SHARC: a Framework for Quality-Conscious Web Archiving

Dimitar Denev, Arturas Mazeika, Marc Spaniol, Gerhard Weikum
**Motivation**

**Web archives**

- **Comprehensive coverage**
- **Gold mine** for analysts
- **Data quality** is an issue
Web Warehouse
Gold Mine for Analysts

News sites

- Politologists, sociologists

Financial sites

- Business analysts

Sport sites

- Journalists

Key Players

- Internet Archive
- European Archive
- National Libraries
Web Warehouse

Poor Quality
Web Warehouse

Poor Quality

Search for a Web Site
Web Warehouse

Poor Quality

Motivation

Model

One Visit Strategies

Visit-revisit strategies

Experiments

Conclusions

A. Mazeika

SHARC: a Framework for Quality-Conscious Web Archiving
Web Warehouse
Poor Quality

Choose Time
Web Warehouse
Poor Quality

Motivation

Model

One Visit Strategies

Visit-revisit strategies

Experiments

Conclusions

Lessons learned

No instantaneous snapshots
Crawls are slow → Changes occur
Unbounded inconsistency

Contributions

First work to address data quality in web archives

Framework

Web archiving model
Optimization strategies
Web Warehouse
Poor Quality

Alemannia Aachen

Match with Munich

Motivation
Model
One Visit Strategies
Visit-revisit strategies
Experiments
Conclusions

Web archiving model
Optimization strategies

Lessons learned
No instantaneous snapshots
Crawls are slow
Changes occur
Unbounded inconsistency

Contributions
First work to address data quality in web archives

A. Mazeika SHARC: a Framework for Quality-Conscious Web Archiving
Web Warehouse

Poor Quality

Match with Munich

A. Mazeika

SHARC: a Framework for Quality-Conscious Web Archiving
Web Warehouse

Poor Quality

Alemannia Aachen - Offizielle Website - Mozilla Firefox

Motivation

Model

One Visit Strategies

Visit-revisit strategies

Experiments

Conclusions

Web archiving model

Optimization strategies

Contributions

First work to address data quality in web archives

Lessons learned

No instantaneous snapshots

Crawls are slow

→ Changes occur

Unbounded inconsistency

Lessons learned

Match with Munich

Match with Leverkusen
Motivation

Model

One Visit Strategies

Visit-revisit strategies

Experiments

Conclusions

Web Warehouse

Poor Quality

Match with Munich

X

Match with Leverkusen

A. Mazeika

SHARC: a Framework for Quality-Conscious Web Archiving
Web Warehouse

Poor Quality

Match with Munich

Match with Leverkusen

A. Mazeika

SHARC: a Framework for Quality-Conscious Web Archiving
Web Warehouse

Poor Quality

Mein Herz gehört Dir!

Yet another match (Borussia)

Match with Munich

Match with Leverkusen

A. Mazeika

SHARC: a Framework for Quality-Conscious Web Archiving
Web Warehouse

Poor Quality

Motivation

Model

One Visit Strategies

Visit-revisit strategies

Experiments

Conclusions

Web archiving model

Optimization strategies

Lessons learned

No instantaneous snapshots

Crawls are slow

→

Changes occur

Unbounded inconsistency

Contributions

First work to address data quality in web archives

Yet another match (Borussia)

Match with Munich

X

Match with Leverkusen

A. Mazeika

SHARC: a Framework for Quality-Conscious Web Archiving
Web Warehouse
Poor Quality

Lessons learned

- No instantaneous snapshots
- Crawls are slow → Changes occur
- Unbounded inconsistency

Contributions

- First work to address data quality in web archives
- Framework
  - Web archiving model
  - Optimization strategies
Outline

1. Motivation
2. Model
3. One Visit Strategies
4. Visit-revisit strategies
5. Experiments
6. Conclusions
Model

Concepts

- Change
- Page Capture
Model

Concepts

Change
Page Capture
Model

Concepts

- Politeness delay
Model

Concepts

- Politeness delay
- Capture interval
Model

The model is illustrated in the diagram, which shows the timeline of page captures and changes. The x-axis represents time, with capture intervals marked as $t_0$, $t_1$, $t_2$, $t_3$, $t_4$, $t_5$. The y-axis represents the pages, labeled as $p_0$, $p_1$, $p_2$, $p_3$, $p_4$, $p_5$. The red crosses indicate changes, and the black dots represent page captures.

