

Creating Permanent Test Collections of Web Pages for Information Extraction Research

Bernhard Pollak & Wolfgang Gatterbauer

presentation by Bernhard Pollak

for the SOFSEM 07

January 2007

It's about ...

Motivation

Problem

Solution

Experiments

*Creating Permanent Test Collections of Web Pages
for Information Extraction Research*



Saving Web Pages

Why saving web pages?

Motivation

Problem

Solution

Experiments

⇒ We **want to**:

- *reproduce*
- *share*
- *compare*

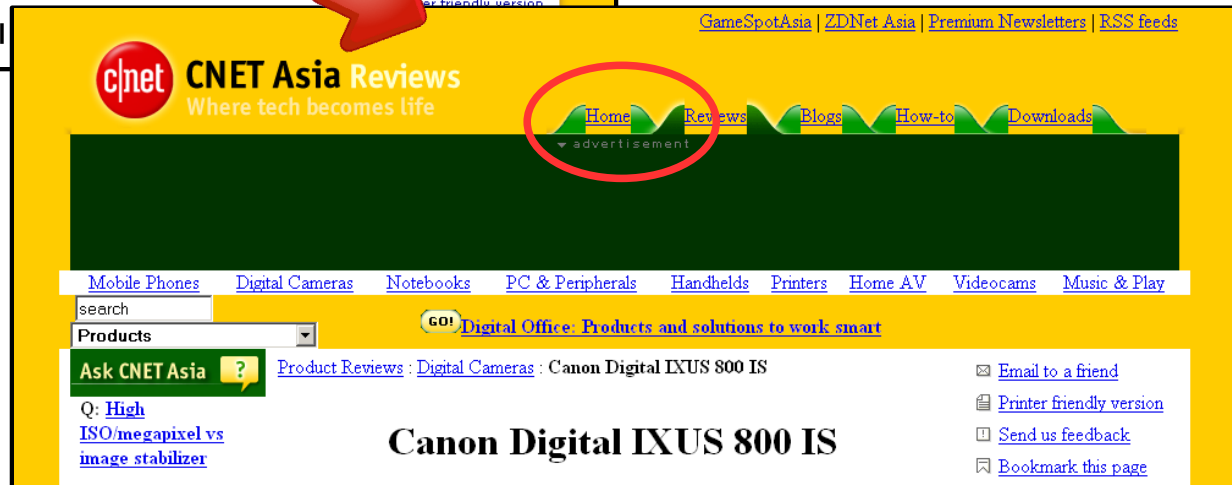
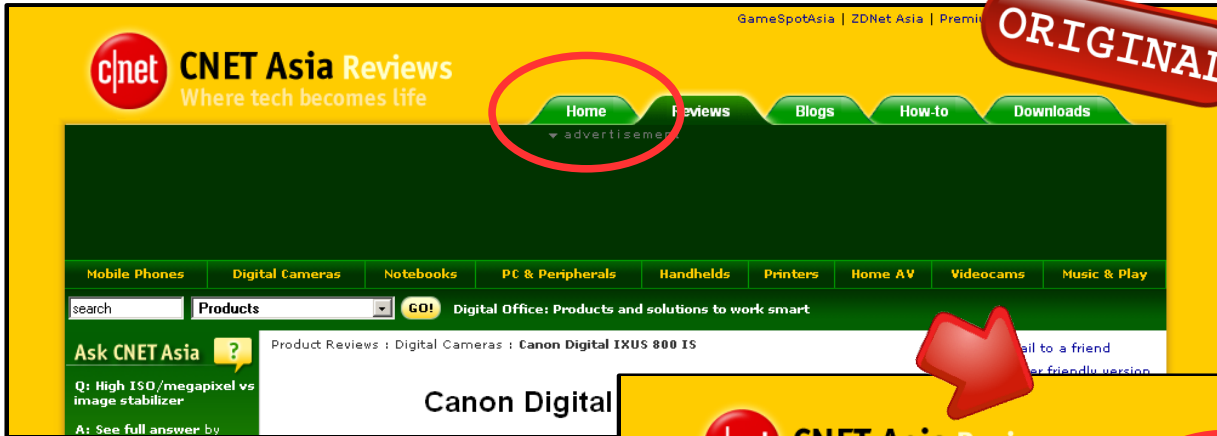
⇒ *AND* because of **visual** web page IE:

- **visual exact** reproduction

404 error: File not found
The URL you requested was not found.

Firefox 2.0 "Save Complete"

Motivation Problem Solution Experiments



visual exact ?

Web Site Downloader: HTTrack 3.33

Motivation Problem Solution Experiments

The image shows a side-by-side comparison of the Expedia.de website. The left side shows the original website with a navigation menu at the top. A red arrow points to a missing menu element in the right side. A red stamp with the word 'ORIGINAL' is placed over the top right of the left screenshot. A blue box at the bottom left contains the text 'The menu is completely lost!'. The right side shows a version of the website where the menu is missing, and the search results are partially obscured by a semi-transparent overlay.

