# Crawling for Temporal Information Retrieval

Gerhard Gossen

2014-11-03

# Introduction

- Today: Only looking at Web IR
- Web IR: First step is collecting documents from the Web
  - no ‘predefined’ collection
- Typically through crawlers
- In this lecture: How does time affect the crawling process?

# Overview

1. Introduction

2. Short Introduction to Crawling

3. Crawling & Time

4. Re-Crawling

5. Temporally Coherent Crawling

# Web Crawling

- Standard method for collection of Web documents
- Using a program called (web) crawler
  - also: harvester, spider, robot
- Uses hyperlinks between documents for discovery
- Store collected documents for further processing (indexing, analysis, . . . )
- Further reading: Christopher Olston and Marc Najork. "Web Crawling". In: *Foundations and Trends in Information Retrieval* 4.3 (2010), pp. 175–246. DOI: 10.1561/1500000017

# Web Crawling Algorithm

## Basic Web Crawling Algorithm

```
WHILE not done
    Get URL to fetch from queue
    Fetch URL
    Parse retrieved document
    Add new URLs from document to queue
    Store document
```

# Additional considerations

**Politeness** Only download one documents every *n* seconds (typically 5-10s) to avoid overloading servers or being blocked

**robots.txt** Obey instructions from Web sites to crawlers

**Parallel crawling** Crawler typically waits for remote servers or network, run in parallel to ensure high throughput (typically 100s of parallel threads)

**Robustness**

- Against malformed input (at every layer: network, TCP/IP, HTTP, content)
- Against spider traps (Web sites that generate infinitely many pages)
- Against spam

**Quality** Resulting collection should have good documents. Criteria are e.g. Relevance, Freshness, Diversity, . . .

# Special Considerations for Time

### What are relevant dimensions of change?

# Special Considerations for Time

### What are relevant dimensions of change?

- Changing content of already crawled pages

# Special Considerations for Time

### What are relevant dimensions of change?

- Changing content of already crawled pages
- Appearance / disappearance of pages

# Special Considerations for Time

### What are relevant dimensions of change?

- Changing content of already crawled pages
- Appearance / disappearance of pages
- Appearance / disappearance of links between pages

# Special Considerations for Time

## What are relevant dimensions of change?

- Changing content of already crawled pages
- Appearance / disappearance of pages
- Appearance / disappearance of links between pages
- Users of IR system change interests

# Content changes

- Individual pages change their content often
  - More than 40% change at least daily [CG00]
- But: No overall pattern, change occurs at different frequencies & time scales (seconds to years)

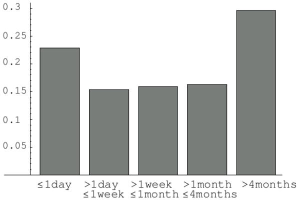


**Figure:** Change rate of Web pages

- Frequency can be modelled as a Poisson process: With

$X(t))$ number of occurrence of a change in $(0, t]$
$\lambda$ change rate

$$Pr\{X(s+t) - X(t) = k\} = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$$

for $k = 0, 1, ...$

# Change types [OP08]

Temporal behavior of page (regions) can be classified as

**static** no changes

**churn** new content supplants old content, e.g., quote of the day

**scroll** new content is appended to old content, e.g., blog entries

# The Web changes [NCO04; Das+07]

- New pages are created at a rate of 8% per week
- During one year 80% of pages disappear
- New links are created at the rate of 25% per week
  - significantly faster than the rate of new page creation
- Links are retired at about the same pace as pages

# Users change

- User interests change
- Goals of IR system maintainer change
- → requires adaptation of crawl strategy

# Re-Crawling Strategies

Crawlers need to balance different considerations:

**Coverage** fetch new pages

**Freshness** find updates of existing pages

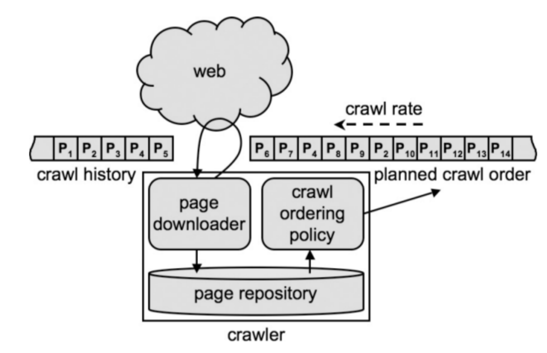


**Figure:** Crawl ordering model [ON10]

# Basic strategies

## Batch crawling

- Stop and restart crawl process periodically
- Each document is only crawled one time per crawl

## Incremental crawling

- Crawling is run continuously
- A document can be crawled multiple times during a crawl
- Crawl frequency can differ between different sites

Batch crawling is easier to implement, incremental more powerful.

