Predicting Web 2.0 Thread Updates

Shawn Tan

22^{nd} November 2012

Contents

Motivation

Why predict updates?

1. Increase in number of sites with discussion threads
2. Keeping up to date with information in discussion threads
3. Incremental crawlers

Motivation

What do we need to do?
Balance:

- Bandwidth consumption
  - Cannot repeatedly visit page at short intervals
- Timeliness of visits
  - Visiting at too large intervals causes data to be not fresh.

Motivation

Contributions

1. Provided evaluation metrics that can be parameterised
2. Proposed methods that perform better than the baseline, and can be employed for making revisit time estimation.

Contents

Revisitation Policies

Coffman and Liu (1997)

- Follows Poisson process $\Rightarrow$ revisit at times proportional to μ is optimal

Empirical evaluations

Performed by Cho and Garcia-Molina

- Showed empirically that the Poisson process model estimates the update processes well (Cho, 1999)
- Proposed different revisiting or refresh policies (Cho & Garcia-Molina, 2003; Cho & Garcia-molina, 2003)

Brewington and Cybenko (2000) show that page updates are not memoryless, so do not strictly follow a Poisson process.

Yang, Cai, Wang, and Huang (2009)

1. Infer a sitemap.
2. Use a linear-regression model to rank when to next visit
3. Linear model used together with the sitemap information to prioritise the request queue in the crawler.

Has the ability to make use of index information to infer changes in threads. Other types of comment systems do not have such indices.

What about content?

Content for prediction

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** of 20

I noticed this issue today when I powered on my 2011 LG LCD. The set has about 2,100 hrs on it 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 resolve 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 when started up?

Content for prediction

Evaluation Metrics

1. Yang et al. (2009) T-score metric, but dependent on bandwidth
2. Weighted sum of probability of misses and false alarms used by segmentation metric in Georgescul, Clark, and Armstrong (2009). Not useful for measuring time differences.

Summary

- Visiting at average update rate may not be suitable for user-generated content.
- Does not make use of content in prediction.
- Evaluation metrics not suitable for comparison

Contents

Requirements

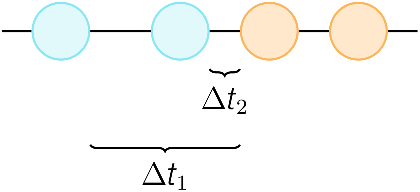
1. Balance:
   - Bandwidth consumption
   - Timeliness of visits
2. Parameterised such that can attribute different importance to both
3. Simple metric to compare across algorithms

Visit/Post ratio

$$\frac{\text{Number of Visits}}{\text{Number of Posts}}$$

- Get an idea of how many number of visits needed before a post is retrieved.

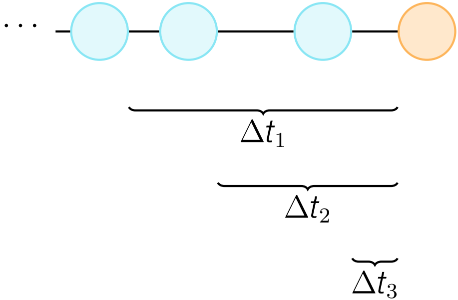
T-score



How do we compare?

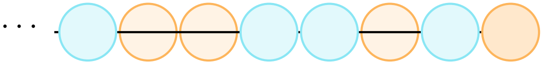
- Two different metrics
  - Visit/Post ratio
  - T-score
- Problem: how to compare?
  - One may be higher than the other
  - T-score / Bandwidth may be less important
  - Not on the same scale of things, don't represent similar units

Worst-case T-score ($T_{\max}$)



1. Assume a visit at the last post.
2. *T*-score in this case would be the maximum possible

Worst-case Visit/Post ratio



1. Assume a discrete time unit (minutes).
2. For every time unit that a post does not appear in, assume a visit appears.
3. Visit/Post ratio in this case would be the maximum possible.

Pr_{error}

- Take ratio:
  - $Pr_{\text{FA}} = \dfrac{\text{P/V-ratio}}{\text{Maximum P/V-ratio}}$
  - $Pr_{\text{miss}} = \dfrac{T}{T_{\text{max}}}$
- Weighted sum of both metrics:

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

- $0 \leq \alpha \leq 1$

Contents

Features

What features?

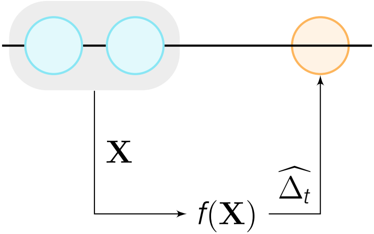
- Single post: features too sparse, too little information
- *w* recent posts: more indicative of current state of thread.

