Exploiting the Social and Semantic Web for guided Web Archiving*

Thomas Risse[1], Stefan Dietze[1], Wim Peters[2], Katerina Doka[3], Yannis Stavrakas[3], and Pierre Senellart[4]

[1] L3S Research Center, Hannover, Germany,
{risse, dietze}@L3S.de
[2] University of Sheffield, UK,
w.peters@dcs.shef.ac.uk
[3] IMIS / RC "ATHENA", Athens, Greece,
katerina@cslab.ece.ntua.gr, yannis@imis.athena-innovation.gr
[4] 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. 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.

Keywords: Web Archiving, Web Crawler, Text Analysis, Social Web

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. The report *Sustainable Economics for a Digital Planet* [1] states that "the first challenge for preservation arises when demand is diffuse or weakly articulated." This is especially the case for non-traditional digital publications, e.g., blogs, collaborative space, or digital lab books. The challenge with new forms of publications is that there can be a lack of alignment between what institutions see as worth preserving, what the owners see as of current value, and the incentive to preserve together with the rapidness at which decisions have to be made. For ephemeral publications such as the Web, this misalignment often results in irreparable loss. Given the deluge of digital information created and this situation of uncertainty, a first necessary step is to be able to respond quickly, even if in a preliminary fashion, by the timely creation

* This work is partly funded by the European Commission under ARCOMEM (ICT 270239).

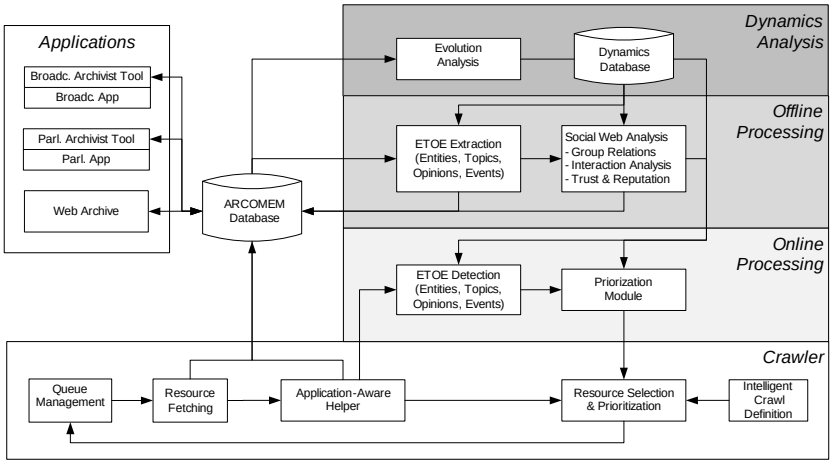

Fig. 1. Overall Architecture

of archives, with minimum overhead enabling more costly preservation actions further down the line. This is the challenge that the ARCOMEM[5] project is addressing.

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. Thus, entities and events are natural candidates for focusing new types of content acquisition processes in preservation as well as for archive enrichment.

The crawler architecture we propose here is the basis for current implementation activities in the ARCOMEM project. Note that the system is only partially implemented at the moment, and we therefore do not present any evaluation.

The rest of the paper is structured as follows. Section 2 gives an overview of the overall architecture and the different processing phases. More details about the content and Social Web analysis as well as crawler guidance are presented in Section 3. We discuss the state of the art in Web archiving and related fields in Section 4. Finally, Section 5 gives conclusions and an outlook to future work.

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 page and its components. These components of a Web page are called *Web objects* and can be the title, a paragraph, an

[5] ARCOMEM – From Collect-All ARchives to COmmunity MEMories, `http://www.arcomem.eu/`

image or a video. Since a thorough analysis of all Web objects 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. 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 four processing levels: the *crawler level*, the *online processing level*, the *offline processing level*, and *dynamics analysis*, that revolve around the ARCOMEM database as depicted in Figure 1. The ARCOMEM database is the focal point for all components involved in crawling and content analysis. It stores all information from the crawl specification over the crawled content to the extracted knowledge. Therefore a scalable and efficient implementation together with a sophisticated data model is necessary. The different processing levels are described below.

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 NLP 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 and preserved are selected in a semi-automatic way.

Dynamics Analysis Level: Finally, a more advanced processing step takes places. It operates on collections of Web objects that have been collected over time in

order to register 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.

3 Analysis for Crawl Guidance and Enrichment

We now describe in more detail the major analyses that are performed at all levels of the ARCOMEM architecture. We discuss over content analysis, analysis of the social Web, data enrichment, and crawler guidance itself. The discussion about dynamics analysis is post-poned to a later publication.

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 (s. 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 occurrence in Web objects. In the online phase, single documents will be analyzed to determine their relevance to the crawl specification.