ORIGINAL

Es stehen keine Angebote auf Ihrem [Merkzettel](#)

1. Suche 2. Reiseziel 3. Hotelauswahl

4. Wählen Sie den Reiseternin für dieses Hotel

Mallorca - Spanien

Sie können die Liste sortieren nach:

Abflughafen	Hinreise	Dauer	Anbieter
München	Sa. 10.02	14 Tage	alltours

Anzahl Reisende: 2 Erwachsene, 1 Kind, 2. Kind

13 - 16 Tage, Hotelkategorie: beliebig

Suche ändern

Es stehen keine Angebote auf Ihrem [Merkzettel](#)

1. Suche 2. Reiseziel 3. Hotelauswahl

4. Wählen Sie den Reiseternin für dieses Hotel

Mallorca - Spanien

Sie können die Liste sortieren nach:

Abflughafen	Hinreise	Dauer	Anbieter	Leistun
München	Sa. 10.02	14 Tage	Veranst Hotelinfos	Doppel Halbpe
Amsterdam (NL)	Sa. 03.02	14 Tage	Veranst Hotelinfos	Doppel Halbpe

Personal Web Archive: WebCite

Motivation Problem Solution Experiments

VOR ←

Wo wollen Sie heute hin?
Von Ort:
Nach Ort:

Fahrplanauskunft - Verkehrsverbund Ost-Region

Start Stadt/Gemeinde

- Haltestelle
- Straße/Hausnummer
- Wichtiger Punkt

Ziel Stadt/Gemeinde

WebCite
[What's this?]

Showing WebCite

Wo wollen Sie heute hin?
Von Ort:
Nach Ort:

Verkehrsverbund Ost-Region (VOR)
Gesellschaft m.b.H.
Mariahilfer Straße 77-79
1060 Wien
Tel.: +43(0)1 526 60 48
e-Mail: office@vor.at

neusiedlŽmobil - nežmo startet!
Seit 12. Dezember 2006 fahren in Neusiedl am See ein Stadtbus ergänzt durch ein
- das neue Angebot wurde neusiedlŽmobil, kurz nežmo benannt.

How to address frame state in URL?

Personal Proxy: WWWOFFLE 2.8e

Motivation Problem Solution Experiments



The image shows a screenshot of the Sacher Online Shop website. The website header includes navigation links: "Hotel Sacher Wien", "Hotel Sacher Salzburg", "Sacher Cafes", and "Original S". The main content area features a large "ONLINE SHOP" banner with a decorative background. Below the banner, there are three product categories: "ORIGINAL SACHER-TORTE", "GLASS / PORCELAIN", and "SOU". A red arrow points from the "ORIGINAL" stamp to the "WWWOFFLE - World Wide Web Offline Explorer - v2.8e" error message box. The error message box contains the following text:

Serverfehler

Der WWWOFFLE Server hat einen Fehler entdeckt:
SSL proxy connection while offline is not allowed.
Das Programm kann diese Anfrage nicht weiter bearbeiten.

WWWOFFLE Fehler, der eigentlich nicht auftreten darf. Es gibt in dieser Situation keine Möglichkeit (ausser...
ren. Falls dieser Fehler mehrfach auftritt und WWWOFFLE korrekt konfiguriert ist und Sie ansonst...
r, willkürliche Systemabstürze, Festplattenfehler, etc.) bitte melden Sie diesen Fehler und die Umstã

HTTPS problem!

The Solution: Saving based on DOM Tree

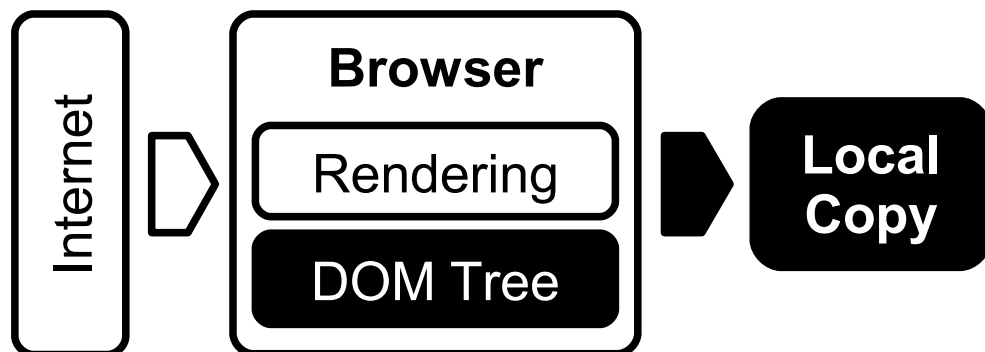
Motivation

Problem

Solution

Experiments

- ⇒ **no conceptual problems**
- ⇒ *a client centric approach*
- ⇒ *dynamic changes are considered*
- ⇒ *decision on visual relevant objects*



```
#document
  HTML
  HTML
    HEAD
    BODY
      #text
      TABLE
        #text
        TBODY
          TR
            #text
            TD
              #text
            TD
            TD
            #text
```