Windowing

$$f(\mathbf{X})$$

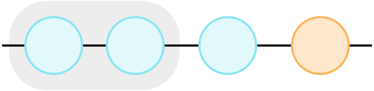
- Posts
- Visits
- Window

Windowing



- Posts
- Visits
- Window

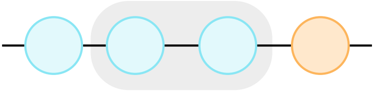
Windowing



$$f(\mathbf{X})$$

- Posts
- Visits
- Window

Windowing



$$f(\mathbf{X})$$

- Posts
- Visits
- Window

Window time intervals ($\mathbf{t}_\Delta$)

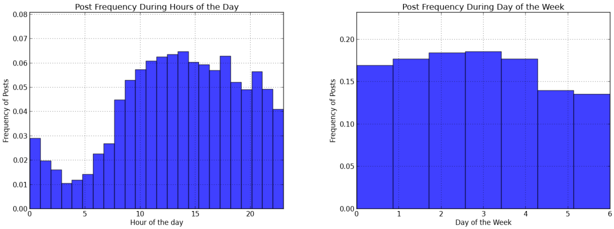
- Time intervals between posts inside window
- Yang et al. (2009)

Time context (t_{ctx})

- Day of week and hour of day (bit vector)

$$\underbrace{[0, 0, \ldots, 1, \ldots 0]}_{\text{length of 24}}$$

- Yang et al. (2009)



Content Features (v)

- Term occurrence counts.
- 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

Algorithms

Average Revisits (BL)

- Baseline
- Average Δ_t in training set.

Support Vector Regression (SVR)

- An extension of using Support Vector Machines for classification
- Advantages in high dimension feature vectors (Drucker, Burges, Kaufman, Smola, & Vapnik, 1997)
- Radial Basis Function (RBF) kernel.

Stochastic Gradient Descent (SGD)

- Linear regression produces poor results – too big, or negative
- Have a function with upper bound (Λ) and lower bound (λ)
- Sigmoid function from neural networks

Stochastic Gradient Descent (SGD)

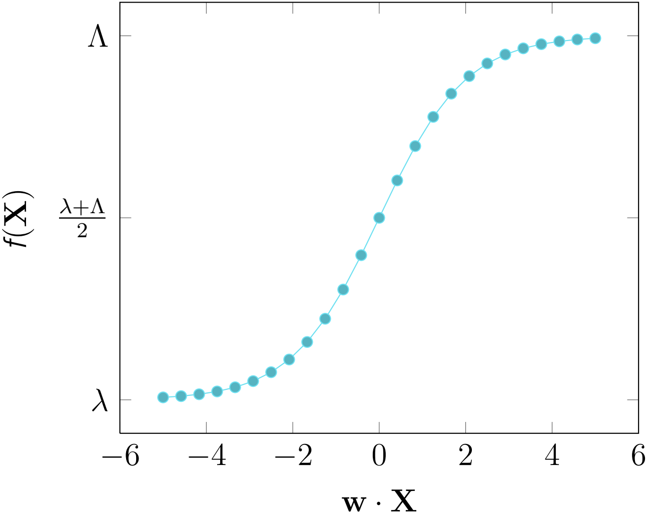
Function to be fitted:

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

Update rule:

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

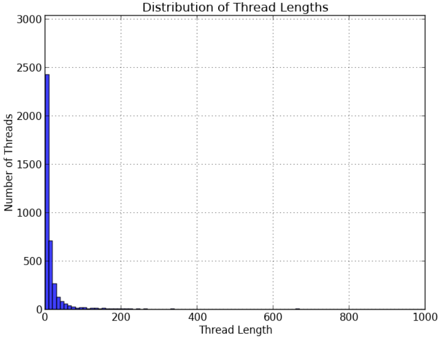
Stochastic Gradient Descent (SGD)



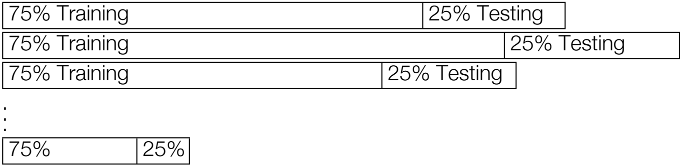
Contents

http://www.avsforum.com/f/

- 4,158 threads
- 1,002,225 posts



Experiment setup



Feature set selection

Tuning set

1. Used threads with 100 - 200 posts
2. Total of 97 threads
3. Use these to perform tuning

Feature set selection

Tuning set

1. Used threads with 100 - 200 posts
2. Total of 97 threads
3. Use these to perform tuning

Using SVR

1. Window size $w = 15$
2. $\mathbf{t}_\Delta + \mathbf{t}_{\text{ctx}} + \mathbf{v}$ gives best results

$\mathbf{t}_\Delta$ Δ_t for all time differences in window

$\mathbf{t}_{\text{ctx}}$ Time context (hour of day, day of week)

$\mathbf{v}$ Word frequencies

Feature set selection

Tuning set

1. Used threads with 100 - 200 posts
2. Total of 97 threads
3. Use these to perform tuning

SGD parameter tuning

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

- Log-scale search
- Range: $5 \cdot 10^{-12} \leq \eta \leq 5 \cdot 10^{-6}$
- Best value: $\eta = 5 \cdot 10^{-8}$

