Efficient Monitoring Algorithm for Fast News Alert

Ka Cheung "Richard" Sia
kcsia@cs.ucla.edu

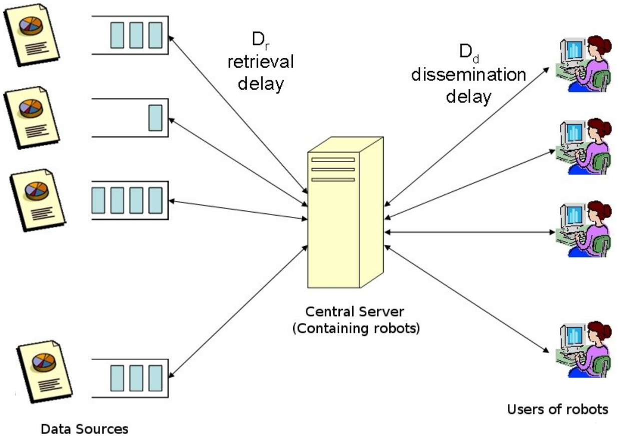
UCLA

Backgroud

- Goal
  - Monitor and collect information from the Web
  - Answer most of users' queries
- Challenges
  - Billions of pages to monitor
  - Information are updated frequently
  - Users want information fresh!

Information aggregator framework

- Server-based monitoring and dissemination



Overview

- Modeling the posting generation process
  - Definition of delay
  - Poisson process

Overview

- Modeling the posting generation process
  - Definition of delay
  - Poisson process
- Crawl scheduling
  - Resource allocation (*how often to contact?*)
  - Retrieval scheduling (*when to contact?*)

Overview

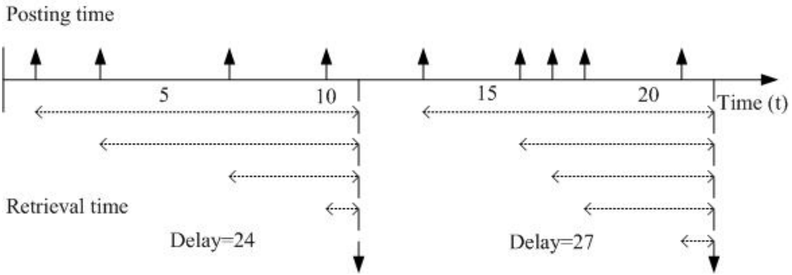
- Modeling the posting generation process
  - Definition of delay
  - Poisson process
- Crawl scheduling
  - Resource allocation (*how often to contact?*)
  - Retrieval scheduling (*when to contact?*)
- The collected data
  - ∼10k RSS (since September 2004)
  - ∼40k Weblogs (since April 2004)

New challenges

- Higher requirement on freshness
- Finer time granularity (will traditional assumption be valid?)

Terminology

- t_i - posting generation time
- τ_j - time of the j^{th} contact
- $D(O) = \sum_{i=1}^{k}(\tau_j - t_i)$, where $t_i \in [\tau_{j-1}, \tau_j]$

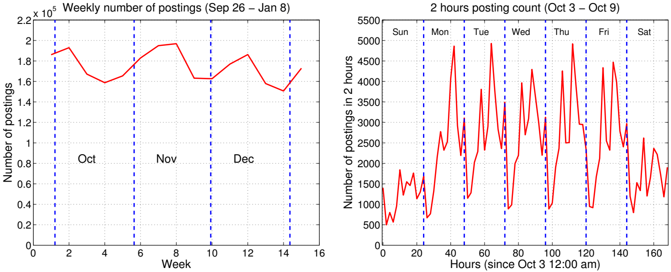


Posting generation model

- Homogeneous Poisson model
  $\lambda(t) = \lambda$ at any t
- Periodic inhomogeneous Poisson model
  $\lambda(t) = \lambda(t - nT)$, $n = 1, 2, \ldots$

Posting generation model

- Homogeneous Poisson model
  $\lambda(t) = \lambda$ at any t
- Periodic inhomogeneous Poisson model
  $\lambda(t) = \lambda(t - nT),\ n = 1, 2, \ldots$