Concepts

- Politeness delay
- Periodic crawls
- Capture interval

Motivation

One Visit Strategies

Visit-revisit strategies

Experiments

Conclusions
Model

Concepts

- Politeness delay
- Capture interval
- Periodic crawls
- Time travel access
Motivation

Model

A. Mazeika SHARC: A Framework for Quality-Conscious Web Archiving
**Model**

**Concepts**
- Politeness delay
- Capture interval
- Periodic crawls
- Time travel access
- Blur
**Motivation**

**Model**

**One Visit Strategies**

**Visit-revisit strategies**

**Experiments**

**Conclusions**

---

**Model**

---

**Concepts**

- Politeness delay
- Capture interval
- Periodic crawls
- Time travel access
- Blur
Model

Concepts

- Politeness delay
- Capture interval
- Periodic crawls
- Time travel access
- Blur
- Observation interval
Model (cont’d)

One visit

- Optimization criteria
  - Blur (stochastic)
  - ##pages (deterministic)

- Assumptions
  - Query time is U(0,1)
  - $p_i$ Poisson process with $\lambda_i$

Visit-revisit

- Data quality solutions
  - SHARC-offline
  - SHARC-online
  - SHARC-revisits
  - SHARC-threshold
Blur

**Definition**

\[ \text{Blur} = \int_0^T \sum_{i=1}^n \left( E \left[ \#\text{changes in } t_i, t_i \right] \right) dt \]
Blur

**Definition**

\[ \text{Blur} = \int_0^T \sum_{i=1}^n \left( \mathbb{E}\left[ \text{#changes in } t_i, t \right] \right) \, dt \]
Blur

Definition

\[
\text{Blur} = \frac{\int_0^T \sum_{i=1}^n \left( E\left[ \text{#changes in} \, t_i, t_n \right] \right) \, dt}{T}
\]
**Blur**

### Definition

\[
\text{Blur} = \sum_{i=1}^{n} \left( \# \text{changes in} [t_i, t] \right) \, dt
\]
Blur

**Definition**

\[
\text{Blur} = \sum_{i=1}^{n} \left( \# \text{changes in}[t_i, t] \right) dt
\]
Blur

Definition

\[ \text{Blur} = \sum_{i=1}^{n} \left( \# \text{changes in } [t_i, t] \right) dt \]
Blur

Definition

$$\text{Blur} = \sum_{i=1}^{n} \left( \# \text{changes in } [t_i, t] \right) dt$$
Blur

Definition

\[ \text{Blur} = \sum_{i=1}^{n} \left( \# \text{changes in}[t_i, t] \right) \, dt \]
Blur

Definition

\[
\text{Blur} = \frac{1}{T} \int_0^T \sum_{i=1}^n \left( \text{#changes in}[t_i, t] \right) \, dt
\]
**Definition**

\[
\text{Blur} = \frac{1}{T} \int_0^T \sum_{i=1}^n \left( \mathbb{E} \left[ \text{#changes in}[t_i, t] \right] \right) \, dt
\]
Blur: the Order Matters

Blur = 4

Blur = 0

Observation Interval

A. Mazeika SHARC: a Framework for Quality-Conscious Web Archiving
SHARC-offline

- Find the best sequence $p_{j_0}, p_{j_1}, \ldots, p_{j_n}$ such that

$$\text{Blur} = \frac{1}{T} \int_0^T \sum_{i=1}^{n} \left( E \left[ \# \text{ changes in } [i\Delta, t] \right] \right) dt = \frac{1}{T} \sum_{i=1}^{n} \lambda_j \omega(i)$$

- $\lambda_j$ - change rate of $p_{j_i}$
- $\omega(i)$ - penalty of position $i$
SHARC-offline