# Batch Crawling Strategies

- Goal is to maximize Weighted Coverage:

$$WC(t) = \sum_{p \in C(t)} w(p),$$

with

| | |
|---|---|
| $t$ | time since start of crawl |
| $C(t)$ | pages crawled until time $t$ |
| $w(p)$ | weight of page $p$, $(0 \leq p \leq 1)$ |

- Main strategy types (ordered by complexity):
  - Breadth-first search
  - Order by in-degree
  - Order by PageRank

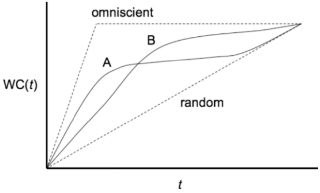


**Figure:** Weighted coverage as a function of time $t$

# Incremental Crawling Strategies

- Goal is to maximize Weighted Freshness:

$$WF(t) = \sum_{p \in C(t)} w(p) \times f(p, t),$$

  with $f(p, t)$: freshness level of page $p$ at time $t$

- Steady state average of WF:

$$\overline{WF} = \lim_{t \to \infty} \frac{1}{t} \int_{o}^{t} WF(t) dt$$

- Trade-off between coverage and freshness: Often treated as a business decision, needs to be tuned towards goals of specific application

# Maximizing Freshness [CG03]

**Model estimation** create a temporal model for each page $p$

**Resource allocation** Given a maximum crawl rate $r$, decide on a revisitation frequency $r(p)$ for each page

**Scheduling** Produce a crawl order that implements the targeted revisitation frequencies as close as possible

# Model estimation

- Create temporal model of temporal behavior of $p$
  - given samples of past content $p$ / pages similar to $p$
- Samples are often not be evenly-spaced
- Content can give hints about change frequency
  - HTTP headers, number of links, depth of page in site
- Similar pages have similar behavior
  - same site
  - similar content
  - similar link structure

# Resource allocation

## Binary Freshness model

$$f(p,t) = \begin{cases} 1 & \text{if old copy is equal to live copy} \\ 0 & \text{otherwise} \end{cases}$$

- Intuitively good strategy: proportional resource allocation
  - assign revisitation frequency proportional to change frequency
- But: uniform resource allocation achieves better average binary freshness
  - assuming equal page weights
- Reason:
  - Pages with high change frequency are stale very often regardless of crawl frequency (A)
  - Pages with lower change frequency can be kept fresh more easily (B)
  - Better to keep several pages of type B fresh than wasting resources on page of type A

# Resource allocation (continued)

## Continuous freshness model

$$age(p,t) = \begin{cases} 0 & \text{if old copy is equal to live copy} \\ a & \text{otherwise} \end{cases}$$

*a* is the amount of time between cached and live copy

- revisitation frequency increases with change frequency
  - *a* increases monotonically, crawler cannot “give up” on a page
- Instead of age, crawler can also consider content changes directly
  - distinguish between long-lived and ephemeral content

