Predicting Web 2.0 Thread Updates

Shawn Tan

Table of Contents

Motivation

- Many sites with thread-based discussion features
- Users post product reviews, feedback

Obtaining such up-to-date information may be vital to companies.

Table of Contents

Refresh policies for incremental crawlers

Many works have used time difference to estimate page updates.

1. Coffman et. al. 1997 analysed the theoretical aspects.
2. Cho and Garcia-Molina trace the change history of 720,000 web pages collected over 4 months.
   2.1 Showed empirically that the Poisson process model estimates the update processes well (Cho et. al. 1999)
   2.2 Proposed different revisiting or refresh policies (Cho et. al. 2003, Garcia-molina et. al. 2003)
3. Also used in Tan et. al. 2007.

Problems with Poisson

The Poisson distribution is memoryless, and in experimental results due to Brewington and Cybenko 2000, the behaviour of site updates are not.

Using Site-level Knowledge

Yang et. al. 2009, attempted to resolve this by

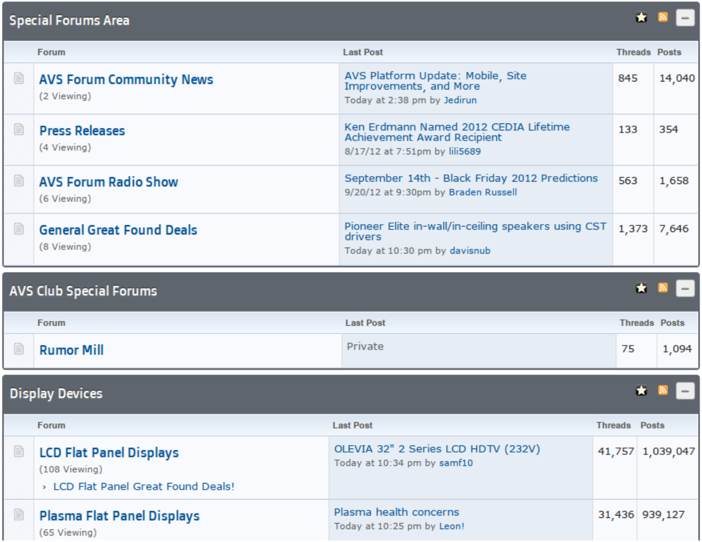
1. Using the list structure of forum sites to infer a sitemap.
2. Use a linear-regression model to predict when the next update to the thread will arrive.
3. Linear model used together with the sitemap information to prioritise the request queue in the crawler.
4. Has the ability to make use of index information to infer changes in threads. Other types of comment systems do not have such indices.

Summary

- Previous work dealt with Web 1.0 sites.
- Did not take into account the content in the posts.
- Evidence to show that using time, while makes reasonable prediction, does not fully model the behaviour of threads.

Table of Contents

avsforum.com



Special Forums Area			
Forum	**Last Post**	**Threads**	**Posts**
AVS Forum Community News (2 Viewing)	AVS Platform Update: Mobile, Site Improvements, and More Today at 2:38 pm by Jedirun	845	14,040
Press Releases (4 Viewing)	Ken Erdmann Named 2012 CEDIA Lifetime Achievement Award Recipient 8/17/12 at 7:51pm by lili5689	133	354
AVS Forum Radio Show (6 Viewing)	September 14th - Black Friday 2012 Predictions 9/20/12 at 9:30pm by Braden Russell	563	1,658
General Great Found Deals (8 Viewing)	Pioneer Elite in-wall/in-ceiling speakers using CST drivers Today at 10:30 pm by davisnub	1,373	7,646

AVS Club Special Forums			
Forum	**Last Post**	**Threads**	**Posts**
Rumor Mill	Private	75	1,094

