Predicting Web 2.0 Thread Updates Progress Update

Shawn Tan

Table of Contents

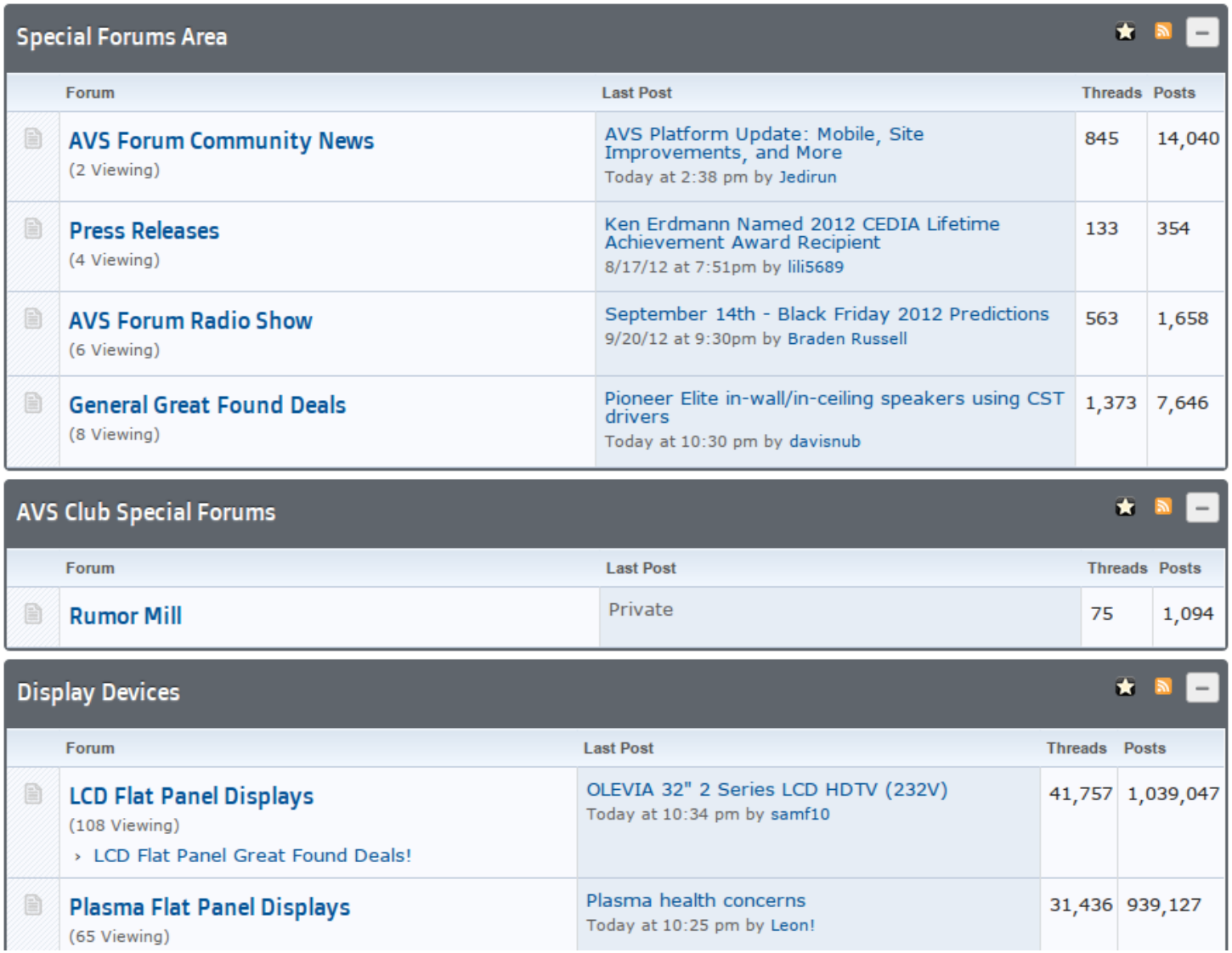
Motivation

- 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