- Find the best sequence $p_{j_0}, p_{j_1}, \ldots, p_{j_n}$ such that

$$\text{Blur} = \frac{1}{T} \int_0^T \sum_{i=1}^n \left( \mathbb{E} \left[ \# \text{ changes in } [i\Delta, t] \right] \right) dt = \frac{1}{T} \sum_{i=1}^n \lambda_{j_i} \omega(i)$$

- $\lambda_{j_i}$ - change rate of $p_{j_i}$
- $\omega(i)$ - penalty of position $i$
SHARC-offline

- Find the best sequence $p_{j_0}, p_{j_1}, \ldots, p_{j_n}$ such that

$$\text{Blur} = \frac{1}{T} \int_0^T \sum_{i=1}^n \left( E\left[ \# \text{ changes in}[i\Delta, t] \right] \right) dt = \frac{1}{T} \sum_{i=1}^n \lambda_{j_i} \omega(i)$$

$\lambda_{j_i}$ - change rate of $p_{j_i}$
$\omega(i)$ - penalty of position $i$
SHARC-offline

- Find the best sequence $p_{j_0}, p_{j_1}, \ldots, p_{j_n}$ such that

$$\text{Blur} = \frac{1}{T} \int_0^T \sum_{i=1}^n \left( \mathbb{E} \left[ \# \text{ changes in}[i\Delta, t] \right] \right) dt = \frac{1}{T} \sum_{i=1}^n \lambda_j \omega(i)$$

$\lambda_j$ - change rate of $p_{j_i}$
$\omega(i)$ - penalty of position $i$
SHARC-offline

- Find the best sequence $p_{j_0}, p_{j_1}, \ldots, p_{j_n}$ such that

$$\text{Blur} = \frac{1}{T} \int_0^T \sum_{i=1}^n \left( \mathbb{E} \left[ \# \text{ changes in } [i\Delta, t] \right] \right) dt = \frac{1}{T} \sum_{i=1}^n \lambda_j \omega(i)$$

- Organ Pipes

$\lambda_j$ - change rate of $p_{j_i}$
$\omega(i)$ - penalty of position $i$
SHARC-offline

- Find the best sequence $p_{j0}, p_{j1}, \ldots, p_{jn}$ such that

$$\text{Blur} = \frac{1}{T} \int_{0}^{T} \sum_{i=1}^{n} \left( \mathbb{E} \left[ \# \text{ changes in} [i\Delta, t] \right] \right) dt = \frac{1}{T} \sum_{i=1}^{n} \lambda_j \omega(i)$$

- Organ Pipes

$\lambda_j$ - change rate of $p_{ji}$
$\omega(i)$ - penalty of position $i$

Theorem

- SHARC-offline is optimal
Motivation

Model

One Visit Strategies

Visit-revisit strategies

Experiments

Conclusions

SHARC-offline (Sharp Archiving of Web-Site Captures)

- Find the best sequence \( p_{j_0}, p_{j_1}, \ldots, p_{j_n} \) such that

\[
\text{Blur} = \frac{1}{T} \int_0^T \sum_{i=1}^n \left( E \left[ \# \text{ changes in}[i\Delta, t] \right] \right) dt = \frac{1}{T} \sum_{i=1}^n \lambda_j \omega(i)
\]

- Organ Pipes

\( \lambda_j \) - change rate of \( p_{j_i} \)

\( \omega(i) \) - penalty of position \( i \)

Theorem

- SHARC-offline is optimal
SHARC-offline \textbf{(Sharp Archiving of Web-Site Captures)}

- Find the best sequence $p_{j_0}, p_{j_1}, \ldots, p_{j_n}$ such that

\[\text{Blur} = \frac{1}{T} \int_0^T \sum_{i=1}^n \left( E \left[ \# \text{ changes in} [i\Delta, t] \right] \right) dt = \frac{1}{T} \sum_{i=1}^n \lambda_{j_i} \omega(i)\]