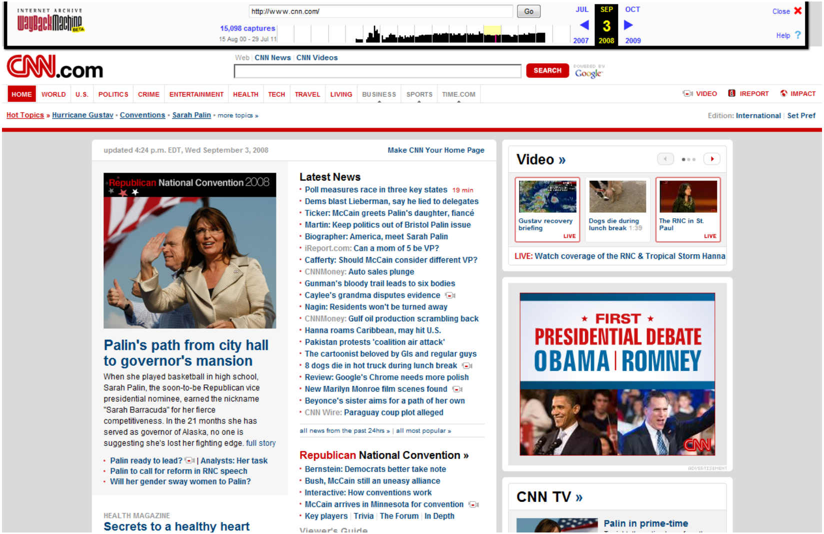
# Scheduling

Goal: Produce a crawl ordering that implements the targeted revisitation frequencies as close as possible

Uniform spacing of downloads of $p$ achieves best results.

# Temporal Coherence in Web Archives [Den+11]

- Web archives provide historical snapshots of Web pages
- Allow navigation in old versions of page
- However: Linked pages, images, scripts are crawled at different times
  - Pages show wrong images
  - Linked pages are from different points in time



INTERNET ARCHIVE WayBackMachine BETA

http://www.cnn.com/ Go

15,098 captures

15 Aug 00 - 29 Jul 11

JUL SEP OCT

2007 3 2008 2009

Close

Help ?

CNN.com

Web | CNN News | CNN Videos

SEARCH

POWERED BY Google

HOME WORLD U.S. POLITICS CRIME ENTERTAINMENT HEALTH TECH TRAVEL LIVING BUSINESS SPORTS TIME.COM VIDEO IREPORT IMPACT

Hot Topics » Hurricane Gustav • Conventions • Sarah Palin • more topics »

Edition: International | Set Pref

updated 4:24 p.m. EDT, Wed September 3, 2008

Make CNN Your Home Page

Republican National Convention 2008

## Palin's path from city hall to governor's mansion

When she played basketball in high school, Sarah Palin, the soon-to-be Republican vice presidential nominee, earned the nickname "Sarah Barracuda" for her fierce competitiveness. In the 21 months she has served as governor of Alaska, no one is suggesting she's lost her fighting edge. full story

- Palin ready to lead? | Analysts: Her task
- Palin to call for reform in RNC speech
- Will her gender sway women to Palin?

HEALTH MAGAZINE

Secrets to a healthy heart

### Latest News

- Poll measures race in three key states 19 min
- Dems blast Lieberman, say he lied to delegates
- Ticker: McCain greets Palin's daughter, fiancé
- Martin: Keep politics out of Bristol Palin issue
- Biographer: America, meet Sarah Palin
- iReport.com: Can a mom of 5 be VP?
- Cafferty: Should McCain consider different VP?
- CNNMoney: Auto sales plunge
- Gunman's bloody trail leads to six bodies
- Caylee's grandma disputes evidence
- Nagin: Residents won't be turned away
- CNNMoney: Gulf oil production scrambling back
- Hanna roams Caribbean, may hit U.S.
- Pakistan protests 'coalition air attack'
- The cartoonist beloved by GIs and regular guys
- 8 dogs die in hot truck during lunch break
- Review: Google's Chrome needs more polish
- New Marilyn Monroe film scenes found
- Beyonce's sister aims for a path of her own
- CNN Wire: Paraguay coup plot alleged

all news from the past 24hrs » | all most popular »

### Republican National Convention »

- Bernstein: Democrats better take note
- Bush, McCain still an uneasy alliance
- Interactive: How conventions work
- McCain arrives in Minnesota for convention
- Key players | Trivia | The Forum | In Depth