Full Evaluation on Dataset

	T-score	Visit/Post	Pr_{error}	
SVR	-938.471 $\pm$ 161.545	1.29 $\pm$ 5.207	-0.00728 $\pm$ 0.00327	$p < 0.05$

Full Evaluation on Dataset

Tuning set

1. Total of 830 threads (Distribution of thread length is long tail)

	T-score	Visit/Post	Pr_{error}	
SGD	-479.093 $\pm$ 391.269	1.13 $\pm$ 5.198	-0.00687 $\pm$ 0.00320	$p < 0.10$

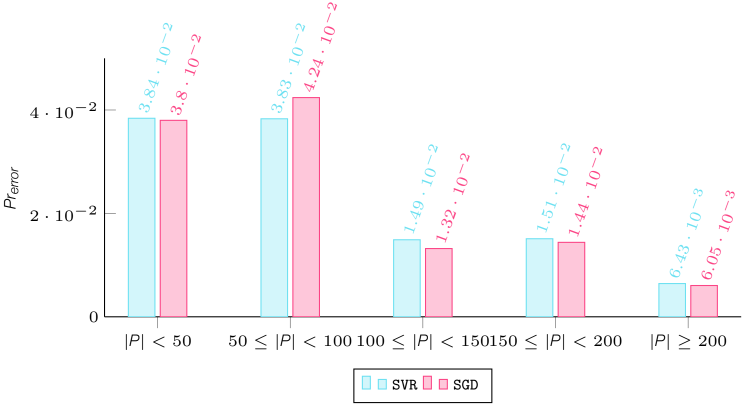
1. SVR method better than baseline.
2. SGD does not perform as well.
3. Incurs some ‘penalty’ on the Visit/Post ratio

What if we breakdown the results by thread length?

Full Evaluation on Dataset

Tuning set

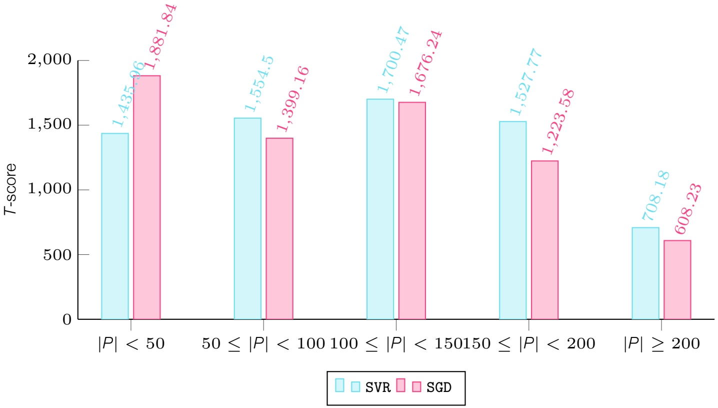
1. Total of 830 threads (Distribution of thread length is long tail)



Full Evaluation on Dataset

Tuning set

1. Total of 830 threads (Distribution of thread length is long tail)



Recommendations & Limitations

Recommendations for incremental crawler:

1. SVR could be used to predict revisitation rates initially, when $|P| < 100$.
2. Later, SGD can be used.

However,

1. Experiment results are based on only one dataset
2. SGD is slow

Contents

Contributions

1. Provided evaluation metrics that can be parameterised
2. Proposed methods that perform better than the baseline, and can be employed for making revisit time estimation.

Topic Modeling

- Separate window content into different topics
- Try to use distribution of Δ_t in different topics to make prediction

Natural Language Processing (NLP)

Usage of more NLP techniques, like for example in Wang, Chen, and Kan (2012)

- Sentiment analysis
- Discourse relation
- Sentence similarity

Thank you! Questions?

Contents

Brewington, B. E., & Cybenko, G. (2000). Keeping up with the changing web. Computer, , 2000, 52–58.

Cho, J. (1999). The evolution of the web and implications for an incremental crawler. Science, , 1999, 1–18.

Cho, J., & Garcia-Molina, H. (2003). Effective page refresh policies for Web crawlers. ACM Transactions on Database Systems, 28(4), December, 2003, 390–426.

Cho, J., & Garcia-molina, H. (2003). Estimating Frequency of Change. (650), 2003.

Coffman, E., & Liu, Z. (1997). Optimal robot scheduling for web search engines. Sophia, , 1997.

Drucker, H., Burges, C., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. Advances in neural information processing systems, , 1997, 155–161.

Georgescul, M., Clark, A., & Armstrong, S. (2009). An analysis of quantitative aspects in the evaluation of thematic segmentation algorithms. …of the 7th SIGdial

Workshop on ..., (July), 2009, 144–151.
Wang, A., Chen, T., & Kan, M. (2012). Re-tweeting from a Linguistic Perspective. NAACL-HLT 2012, , 2012.
Yang, J., Cai, R., Wang, C., & Huang, H. (2009). Incorporating site-level knowledge for incremental crawling of web forums: A list-wise strategy. on Knowledge, , 2009, 1375–1383.