PUBLISH ON DEMAND

TweakCorps: Re-Targeting Existing Webpages for Diverse Devices and Users

FALL 2013

UPDATE

mediaX
STANFORD UNIVERSITY

mediaX
at Stanford University

mediaX connects businesses with Stanford University's world-renowned faculty to study new ways for people and technology to intersect.

We are the industry-affiliate program to Stanford's H-STAR Institute. We help our members explore how the thoughtful use of technology can impact a range of fields, from entertainment to learning to commerce. Together, we're researching innovative ways for people to collaborate, communicate and interact with the information, products, and industries of tomorrow.

TweakCorps

Re-targeting Existing Webpages for Diverse Devices and Users

Research Team: Scott Klemmer, Associate Professor, Computer Science; Maxine Lim, Graduate Student, Computer Science.

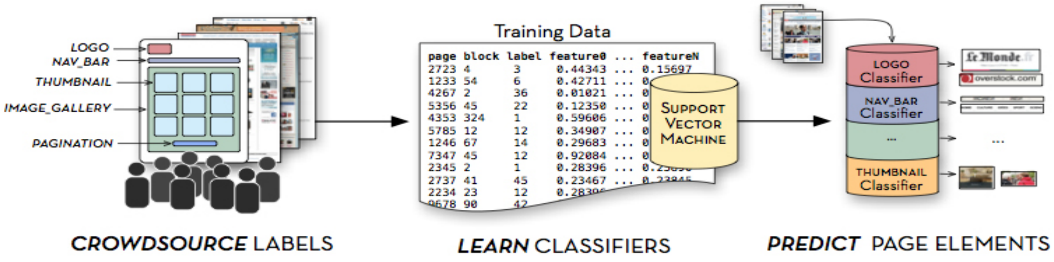


Figure 1: The pipeline for learning structural semantic classifiers for the Web. First, a large set of labeled page elements are collected from online workers. Next, these labels are used to train a set of regularized support vector classification SVMs. These classifiers are then used to identify semantic elements in new pages.

Project Overview

The research team has developed a machine-learning algorithm that can label web pages. The algorithm takes the structure of pages, conducts design mining and provides the opportunity to learn about design elements from multiple sources.

While designers would prefer to design for as few devices as possible, the incentive of hardware manufacturers is to offer a "new market." The team worked on this learning algorithm to solve the tension between hardware manufactures, who seek to develop new systems, and designers, who prefer to not make too many changes to their web design work when new devices are developed.

The team's algorithm enables a new kind of design-based machine learning, that can stream structured visual descriptors for page elements from a central repository. The algorithm also allows data to be collected and integrated with the repository for supervised learning applications like crowd-sourcing.

The algorithm takes the structure of the web pages, and analyzes them. It is composed of five integrated components:

- A web crawler
- A proxy server
- The data store
- The post-process
- The API

For this project, the team focused on one popular class of semantic identifiers: Those concerned with structure – or information architecture – of a page. The team explored a different tactic for adding structural semantics to web pages; with accurate learning classifiers, pages can be semantified automatically, in a post-hoc fashion, decoupled from the design and authoring process. To this end, the team presents a classification method based on support vector machines (i.e., a supervised learning model with algorithms that analyze data and recognize patterns), trained on a large collection of human-labeled page elements and employing a feature space comprised of visual, structural, and rendered-time page properties (Lim, Kumar, Torres, Talton, Satyanarayan, Klemmer, 2013) (see Figure 1).

The team took a crowd-sourced approach and recruited 400 participants on Amazon's Mechanical Turk, who collected more than 21,000 semantic labels over a corpus of over 1,400 web pages. The team used labels to determine the set of classifier and provide training data for machine learning (Lim, Kumar, Torres, Talton, Satyanarayan, Klemmer, 2013). Every participant applied semantic labels to at least 10 elements on each of five pages. The pages used in the study were drawn from the Webzeitgeist design repository, which provides visual segmentations and page features for more than 100,000 webpages. Webzeitgeist was a platform developed by the team to help users understand design demographics, automate design curation, and support new data-driven design interactions (See: http://vis.stanford.edu/papers/webzeitgeist)

To more thoroughly understand how labels relate to one another, the team created a co-currence matrix for the 85 most-frequent labels, each of which was used 20 or more times. To evaluate the feasibility of learning structural semantics from data, the team trained binary SVM classifiers for the study's 40 most frequent labels.

