away lyrics
6. Grandma's Feather Bed lyrics	13. Poems, Prayers And Promises lyrics
7. Follow Me lyrics	14. Rhymes And Reasons lyrics

[1] 2 3 4 5 6 7 8 9 > >>

Yellow Submarine
Penny Lane

- Soundtrack Lyrics
- Lyred.com Lyrics
- Starpuisee.com
- Metro Lyrics
- Lyrics Time
- LyricsSpy.com
- Lyrics Database
- Country Lyrics & Tabs
- TagPower.com
- New Lyrics
- LyricsStation.com
- PopLyrics.net
- MP3Songs.org.uk
- Lyrics Mansion
- Java Games

June 19, 2007

11

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 counts
 - 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”*

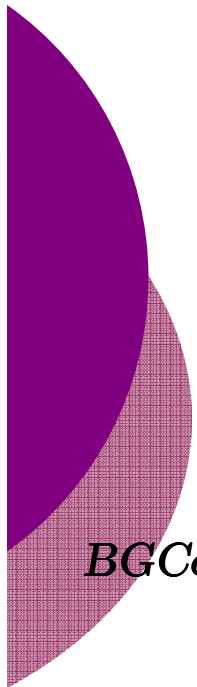


Automatically labeling data

- *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

The screenshot shows a website for 'BEST CASINOS'. At the top, there is a navigation bar with 'Articles - Tell a Friend - Bookmark'. Below that is a banner for 'Silver DOLLAR CASINO' with 'UP TO \$110' and a large 'CLICK HERE!' button. A secondary navigation bar includes 'Home | Best Casinos | Slot Machines | Video Poker | Table Games'. The main content area is titled 'Casino Reviews' and 'Reviews of the Top Online Casinos'. It features two review cards: 'Monaco Gold Casino' with a 5-star rating and 'All Poker Casino' with a 2-star rating. A red sidebar on the left contains 'Specials!' with promotional text. The footer contains a list of links: 'Newsletter - Webmasters - Partners - Disclaimer - Privacy Policy - Contact Us', a grid of game-related links, and a row of utility links: 'Find Credit Cards | How to Lookup People | Free Credit History Check | Insurance Broker | Play Blackjack Online'. Copyright information for 'FindBestCasinos.com' is at the bottom.



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”

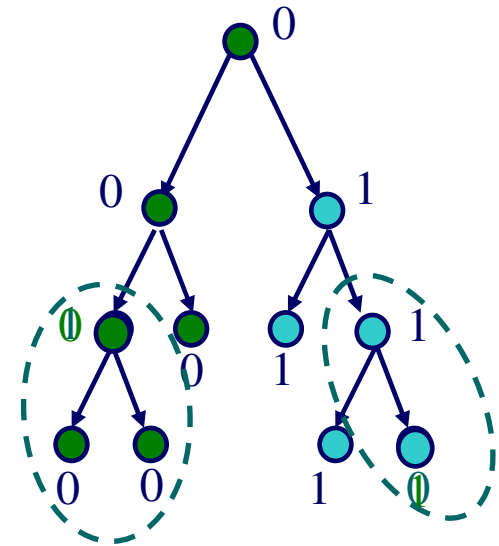
The screenshot shows a website with a blue header containing the text "BEST CASINOS" and "Articles - Tell a Friend - Bookmark". Below the header is a yellow banner with a "Silver DOLLAR CASINO" logo, "UP TO \$110", and a large "CLICK HERE!" button. A navigation bar below the banner lists "Home | Best Casinos | Slot Machines | Video Poker | Table Games". The main content area is titled "Casino Reviews" and "Reviews of the Top Online Casinos". It features two review cards: "Monaco Gold Casino" (4 stars) and "All Poker Casino" (2 stars). A red sidebar on the left contains "Specials!" with offers like "Get a 100% match up to \$75 at Spin Palace!" and "Get a 100% free bonus up to \$200!". The footer, enclosed in a red border, contains links for "Newsletter - Webmasters - Partners - Disclaimer - Privacy Policy - Contact Us", a grid of game-specific links (Baccarat, Black Jack, etc.), "Find Credit Cards | How to Lookup People | Free Credit History Check | Insurance Broker | Play Blackjack Online", and copyright information for "FindBestCasinos.com".

*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 $x(1), \dots, x(n)$ for nodes in a tree, find smoothed scores $y(1), \dots, y(n)$ such that $i = \text{parent}(j)$ implies $y(i) \leq y(j)$ and minimize

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



The algorithm

- *Lemma: For L_1 distance, each y_i must equal some x_j*
- *The optimal solution found using a dynamic program*
 - *Complexity: $O(n^2 \log n)$, $n = \text{tree size}$*
 - *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

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

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



Application: Duplicate detection

	<i>Total pairs</i>	<i>Page-level</i>	<i>Site-level</i>	<i>Full text</i>
<i>Dups</i>	1711	1299 (76%)	730 (42.7%)	529 (30.9%)
<i>Non-dups</i>	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 content*
- *PageLevel detects more templates than SiteLevel*

<p>UNIVERSITY OF PHOENIX ONLINE</p>	<p>BACHELOR'S DEGREES</p> <ul style="list-style-type: none"> 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 	<p>NEW Bachelor of Science in Communications</p> <p>MASTER'S DEGREES</p> <ul style="list-style-type: none"> Master of Arts in Education Master of Business Administration Master of Science in Nursing 	<p>LEARN MORE →</p>

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](#) [Annie's 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 Virginia
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

***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

Advance Search For your Website?

Browse the web faster. Get Firefox with Google Toolbar.

Google™

- 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 | 8. Sunshine on My Shoulders lyrics |
| 2. Take Me Home, Country Roads lyrics | 9. Annie's Other Song lyrics |
| 3. Leaving on a Jet Plane lyrics | 10. I'm Sorry lyrics |
| 4. Annie's Song lyrics | 11. Scotsman lyrics |
| 5. Rocky Mountain High lyrics | 12. Fly Away lyrics |
| 6. Grandma's Feather Bed lyrics | 13. Poems, Prayers And Promises lyrics |
| 7. Follow Me lyrics | 14. Rhymes And Reasons lyrics |

email to friend

Receipt email Sender email

Message

- Popular John Denver lyrics**
- | | |
|---------------------------------------|--|
| 1. Perhaps Love lyrics | 8. Sunshine on My Shoulders lyrics |
| 2. Take Me Home, Country Roads lyrics | 9. Annie's Other Song lyrics |
| 3. Leaving on a Jet Plane lyrics | 10. I'm Sorry lyrics |
| 4. Annie's Song lyrics | 11. Scotsman lyrics |
| 5. Rocky Mountain High lyrics | 12. Fly Away lyrics |
| 6. Grandma's Feather Bed lyrics | 13. Poems, Prayers And Promises lyrics |
| 7. Follow Me lyrics | 14. Rhymes And Reasons lyrics |

- 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	Mp3	Poly	Mono	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