Structure of the Lecture

- Introduction
  - Short Excurse: Preservation and Long-time Archiving
  - Motivation for Web Archiving
  - Web Archiving at a Glance
  - Web Archiving Challenges
- Web Archiving Methods and Technologies
- Current Research in Web archiving
Excurse: Preservation and Long-time Archiving

• purpose of preservation:*
  • to ensure protection of information of enduring value
  • for access by present and future generations

• considering long time frames (> 50 years)
• includes dealing with:
  • material deterioration (e.g. battling decay of acid-based paper, nitrate film, photos)
  • storage conditions (temperature and humidity control, disaster prevention)
  • organizational issues
  • political and management issues

Excurse: Preservation in the Digital Age

things should get easier:

• fast and lossless copying
• inexpensive storage with sinking prices
• less physical storage requirements (space)
• digitization as a means of preservation

but:

• faster media deterioration
• fast obsolescence in retrieval and playback technologies
• new challenges due to the medium
Why Archiving the Web? And why Not?

• central and growing role of Web in everyday life
  → reflection of current societies and their processes (Web as cultural heritage artifact)
  → worth preserving for future generations

however, there are also some counter-arguments:

• Is Web content worth archiving?
  • no quality control (as compared to traditional publishing)
  • ephemeral by nature
  • but: see above

• Is the Web not self-preserving for relevant content?
  • Idea: good content will stay anyway, unimportant things will disappear
  • but: long-term survival of content does also depend on organizational issues, upcoming of new content, technical environments, ...

• Is Web archiving feasible?
  • fast growth and evolution of the Web makes archiving a big challenge
  • but: Web scale solutions exist for Web search; storage still decreases in cost;
Web Archiving at a glance

based on

- (Web) search technology (crawling for building a search index)
- Internet and Web protocols (HTTP, HTML, etc.)

Basic process

Starting point: Seed list of URLs

Step 1: get Web page pointed to by first URL in seed list

Step 2: parse content and collect pointers to associated objects:
  - hyperlinks in page
  - embedded objects (images, documents)

Step 3: store content in archive (possibly after modification)

Step 4: fetch images and embedded objects and store them

Step 5: add identified links to seed list

Step 6: repeat from step 1
Web Archiving Challenges

• Content selection: What to preserve?
  • how to create the seed list
  • which links to follow
  • when to stop

• Web content acquisition: How to get the content?
  • nature of Web content (Collection of Web resources)
  • ...
Nature of Web Content

Model for Web content
- Web as a collection of Web resources
- Web resource as black box
- delivers different instantiations upon requests;
  depending on dynamic generation, session IDs, request parameters, cookies, etc.

Resulting Web archiving challenges
- large, potentially unlimited number of instantiations
- Web resource cannot be directly archived

Solutions:
- archiving of samples (impression of the Web resource to user)
- Hidden Web archiving (see later slides)
Web Archiving Challenges

- Content selection: What to preserve?
  - ...

- Web content acquisition: How to get the content?
  - ...
    - dealing with heterogeneous, evolving and complex content types (e.g. videos, streaming, active content)
    - dealing with embedded applications
    - dealing with the Hidden Web
    - archiving Social Web content
    - recognition of duplicates
    - copyright and privacy
    - capturing change in pages
Web Archiving Challenges – cont.

• Archive content storage and organization
  • How to store the collected content
  • managing different snapshots

• Web archive quality
  • avoiding spam and redundancy
  • archive completeness
  • achieving snapshot coherence

• Access and long-term usability
  • Web archiving user interfaces
  • dealing with evolution
  • long-term usability
Structure of the Lecture

• Introduction
  • Short Excurs: Preservation and Long-time Archiving
  • Motivation for Web Archiving
  • Web Archiving at a Glance
  • Web Archiving Challenges

• Web Archiving Methods and Technologies
  • Archiving Method Classification
  • Archiving the Hidden Web
  • Web Archive Access

• Current Research in Web archiving
Web Archiving Technologies and Methods

- No single Web archiving methods that is adequate for the full variety of Web publishing settings and type of Web Archive
- Variety of Web archiving methods exist

- Classification of Archiving Methods:
  - Acquisition Approach
    - Client side Archiving
    - Transaction Archiving
    - Server-side archiving
  - Organization & Storage
    - Local file system served Archives
    - Web served Archives
    - Non Web Archives
  - Crawling Strategy
    - Intensive archiving
    - Extensive Archiving
  - Archive Scope
    - Site-centric Archive
    - tropic-centric Archive
    - Domain-centric Archive
Classification I: Acquisition approach

