Web Archiving

Andreas Rauber
Department of Software Technology and Interactive Systems
Vienna University of Technology
http://www.ifs.tuwien.ac.at/~andi

Motivation

- Unisono:
  Internet ist DIE Basis für Informationsaustausch, Kommunikation, Dokumentation, Kulturgut, wissenschaftliche Informationen, …
- aber
  Information im Internet ist flüchtig
  durchschnittliche Lebensdauer von Webdokumenten nur wenige Tage bis Wochen
- Digital Dark Age?
- Webarchiving zum Aufbau von Sammlungen des WWW

Überblick

- Webarchiving-Initiativen
- Methoden & Tools
  - Datensammlung
  - Speicherung, Verwaltung und Zugriff
  - Tools und Services
- Ethische Aspekte
- Forschungsfragen und weiterführende Infos
- Zusammenfassung

Webarchiving-Initiativen

- Internet Archive
  - www.archive.org
  - erstes Webarchive im Jahre 1996
  - ursprünglich nur html/txt Seiten
  - später Bilder, Videos, andere Inhalte
  - mittlerweile auch Spezial-sammlungen
  - Daten primär von Search Engine Crawler (Alexa)
    --> nicht "archival quality"
  - weltweit, shallow

Webarchiving-Initiativen

- KulturarW3
  - kw3.kb.se
  - ebenfalls ab 1996
  - nationales Web Archiv: .se
  - "vollständige" "Snapshots"
  - Datensammlung ursprünglich mit Combine Crawler
  - Archivierung auf Tape Roboter

Webarchiving-Initiativen

- Pandora
  - pandora.nla.gov.au
  - Seit 1996
  - manuelle Sammlung von Internetpublikation
  - ursprünglich rein manueller Sammlungsansatz
  - mittlerweile ergänzt durch Crawling
  - derzeit ca.
    14.500 Publikationen (1.6TB)

Webarchiving-Initiativen

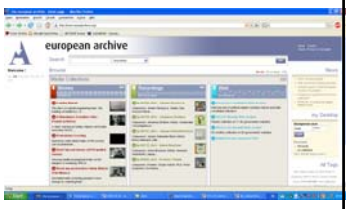
- AOLA - Austria On-line Archiv
  - www.ifs.tuwien.ac.at/~aola
  - 2000/2001
  - nationales Webarchiv, (unvollständige) Snapshots als Pilotstudie
  - Vergleich von Nedlib und Combine Crawler
  - kurzfristig 2-größtes europäisches Webarchiv
  - Start: 1. 1. 2008 basierend auf neuen Technologien und breiten Sammlungsstrategien

Webarchiving-Initiativen

- Netarchive.dk
  - www.netarchive.dk
  - ab 2005
  - nationales dänisches Webarchiv
  - Kooperation Royal Library und Staatsbibliothek
  - Heritrix Crawler
  - ca. 700.000 Domänen, 36TB Daten (2007)

Webarchiving-Initiativen

- European Web Archive
  - www.europarchive.org
  - Amsterdam, Paris
  - Heritrix Crawler
  - 250 TB storage
  - frei zugänglich
  - enge Kooperation mit dem Internet Archive
  - Crawls on-demand für Italien (snapshot 2006), UK Government pages,…



Webarchiving-Initiativen

- International Internet Preservation Coalition (IIPC)
  - netpreserve.org
  - gegründet 2003
  - Gruppe von Nationalbibliotheken
  - Koordinierung der Forschungs- und Entwicklungsaktivitäten
  - Entwicklung z.B. eines eigenen Archival Crawlers (Heritrix)
  - ARC / WARC Format als (de-facto) Speicherungs-Standard