A crawl campaign is described by a crawl specification given by the archivist. This specification consists of, in addition to other parameters, a search string where the archivist specifies in their own words the semantic focus of the crawl campaign. The search string is a combination of entities, topics, and events, plus free terms. Since it will not always be possible to literally match the search string with the content of a Web page, it is important to learn from an initial set of pages how the search string will be represented on real pages. This analysis will be done in the offline phase since it requires a collection of Web pages and is computationally more expensive. The result of this analysis is used in the online phase to derive the relevance of a page with respect 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.

The results of the Social Web analysis can also be leveraged in the contextualization process to further enrich the Web objects, e.g., if the object is tweeted by many nature experts it may be a good candidate for nature topics. Furthermore, the similarity and overlap between the provided objects and objects already seen before is established in order to interlink those that are discussing the same event, activity, or entity, improving the contextualization of the involved Web objects.

Data Enrichment and Consolidation. Data extracted via dynamics analysis and content analysis is heterogeneous. For instance, during one particular cycle, the

text analysis component might detect an entity from the term "Ireland", while during later cycles, entities based on the term "Republic of Ireland" or the German term "Irland" might be extracted. These would all be classified as entities of type *arco:Location* and correctly stored in the ARCOMEM data store as separate entities described according to the ARCOMEM RDF schema. Data enrichment and consolidation follows three aims: (a) enrich existing entities with related publicly available knowledge; (b) disambiguation and (c) identify data correlations such as the ones above. (a), (b) and (c) exploit publicly available data from the Linked Open Data cloud[6] which offers a vast amount of data of both domain-specific and domain-independent nature.

Crawler Guidance. As shown on the bottom part of Figure 1, the crawler used in the ARCOMEM project includes a number of functionalities that are not found in traditional Web crawlers. First, we replace the traditional crawl definition by an *intelligent crawl definition*, which allows the specification of relevance scores and the referencing of the particular kinds of Web applications and ETOEs that define the scope of the archiving task. *Queue management* functions similarly as in a traditional architecture, but 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. This is a critical phase for using clues from the Social Web to crawl content, because, depending on the kind of Web application that is being crawled, the kind of relevant crawling actions and Web objects to be extracted vary dramatically.

Crawling actions thus obtained are sent for further analysis and ranking to online phase modules. 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 Related Work

Since 1996, several projects have pursued Web archiving (e.g., [2]). The Heritrix crawler [3], jointly developed by several Scandinavian national libraries and the Internet Archive through the International Internet Preservation Consortium (IIPC), is a mature and efficient tool for large-scale, archival-quality crawling.

The method of choice for memory institutions is client-side archiving based on crawling. This method is derived from search engine crawl, and has been evolved

[6] `http://lod-cloud.net/`

by the archiving community to achieve a greater completeness of capture and a reduction of temporal coherence of crawls. These two requirements follow from the fact that, for Web archiving, crawlers are used to build collections and not only to index [4]. These issues were addressed in the European project LiWA (Living Web Archives)[7].

The task of crawl prioritization and focusing is the step in the crawl processing chain which combines the different analysis results and the crawl specification for filtering and ranking the URLs of a seed list. A number of strategies such as breadth-first, back link count and PageRank exist for this. PageRank and breadth-first are good strategies to crawl "important" content on the Web [5], but since these generic approaches do not cover specific information needs, focused or topical crawls have been developed [6]. However, these approaches have only a vague notion of topicality and do not address event-based crawling.

5 Conclusions & Future Work

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 shallower 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.

The implementation of the presented architecture is underway. Parts of the system are built upon existing technologies while other, like the Social Web analysis, need to be developed from scratch. Also, a number of research questions need to be addressed. For example the typically limited set of reference pages and the limited time to detect topics, entities, and events during crawling are open issues. Also how the different extracted information, interaction patterns, etc., can be combined for prioritizing URLs is currently an open question.

References

1. Blue Ribbon Task Force on Sustainable Digital Preservation and Access: Sustainable economics for a digital planet, ensuring long-term access to digital information. http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf (2010)
2. Arvidson, A., Lettenström, F.: The Kulturarw project – the Swedish royal Web archive. Electronic library **16**(2) (1998)
3. Mohr, G., Kimpton, M., Stack, M., Ranitovic, I.: Introduction to Heritrix, an archival quality Web crawler. In: 4th International Web Archiving Workshop. (2004)
4. Masanès, J.: Web archiving. Springer (2006)
5. Baeza-Yates, R., Castillo, C., Marin, M., Rodriguez, A.: Crawling a country: better strategies than breadth-first for Web page ordering. In: 14th WWW Conf. (2005)
6. Menczer, F., Pant, G., Srinivasan, P.: Topical Web crawlers: Evaluating adaptive algorithms. ACM Trans. Internet Technol. **4** (2004) 378–419

[7] http://www.liwa-project.eu/