Challenges

- Webpages have different elements and are made by different people. There is no consistency in data. This raises the question: How can you deal with unruly data?
- On the web, the main challenge is building systems that can deal with messy, badly formatted data.
- Designs have more and different types of structures.

Next Steps

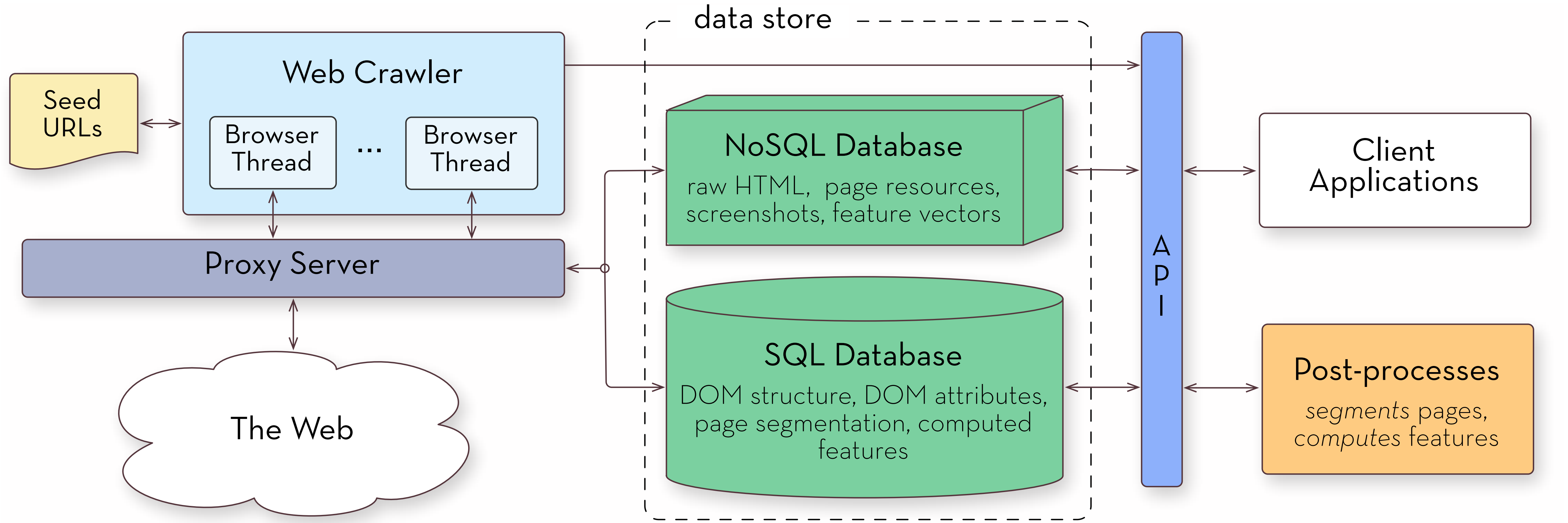
- Maxine Lin will submit the "Learning Structural Semantics for the Web" paper and will present a live demo of the prototype's search capability.
- The team would also like to get a preliminary version of a responsive design (short-term horizon).
- Continuing work that was been underway for 2 1/2 years, the team plans to improve accuracy of the classifiers and have search on Webzeitgeist ready by the summer.

