ARCOMEM - From Collect-All ARchives to COmmunity MEMories*

Thomas Risse
L3S Research Center
University of Hanover
Germany
risse@L3S.de

Wim Peters
University of Sheffield
UK
w.peters@dcs.shef.ac.uk

ABSTRACT

The ARCOMEM project is about memory institutions like archives, museums and libraries in the age of the Social Web. Social media are becoming more and more pervasive in all areas of life. ARCOMEM's aim is to help to transform archives into collective memories that are more tightly integrated with their community of users and to exploit Web 2.0 and the wisdom of crowds to make Web archiving a more selective and meaning-based process. ARCOMEM (FP7-IST-270239) is an Integrating Project in the FP7 program of the European Commission, which involves twelve partners from academia, industry and public sector. The project will run from January 1, 2011 to December 31, 2013.

Categories and Subject Descriptors

D.2.11 [**Software Architectures**]: Domain-specific architectures; H.3.6 [**Library Automation**]: Large text archives

General Terms

Algorithms, Design, Languages

Keywords

Web Archiving, Web Crawler, Architecture, Text Analysis, Social Web

1. INTRODUCTION

ARCOMEM[1] (FP7-IST-270239) - From Collect-All ARchives to COmmunity MEMories - is about memory institutions like archives, museums and libraries in the age of the social web. Social media are becoming more and more pervasive in all areas of life. ARCOMEM's aim is to help to transform archives into collective memories that are more tightly integrated with their community of users and to exploit Web 2.0 and the wisdom of crowds to make web archiving a more selective and meaning-based process.

The BRTF report on Sustainable Economics for a Digital Planet [3] 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 a current value and the incentive to preserve as well as the rapidness at which decisions have to be made. For ephemeral publications 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 preliminary, by the timely creation of archives, with minimum overhead associated that would support later engagement in more costly preservation actions. This is the challenge that ARCOMEM is addressing, relying on the Wisdom of the Crowds for intelligent content appraisal, selection, contextualization and preservation.

ARCOMEM is an Integrating Project in the FP7 program of the European Commission, which involves twelve partners from academia, industry and public sector. The project will run from January 1, 2011 to December 31, 2013. In this paper we will present a project overview, the expected outcome and the general approach. Furthermore we will describe a demonstrator, which shows a core component of the upcoming system.

2. PROJECT GOALS

The goal of the ARCOMEM project is to develop methods and tools for transforming digital archives into community memories based on novel socially-aware and socially-driven preservation models. This will be done (a) by leveraging the Wisdom of the Crowds reflected in the rich context and reflective information in the Social Web for driving innovative, concise and socially-aware content appraisal and selection processes for preservation, taking events, entities and topics as seeds, and by encapsulating this functionality into an adaptive decision support tool for the archivist, and (b) and by using Social Web contextualization as well as extracted information on events, topics, and entities for creating richer and socially contextualized digital archives.

The Social Web not only provides a rich source of user generated content. It also contextualizes content and reflects content understanding and appraisal within society. This is done by interlinking, discussing, commenting, rating, referencing, and reusing content. The ARCOMEM project will analyze and mine this rich social tapestry to find clues for deciding what should be preserved (based on its reflection in the Social Web), how to contextualize content within digital archives based on their Social Web context, and how to best preserve this context. The Social Web based contextualization will be complemented by exploring topic-centered, event-centred and entity-centred processes for content appraisal and acquisition as well as rich preservation.

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

[1] http://www.arcomem.eu

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

2.1 Scientific and Technological Objectives

To achieve its goal, the ARCOMEM project will pursue the following scientific and technological objectives.

1. Social Web analysis and Web mining, which includes effective methods for the analysis of Social Web content, analysis of community structures, discovery of evidence for content appraisal, analysis of trust and provenance, and scalability of analysis methods. Key research topics are the search for regularities in the way social networks evolve over time [10] and the detection of leaders and followers in a social network [8];

2. Event detection and consolidation, which includes information extraction technologies for detection of events (e.g. [13]) and for detecting (multimedia) entities related to events [7, 12]; methods for consolidating event, entity and topic information within and between archives, models for events, covering different levels of granularity, and their relations;

3. Perspective, Opinion, and Sentiment detection, which includes scalable methods for detecting and analyzing opinions (e.g. [9]), perspectives taken, and sentiments expressed in the Web and especially Social Web content;

