Archive Fidelity

The first problem area is the archive’s fidelity and authenticity to the original. Fidelity comprises the ability to capture all types of content. Current crawlers fail to capture all Web content, because the current Web includes much more than simple HTML pages as dynamically created pages, multimedia content that is delivered using media-specific streaming protocols; hidden Web content that resides in data repositories and content-management systems behind Web site portals.

In this context LIWA develops a new crawler based on executing pages. This method requires three specificities:

The first is to run an execution environment (HTML plus JavaScript, Flash etc.) in a controlled manner so that discoverable links can be extracted systematically.

The second challenge is to encapsulate these headless browsers in a crawler-like workflow, with the purpose of systematically exploring all the branches of the web graph.

The last but not the least of the challenges, is to optimize this process so that it can scale to the size required for archiving sites.

This three challenges have been implemented in the new crawler that one of the partner has developed (Hanzo Archives Ltd) and it is already used in production by them to archive a wide range of sites that can’t be archived by pre-existing crawlers, as well as in testing by another of the LiWA partner, the Internet Memory Foundation.

Some tests on small scale have been made to compare results between 2 methods. Analyses show significant improvements in the quality of the crawls. It is worth noting that these improvements are obtained on small-scale crawls. Indeed, the link extractor increases processing time, however it is largely compensated by the fact that it saves human operator’s time.

The screenshot below shows the difference between the 2 methods. A visual control shows clearly that one capture is more complete than the other one.
Archiving Rich Media Content

As part of the new technologies for Web archiving developed in the LiWA project a specific module was designed to enhance the capturing capabilities of the crawler for different multimedia content types specifically when served via streaming.

The LiWA Rich Media Capture module delegates the multimedia content retrieval to an external application (such as MPlayer or FLVStreamer) that is able to handle a larger spectrum of transfer protocols than Heritrix.

The module is constructed as an external plugin for Heritrix. Using this approach, the identification and retrieval of streams is completely decoupled, allowing the use of more efficient tools to analyze video and audio content. At the same time, using the external tools helps in reducing the burden on the crawling process.

At IM, the LIWA video capture module is used on a daily basis to fetch video served with streaming server. It has significantly improved the quality of archiving for video-centric sites (like broadcasters and TV sites for instance) but also for mainstream sites that use video hosting services (like Youtube).

An example is presented below with a conference video between James Cameron (UK Prime Minister) and Mark Zukerberg (founder of Facebook) on Youtube.

The screenshot compares the archive and the online video. On left online video and on right the archive video, the only difference is the archive’s video player.

How can an existing, quality focus, web archiving workflow be improved?

The National Library of the Czech Republic (NLP) has been building its web archive since 2000. The archive is focused on “bohemical” web resources, i.e. websites that are related to the Czech Republic or its people.

The NLP strives to control the quality of archived websites from selective harvests. This manual, laborious and time-consuming process, often refers to quality assurance or simply QA. It basically requires visual inspection of all harvested websites by the project staff.

Archive coherence has brought a surprising and unexpected by-product in this respect: the results of a crawl’s temporal coherence analysis can be used to generate a graphical representation of a website in the form of nodes representing individual pages and links between them (see Figure 4).

The colors of the nodes indicate whether pages have changed during the crawl (or, more precisely, between the initial crawl and a re-crawl).

Some of these graphs are in fact little pieces of web art, but more importantly, they have a potential to reveal some irregularities or problems, such as crawler traps or missing pages. In addition, the distribution of the nodes’ colour in the graph can indicate a rough estimate of the rate of change of the website.

Using the graphs for QA could bring some benefits as they can alert curators to the existence of crawler traps and other quality issues, which are easy to miss during manual QA. These Graphs are inserted into the QA workflow, thus allowing the curators a quick visual inspection. Clicking on the graph will bring up an interactive version of it from Figure 5, which the graph can be navigated and zoomed. Hovering over a node displays the node’s URL.

LiWA Tools on Open source

LiWA released most of its developed components and tools in open-source: LiwaCurvatureModule, Semantic Analyzer, rich- media-capture-plugin.

All LiWA softwares are grouped under the “liwa-technologies” project on Google code: http://code.google.com/p/liwa-technologies/.
Internet archives are becoming more and more concerned about spam in view of the fact that, under different measurement and estimates, roughly 10% of the Websites and 20% of the individual HTML pages constitute spam. The objective is to reduce the amount of fake content the archive has to deal with. The toolkit helps prioritize crawls by automatically detecting content of value and exclude artificially generated manipulative and useless content.

With the illustration over 100,000 pages WEBSPAM-UK2007 data along with 7 previous monthly snapshots of the .uk domain, we have investigated the tradeoff between feature generation and spam classification accuracy. We proposed graph similarity based on temporal features, which aim to capture the nature of linkage change of the neighborhood of hosts. Our features achieve better performance than previously published methods; however, when combining them with the public link-based feature set we get only marginal performance gain.

By our experiments it has turned out that the appropriate choice of the machine learning techniques is probably more important than devising new complex features. We have managed to compile a minimal feature set that can be computed incrementally very quickly to allow intercepting spam near crawl time. Our results open the possibility for spam filtering practice in web archives that are mainly concerned about their resources waste and would require fast reacting filters.

To continue

Although LiWA is first and foremost a research project (the first of its kind in Web Archiving in the world), the results are already very valuable to practitioners of the field.

During the LiWA project many new approaches have been developed to address major issues in Web archiving and archive accessibility. LiWA can be seen as a starting point for a number of new activities in the field of Web archiving and Web preservation. The sheer size, complexity and dynamics of the Web make high quality Web archiving still an expensive and time-consuming challenge. Therefore new crawling strategies are necessary that focus on content completeness in term of opinions, topics or entities etc..

The new Integrated Project AR-COMEM (From Collect-All Archives to Community Memories) leverages the Social Web for content appraisal and selection.

Beside preservation, a deeper understanding of the Internet content characteristics (size, distribution, form, structure, evolution, dynamic) is also necessary in many areas of today’s science. The European funded project LAWA (Longitudinal Analytics of Web Archive data) builds an experimental testbed for large-scale data analytics. Its focus is on developing a sustainable infrastructure, scalable methods, and easily usable software tools for aggregating, querying, and analysing heterogeneous data at Internet scale. Particular emphasis is given to longitudinal data analysis along the time dimension for Web data that has been crawled over extended time periods.

Dissemination results

Documentation and scientific publications have been submitted to high-level conferences, workshops and publishers in the field. One of the challenges for LiWA was that Web Archiving is a new field of research and had no established regular scientific publisher. In order to structure this, each year, a dedicated session for the LiWA project has been organized at the International Web Archiving Workshop, a workshop with a traditionally strong practical orientation (as opposed to a scientific one).