WZG & Structural Semantics

A Few Details Explained

Obtaining Structural Data

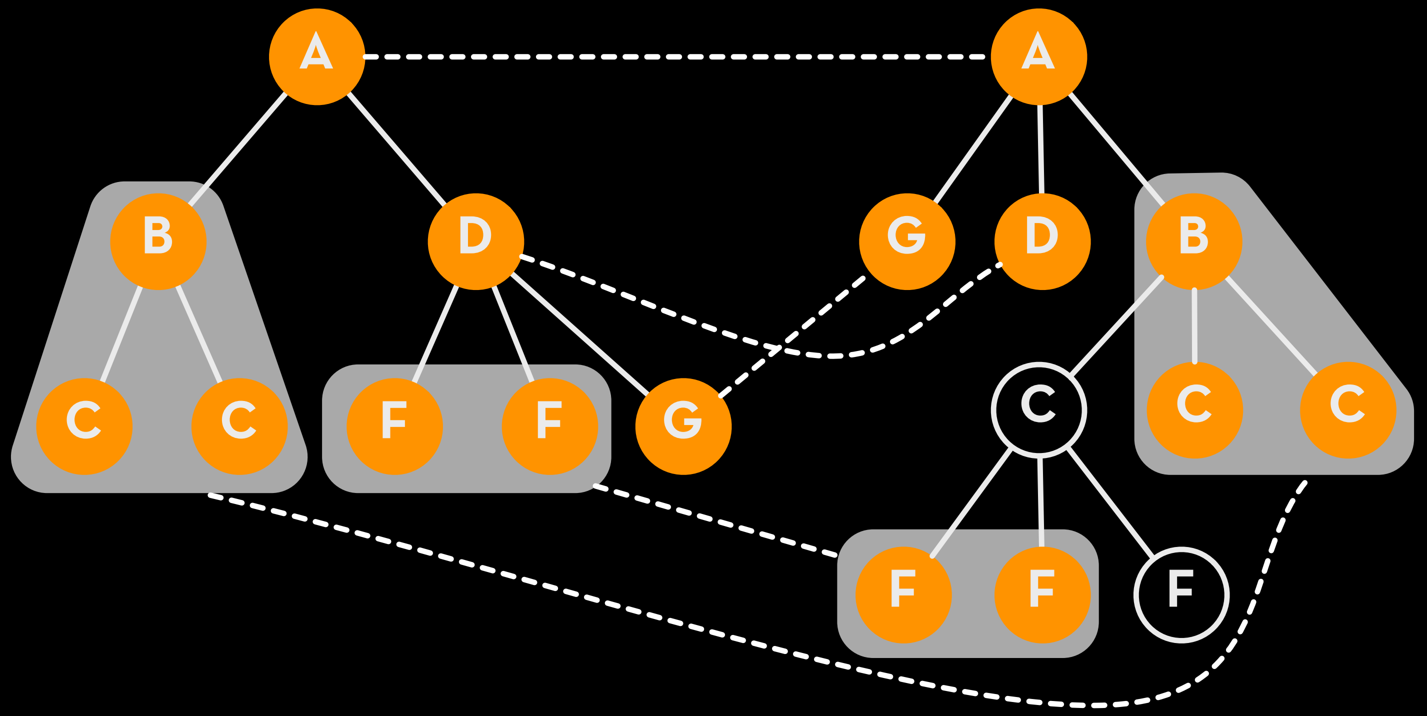
Crawler loads page and computes DOM.



data store
Seed URLs
Web Crawler
Browser Thread
...
Browser Thread
Proxy Server
The Web
NoSQL Database
raw HTML, page resources, screenshots, feature vectors
SQL Database
DOM structure, DOM attributes, page segmentation, computed features
API
Client Applications
Post-processes
segments pages, *computes* features

Obtaining Structural Data

Use flexible tree-matching algorithm to discard nodes that don't contribute to visual appearance.



relaxes rigid ancestry constraint

flexible semantic, sibling, and ancestry cost model

Obtaining Structural Data

Features are computed and stored in the database.