4. Concise content purging, which includes detection of duplicates and near-duplicates and an adequate reflection of content diversity with respect to textual content, images, and opinions;

5. Intelligent adaptive decision support, which includes methods for combining and reasoning about input from social Web analysis, diversity and coverage aspects, extracted information, domain knowledge and heuristics, etc.; methods for adapting the decision strategies to inputs received;

6. Advanced Web Crawling, which includes the integration of event-centered and entity-centered strategies, the use of social Web clues in crawling decisions and methods for translating by example and descriptive crawling specifications into crawling strategies;

7. Approaches for "semantic preservation", which includes methods for enabling long-term interpretability of the archive content; methods for preserving the original context of perception and discourse in a semantic way; methods for dealing with evolution on the semantic layer.

2.2 Application Scenarios

ARCOMEM is examining two application domains of the Social Web, whose applications scenarios will be used for validating and showcasing the ARCOMEM technology.

Parliament Application: Parliament libraries provide Members of Parliament and their assistants, journalists, political analysts and researchers' information, research and documentation for parliamentary issues. Beside traditional publications the Web and the Social Web plays an increasingly important role as an information source since it provides important and crucial background information, reactions, the comments made by the general public. It is in the interests of the parliaments to create a platform for preserving, managing, mining and analyzing all the information provided in the social media. For gathering the information from the web the crawl specification consists of a number of parameters like search string with events, locations, entities, etc., an initial seed lists, social media categories, etc.

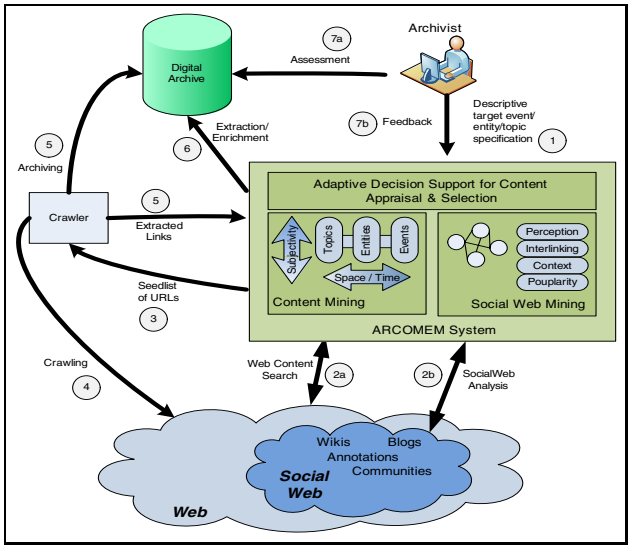

Figure 1: Overview of the ARCOMEM system

Broadcaster Application: Due to the increasing importance of the web and social web, journalists will in future not be able anymore to only use reliable sources like news agencies, PR-material or libraries. User generated content will become another important information source. This shift in importance is also the case when own events should be documented and their impact should be analyzed.

In both cases it is important that the user generated content stays accessible even if the original source disappears. In addition they need support in the verification of the information. Therefore context information such as the user, the event, related links or entities mentioned on the Web pages need to be preserved as well. The final Web archive will allow an effective use of the content even decades later.

3. MAIN OUTCOMES

The main expected outcome of the ARCOMEM project is the ARCOMEM system, which will consist of the ARCOMEM socially-aware content appraisal and selection tool and components for archive enrichment and contextualization. This main outcome is depicted in the central box of Figure 1 together with the main information flows.

On the level of models and approaches, the outcome of the ARCOMEM project will be a novel model and approach for socially-aware and socially-driven preservation, including an innovative model for Social Web driven content appraisal and content selection, enabling socially-aware content prioritization, and "by example" - by providing a set of reference pages - and descriptive content selection based on topics, events and entities. It also includes an approach for enriching and contextualizing archive content based on a combination of Social Web context, related events and entities as well as topics, perspectives and temporal/evolutionary aspects aiming for semantic preservation by facilitating short-term, mid-term and long-term archive content use and interpretation.

