The ARCOMEM Approach for Social and Semantic Driven Web Archiving

Thomas Risse¹, Wim Peters², and Pierre Senellart³

¹ L3S Research Center, Hannover, Germany, risse@L3S.de
² University of Sheffield, UK, w.peters@dcs.shef.ac.uk
³ Institut Mines–Télécom; Télécom ParisTech; CNRS LTCI, Paris, France, pierre.senellart@telecom-paristech.fr

Abstract. The constantly growing amount of Web content and the success of the Social Web lead to increasing needs for Web archiving. These needs go beyond the pure preservation of Web pages. Web archives are turning into “community memories” that aim at building a better understanding of the public view on e.g. celebrities, court decisions and other events. Due to the size of the Web, the traditional “collect-all” strategy is in many cases not the best method to build Web archives. In this paper we present the ARCOMEM architecture that uses semantic information such as entities, topics, and events complemented with information from the social Web to guide a novel Web crawler. The resulting archives are automatically enriched with semantic meta-information to ease the access and allow retrieval based on conditions that involve high-level concepts.

1 Introduction

Given the ever increasing importance of the World Wide Web as a source of information, adequate Web archiving and preservation has become a cultural necessity in preserving knowledge. Due to the sheer volume of content, taking arbitrary snapshots is not an appropriate solution. Instead it is necessary to build focused archives with valuable content in an efficient way.

A pivotal factor for enabling next-generation Web archives is crawling. Crawlers are complex programs that nevertheless implement a simple process: follow links and retrieve Web pages. In the ARCOMEM approach, however, crawling is much more complex, as it is enriched with functionality dealing with novel requirements. Instead of following a “collect-all” strategy, archival organizations are trying to build community memories that reflect the diversity of information people are interested in. Community memories largely revolve around events and the entities related to them such as persons, organizations, and locations. Archives can also be built based on content types, types of Web applications, or content creation times. Social Web content allows in addition to select content based on age or gender of persons (in case the information is provided).

Current Web crawler technology is mainly inspired or based on crawlers for search engines. Therefore they have limited or no notion of topics, entities, events,
or the Social Web context. In this article we want to give an overview about
the approach and architecture of the ARCOMEM Web crawler which addresses
the special needs of Web archiving organizations. This new crawler generation
will analyze and mine the rich social tapestry of the Social Web to find clues for
deciding what should be preserved (based on its reflection in the Social Web), to
contextualize content within digital archives based on their Social Web context,
and determine how to best preserve this context. Contextualization based on the
Social Web will be complemented by exploring topic-centered, event-centered,
and entity-centered processes for content appraisal and acquisition, as well as for
rich preservation.

2 Approach & Architecture

The goal for the development of the ARCOMEM crawler architecture is
to implement a socially aware and semantic-driven preservation model. This
requires thorough analysis of the crawled Web content. Since a thorough analysis
of all Web content is time-consuming, the traditional way of Web crawling and
archiving is no longer working. Therefore the ARCOMEM crawl principle is to
start with a semantically enhanced crawl specification that extends traditional
URL-based seed lists with semantic information about entities, topics, or events.
It also allows to focus the crawl by other properties like content type, Social Web
user properties, etc. This crawl specification is complemented by a small reference
crawl to learn more about the crawl topic and intention of the archivist. The
combination of the original crawl specification with the extracted information from
the reference crawl is called the *intelligent crawl specification*. This specification, together with relatively simple semantic and social signals, is used to guide a broad crawl that is followed by a thorough analysis of the crawled content. Based on this analysis a semi-automatic selection of the content for the final archive is carried out.

The translation of these steps into the ARCOMEM system architecture foresees the following five processing levels:

**Crawling Level:** At this level, the system decides and fetches the relevant Web objects as those initially defined by the archivists, and later refined by both the archivists and the online processing modules. The crawling level includes, besides the traditional crawler and its decision modules, some important data cleaning, annotation, and extraction steps.

**Online Processing Level:** The online processing is tightly connected with the crawling level. At this level a number of semantic and social signals such as information about persons, locations, or social structure taken from the intelligent crawl specification are used to prioritize the crawler processing queue. Due to the near-real-time requirements, only time-efficient analysis can be performed, while complex analysis tasks are moved to the offline phase.