POSITION

[*absolute, fractional*] **X position**
[*absolute, fractional*] **Y position**
percent overlap with [*left, top*] of page

CONTENT

number of [**images, links, words**]

VISION

GIST features
[*average, most frequent*] **RGB color**
[*number, percent*] **edge pixels**

DIMENSION

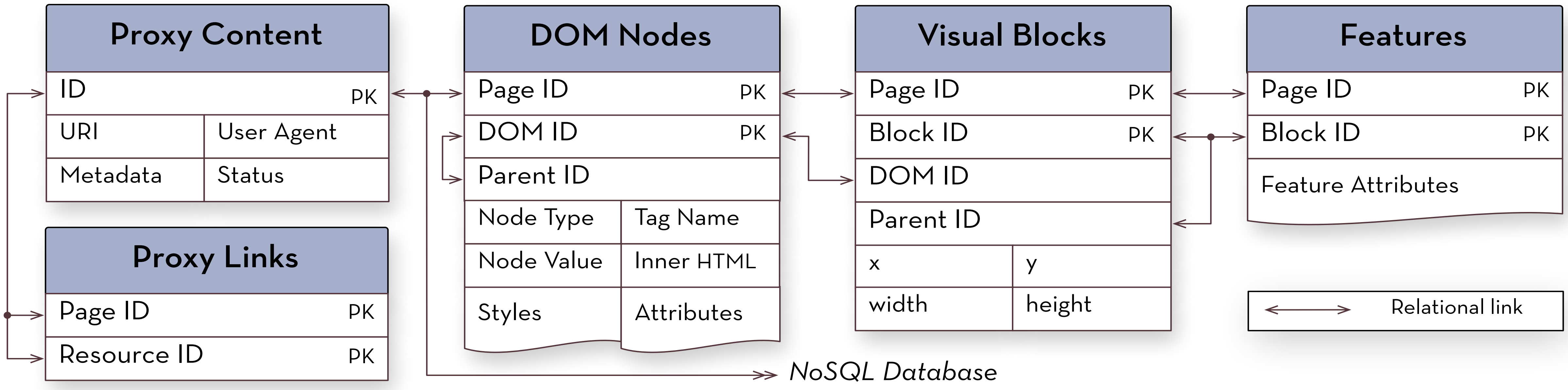
area, height, width, aspect ratio
fractional **area** w.r.t. [*parent, page*]
fractional **height** w.r.t. [*parent, page*]
fraction **width** w.r.t. [*parent, page*]

STRUCTURE

number of [**children, siblings**]
[*absolute, fractional*] **sibling order**
[*absolute, fractional*] **tree level**

Accessing the Data

Through our API, applications can access the vector that enumerates properties for each element.



Proxy Content

ID	PK
URI	User Agent
Metadata	Status

Proxy Links

Page ID	PK
Resource ID	PK

DOM Nodes

Page ID	PK
DOM ID	PK
Parent ID	
Node Type	Tag Name
Node Value	Inner HTML
Styles	Attributes

Visual Blocks

Page ID	PK
Block ID	PK
DOM ID	
Parent ID	
x	y
width	height

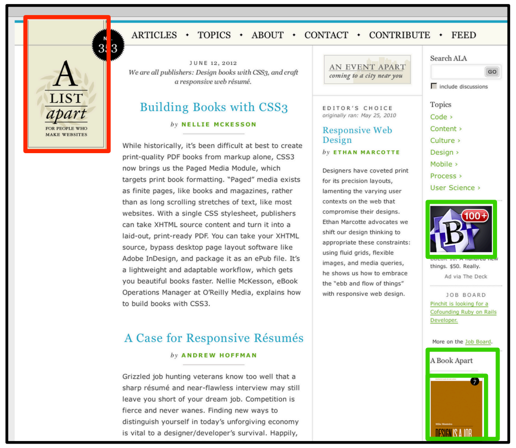
Features

Page ID	PK
Block ID	PK
Feature Attributes	

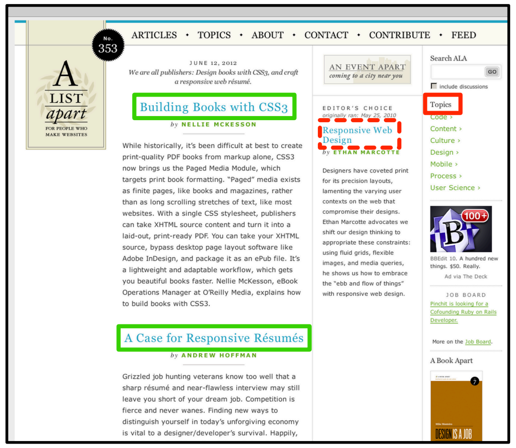
Training Structural Classifiers

Using WZG features, we trained classifiers for semantic elements.