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

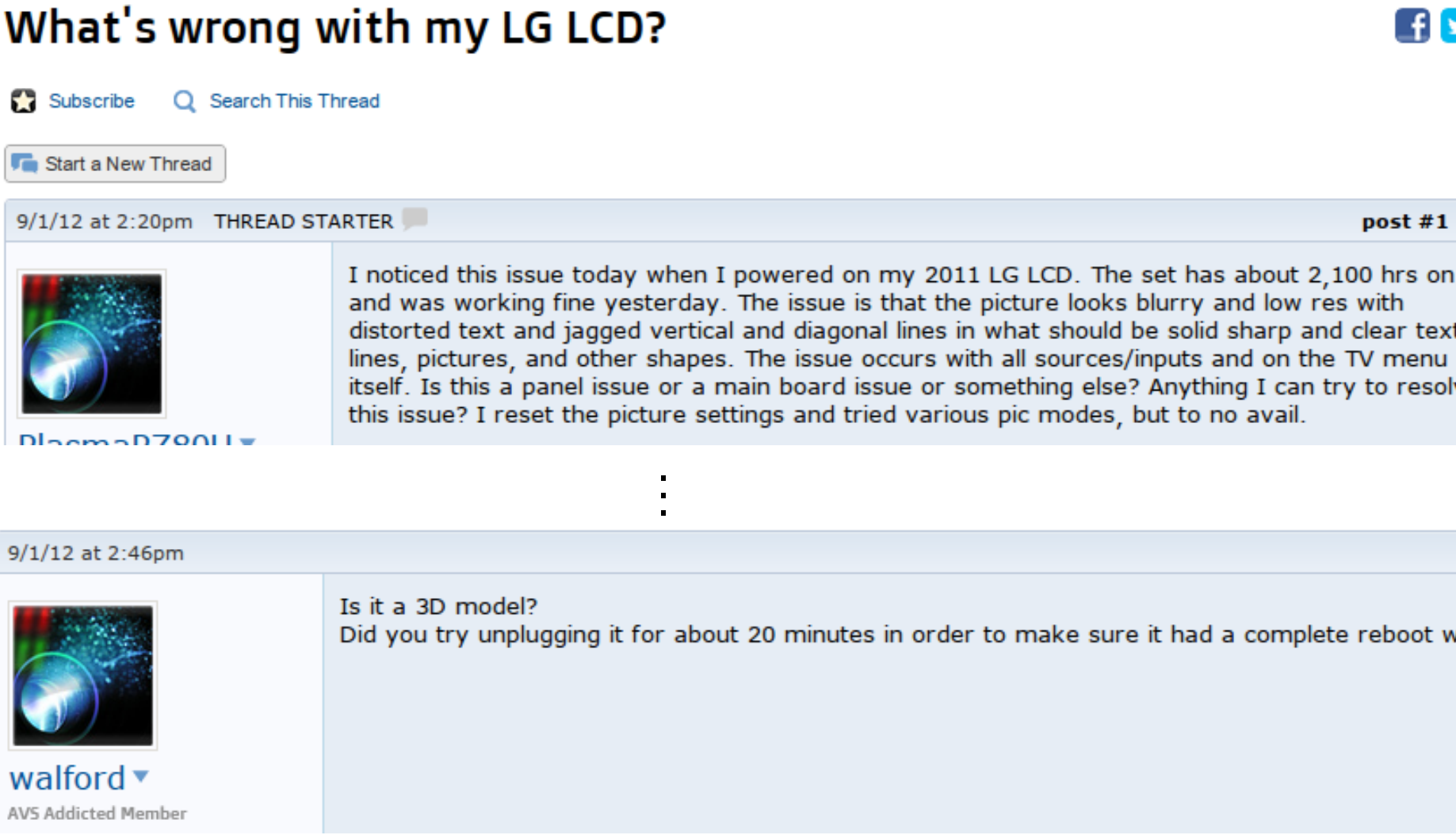
Forum	Last Post	Threads	Posts
Rumor Mill	Private	75	1,094