Viewer's Guide

### Video »

Gustav recovery briefing LIVE

Dogs die during lunch break 1:39

The RNC in St. Paul LIVE

LIVE: Watch coverage of the RNC & Tropical Storm Hanna

★ FIRST ★ PRESIDENTIAL DEBATE OBAMA | ROMNEY

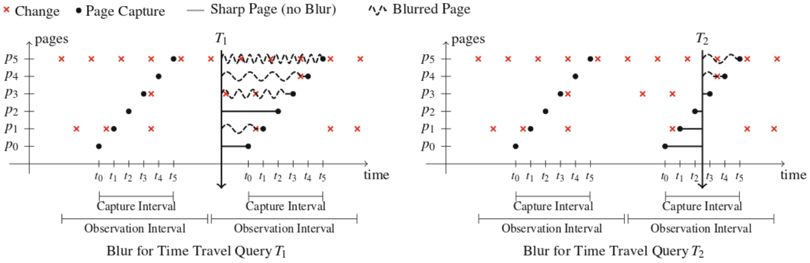
CNN

ADVERTISEMENT

### CNN TV »

Palin in prime-time

# Sharp and blurred pages



- Web Archive user accesses documents through *time-travel queries* for timepoints $T_1$ and $T_2$
- Archive is incomplete, usually retrieves documents that are temporally closest to queried time
  - *observation interval* is the time interval where a given page is returned for a query
- Retrieved pages differ from actual Web pages at that time

# Measuring Temporal Coherence

## Blur

- The blur of a Web page *pi* captured at $t_i$ is the expected number of changes between $t_i$ and query time $t$, averaged over observation interval $[0, \Delta n]$:

  $$B(p_i, t_i, n, \Delta) = \frac{1}{n\Delta} \int_0^{n\Delta} \lambda_i \cdot |t - t_i| dt = \frac{\lambda_i \omega(t_i, n, \Delta)}{n\Delta}$$

  where

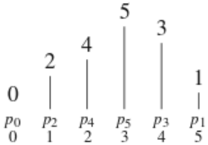
  $$\omega(t_i, n, \Delta) = t_i^2 - t_i n\Delta + \frac{(n\Delta)^2}{2}$$

  is the *download schedule penalty*.
- The blur of an Archive is the sum of the blur values of individual pages

$\omega(t_i, n, \Delta)$ can be interpreted as the penalty of downloading page $\mathrm{p_i}$ at time $\mathrm{t_i}$.

# Optimal crawl strategy

- Based on formalization of blur we can infer best crawl strategy
  - Depends on change frequency
  - Pages with highest frequency are scheduled in the middle of the crawl
  - "organ pipe" arrangement
  - Proof in paper
- Optimal strategy for known change frequencies
- Extension towards online algorithm possible



# Conclusion

1. Introduction
2. Short Introduction to Crawling
3. Crawling & Time
4. Re-Crawling
5. Temporally Coherent Crawling

# Project

## Re-crawling strategies based on content

- Different page types have different temporal behaviors
  - Home page vs blog archive vs. news feed
- Categorize a given page into such a category
- Task: Implementation and evaluation on provided test set

# References I

[CG00] Junghoo Cho and Hector Garcia-Molina. “The Evolution of the Web and Implications for an Incremental Crawler”. In: *VLDB 2000*. 2000, pp. 200–209.

[CG03] Junghoo Cho and Hector Garcia-molina. “Effective page refresh policies for web crawlers”. In: *ACM Transactions on Database Systems* 28 (2003), p. 2003.

[Das+07] Anirban Dasgupta et al. “The Discoverability of the Web”. In: *WWW '07*. 2007. DOI: 10.1145/1242572.1242630.

[Den+11] Dimitar Denev et al. “The SHARC Framework for Data Quality in Web Archiving”. In: *The VLDB Journal* 20.2 (Apr. 2011), pp. 183–207. DOI: 10.1007/s00778-011-0219-9.

# References II

[NCO04] Alexandros Ntoulas, Junghoo Cho, and Christopher Olston. "What's New on the Web?: The Evolution of the Web from a Search Engine Perspective". In: *WWW '04*. 2004. DOI: 10.1145/988672.988674.

[ON10] Christopher Olston and Marc Najork. "Web Crawling". In: *Foundations and Trends in Information Retrieval* 4.3 (2010), pp. 175–246. DOI: 10.1561/1500000017.

[OP08] Christopher Olston and Sandeep Pandey. "Recrawl Scheduling Based on Information Longevity". In: *WWW '08*. ACM, 2008, pp. 437–446. DOI: 10.1145/1367497.1367557.