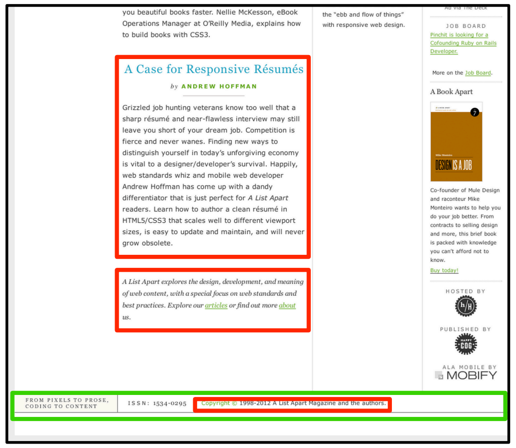
ADVERTISEMENT


ARTICLES • TOPICS • ABOUT • CONTACT • CONTRIBUTE • FEED
A LIST apart
Building Books with CSS3
A Case for Responsive Résumés

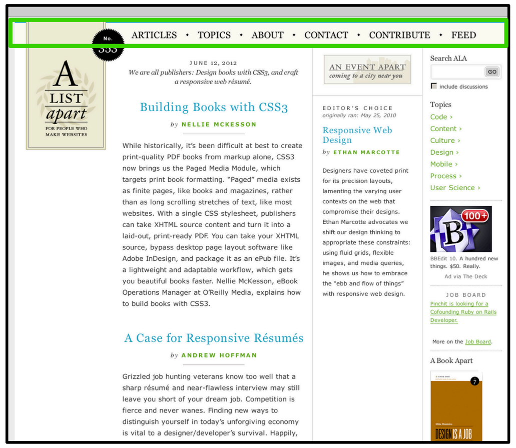
ARTICLE TITLE


ARTICLES • TOPICS • ABOUT • CONTACT • CONTRIBUTE • FEED
Building Books with CSS3
Responsive Web Design
A Case for Responsive Résumés

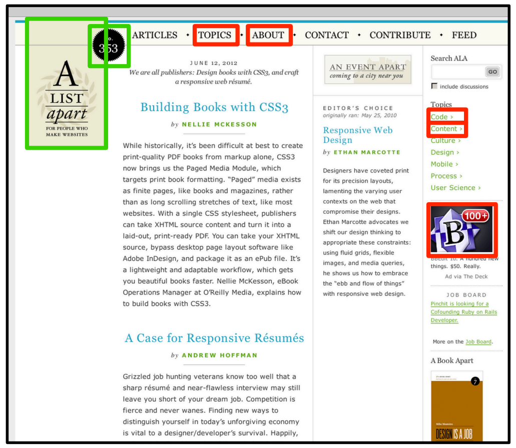
FOOTER


A Case for Responsive Résumés
FROM PIXELS TO PROSE, CODING TO CONTENT
ISSN: 1534-0295
Copyright © 1998-2012 A List Apart Magazine and the authors.