- Organ Pipes

- Theorem
  \textit{SHARC-offline is optimal}

- Assumptions
  \textit{All pages of the site}
SHARC-online

Input:
- O: downloaded list
- E: detected list sorted by $\lambda$s

Algorithm:
- Ascending phase
  - download the coldest
  - until $|O| + |E| > n/2$
- Balancing phase
  - download $(|O| + 1)$st coldest
  - until $|O| > n/2$
- Descending phase
  - download the hottest

Example of a Web Graph

Assumptions
- Prediction $\lambda_i$ given $p_i$
- Estimation $n$
SHARC-online

**Input:**
- O: downloaded list
- E: detected list sorted by λs

**Algorithm:**
- **Ascending phase**
  - download the coldest
  - until \(|O| + |E| > n/2\)
- **Balancing phase**
  - download \((|O| + 1)\)st coldest
  - until \(|O| > n/2\)
- **Descending phase**
  - download the hottest

**Example of a Web Graph**

**Assumptions**
- Prediction \(\lambda_i\) given \(p_i\)
- Estimation \(n\)
**SHARC-online**

**Example**

**Online organ pipes**
- Downloaded
- Detected

**Web graph**

**Algorithm**
- Ascending phase
- Balancing phase
- Descending phase
SHARC-online

Example

Online organ pipes

<table>
<thead>
<tr>
<th>Downloaded</th>
<th>Detected</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
</tr>
<tr>
<td>(\lambda_0)</td>
<td></td>
</tr>
</tbody>
</table>

Web graph

- \(\lambda_0\)
- \(\lambda_1\)
- \(\lambda_2\)
- \(\lambda_3\)
- \(\lambda_4\)
- \(\lambda_5\)

Algorithm

- Ascending phase
- Balancing phase
- Descending phase
SHARC-online

Example

Online organ pipes

<table>
<thead>
<tr>
<th>Downloaded</th>
<th>Detected</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
</tr>
<tr>
<td>$\lambda_0$</td>
<td></td>
</tr>
</tbody>
</table>

Web graph

$\lambda_0 \rightarrow \lambda_1 \rightarrow \lambda_2$

Algorithm

- Ascending phase
- Balancing phase
- Descending phase
SHARC-online

Example

Online organ pipes

<table>
<thead>
<tr>
<th>Downloaded</th>
<th>Detected</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>$\lambda_0$</td>
<td>$\lambda_4$</td>
</tr>
<tr>
<td>$\lambda_1$</td>
<td>$\lambda_3$</td>
</tr>
</tbody>
</table>

Web graph

$\lambda_0 \to \lambda_1 \to \lambda_3 \to \lambda_4 \to \lambda_5 \to \lambda_2$

Algorithm

- Ascending phase
- Balancing phase
- Descending phase
**SHARC-online**

**Example**

### Online organ pipes

<table>
<thead>
<tr>
<th>Downloaded</th>
<th>Detected</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>λ₀</td>
<td>λ₁</td>
</tr>
<tr>
<td>λ₁</td>
<td>2</td>
</tr>
<tr>
<td>λ₄</td>
<td></td>
</tr>
</tbody>
</table>

### Web graph

- \(\lambda₀\)
- \(\lambda₁\)
- \(\lambda₂\)
- \(\lambda₃\)
- \(\lambda₄\)
- \(\lambda₅\)

### Algorithm

- **Ascending phase**
- **Balancing phase**
- **Descending phase**
Motivation

Model

One Visit Strategies

Visit-revisit strategies

Experiments

Conclusions

SHARC-online

Example

Online organ pipes

<table>
<thead>
<tr>
<th>Downloaded</th>
<th>Detected</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

\[ 0 \quad 1 \quad 4 \quad 3 \quad \lambda_0 \quad \lambda_1 \quad \lambda_4 \quad \lambda_3 \quad 2 \quad \lambda_2 \]

Web graph

\[ \lambda_0 \rightarrow \lambda_1 \rightarrow \lambda_3 \rightarrow \lambda_4 \rightarrow \lambda_5 \]

Algorithm

- Ascending phase
- Balancing phase
- Descending phase
Motivation
Model
One Visit Strategies
Visit-revisit strategies
Experiments
Conclusions