Webarchiving-Initiativen

- Weitere Initiativen
  - mittlerweile unzählige nationale Aktivitäten
  - zahlreiche Sondersammlungen (Wahlen, Sportevents, regionale Sammlungen, Katastrophen, ...)
  - Fast alle europäischen Länder, USA, Australien, Singapore, Japan, ...)
  - Schwerpunkt meist auf Sammlung und Organisation
  - Wenig Aktivitäten zur Nutzung (rechtliche Probleme)
  - Kaum Aktivitäten zur Langzeitbewahrung

Überblick

- Webarchiving-Initiativen
- Methoden & Tools
  - Datensammlung
  - Speicherung, Verwaltung und Zugriff
  - Tools und Services
- Ethische Aspekte
- Forschungsfragen und weiterführende Infos
- Zusammenfassung

Sammlungsstrategien

- Sammlung von Daten aus dem Web
- Crawling / Harvesting vs. Submission
  - Crawling: Verfolgen von Linkstrukturen ausgehend von Seed-URLs
  - Submission: manuelles Einpflegen bzw. Abgabe-Interface
- Verschiedene Strategien
  - manuelle Sammlung
  - Submission
  - Snapshot Crawling
  - Event Harvesting / Focused Crawls
  - Selective Harvesting
  - Sonderformen (Datenbanken, Interaktive Elemente)

Sammlungsstrategien

- Manuelle Sammlung / Submission
  - Bestimmte Dateien werden gezielt ins Archiv aufgenommen
  - z.B. Publikationen, Videos, Nachrichtenartikel
  - Webseiten aus Datenbanken hinter Query-Interface (Deep Web)
  - Eingepflegt in manuell gepflegten Datenbestand
  - hohe Qualität bei Metadatenerfassung
  - nur geringer Umfang
  - Kernfrage: was ist wichtig im Web?

Sammlungsstrategien

- Snapshot Crawls
  - verbreiteste Strategie
  - Crawling z.B. des gesamten nationalen Webspace
  - Frage: was ist der "nationale Webspace"?
    - nationale Domäne
    - Rechner im Land (.com, .net, .org, ...)
    - Website mit Inhalten, die ein Land betreffen („Austriaca")
  - Snapshot = Langzeitbelichtung (Crawldauer: mehrere Monate)
  - 1-4 Snapshots pro Jahr
  - Herausforderungen -> Überwachung/Kontrolle
    - Spidertraps
    - Seiten mit nicht automatisch verfolgbarer Linkstruktur (Flash)
    - Fehlerhafte Seiten
    - Deep Web
    - Seiten, die Funktionen auslösen
    - Robot Exclusion Protokolle und Passwörter
    - enorme Datenmengen (TB-Bereich)

Sammlungsstrategien

- Continuous Crawling
  - ähnlich Snapshot Crawls
  - aber: Seiten werden öfter als einmal pro Crawl besucht und bei Veränderung des Inhalts erneut ins Archiv übernommen (ähnlich Suchmaschinen)
    - bei jedem Antreffen eines Links
    - entsprechend Änderungshäufigkeit
    - Priorisierung möglich
  - Hohes Datentransfervolumen
  - Herausforderungen -> Überwachung/Kontrolle
    - identisch mit jenen für snapshot crawls
    - zusätzlich: wann hat sich eine Seite verändert? (Inhalt vs. Werbebanner vs. Visitor-Counter,...)

Sammlungsstrategien

- Event Harvesting / Focused Crawls
  - Snapshot Crawling für fokussierte Subsets des Web
  - Ereignisse, Katastrophen, Wahlen, Sportevents, ...
  - Regionale oder anderweitig thematische Sammlungen
  - kleine, kontrollierte Sammlungen von Seed-URLs, erstellt teilweise unter Verwendung von Suchmaschinen
  - diese dann täglich/wöchentlich gesammelt
  - teilweise im vorhinein geplant (Wahlen, Hochzeiten, Sport,..) teilweise tagesaktuell bei Bedarf initiiert (9/11, dän.Karikaturenstreit)
  - Tunneling & fortgeschrittene IR-Technologien