- acquisition = technical means used to get the content into the archive
- Most common method: Client-side archiving (e.g. Heritrix Crawler)
- for Web server the archiving crawler is a client like any other
- Web pages are fetched via HTTP and stored; links are extracted to find further related pages (see also slide 6);
- based on adapted crawling technology from Web search engines

Advantages:
- simplicity and scalability
- close to how the user sees the Web
- re-use of existing technology (with adaptation)

Disadvantages/Challenges:
- difficulties with hidden Web capturing
- special heuristic methods required for extracting dynamically generated links, links in scripts, links in code, links other media types (incomplete, high adaptation costs)
- problems with authentication, complex request parameters, etc.
- overload of Web servers (politeness rules) has to be avoided
Classification I: Acquisition approach

Alternative Method: Transaction archiving
• inserting a listener for Web traffic of the Web site to be archived (e.g. Page Vault System)
• archiving all unique request and response pairs (request sent by user + page/content delivered)

Advantages:
• archives all seen Web resource instantiations (also including hidden Web content)
• best fit for internal Web archiving

Disadvantages/Challenges:
• requires agreement and collaboration of server’s owner (scalability!)
• adequate methods for deciding about unique and duplicate content still required
Classification I: Acquisition approach

Alternative Method: server-side archiving
- directly copy files, data structures etc. from the server (without using http)

Advantage:
- archiving/copying process is relatively simple
- can help in archiving resource that are not easily (or not at all) accessible to crawlers (see Hidden Web)

Disadvantages/challenges:
- requires collaboration with site owners (lack of scalability to general Web content)
- difficult to make the Web source run again in the archive environment (system dependencies)
Web Archiving Method

Acquisition Approach
- Client side Archiving
  - Transaction Archiving
  - Server-side archiving
- Organization & Storage
  - Local file system served Archives
  - Web served Archives
  - Non Web Archives
- Crawling Strategy
  - Intensive archiving
  - Extensive Archiving
- Archive Scope
  - Site-centric Archive
  - tropic-centric Archive
  - Domain-centric Archive
Classification II: Organization & Storage

Local File system served archives

- create a copy of the Web site’s files and structure in the local file system (“file” prefix)
- navigate like in the Web
- see e.g. HTTrack tool

Advantages:

- easy to implement method
- use of standard browser for Web archive access
- low entrance barrier for Web archive operation

Disadvantages/challenges

- replacement of absolute by relative path required, creation of new names for dynamically created content;
- limitations of hierarchical structure: no direct systematic support for versions of sites, temporal access (crucial for Web archives)
- limitations of file systems for very large numbers of files (Web archives may contain billions of files)

→ adequate for institutional to corporate site archiving, not to be used for middle to large scale Web archives
Classification II: Organization & Storage

Web served Archives

• Web pages are stored as they are crawled in a container file plus further metadata (standard: WARC file);

• additional infrastructure for accessing Web archive:
  • Index structure for translating URL into container file offset for direct access
  • Web server for answering requests
  • methods for re-directing links within the archived page to point into the archive again (possible solutions: script in page or use of proxy)

Advantages:

• scalability (proven for 500 Terabyte Web archives in Wayback machine)
• higher faithfulness to original (no renaming, no changes of links)
• easier to support temporal aspects, migration and archive content delivery (compared to local file system)

Disadvantages/Challenges

• additional infrastructure required
• dynamically created links and scripts may lead out of the archive environment
  → adequate for medium to large Web archives, also usable for small archives
Classification II: Organization & Storage

Just for completeness: Non Web Archives

• archiving in forms that do not rely on hypertext, e.g. creating a PDF document from a Web site

• mainly used for formats that have not been originally created in the Web context, e.g. publication catalogues
Classification III: Archiving Strategy

• basis: identified links can point a) within same Web site, b) to new site
• typically a perimeter is given for limiting overall depth of crawling

Intensive archiving:
• preference for following links within single Web sites (depth first search)
• aims for vertical completeness
• adequate especially for Site-centric archiving

Extensive archiving:
• preference for covering many different sites, deep covering of individual sites secondary (breadth first search)
• aims for horizontal completeness
• adequate especially for topic-centric archiving (used e.g. in Internet Archive)
Classification IV: Web Archiving Scope

• Site-centric archiving
  • archiving an individual Web site
  • increasingly important for Web sites of companies and large organizations

• Topic-centric archiving
  • archiving of relevant Web content related to one topic e.g. a research topics, an election process, etc.
  • manual or semi-automatic selection of relevant sites/pages: e.g. via set of experts