Display Devices			
Forum	**Last Post**	**Threads**	**Posts**
LCD Flat Panel Displays (108 Viewing) › LCD Flat Panel Great Found Deals!	OLEVIA 32" 2 Series LCD HDTV (232V) Today at 10:34 pm by samf10	41,757	1,039,047
Plasma Flat Panel Displays (65 Viewing)	Plasma health concerns Today at 10:25 pm by Leon!	31,436	939,127

User-centric threads

My First Ever DIY Sub

Subscribe   Search This Thread

Start a New Thread

9/25/12 at 4:31am THREAD STARTER   **post #1** of 14

mfrey0118
Amateur A/V Junkie
Advanced Member
offline
519 Posts. Joined 3/2011

Hi everyone!

OK, so long story short, I had an Onkyo HT-S5400, then upgraded the receiver to a 609, upgraded all my speakers, EXCEPT the sub and here we are.

Was dead set on a pre-fab sub from Lava, BIC, or Klipsch. Finally decided to go another route and build my own.

This is what I am working with:

TC Sounds Epic 12" DVC (500w RMS @ 2+2 ohms) sub
Dayton Audio SPA500 amp (540w RMS @ 4 Ohms)

Unfortunately, I have zero box building skills. My room is big too, about 35' x 17' total, kitchen and LR shared, no separating wall, ceiling that goes from 9ft up to like 15ft.

Also, I am on a limited budget. I've already purchased the amp, and I will be ordering the sub early next week.

I figured 3/4" MDF is a good place to start. I also have a fully activated copy of BassBox Pro 6.

⋮

9/25/12 at 10:18am THREAD STARTER

mfrey0118
Amateur A/V Junkie
Advanced Member
offline
519 Posts. Joined 3/2011

Guys thanks so much for lending your expertise...this is good stuff...

9/25/12 at 12:30pm

NicksHitachi
Winning!
AVS Special Member
offline
2,316 Posts. Joined 7/2007
Location: Wilmington, NC

Horn or ported.

As big and tuned as low as you can accomodate.

You'll need EQ for High Pass.

Corner Load

Sell the dayton and get a pro-amp(prefereably one with EQ) you'll not even wake up the TC with

9/26/12 at 6:12am THREAD STARTER

Quote:

Originally Posted by **NicksHitachi**

Questions

What's wrong with my LG LCD?

Subscribe Search This Thread

Start a New Thread

9/1/12 at 2:20pm THREAD STARTER **post #1**

I noticed this issue today when I powered on my 2011 LG LCD. The set has about 2,100 hrs on
and was working fine yesterday. The issue is that the picture looks blurry and low res with
distorted text and jagged vertical and diagonal lines in what should be solid sharp and clear text
lines, pictures, and other shapes. The issue occurs with all sources/inputs and on the TV menu
itself. Is this a panel issue or a main board issue or something else? Anything I can try to resol
this issue? I reset the picture settings and tried various pic modes, but to no avail.

⋮

9/1/12 at 2:46pm

walford
AVS Addicted Member

Is it a 3D model?
Did you try unplugging it for about 20 minutes in order to make sure it had a complete reboot w

Mentions

10/1/10 at 6:12pm

Elkhunter
Senior Member
offline
315 Posts. Joined 7/2008

rdjam:

Wouldn't a 1.4a AVR with 2 simultaneous HDMI outpu

I have an Yamaha RX-A3000 on order (due next Thu

TIA

10/1/10 at 6:24pm THREAD STARTER

rdjam
New toy The Darblet!
AVS Gold Club
offline
9,716 Posts. Joined 3/2005
Location: Miami, FL

Quote:

Originally Posted by **Elkhunter**

rdjam:

Wouldn't a 1.4a AVR with 2 simultaneous HDMI ou

I have an Yamaha RX-A3000 on order (due next

TIA

That should be do-able. Don't have one yet but can't

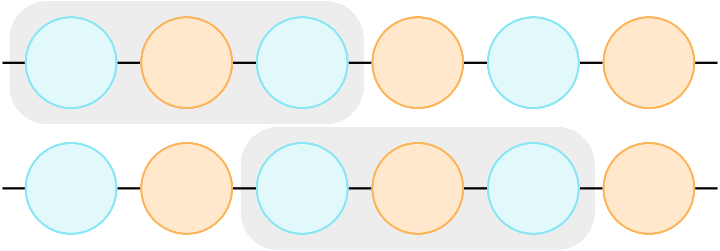
However, I was planning to have one output for my projectors.