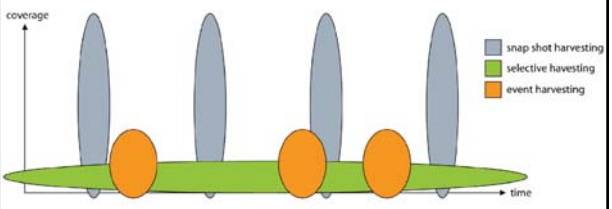
Sammlungsstrategien

- Selective Harvesting
  - Sonderform des Focused Crawls
  - Sammlung einer einzelnen, spezifischen Website in regelmäßigen Abständen
  - hauptsächlich für Periodika
  - spezifische Anpassungen zum Data Cleansing
  - qualitativ sehr hochwertige Sammlung, inkl. Metadaten
  - HTTRACK

Sammlungsstrategien

- Sonderformen
  - spezielle Lösungen für spezifische Aufgaben
  - z.B. Extraktion von Datenbank-Informationen in XML
  - z.B. Session Filming
  - Ergänzung bzw. spezielle Websammlungen
- Kombinationsstrategien
  - von den meisten Archiven angewandt
  - meist Snapshot Crawling (z.B. 1-2 mal pro Jahr)
    + Focused Crawls für events (3-6 pro Jahr)
    + Selective Harvesting für Publikationen
    + Manual Collection evtl. für spezifische Webinhalte

Sammlungsstrategien



- Kombination der einzelnen Harvesting-Arten zum Aufbau eines Webarchivs
  (Abb. aus: Bjarne Andersen: The DK-domain: in words and figures. Technical Report, Februar 2006.
  http://netarchive.dk/publikationer/DFrevy_english.pdf)

Speicherung,Verwaltung und Zugriff

- Speicherung
  - ARC / WARC format
  - ISO TC46/SC4 bearbeitet dzt. Standardisierung
  - XML Containerformat, beinhaltet jeweils 100MB an Daten
  - Speicherung auf hochperformanten Clustern („Petabox“)
  - Back-ups auf Tape
  - früher Cluster von low-cost PCs - Verwaltung zu komplex
  - Verwaltung
- meist Eigenentwicklungen
  - Anpassung an Speicher-Infrasktruktur
  - Lastverteilung der Crawler
  - Indizierung
  - Netarchive-SW seit Juli 2007 verfügbar

Speicherung,Verwaltung und Zugriff

- Zugriff
  - die wenigsten Archive bieten derzeit öffentlichen Zugang
  - Ausnahme: Internet Archive, European Web Archive: Wayback Machine
  - Eingabe einer URL, Auflistung der verfügbaren Seiten in Zeitleiste
  - Copyright, Privacy
  - Wera, NutchWAX zur Volltextsuche

Softwaretools

- Heritrix
  - crawler.archive.org
  - archival quality crawler
  - entwickelt vom Internet Archive in Kooperation mit dem IIPC
  - GNU public license
  - hochgradig skalierend, stabil
  - Weiterentwicklung mit Einbau höherer Intelligenz
    - Erkennung von Duplikaten
    - flexiblere Gestaltung der Crawls

Softwaretools

- Web Curator Tool - WCT
  - webcurator.sourceforge.net
  - entwickelt von Sytec Resources, im Auftrag der British Library und der Nationalbibliothek von Neuseeland
  - Web-Interface zu Heritrix
  - Speziell für Selective Harvesting und Focused Crawls
  - Erstellung von thematischen Listen von Websites

Softwaretools

- Netarchive Suite
  - netarchive.dk/suite
  - entwickelt seit 2004, seit Juli 2007 Open-source verfügbar
  - Dient zur Planung und Durchführung von Harvestingaktivitäten mit Heritrix
  - unterstützt bit-level preservation (redundante Speicherung und Prüfung)