Expected retrieval delay

- Inhomogeneous Poisson model
  rate - $\lambda(t)$
  retrieval time - τ_{j-1}, τ_j

  expected delay - $\int_{\tau_{j-1}}^{\tau_j} \lambda(t)(\tau_j - t)dt$

Expected retrieval delay

- Inhomogeneous Poisson model
  rate - $\lambda(t)$
  retrieval time - τ_{j-1}, τ_j

  expected delay - $\int_{\tau_{j-1}}^{\tau_j} \lambda(t)(\tau_j - t)dt$

- Homoegeneous Poisson model
  expected delay - $\frac{\lambda(\tau_j - \tau_{j-1})^2}{2}$

Objective

Maximize resource utlilization to provide timely informaiton.

Objective

Maximize resource utlilization to provide timely informaiton.

- Resource allocation
  How often to contact data sources?

Objective

Maximize resource utlilization to provide timely informaiton.

- Resource allocation
  How often to contact data sources?
- Retrieval scheduling
  When to contact data sources within a day?

Resource allocation

- Consider n data source $O_1, \ldots, O_n$
  - λ_i - posting rate of O_i
  - w_i - weight of O_i (how important)
  - N - total number of retrievals per day
  - m_i - number of retrievals per day allocated to O_i

Resource allocation

- Consider n data source $O_1, \ldots, O_n$
  - λ_i - posting rate of O_i
  - w_i - weight of O_i (how important)
  - N - total number of retrievals per day
  - m_i - number of retrievals per day allocated to O_i
- Optimal allocation

$$m_i \propto \sqrt{w_i \lambda_i}$$

Retrieval scheduling

m retrieval(s) per day is allocated for data source O, how should we schedule these m retrievals?

- $m = 1$
- $m > 1$

Multiple retrievals per period

m retrievals per period are allocated, when scheduled at time $\tau_1, \ldots, \tau_m$, the expected delay is:

$$D(O) = \sum_{i=1}^{m} \int_{\tau_i}^{\tau_{i+1}} \lambda(t)(\tau_{i+1} - t)dt$$

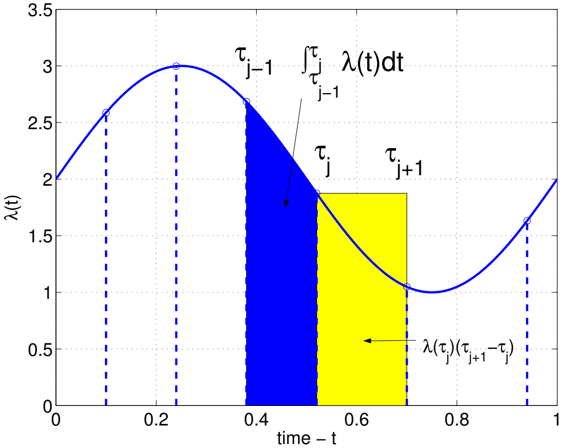
$$\tau_{m+1} = T + \tau_1$$

Criteria for optimality

$$\lambda(\tau_j)(\tau_{j+1} - \tau_j) = \int_{\tau_{j-1}}^{\tau_j} \lambda(t)dt$$

Multiple retrievals per period

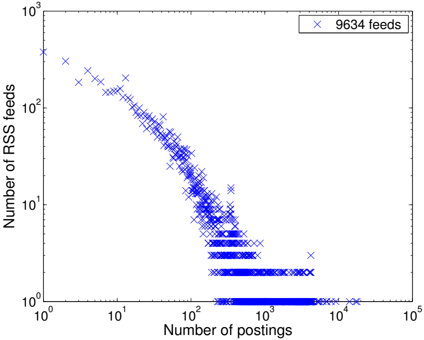
Example: $\lambda(t) = 2 + 2\sin(2\pi t)$