**SHARC-online**

**Example**

**Online organ pipes**

<table>
<thead>
<tr>
<th>Downloaded</th>
<th>Detected</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>λ₀</td>
<td>λ₅</td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>λ₁</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>λ₄</td>
<td>λ₃</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>λ₃</td>
<td>λ₂</td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>λ₂</td>
<td></td>
</tr>
</tbody>
</table>

**Web graph**

- λ₀
- λ₁
- λ₂
- λ₃
- λ₄
- λ₅

**Algorithm**

- Ascending phase
- Balancing phase
- Descending phase
SHARC-online

Example

Online organ pipes

<table>
<thead>
<tr>
<th>Downloaded</th>
<th>Detected</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>(\lambda_0)</td>
</tr>
<tr>
<td>1</td>
<td>(\lambda_1)</td>
</tr>
<tr>
<td>4</td>
<td>(\lambda_4)</td>
</tr>
<tr>
<td>3</td>
<td>(\lambda_3)</td>
</tr>
<tr>
<td>2</td>
<td>(\lambda_2)</td>
</tr>
<tr>
<td>5</td>
<td>(\lambda_5)</td>
</tr>
</tbody>
</table>

\[\lambda_0 \rightarrow \lambda_1 \rightarrow \lambda_4 \rightarrow \lambda_3 \rightarrow \lambda_2 \rightarrow \lambda_5\]

Web graph

Algorithm

- Ascending phase
- Balancing phase
- Descending phase
Visit-revisit Strategies

- Minimize blur (two downloads for every page $p$)
- Reason deterministically about the common interval
  - $p$ is *sharp* if it did not change
  - maximize number of sharp pages ($\#\#\text{pages}$)
**SHARC-revisits**

- Minimizes blur
- Similar to the one visit case
- Two organ pipes

**SHARC-revisits schedule**

\[
\lambda_0 \rightarrow \lambda_1 \rightarrow \lambda_2 \rightarrow \lambda_3 \rightarrow \lambda_4 \rightarrow \lambda_5
\]

visits

revisits

**SHARC-threshold**

- Maximize number of pages
- Give up hopeless pages
- Bigger \( \lambda \rightarrow \) shorter interval

**SHARC-threshold schedule**

\[
p_0 \rightarrow p_1 \rightarrow p_2 \rightarrow p_3 \rightarrow p_4 \rightarrow p_5
\]
Experiments

Datasets

Data Sets

- MPI Web site: http://www.mpi-inf.mpg.de
  - 60682 pages
  - 14 daily captures
- Synthetic dataset
  - 1024 pages,
  - $\lambda_i = \lambda_{n-i} = 1/(i^{skew})$
  - $skew = 1.75$
Experiments: One Visit

Crawl Duration, MPI

Conclusions

- SHARC-online $\approx$ SHARC-offline
- SHARC better in longer crawls
Experiments: One Visit

Prediction error, synthetic dataset

Conclusions

- SHARC strategies are better for at least 50% correct change rates
## Experiments: Visit-Revisit

### pages, synthetic dataset

<table>
<thead>
<tr>
<th>Strategy</th>
<th># pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>SHARC-revisits</td>
<td>770</td>
</tr>
<tr>
<td>SHARC-threshold</td>
<td>874</td>
</tr>
<tr>
<td>Breadth-first</td>
<td>766</td>
</tr>
<tr>
<td>Depth-first</td>
<td>776</td>
</tr>
<tr>
<td>Olston</td>
<td>782</td>
</tr>
<tr>
<td>Hottest-first</td>
<td>765</td>
</tr>
<tr>
<td>Coldest-first</td>
<td>776</td>
</tr>
</tbody>
</table>

**Conclusions**

- SHARC-threshold has most sharp pages
Conclusions and Future Work

Conclusions

- Model of Web archiving
- Blur, $$\#\#$$ pages quality metrics
- SHARC framework
  - Stochastic strategies (Sociologists, historians, etc)
  - Deterministic strategies (Layers, IP specialists, etc)

Future Work

- Partial crawls when absence of changes is a must
- Combine blur with $$\#\#$$ pages