Softwaretools

- NutchWAX
  - www.nutch.org
  - entwickelt vom Nordic Web Archive, Internet Archive und IIPC
  - Indexer für ARC Dateien
  - erlaubt Volltextsuche im Internetarchiv
- Wayback Machine
  - archive-access.sourceforge.net/projects/wayback
  - Navigation im Webarchiv auf Zeitleiste
  - java-basiert
- Wera
  - archive-access.sourceforge.net/projects/wera
  - php-basiertes Interface, das auf NutchWAX aufbaut
  - ähnlich Wayback Machine

Überblick

- Webarchiving-Initiativen
- Methoden & Tools
  - Datensammlung
  - Speicherung, Verwaltung und Zugriff
  - Tools und Services
- Ethische Aspekte
- Forschungsfragen und weiterführende Infos
- Zusammenfassung

Ethics & Web Archiving

- Web archiving is an essential activity to ensure valuable content is being preserved
- Web Archives contain a wealth of extremely valuable information

But:

- Currently most archives are closed to public
- Mostly due to legal reasons
- Need a legal solution

Is this all?

(based on: Rauber Andreas, Kaiser Max und Wachter Bernhard: Ethical Issues in Web Archive Creation and Usage – Towards a Research Agenda. In: Proceedings of the 8th International Web Archiving Workshop, Aalborg, Dänemark. 2008.)

Ethics & Web Archiving

- What should such a legal solution look like?
- Is it only a legal problem?

- There are things that are legal, but ethically dubious
- There are things that are illegal, but ethically acceptable

- Privacy is an essential good
- Most countries are increasingly privacy-aware
- Are there ethical concerns, and if so
  - Are we aware of them?
  - Can we do something to address them?
- Is there a danger of a moratorium on Web Archiving?

Ethics & Web Archiving

- Hypothesis: there are a number of potentially ethically sensitive issues related to Web Archiving, and particularly to access provision

- In order to be able to unlock the value of Web Archives we need to
  - understand them
  - try to address them
  - and, provided we have appropriate technical solutions, have them reflected in the legal regulations
- Try to start this process via some (incomplete) considerations

Overview

- Introduction: do we have ethical issues?
- Assumptions underlying & motivating Web Archiving
- Research questions to assist in solving a privacy dilemma
- A quick glimpse at a case study: automatically identifying private content
- Conclusions and next steps

Assumptions underlying Web Archiving

Assumptions and a number of questions:

- The Web is a new publication medium?
- The ephemeral nature of Web pages is a “design fault”?
- A Web Archive is merely a collection of publicly available information

Assumptions underlying Web Archiving

- The Web is a new publication medium?
  - Are people “publishing” (conscious decision, effort invested,…)
  - If so, are they aware of it?
  - Are kids allowed to publish?
  - Which parts of the Web are publishing, which are communication? (ako chatting-in-the-bus?)
  - Do we have a choice of NOT putting some things on the Web?

Assumptions underlying Web Archiving

- The ephemeral nature of Web pages is a “design fault”?
  - Post-it notes are based on a “faulty” glue -> should we put real glue onto them?
  - If the Web is a publication medium: may there be some who use it as such BECAUSE it is ephemeral? (art, temporary announcements, CV, …)
  - Does being ephemeral make it more a communication medium in the perception of some people?
  - Does society need en ephemeral way of communicating with larger communities in an ephemeral manner? (speaker’s corner, graffitti, …)

Assumptions underlying Web Archiving

- A Web Archive is merely a collection of publicly available information
  - True, but what about Holism? (The whole is more than the sum of it’s parts)
  - Does the ease of use, or the new possibilities of use, change the nature of an information collection? (full-text search, semantic analysis, IR as opposed to conventional archive catalogs)
  - Specialized person profile search engines, used by HR departments (special profile generation services to counter-act this)
  - Technical possibilities will increase in the future (video analysis, semantic analysis, reasoning, …)