• Domain-centric archiving
  • use of upper level domains of DNS to select content: e.g. .jp, .de, .gov or second level domains
  • for larger more systematic archives
  • easy selection criterion for crawling
Web Archive Quality

Quality factors

• completeness:
  • according to defined goals (intensive vs. extensive archiving, specified perimeter)
  • capturing of embedded objects, identified links
• ability to render the original form (navigation, user interaction)
• snapshot coherence:
  • politeness rules: imposing fixed delay between subsequent requests
  • slows down archiving process (up to several days)
  • may lead to incoherent site archives
  • methods for analyzing and improving coherence required (see current research part)
Structure of the Lecture

• Introduction
  • Short Excursus: Preservation and Long-time Archiving
  • Motivation for Web Archiving
  • Web Archiving at a Glance
  • Web Archiving Challenges

• **Web Archiving Methods and Technologies**
  • Archiving Method Classification
  • Archiving the Hidden Web
  • Web Archive Access

• Current Research in Web archiving
Archiving the Hidden Web (aka.
part of the Web
important example for archiving: document or image
collections or
borderline not conceptual, but depending on technology (see example link detection in Flash)
large part of the overall Web
→ archiving of the difficult
Archiving the Hidden Web – Methods

1. Client-side archiving (only partially possible)

Method Overview

- detect relevant HTML forms (search forms)
  - use heuristics to distinguish from other types of forms
- extract and interpret query fields
  - rely on typical layouts to find labels
  - compare with known labels for interpretation
  - use of regularities in forms (e.g. frequently used attributes such as title, keyword, price)
- learn to fill them in and fetch resulting content
  - generation of requests with good coverage (e.g. time periods)
  - use of fields with limited domains (e.g. Zip codes, dates)
  - use of vocabularies learned from other contexts (author lists, keyword lists, etc.)
  - use of first query results for generating further queries (query-based sampling)
- approach limited in case fields are too open or undefined
Archiving the Hidden Web – Methods cont.

2. Crawler-Server Collaboration

- idea: content provider provides additional means to enable crawling (archiving) of hidden content

**Methods:**

- **Hidden link pages:**
  - pages with links to all individual objects in the collection
  - adequate robot directives e.g. "noindex, follow"
  - requires adequate linking schema for objects of collection → crawling by standard technology, also indexing for Web search

- **Standardized access services and protocols:**
  - exposes collection metadata via HTTP using XML syntax
  - can communicate with OAI server
  - also supports delivery of metadata, collection listings, querying by date, etc.

- **drawback:** OAI-MHP has to be implemented by content provider

---

Web Archiving, November 30, 2010

29
Archiving the Hidden Web – Methods cont.

3. Sever-side archiving
   • focus on creation of rich archives
   • actions from content provider required

Possible Method (used by Bibliotheque Nationale de France)
starting point: collection to be archived, metadata database with information describing the collection objects
   • mapping of metadata database to schema supported by the archive (possibly tool supported)
   • creation of an XML version of the metadata database based on mapping
   • adaptation of linking schema metadata \(\rightarrow\) digital object
   • storage of XML version and collection objects
   • inclusion of an HTML form to query the collection (ensuring accessibility)
Structure of the Lecture

- Introduction
  - Short Excursus: Preservation and Long-time Archiving
  - Motivation for Web Archiving
  - Web Archiving at a Glance
  - Web Archiving Challenges
- Web Archiving Methods and Technologies
  - Archiving Method Classification
  - Archiving the Hidden Web
  - Web Archive Access
- Current Research in Web archiving
Web Archives Access Example: WayBack machine

- browser for the content archived by the Internet Archive (15 Billion pages)
- online available at: http://www.archive.org/

- given an URL shows the archived versions of the site in a time line
- considered time range can be restricted
### Archived Results from Jan 01, 1996 - latest

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2 pages</td>
<td>7 pages</td>
<td>7 pages</td>
<td>22 pages</td>
<td>25 pages</td>
<td>68 pages</td>
<td>21 pages</td>
<td>5 pages</td>
<td>0 pages</td>
<td>1 pages</td>
<td>0 pages</td>
<td>5 pages</td>
<td>5 pages</td>
<td>3 pages</td>
</tr>
</tbody>
</table>

* denotes when site was updated.