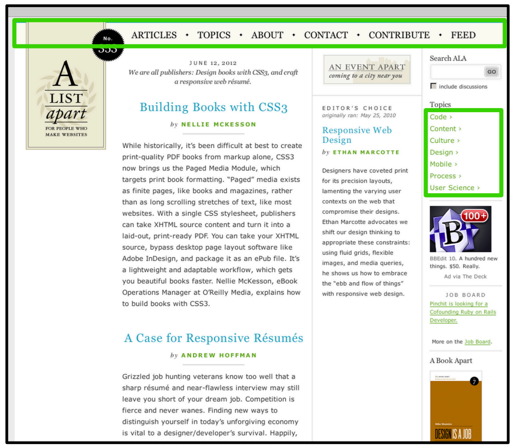
HEADER


ARTICLES • TOPICS • ABOUT • CONTACT • CONTRIBUTE • FEED
Building Books with CSS3
A Case for Responsive Résumés

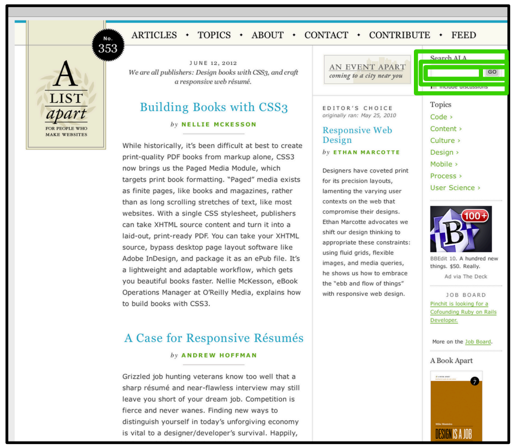
LOGO



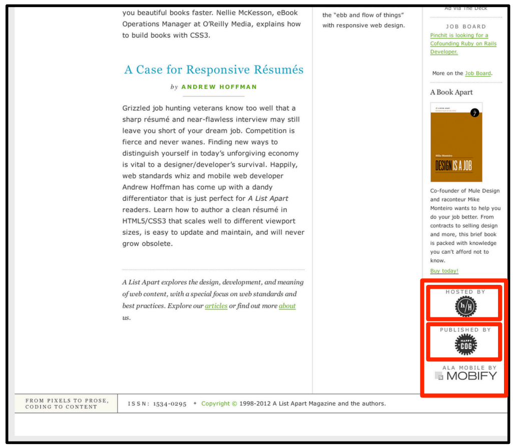