Overview

- Introduction: do we have ethical issues?
- Assumptions underlying & motivating Web Archiving
- Research questions to assist in solving a privacy dilemma
- A quick glimpse at a case study: automatically identifying private content
- Conclusions and next steps

Research questions

- What are the *ethical constraints*, and how they can be more precisely *defined or formalized*,
- Which *approaches* users of Web archives with *potentially dubious intentions* might employ to obtain information that should not be provided by privacy-respecting archives,
- In how far *technological solutions* such as query analysis, machine learning and data mining can help in identifying potentially harmful queries, potentially incriminating content on Web pages, information worth of protection, or combinations thereof,
- How *legal regulations* might be formulated in order to allow (partial) access to Web archive content in a save, ethically correct, and useful manner

Research questions

- Formalizing ethical constraints and risks
  - Risk analysis and threat scenarios
  - Types of information
  - Types of creators
  - Types of Web usage
  - Aggregation of information
  - Risk of NOT providing certain information

Research questions

- Understanding potentially dubious usage
  - Who may be misusing a Web Archive?
  - For what purpose may it be misused?
  - How would the misuse look like?
  - What is the potential damage?
  - Can we detect it before misuse happens?
  - Can we put up barriers to misuse – and what would be the impact of these on normal usage
  - Is the analysis of usage ethically questionable as well?

Research questions

- Technological approaches to solutions
  - Content mining to identify sensible information? (sites, pages, paragraphs, text tokens, …)
  - Query analysis to identify patterns?
  - Site analysis to identify creators & purpose of creation? (children, private comment / communication, ephemeral information)
  - Technical means of controlling access and means of access?

Research questions

- Legal solutions
  - Which legal solutions are also technically feasible?
  - How can policies be formulated, enacted and controlled?
  - How good a solution is “good enough” to be ethically acceptable, and thus legally safe to specify and act upon?

Overview

Case Study

- Identifying potentially private information segments
- Proof of concept, meta-information
- NOT as basis for limiting/blocking access or excluding from archive, etc.
- Approach:
  - Take text documents
    - Pages
    - Paragraphs
  - Identify which ones are potentially private
  - Train a classifier (SVM, Bayesian Networks, …)
- Similar to Genre Classification

Case Study

Two text collections:

- Santini 7-genre corpus:
  - 1400 documents (200 per genre, balanced corpus)
  - *Blog, Personal Homepage,* E-Shop, FAQ, Listings, Newspaper Frontpage and Search Page
- Product review pages
  - 50 pages amazon top-100-books 2007
  - 20 pages amazon video reviews
  - Analyzed on paragraph level:
    - description / review
    - description + professional review / "private" review
  - Using model trained on Santini 7-genre-corpus

Case Study

Features:

- Text statistics (averages)
  - # words, +syllables, word-length
- Tokens (rel. frequ.)
  - word-, symbol-, space-, number-, punctuation-, control-tokens
- Readability Indices
  - Flesh Reading Ease, Flesh-Kincaid Grade Level
- Look-ups (rel. frequ.)
  - places, time/date, person names, currencies, addresses, company names, abbreviations, telephone/fax numbers
- Part-of-Speech (rel. frequ.)
  - nouns, verbs, adjectives, adverbs, modal verbs, conjugations, pronouns, personal pronouns, articles, prepositions, exclamations, list elements
- Presentation (rel. frequ.)
  - links, headings, paragraphs, text formatting, lists, tables, graphics, frames, forms, multimedia elements

Case Study

Results:

- Classification Santinin 7-genre corpus using SVM

Genre	Precision	Recall	F2
Blog	93,17	87,22	90,01
Homepage	89,59	48,33	62,79
Private total	91,38	67,78	76,40

- Classification product pages, Naïve Bayes, 7-g model

Corpus	Precision	Recall	F2
top-100-books	77,91	84,07	80,87
top-100-books-b	78,27	85,75	81,84
video games	87,66	96,02	91,65