Display Devices

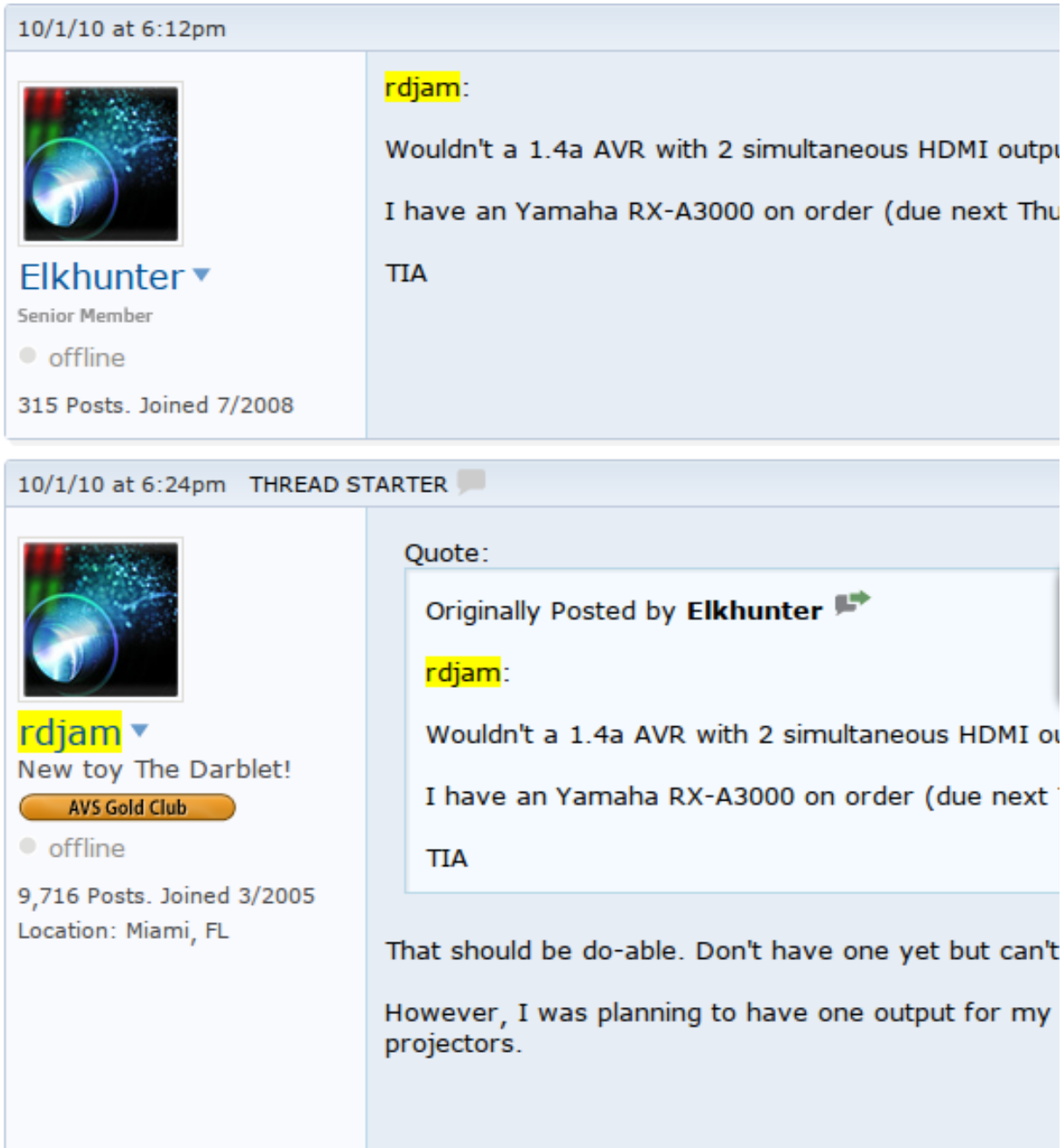
Forum	Last Post	Threads	Posts
LCD Flat Panel Displays (108 Viewing) › LCD Flat Panel Great Found Deals!	OLEVIA 32" 2 Series LCD HDTV (232V) Today at 10:34 pm by samf10	41,757	1,039,047
Plasma Flat Panel Displays (65 Viewing)	Plasma health concerns Today at 10:25 pm by Leon!	31,436	939,127