Our Solution: WebPageDump

Motivation

Problem

Solution

Experiments

- ⇒ based on the *Scrapbook DOM Saver* extension
- ⇒ adapted for IE researchers needs
- ⇒ AND: ***improved quality of local copy***

But: **How** to measure the **visual exactness** of web pages?

Experimental Setup

Motivation Problem Solution Experiments

⇒ Image Comparison

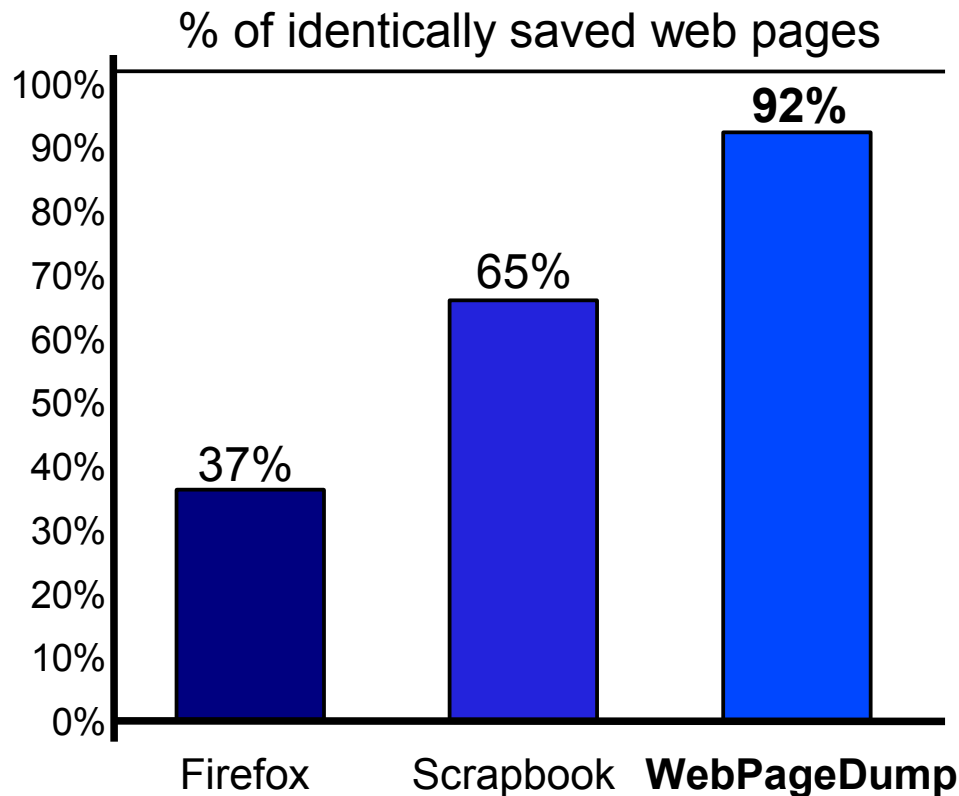
- using *web page image*
- PNG Format (compressed)

⇒ Strong correlation

- between *pixel differences* and *compressed file size*

The screenshot shows the 'webpagedump' project website. It features a navigation bar with links for 'introduction', 'using', 'limitations', and 'publication'. The main content is divided into four sections: 'introduction', 'using', 'limitations', and 'publication'. The 'introduction' section describes the tool as a Firefox extension for saving web pages. The 'using' section lists various command-line options for saving and processing pages. The 'limitations' section discusses the tool's reliance on DOM tree processing and its handling of certain HTML elements. The 'publication' section includes a disclaimer and an abstract of a research paper.

Experimental Results



⇒ 450 web pages

⇒ Tested domains:

- arabic
- chinese
- various languages
- random
- digital cameras

reduction of faults
from 35% to 8%

Thank you for your attention

Paper:

Bernhard Pollak, Wolfgang Gatterbauer. Creating Permanent Test Collections of Web Pages for Information Extraction Research. In Proceedings of SOFSEM 2007 Student Research Forum, January 2007.

WebPageDump:

<http://www.dbai.tuwien.ac.at/user/pollak/webpagedump>

Additional References:

HTTrack: <http://www.httrack.com>

WWWOFFL: <http://www.gedanken.demon.co.uk/wwwoffle>

WebCite: <http://www.webcitation.org>

Scrapbook: <http://amb.vis.ne.jp/mozilla/scrapbook>