Table of Contents

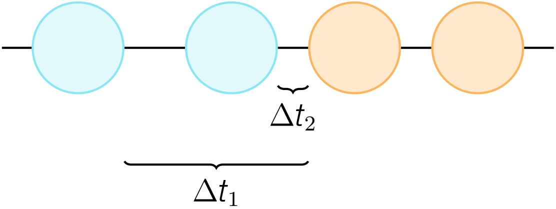
Requirements

- Balance of freshness and bandwidth usage.
- Penalise when using too much bandwidth (visiting the site too much).
- Penalise when “database” not fresh (visiting the site too little).

Events



T-score



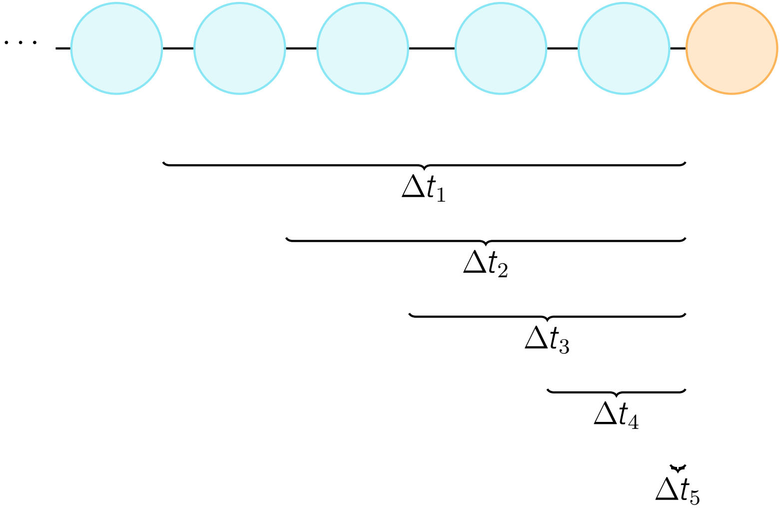
$$T = \frac{1}{|P|}\sum_{i=1}^{|P|}\Delta t_i$$

From Yang et. al. 2009

Visit/Post ratio

Number of visits per post, keep the T-score in check.

Normalised T-score



Normalised Visit/Post ratio

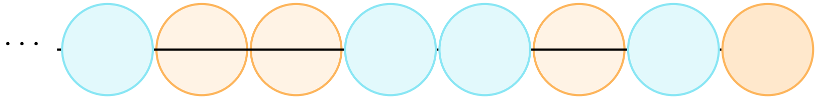


Table of Contents

Experiment Setup

75% Training	25% Testing
75% Training	25% Testing
75% Training	25% Testing
⋮	
75%	25%

For parameter tuning:

1. Threads from 100 to 1000 posts
2. 107 Threads in total

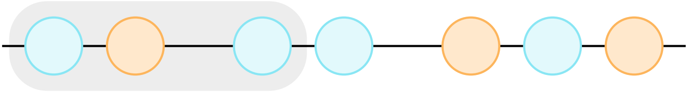
Baseline

Take the average Δ_t from training set, and use that as the revisit time.

	Pr_{error}	T-score	Visit/Post
Average	0.501 ± 0.001	1764.474 ± 267.227	18.117 ± 7.290

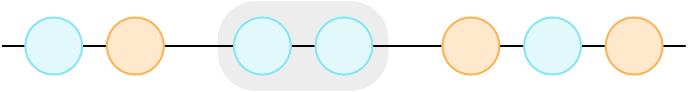
Windowing

Use features from windows of posts. Number of posts in window given by w.