User-centric threads

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
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
walford
AVS Addicted Member

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

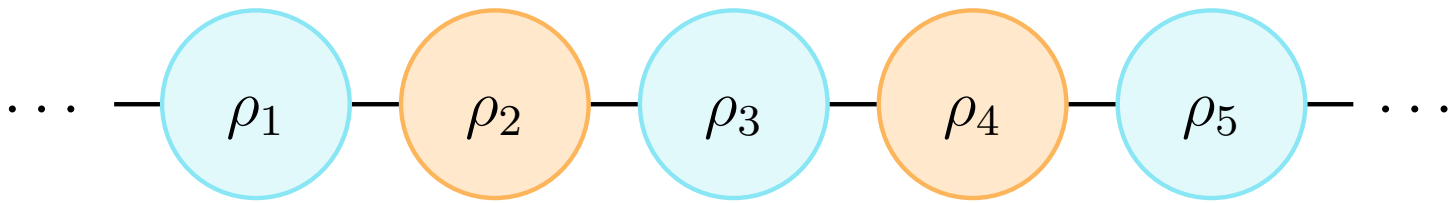
However, I was planning to have one output for my projectors.

Table of Contents

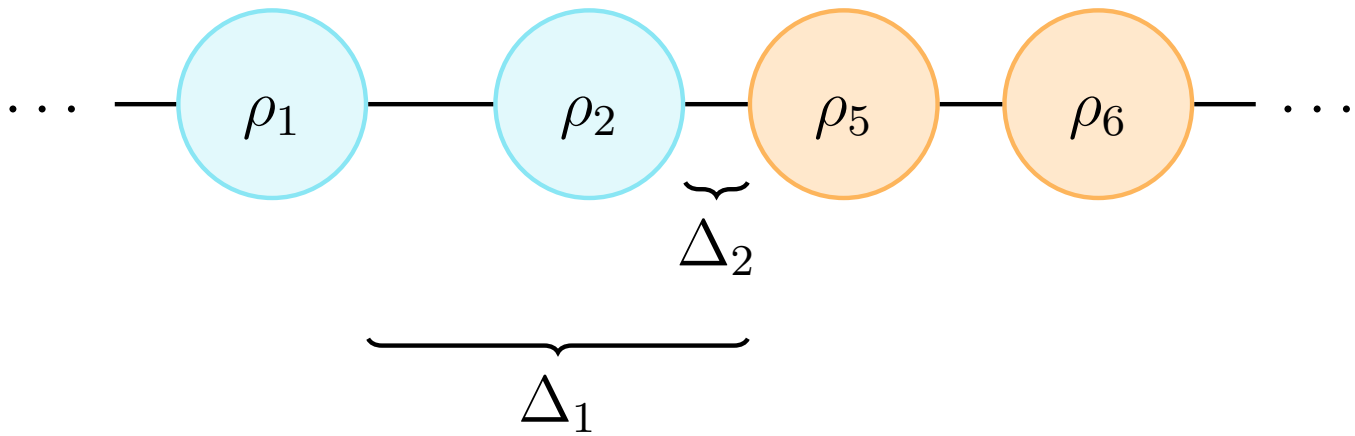
Requirements

- 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



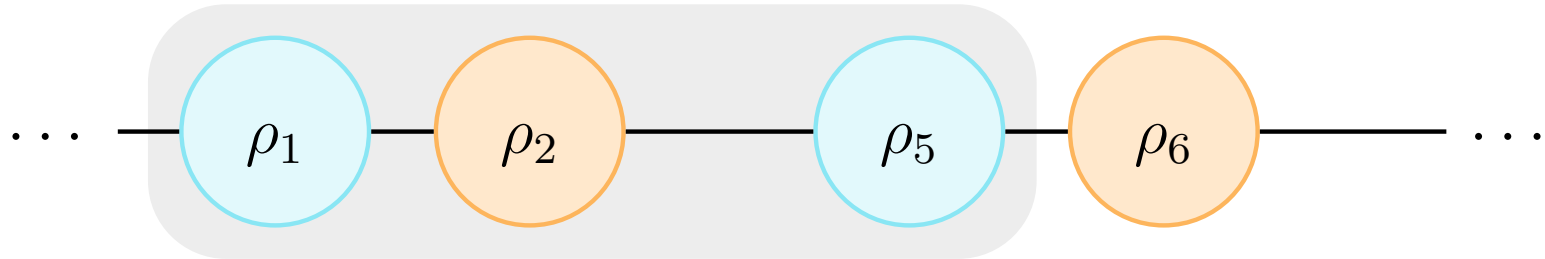
$$T = \frac{1}{N} \sum_{i \in posts} \Delta_i$$