**Offline Processing Level:** At this level, most of the basic processing over the data takes place. The offline, fully-featured, versions of the entity, topics, opinions, and events analysis (ETOE analysis) and the analysis of the social contents operate over the cleansed data from the crawl that are stored in the ARCOMEM database. These processing tools perform linguistic, machine learning, and other information extraction methods in order to provide a rich set of metadata annotations that are interlinked with the original data. The respective annotations are stored back in the ARCOMEM database and are available for further processing and information mining. After all the relevant processing has taken place, the Web pages to be archived are selected in a semi-automatic way.

**Cross Crawl Processing:** Finally, a more advanced processing step takes place that operates across different crawls and crawl campaigns. This allows analysis in the content that requires longer time spans but also to work on a larger set of Web objects of a specific kind if a single crawl is not able to collect enough of them.

The analysis over longer time spans allows registering the evolution of various aspects identified by the ETOE and Web analysis components. As such, it produces aggregate results that pertain to a group archive of objects, rather than to particular instances.

**Applications:** ARCOMEM foresees different ways for users to interact with the system. The Crawler Cockpit allows archivist to specify or modify crawl specifications and do the quality assurance. The Crawler Cockpit is also used to create the final Web archives. Based on a relevance analysis, a semi-automatic method proposes to the archivist relevant Web pages from the ARCOMEM database that should be preserved. The selected content including all semantic enrichments will be transferred to the WARC files for preservation.
The SARA end-user applications allow users to search the archives by domain, time and keywords. Furthermore, browsing the archives via different facets like topics, events, and entities, and visualizing the sentiments of Social Web postings complement the end user application. However, the applications are not limited to the described examples. The ARCOMEM system is open to any kind of application that wants to use it.

3 Analysis for Crawl Guidance and Archive Building

**Content Analysis.** The aim of this module is the extraction and detection of informational elements called ETOEs (Entities, Topics, Opinions, and Events) from Web pages (see Section 2). The ETOE extraction takes place in the offline phase and processes a collection of Web pages. The results of the offline ETOE extractions are used to (1) get a better understanding of the crawl specification and (2) populate the ARCOMEM database with structured data about ETOEs and their occurrences in Web objects. In the online phase, single documents will be analyzed to determine their relevance to the crawl specification.

**Social Web Analysis.** The aim of the Social Web analysis is to leverage the Social Web to contextualize content and information to be preserved, and to support the crawler guidance. In social networks users are discussing and reflecting about all kinds of topics, events, and persons. By doing so, they regularly post links to other relevant Web pages or Social Web content. As these links are recommendations of individuals in the context of their social online activities they are highly relevant for preservation. However, since users are unknown and anonymous it is necessary to derive their reputation and trustworthiness in the social community during the Social Web analysis.

**Crawler Guidance.** In ARCOMEM we replaced the traditional crawl definition by an intelligent crawl definition as presented in Section 2. The classical page fetching module is replaced by some more elaborate resource fetching component able to retrieve resources that are not just accessible by a simple HTTP GET request (but by a succession of such requests, or by a POST request, or by the use of an API), or individual Web objects inside a Web page (e.g., blog posts, individual comments, etc.).

After a resource (for instance a Web page) is fetched, an application-aware helper module is used in place of the usual link extraction function, to identify the Web application currently being crawled, decide on and categorize crawling actions (e.g., URL fetching, using an API) that can be performed on this particular Web application, and the kind of Web objects that can be extracted.

Crawling actions thus obtained are sent for further analysis and ranking to modules of the online phase. They are then filtered and prioritized by a resource selection & prioritization module using both intelligent crawling definition and feedback from online analysis modules to prioritize the crawl. Semantic analysis can thus make an impact on crawl guidance: for example, if a topic relevant to the intelligent crawl specification is found in the anchor text of a link to an external Web site, this link may be prioritized over others on the same page.
4 Conclusions

In this paper we presented the approach we follow to develop a social and semantic aware Web crawler for creating Web archives as community memories that revolve around events and the entities related to them. The need to make decisions during the crawl process with only a limited amount of information raises a number of issues. The division into different processing phases allows us to separate the initial complex extraction of events and entities from their faster but more shallow detection at crawl time. Furthermore, it allows in the offline phase to learn more about particular events and topics the archivist is interested in and to get more insights about trustful content on the Social Web.