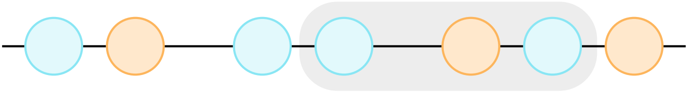
Windowing

Use features from windows of posts. Number of posts in window given by w.



Windowing

Use features from windows of posts. Number of posts in window given by w.



Window-based average

Take the average Δ_t from ~~training set~~ the previous window, and use that as the revisit time.

	T-score	Visit/Post	Pr_{error}
$w = 5$	6420 ± 1000	16.6 ± 7	0.028 ± 0.005
$w = 10$	4580 ± 700	17.4 ± 8	0.028 ± 0.005
$w = 15$	3830 ± 600	18.5 ± 9	0.021 ± 0.003
$w = 20$	3340 ± 400	18.3 ± 9	0.022 ± 0.004

Performs worse than the simple average baseline.

Support Vector Regression

Using only the window's Δ_t as features.

	T-score	Visit/Post	Pr_{error}
$w = 5$	1537.682 ± 234.658	18.056 ± 7.585	0.018 ± 0.004
$w = 10$	1485.157 ± 198.664	18.523 ± 8.028	0.019 ± 0.004
$w = 15$	1433.771 ± 185.080	19.396 ± 8.896	0.016 ± 0.003
$w = 20$	1577.639 ± 229.482	19.037 ± 8.690	0.019 ± 0.004

Content-based features

Count of individual tokens used:

1. Text is stemmed, stopwords removed
2. Occurences of usernames are replaced with ‘#USER#’
3. Occurences of tokens with mixtures of alphabets and numbers are replaced with ‘#MODEL#’
4. Univariate regression tests used to select features

Time-context

1. Hour of the day
2. Day of the week

Represented as bit vectors

Content features only

Using only the content features (stemmed word frequency counts).

	T-score	Visit/Post	Pr_{error}
$w = 5$	1593.380 ± 237.070	18.007 ± 7.585	0.019 ± 0.004
$w = 10$	1546.839 ± 198.243	18.493 ± 8.030	0.023 ± 0.006
$w = 15$	1491.695 ± 187.589	19.359 ± 8.899	0.021 ± 0.005
$w = 20$	1645.177 ± 232.365	19.017 ± 8.694	0.024 ± 0.005

Worse than the time difference approach, would using both sets of features help?

Content features $+\Delta_t+$ time-context

	T-score	Visit/Post	Pr_{error}
$w = 5$	1537.673 ± 234.657	18.056 ± 7.585	0.018 ± 0.004
$w = 10$	1485.137 ± 198.662	18.523 ± 8.028	0.019 ± 0.004
$w = 15$	1433.762 ± 185.078	19.396 ± 8.896	0.016 ± 0.003
$w = 20$	1577.639 ± 229.482	19.037 ± 8.690	0.019 ± 0.004

Discounted Sum

Discounted sum of feature vectors from previous windows.

$$\mathbf{X}'_t = \mathbf{X}_t + \alpha \mathbf{X}'_{t-1}$$

Where $0 \geq \gamma > 1$. Here we use only the word count as before.