From Yang et. al. 2009

Visit/Post ratio

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

Pr_{error}



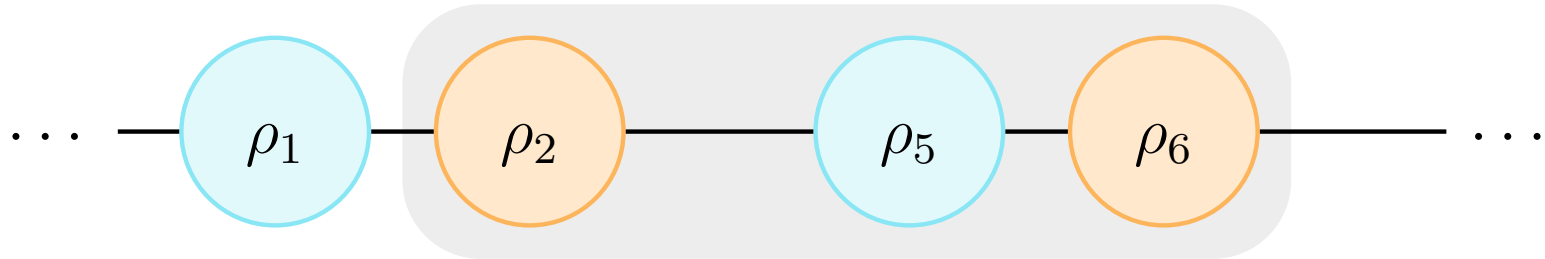
1. Pr_{fa} More visits than posts, false alarm.
2. Pr_{miss} More posts than visits, miss.

Weighted average use as error metric.

$$Pr_{error} = \alpha Pr_{fa} + (1 - \alpha) Pr_{miss}$$

Georgescul et. al. 2006

Pr_{error}



1. Pr_{fa} More visits than posts, false alarm.
2. Pr_{miss} More posts than visits, miss.

Weighted average use as error metric.

