Efficient Monitoring Algorithm for Fast News Alert

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

![Diagram of server-based monitoring and dissemination](image-url)
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?*)

The collected data

- ∼10k RSS (since September 2004)
- ∼40k Weblogs (since April 2004)
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
  - \(\sim 10k\) RSS (since September 2004)
  - \(\sim 40k\) Weblogs (since April 2004)
New challenges

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

- $t_i$ - posting generation time
- $\tau_j$ - time of the $j^{th}$ contact

$$D(O) = \sum_{i=1}^{k} (\tau_j - t_i), \text{ where } t_i \in [\tau_{j-1}, \tau_j]$$
Posting generation model

- Homogeneous Poisson model
  \[ \lambda(t) = \lambda \text{ at any } t \]

- Periodic inhomogeneous Poisson model
  \[ \lambda(t) = \lambda(t - nT), \quad n = 1, 2, \ldots \]
Posting generation model

- **Homogeneous Poisson model**
  \[ \lambda(t) = \lambda \text{ at any } t \]

- **Periodic inhomogeneous Poisson model**
  \[ \lambda(t) = \lambda(t - nT), \ n = 1, 2, \ldots \]

---

**Weekly number of postings (Sep 26 – Jan 8)**

- **2 hours posting count (Oct 3 – Oct 9)**
Expected retrieval delay

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

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 - $\tau_{j-1}, \tau_j$

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

- Homogeneous Poisson model
  expected delay - $\frac{\lambda(\tau_j - \tau_{j-1})^2}{2}$
Maximize resource utilization to provide timely information.
Objective

Maximize resource utilization to provide timely information.

- Resource allocation
  - How often to contact data sources?
Maximize resource utilization to provide timely information.

- Resource allocation
  How often to contact data sources?

- Retrieval scheduling
  When to contact data sources within a day?
Consider $n$ data source $O_1, \ldots, O_n$

- $\lambda_i$ - 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 $O_i$
Consider \( n \) data source \( O_1, \ldots, O_n \)
- \( \lambda_i \) - 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

\[
m_i \propto \sqrt{w_i \lambda_i}
\]
$m$ 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

- ~10k RSS feeds from Sep 21 - Dec 20 2004
- Characteristics of posting generation
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
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
1. Even scheduling
2. Retrieval scheduling only
3. Resource allocation only
4. Combined
## Performance

<table>
<thead>
<tr>
<th>strategy</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>average delay (in min)</td>
<td>645</td>
<td>581</td>
<td>433</td>
<td>395</td>
</tr>
<tr>
<td>max delay (in min)</td>
<td>1440</td>
<td>1440</td>
<td>9120</td>
<td>10073</td>
</tr>
<tr>
<td>standard deviation</td>
<td>392</td>
<td>405</td>
<td>542</td>
<td>560</td>
</tr>
</tbody>
</table>

Statistics breakdown of posting delay using one retrieval per day.
Summary

- Efficient Monitoring
  - Resource allocation
  - Retrieval scheduling
  - Include user access pattern (extension)

- Data
  - 1 year of weblogs and half year of RSS data
  - For prototype testing