Material typically becomes available here 6 months or more after collection, with some exceptions. See FAQ.
Example:
www.stern.de

December 21, 1996

- stern Cockpit
- applet no longer running
Example:
www.stern.de

February 8, 1999

- still missing pictures
- part of the links is not working
Example:
www.stern.de

- August + October, 2009
- mixed quality
Structure of the Lecture

• Introduction

• Web Archiving Methods and Technologies

• Current Research in Web archiving
  • Current Research Projects
  • Web Spam
  • Terminology Evolution
  • Temporal Coherence
  • Research Papers
Current Research in Web Archiving

• Web archiving is still relatively new area
• requires a lot of engineering as well as research

Examples Projects

• European project ARCOMEM (to be started in January of 2011)
• European project LiWA (Living WEB Archives)
ARCOMEM

From Collect-All Archives to Community Memories – Leveraging the Wisdom of the Crowds for Intelligent Preservation

- large European Project on Web Archiving in the context of the Social Web
- collaboration with Yahoo! Research, European Archive, University of Trento, University of Southampton, University of Sheffield, SWR, Deutsche Welle, Austrian and Greek Parliament
- Start in January 2011

Goals:
- use Wisdom of the Crowds as well as relationships to entities and events to decide upon what should go into the archive
- enrich archive content with information on events, entities and information gathered from the Social Web to go from archives to Community memories
- enable “by example” content selection for archives as well as collaborative archive creation inspired by mechanisms of the Social Web
ARCOMEM – cont.

Research topics:
- Social Web analysis and Web mining
- advanced crawling techniques
- event detection and consolidation
- perspective, opinion, and sentiment detection,
- approaches for “semantic” preservation
- …

Two applications:
- social web archiving for broadcasters
- social web archiving for political discussions
Motivation

Role of Web:

- providing information and services for seemingly all domains
- reflecting all types of events, opinions, and developments within society, science, politics, environment, business, etc.
- giving room for the articulation for a multitude of stakeholders

→ Archiving this quickly changing multifaceted information space has becomes a relevant issue for cultural heritage

Web archiving imposes various challenges:

- Inherent ephemeral character
- Hidden Web
- Social Web
- Preservation
- New types of content
- Change & Evolution

...
LiWA Goal
Next generation Web Archiving technology for:
- high Quality Web Archives
- long-term Archive usability

→ From Web page storage to “Living Web Archives“
LiWA Objectives: Archive Fidelity

Next generation Web Archiving methods and tools:

- enhancing Archive Fidelity and authenticity by
  - capturing all types of content
  - capturing of Hidden Web
  - detecting traps
LiWA Objectives: Archive Fidelity

Next generation Web Archiving methods and tools:
• enhance Archive Fidelity and authenticity
  ▫ capture all types of content
  ▫ detect traps
  ▫ filtering Web spam
  ▫ filtering noise
LiWA Objectives: Archive Coherence

Next generation Web Archiving methods and tools:

• enhance Archive Fidelity and authenticity
• improve Archive Coherence and Integrity
  ▫ deal with issues of temporal Web construction
  ▫ identify, analyse and repair temporal gaps
  ▫ consistent Web archive federation
LiWA Objectives: Archive Interpretability

Next generation Web Archiving methods and tools:

- **enhance Archive Fidelity and authenticity**
- **improve Archive Coherence and Integrity**
- **facilitate (long-term) Archive Interpretability**
  - dealing with terminology evolution
  - handling semantic evolution
  - preparing for evolution aware access support
LiWA modules in Web archiving workflow
Structure of the Lecture

• Introduction
• Web Archiving Methods and Technologies
• Current Research in Web archiving
  • Current Research Projects
  • Web Spam
  • Terminology Evolution
  • Temporal Coherence
  • Research Papers
Web spam: for (or against) search engines

Compute the out degree

On the Feasibility of Low-rank Approximation for Personalized PageRank

File Format: PDF/Adobe Acrobat - View as HTML transition matrix of the Web graph for computing personalized PageRank. Out-degree. Hence the base of links ...

http://www.ilab.sztaki.hu/~stamas/publications/benczur05low_rank_ppr.pdf Cached - Similar pages

Schools for pharmacy phh mortgage cendant songs ring tones community credit union houston philadelphia penn s settlement hawaii insurance commissioner debt coverage ratios auto loan refinance classic video games online wh health insurance long beach schools financial credit union insurance umbrella policy disaster unemployment insur mag mutual insurance company debit & credit chevron gas credit card money affiliate car loan application paradis casino photos progressive insurance claims office halloween bingo sheet binion world poker open pharmacy mass services credit union mortgage rates outlook cover insurance arts administration degree credit counseling governr lose weight casino star odds against 7 even party poker ipo