$$Pr_{error} = \alpha Pr_{fa} + (1 - \alpha) Pr_{miss}$$

Georgescul et. al. 2006

Table of Contents

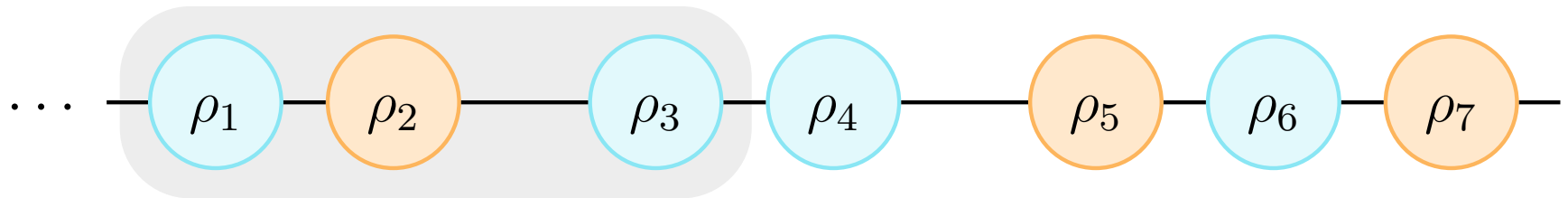
Baseline

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

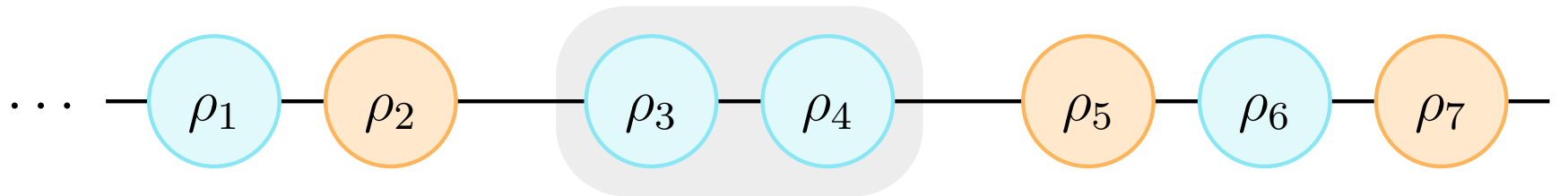
Windowing

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



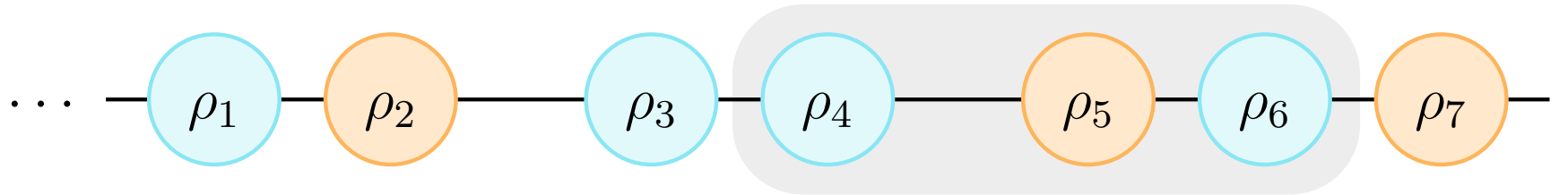
Windowing

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.

	Pr_{error}	T-score	Visit/Post
$w = 1$	0.504 ± 0.003	18862.320 ± 4267.812	16.142 ± 7.049
$w = 5$	0.502 ± 0.003	6418.208 ± 962.716	16.464 ± 7.386
$w = 10$	0.504 ± 0.003	4598.955 ± 682.458	17.291 ± 7.872
$w = 15$	0.504 ± 0.003	3833.605 ± 600.824	18.337 ± 8.727
$w = 20$	0.504 ± 0.003	3340.929 ± 444.908	18.102 ± 8.541

Performs worse than the simple average baseline.

Support Vector Regression

Using only the window's Δ_t as features.

	Pr_{error}	T-score	Visit/Post
$w = 1$	0.498 ± 0.002	1576.082 ± 253.300	18.267 ± 7.290
$w = 5$	0.498 ± 0.002	1541.595 ± 232.272	17.907 ± 7.508
$w = 10$	0.499 ± 0.002	1488.688 ± 196.648	18.371 ± 7.947
$w = 15$	0.500 ± 0.002	1443.138 ± 183.408	19.234 ± 8.805
$w = 20$	0.499 ± 0.001	1584.171 ± 227.209	18.880 ± 8.602

Performs better than baseline, but, what happens if we use content?

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).