Case Study

Results:

Conclusions

- Web Archives contain valuable information that would be very useful to make freely available to the public
- We do believe there are ethical issues with (providing access to) Web Archives
- We need a legal solution to these problems
- For a legal solution we need an understanding of
  - The problems and challenges
  - Risks and their consequences
  - Technical means to automatically counter these and to enforce legal regulations
- Otherwise we may face the risk of having a moratorium / legal regulation stopping Web Archiving all together….
  ….or at least feel very bad continuing doing it

Forschungsfragen

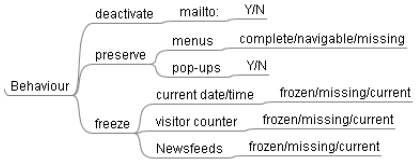
- Data und Crawl Management
- Spider traps, non-traditional links, Social Web
- Management der Datenvolumina, Back-up
- Indexing und Suche
- Ethische Aspekte
- **Preservation**

Forschungsfragen

- Preservation
  - derzeit fast nur Datensammlung
  - Bit-stream Preservation, kaum logische Preservation
  - Datenvolumen
  - Anzahl verschiedener Fileformate, Korrektheit
  - Significant Properties? Authenticity?
    (Wie sieht eine Webseite aus?)
  - Bereitschaft, gewissen Verlust zu tolerieren

Forschungsfragen

Bsp.: Preservation – Web Archives - Behaviour



- Visitor counter and similar functionalities can be
  - Frozen at harvesting time
  - Omitted
  - Remain operational, i.e. the counter will be increased upon archival calls
    (is this desired? count? demonstrate functionality?)

Weiterführende Infos

- IWAW: International Web Archiving Workshop
  - www.iwaw.net
  - jährliche Fachtagung mit aktuellen Forschungsarbeiten sowie Best-Practice Präsentationen
  - seit 2001, meist in Kombination mit ECDL,
    2007 gemeinsam mit JCDL
  - voraussichtlich 2. Oktober 2009, Korfu, Griechenland

Weiterführende Infos

- Brewster Kahle. Preserving the internet. *Scientific American*, March 1997.
- Adrian Brown: Archiving Websites: A Practical Guide for Information Management Professionals. Facet Publishing. Juli 2006.
- Julien Masanes (Editor): Web Archiving. Springer Verlag. September 2006.
- Andreas Rauber, Hans Liegmann: Web-Archivierung zur Langzeiterhaltung von Internet-Dokumenten. In: Heike Neuroth (Editor): nestor-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung, 2. Ausgabe, erscheint Sommer 2009

Zusamenfassung

- Durchschnittliche Lebensdauer eines Web-Dokuments nur wenige Tage bis Wochen
- Rettung und Bewahrung der Daten - Digital Dark Age?
- Unterschiedliche Strategien zur Sammlung
  -> meist Kombination
- Mittlerweile zahlreiche Tools verfügbar
- sorgfältige Planung - erhebliche Größe!
- Schwerpunkt der Aktivitäten derzeit:
  Sammlung, nicht Langzeitarchivierung
- Verantwortungsbewusster Umgang mit Daten notwendig

Übung

- **Aufgaben**
  - Welche Metadaten würden Sie bei einer archivierten Netzressource erfassen? Welche davon könnten automatisiert erfasst werden?
  - Was sehen Sie als Gesichtspunkte der Authentizität von Netzressourcen an? Wie kann diesen Rechnung getragen werden?
  - Sie archivieren das Internetangebot Ihrer / einer Institution. Würden Sie grundsätzlich das ganze Angebot bewahren oder eventuell auch nur einen Teil davon? Warum? Welche Kriterien Sie bei der Wahl der Zeitintervalle heranziehen?
  - Definieren sie eine (kombinierte) Sammlungsstrategie für den Aufbau eines nationalen Webarchivs