Compute the out d4egree compute thne out degree compute the out degree compute the 0ut degree compute the out degree compute the out degree compute the out degree compute the out degree compute the out degree compute the out degree compute the out degree compute the out degree compute the out degree compute the out degree compute the out degree compute the out degree compute the oujt degree compute the oujt degree compute the out degree compute the out degree compute the out degree compute the out degree compute the out degree compute the out degree compute the out degree compute the out degree compute the out degree compute the out degree compute the out degree compute the out degree compute the out degree compute the ou
Web Spam: indexing vs. archiving

Primary target: search engines, manipulate ranking
As side effect, we also archive spam
But very costly if:

- traps crawler
- 10+% sites
- near 20% HTML pages

not fought against:

- Reputable 70.0%
- Spam 16.5%
- Non-existent 7.9%
- Ad 3.7%
- Weborg 0.8%
- Unknown 0.4%
- Alias 0.3%
- Empty 0.4%

2004 .de crawl courtesy: T. Suel
Filter technology: Know your neighbor

Honest pages rarely point to spam
Spam cites many, many spam

1. Predicted spamicity $p(v)$ for all pages
2. Target page $u$, new feature $f(u)$ by neighbor $p(v)$ aggregation
3. Reclassification by adding the new feature

Siklósi, Benczúr et al.,
Web Spam Hunting @ Budapest, AIRWeb 2008
Structure of the Lecture

• Introduction
• Web Archiving Methods and Technologies
  • Current Research in Web archiving
    • Current Research Projects
    • Web Spam
    • Terminology Evolution
    • Temporal Coherence
    • Research Papers
Terminology evolution in Long Term Archives

St. Piter Burh → St. Petersburg → Petrograd → Leningrad → St. Petersburg


St. Piter Burh (1703)
St. Petersburg (1703-1914)
Petrograd (1914-1924)
Leningrad (1924-1991)
St. Petersburg (1991-present)
Process for automatic detection of evolution

Terminology Extraction

- Co-occurrence Graph
- Clustering to find word sense

Detection of word senses from one collection at one period in time

Detection of term evolution

- Cluster Evolution
- Term Evolution

- Term Concept Graph

- Part of speech tagging
- Lemmatization
- Stopword removal
- Dictionary

- Sliding window
- Grammatical relations
- Graph

- Different clustering techniques
- Different similarity measures
- Clusters / Concepts
Structure of the Lecture

• Introduction
• Web Archiving Methods and Technologies
  • Current Research in Web archiving
    • Current Research Projects
    • Web Spam
    • Terminology Evolution
    • Temporal Coherence
    • Research Papers
Motivation

“Easy” for a human to recognize (in-)coherence
“Tough” for a machine to evaluate (in-)coherence (immediately)

- Requires semantic analysis of contents
- Reliable last-modified stamps

[cf. Spaniol et al.: “’Catch me if you can’: Visual Analysis of Coherence Defects in Web Archiving”, IWAW 2009]

⇒ Double access of contents for coherence analysis

Reference:
- as of 13/02/2007
- as of 29/01/2007
- as of 17/02/2007
- as of 19/02/2007
Best-Effort Coherence by Example

Observation Interval

Crawl Interval

Observation Interval

Blur = 5

Blur = 2

Blur = 1
Best-Effort Coherence by Example

Observation Interval

Crawl Interval

Observation Interval

p₁

Blur = 4

p₂

Blur = 1

p₃

p₄

Blur = 0

p₅
Structure of the Lecture

• Introduction
• Web Archiving Methods and Technologies
• Current Research in Web archiving
  • Current Research Projects
  • Web Spam
  • Terminology Evolution
  • Temporal Coherence
• Research Papers
Papers:


Spam is difficult to detect automatically, but humans are quite good at it.

Idea: Start with a small, human generated set of good pages and propagate the trust of this set using a pagerankish algorithm.

Basic assumption: Good pages point mostly only to good pages, but rarely to bad ones.
Query expansion of named entities (i.e. persons, roles, ...) can be employed in order to increase retrieval effectiveness.

There are time dependent and time independent synonyms for such entities.

On monthly snapshots of wikipedia do:
1. Named entity recognition and synonym extraction specific for Wiki
2. Improving time of synonyms using a model for temporal...

Web pages have to be crawled in a “polite” manner, so crawling can take weeks.

SHARC assumes change rates of Web pages can be statistically predicted based on page types, directory depths, and URL names.

Presents four strategies to achieve an optimal download schedule to maximize “sharpness” of the crawls.
Thanks!