NAVIGATION BAR



SEARCH

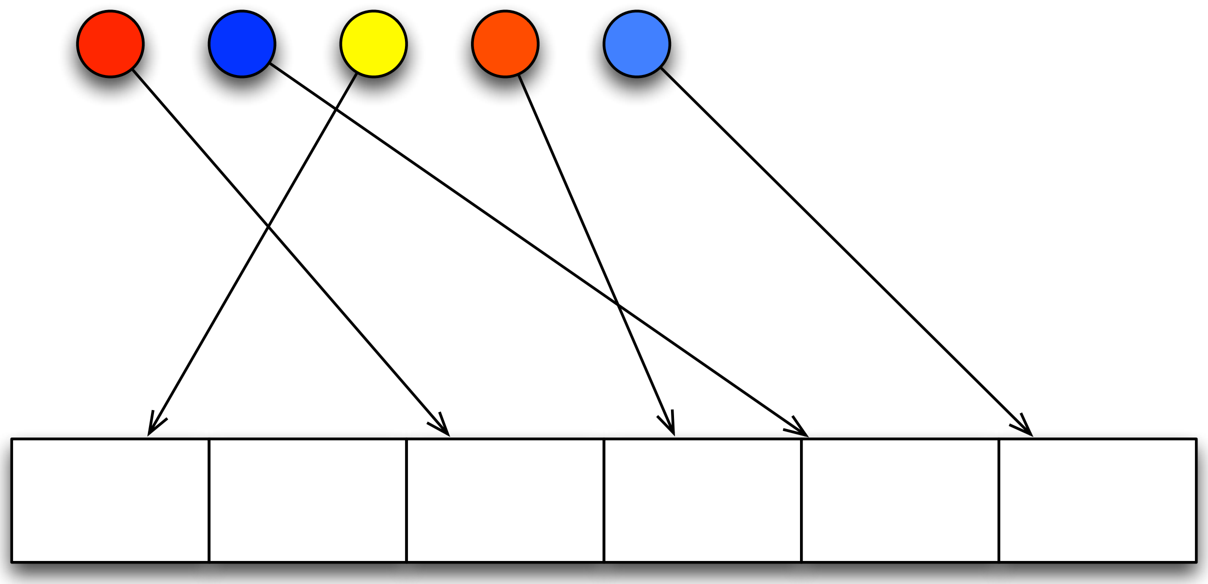


SOCIAL MEDIA

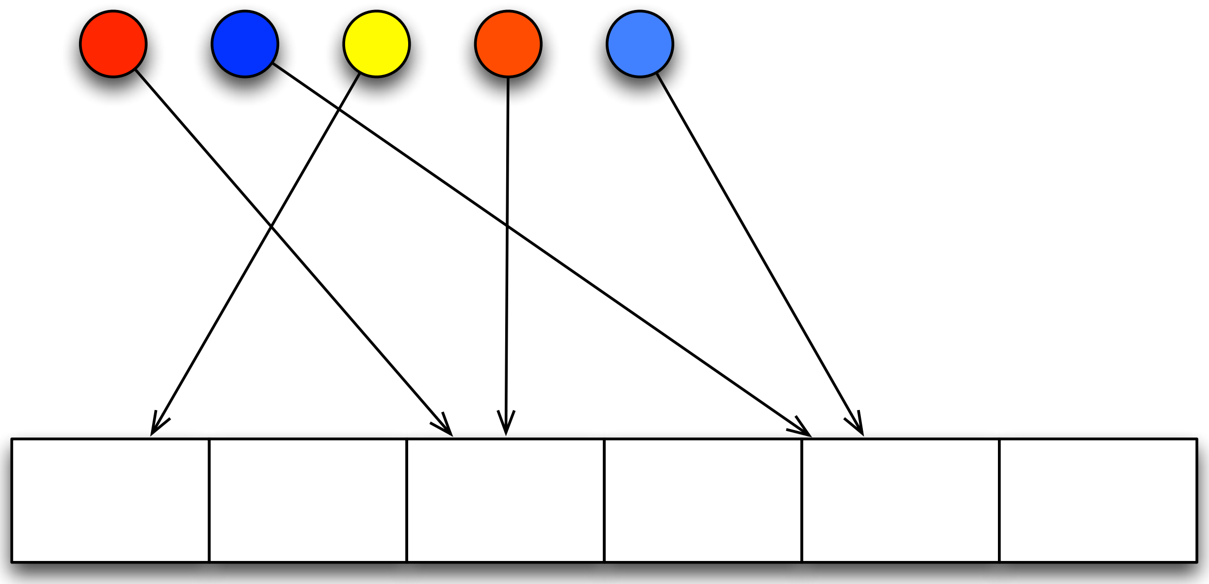


Search Engine: Limitations

Supporting LSH requires a large amount of RAM for real-time queries, upward of 200GB.