	T-score	Visit/Post	Pr_{error}
$\alpha = 1$	1433.761 ± 185.078	19.396 ± 8.896	0.002 ± 0.000
$\alpha = 2$	1433.759 ± 185.078	19.396 ± 8.896	0.002 ± 0.000
$\alpha = 3$	1433.757 ± 185.078	19.396 ± 8.896	0.002 ± 0.000
$\alpha = 4$	1433.755 ± 185.077	19.396 ± 8.896	0.002 ± 0.000
$\alpha = 5$	1433.755 ± 185.077	19.396 ± 8.896	0.002 ± 0.000
$\alpha = 6$	1433.755 ± 185.077	19.396 ± 8.896	0.002 ± 0.000
$\alpha = 7$	1433.755 ± 185.077	19.396 ± 8.896	0.002 ± 0.000
$\alpha = 8$	1433.755 ± 185.077	19.396 ± 8.896	0.002 ± 0.000
$\alpha = 9$	1433.746 ± 185.076	19.396 ± 8.896	0.002 ± 0.000

Stochastic Gradient Descent

Function to be fitted:

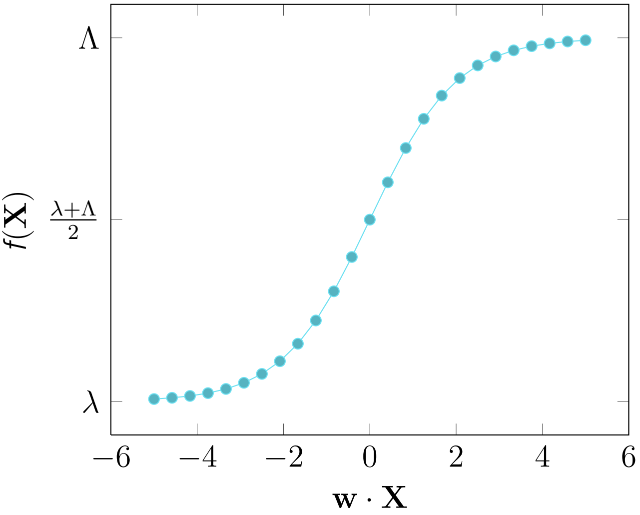
$$f(\mathbf{X}) = \frac{\Lambda - \lambda}{1 + e^{\mathbf{w} \cdot \mathbf{X}}} + \lambda$$

Update rule:

$$\Delta \mathbf{w}_i = \eta \underbrace{\left(\widehat{\Delta_t} - \Delta_t \right)}_{\text{error term}} \underbrace{\left(f(\mathbf{X})(1 - f(\mathbf{X})) \right)}_{\text{gradient}} \mathbf{X}_i$$

Update rule is used everytime a new post and time interval is observed.

Scaled Sigmoid Function



SGD results

	T-score	Visit/Post
$\eta = 5 \cdot 10^{-5}$	1595.563	19.097
$\eta = 5 \cdot 10^{-6}$	1525.705	19.122
$\eta = 5 \cdot 10^{-7}$	1440.440	19.121
$\eta = 5 \cdot 10^{-8}$	1407.172	19.108
$\eta = 5 \cdot 10^{-9}$	1416.182	19.110
$\eta = 5 \cdot 10^{-10}$	1451.729	19.106
$\eta = 5 \cdot 10^{-11}$	1482.868	19.104
$\eta = 5 \cdot 10^{-12}$	1487.555	19.104

Wilcoxon's signed-rank test

Cannot use Student's t test, since distribution of T-scores is not normal

- Non-parametric statistical hypothesis test
- Data paired, and come from same population
- H_0 : No difference between my model and baseline (average time revisitation)
- H_1 : My model performs better than baseline

Results of Paired Tests

T-score	-381.885 $\pm$ 153.673
Visit/Post	1.053 $\pm$ 1.749
Single-sided test	0.0486

Analysis of Data

What words are important, and how do they change over time?
How does the number of posts used to train the model affect the result?

Limitations

1. Experiments only performed on one forum
2. (In)Correctness of model
3. Stochastic Gradient Descent is slow

Future Work

1. Incorporate decay into model
   - SVM with adaptive parameters (Cao & Tay, 2003)
2. NLP techniques
   - Use features from Wang, Chen, and Kan (2012)
3. Lexical Chaining
   - Wang and McCarthy (2011)

Questions? Suggestions?