	Pr_{error}	T-score	Visit/Post
$w = 1$	0.496 ± 0.002	1649.606 ± 262.578	18.255 ± 7.292
$w = 5$	0.495 ± 0.001	1596.220 ± 234.643	17.859 ± 7.508
$w = 10$	0.498 ± 0.001	1554.391 ± 196.343	18.341 ± 7.949
$w = 15$	0.497 ± 0.001	1500.391 ± 185.857	19.197 ± 8.808
$w = 20$	0.494 ± 0.002	1653.162 ± 230.106	18.859 ± 8.606

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

Content features $+\Delta_t+$ time-context

	Pr_{error}	T-score	Visit/Post
$w = 1$	0.498 ± 0.002	1537.992 ± 251.250	18.272 ± 7.291
$w = 5$	0.498 ± 0.002	1541.587 ± 232.271	17.907 ± 7.508
$w = 10$	0.499 ± 0.002	1488.669 ± 196.646	18.371 ± 7.947
$w = 15$	0.500 ± 0.002	1443.130 ± 183.407	19.234 ± 8.805
$w = 20$	0.499 ± 0.001	1584.171 ± 227.209	18.880 ± 8.602

Improved performance by an hour on average, still nothing significant.

Table of Contents

Discounted Sum

Discounted sum of feature vectors from previous windows.

$$\mathbf{X}_t = \mathbf{v}_t + \gamma \mathbf{X}_{t-1}$$

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

	Pr_{error}	T-score	Visit/Post
$\alpha = 0.1$	0.500 $\pm$ 0.002	1443.129 $\pm$ 183.407	19.234 $\pm$ 8.805
$\alpha = 0.2$	0.500 $\pm$ 0.002	1443.127 $\pm$ 183.407	19.234 $\pm$ 8.805
$\alpha = 0.3$	0.500 $\pm$ 0.002	1443.126 $\pm$ 183.407	19.234 $\pm$ 8.805
$\alpha = 0.4$	0.500 $\pm$ 0.002	1443.124 $\pm$ 183.406	19.234 $\pm$ 8.805
$\alpha = 0.5$	0.500 $\pm$ 0.002	1443.121 $\pm$ 183.406	19.234 $\pm$ 8.805
$\alpha = 0.6$	0.500 $\pm$ 0.002	1443.119 $\pm$ 183.406	19.234 $\pm$ 8.805
$\alpha = 0.7$	0.500 $\pm$ 0.002	1443.116 $\pm$ 183.405	19.234 $\pm$ 8.805
$\alpha = 0.8$	0.500 $\pm$ 0.002	1443.112 $\pm$ 183.405	19.234 $\pm$ 8.805
$\alpha = 0.9$	0.500 $\pm$ 0.002	1443.107 $\pm$ 183.404	19.234 $\pm$ 8.805

Stochastic Gradient Descent

Function to be fitted:

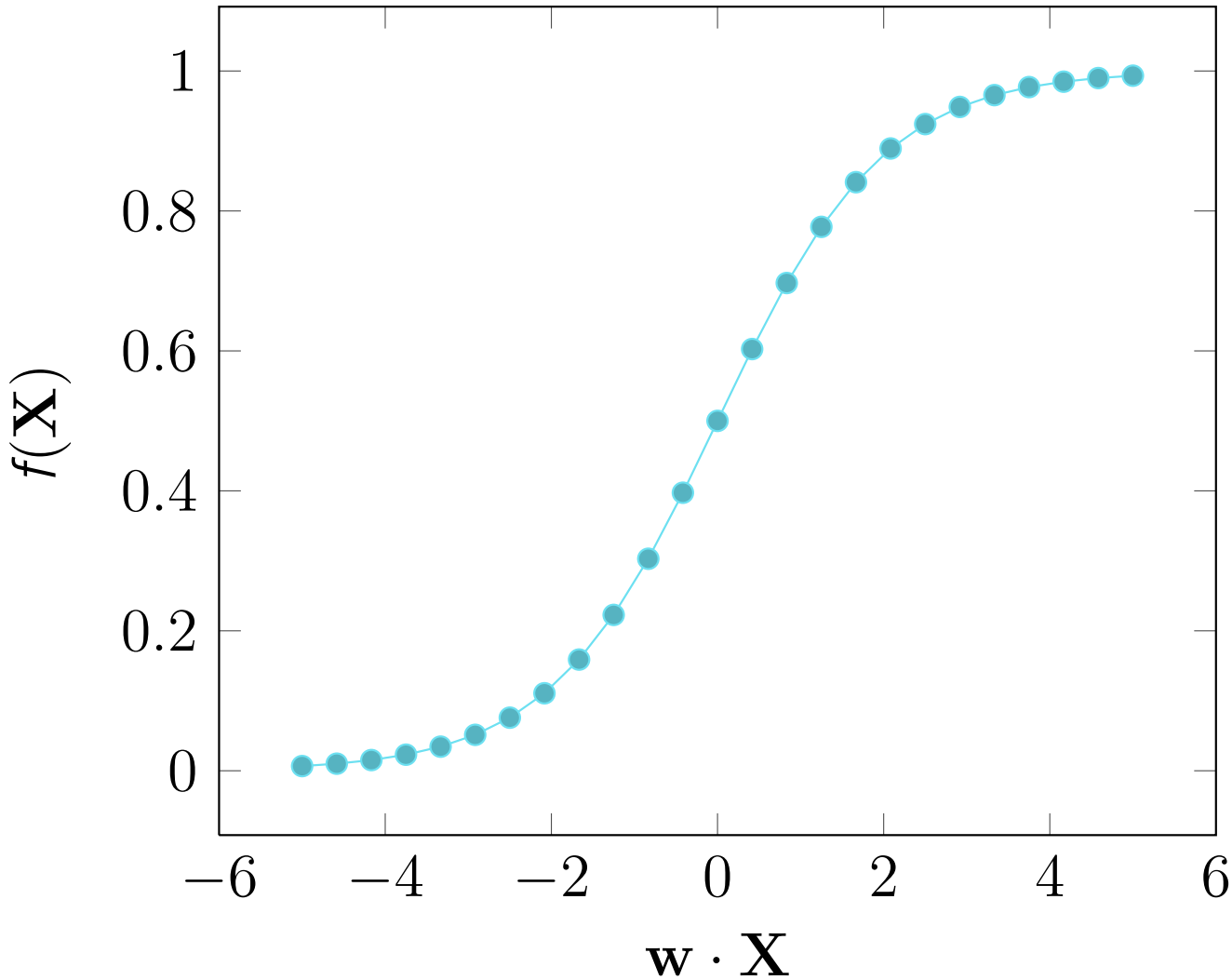
$$f(\mathbf{X}) = \frac{\Lambda - \lambda}{1 + e^{\mathbf{w}\cdot\mathbf{X}}} + \lambda$$

Update rule:

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

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

Scaled Sigmoid Function



SGD results

With the right η it did comparably well against previous methods, but nothing significantly better.
I also tried $\eta = 5 \cdot 10^{-1}$ to $\eta = 5 \cdot 10^{-4}$ but resulted in buffer overflow when calculating the exponent.

	Pr_{error}	T-score	Visit/Post
$\eta = 5 \cdot 10^{-5}$	0.499	1595.563	19.097
$\eta = 5 \cdot 10^{-6}$	0.501	1525.705	19.122
$\eta = 5 \cdot 10^{-7}$	0.502	1440.440	19.121
$\eta = 5 \cdot 10^{-8}$	0.501	1407.172	19.108
$\eta = 5 \cdot 10^{-9}$	0.502	1416.182	19.110
$\eta = 5 \cdot 10^{-10}$	0.501	1451.729	19.106
$\eta = 5 \cdot 10^{-11}$	0.501	1482.868	19.104
$\eta = 5 \cdot 10^{-12}$	0.501	1487.555	19.104

Is there a name for this?

Work in progress...

1. Another (better) metric for prediction models.
2. Better way to present *T*-scores and Post/Visit ratios.
3. Types of discussion (topic modeling?) and relationships with time intervals
4. Full-scale evaluation for entire forum
5. More datasets!