Experiment

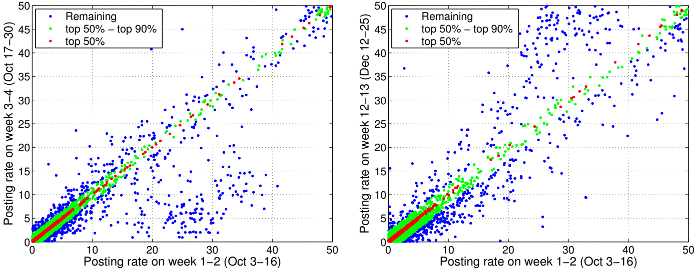
- $\sim$10k RSS feeds from Sep 21 - Dec 20 2004
- Characteristics of posting generation

Distribution of posting rate



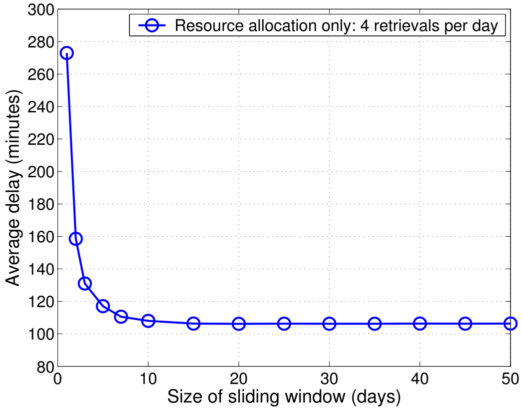
- 9634 RSS feeds are used
- Power-law distribution

Is posting rate stable and predictable?



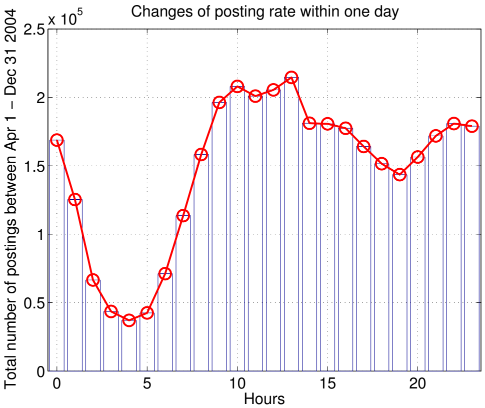
- The closer to diagonal, the more the stability and predictability
- red - top 50%, green - top 80%, blue - rest

How much history to keep?



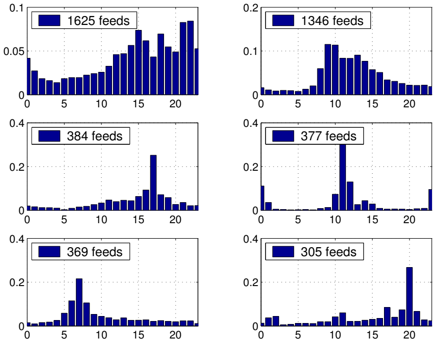
- Reallocate resource everyday
- 2 weeks is a good choice

What is the posting pattern?



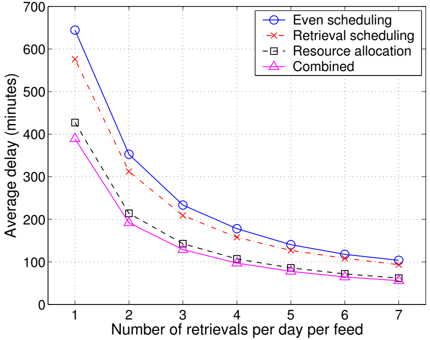
- Periodic (daily pattern)
- inactive at night

What are the individual pattern?



- K-mean clustering
- Optimize for different patterns

Performance



1. Even scheduling
2. Retrieval scheduling only
3. Resource allocation only
4. Combined

Performance

strategy	1	2	3	4
average delay (in min)	645	**581**	**433**	395
max delay (in min)	1440	**1440**	**9120**	10073
standard deviation	392	405	542	560

Statistics breakdown of posting delay using one retrieval per day.

Summary

- Efficient Monitoring
  - Resource allocation
  - Retrieval scheduling
  - $\rightarrow$ Include user access pattern (extension)
- Data
  - 1 year of weblogs and half year of RSS data
  - For prototype testing