General Hashing

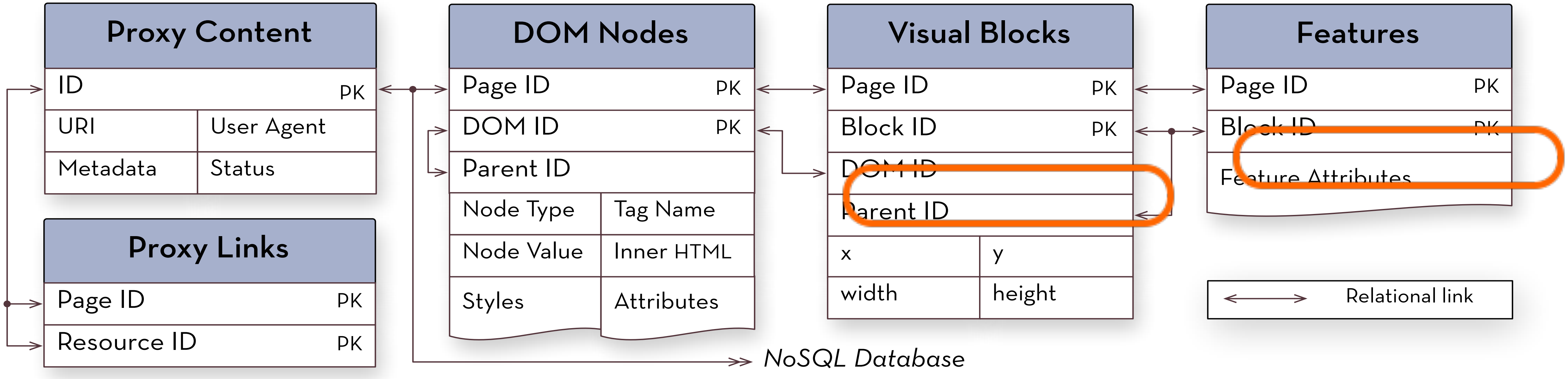


Locality-Sensitive Hashing

Cropping and loading screenshots also limits performance.

Graph DB

For certain queries, a graph database can increase query performance.



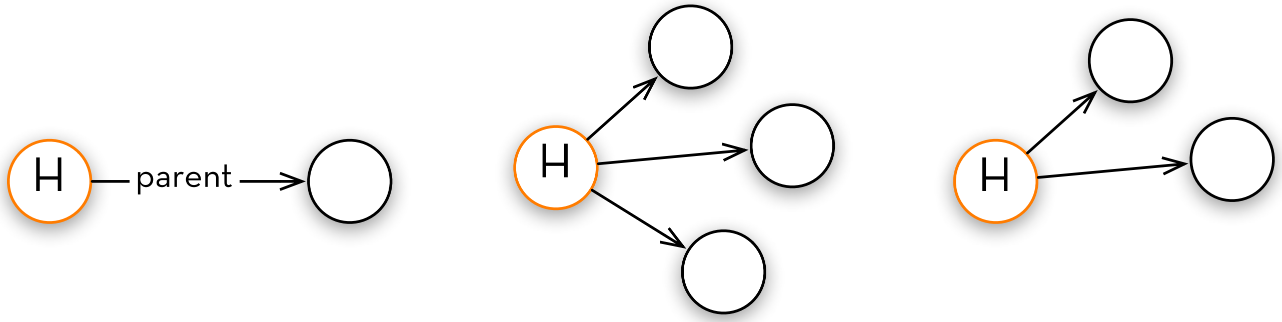
Proxy Content	
ID PK	
URI	User Agent
Metadata	Status

Proxy Links
Page ID PK
Resource ID PK

DOM Nodes	
Page ID PK	
DOM ID PK	
Parent ID	
Node Type	Tag Name
Node Value	Inner HTML
Styles	Attributes

Visual Blocks	
Page ID PK	
Block ID PK	
DOM ID	
Parent ID	
x	y
width	height

Features
Page ID PK
Block ID PK
Feature Attributes



Additional Reading:

Statement of the Publish On Demand Research Theme
http://mediax.stanford.edu/POD/concept

For more information:

- membership
- research themes
- events (conferences, seminars, workshops etc.)

Please visit our website - *http://mediax.stanford.edu*

Like us on Facebook -
https://www.facebook.com/MediaXatStanford
Follow us on Twitter -
https://twitter.com/mediaXStanford
Join us on LinkedIn -
http://www.linkedin.com (search for MediaX at Stanford)
Watch us on YouTube -
http://www.youtube.com/user/mediaxstanford

or contact:

Martha Russell, Executive Director - marthar@stanford.edu

Jason Wilmot, Communications Manager - jwilmot@stanford.edu

Adelaide Dawes, Program Manager - adelaide@stanford.edu

Acknowledgements

Many thanks to mediaX Members and Partners, whose contributions have supported this Research.

Special thanks to the professors and researchers who provided these research updates. Thanks also to Addy Dawes, Program Manager, and Jason Wilmot, mediaX Communications, for editing and putting this report into the format you are reading.