On the level of tools and applications the expected outcome of the ARCOMEM project is the ARCOMEM socially-aware content appraisal and selection tool, and two exemplary prototypical applications built on top of the ARCOMEM System in the broadcasting and political domains described in section 2.2. These tools build on the integrated ARCOMEM system, and will support the socially-

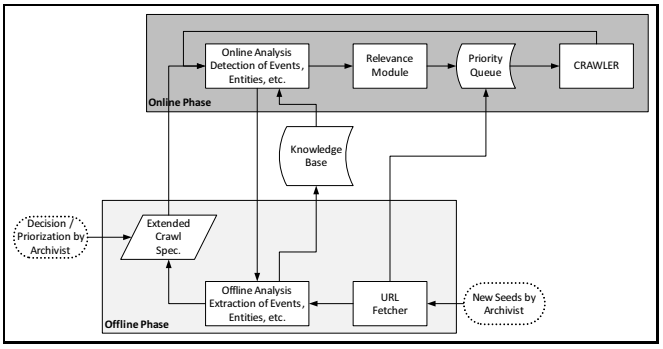

Figure 2: ARCOMEM Online and Offline Processing

aware, intelligent, adaptive and autonomous content appraisal and selection process. This automated process can be triggered in a "by example" or descriptive fashion; the complex content appraisal and selection decisions making will be supported based on reasoning about the diverse extracted evidence (based on social context, perspectives, topics, events, and entities) and following a strategy of concise coverage. Archivists will be enabled to monitor and interact with the process, using the feedback received to adapt the content appraisal and selection strategy of the tool for preservation tasks.

Furthermore, in the ARCOMEM project, the digital archive content is contextualized with information from the Social Web and enriched with information on related events and entities as well as on topics and perspectives taken. The semantic knowledge inferred from the extracted information is embedded into the formats used for preserving the content preparing enriched preservation objects for future use and triggering the digitization of similar content.

4. APPROACH & ARCHITECTURE

4.1 Overall Approach

Archivists will be able to trigger interactive and intelligent content appraisal and selection processes in two ways: either "by example" or by a high-level description of relevant entities, topics and events. Intelligent and adaptive decision support for this will be based on combining and reasoning about the extracted information and inferring semantic knowledge, combining logic reasoning with adaptive content selection strategies and heuristics.

The system is built around two loops: content selection and content enrichment. The **content selection** loop aims at content filtering based on community reflection and appraisal. Social Web content will be analysed regarding the interlinking, context and popularity of web content, regarding events, topics and entities. These results are used for building the seed lists to be used by existing Web crawlers. Within the **content enrichment** loop, newly crawled pages will be analysed for topics, entities, events, perspectives, Social Web context and evolutionary aspects in order to link them together by means of the events and entities. In the following we will focus on the **content selection loop**.

4.2 Architecture

The main tasks of a Web crawler are to download a Web page and to extract links from that page to find more pages to crawl. An intelligent filtering and ranking of links enables focusing of the crawls. We will combine a breadth-first strategy with a semantic ranking that takes into account events, topics, opinions and entities (ETOEs). The extracted links are weighted according to the relevance of the page to the semantically rich crawl specification. The general architecture is depicted in Figure 2.

The whole process is divided into an online and offline phase. The online phase focuses on the crawl task itself and the guiding of the crawler, while the offline phase is used to analyze the crawl results and the crawl specification to setup a knowledge base for the online decision making.

Offline Phase: To bootstrap a new crawl campaign, the archivist specifies a crawl by giving an initial seed list complemented with some information about events, entities and topics. e.g. [event: "Rock am Ring"], [Band: "Coldplay"], [Location: "Nürburgring"]. The idea behind the following process is that the archivist is not able to give a full crawl specification as they cannot be fully aware of how the events, topics, etc. they are interested in are represented on the web. Therefore the crawler needs to help the archivist to improve the specification.

The initial seed list is used by the **URL Fetcher** to initiate a reference crawl. This reference crawl will be analyzed by the **offline analysis component** to extract ETOEs, which are used to derive an extended crawl specification. In this step the archivists need to assess the relevance of the extracted information to the envisioned crawl. They have the possibility to weight the information and also to explicitly exclude some of it from the crawl. The resulting **extended crawl specification** is handed over to the online phase.

In addition to the extended crawl specification, a **knowledge base** will be built, in order to provide additional information such as more detailed descriptions of events or entities, different lexical forms or other disambiguation information. The offline phase will be called regularly from the online phase to further improve the crawl specification and the knowledge base.

Online Phase: The online analysis component receives newly crawled pages from the crawler and the extended crawl specification from the offline phase. Due to the necessary high crawl frequency, the processing time and decision making for a single page should take no longer then a second. Therefore complex analysis like extracting new ETOEs is not possible. Instead, the analysis component will rely on the information in the knowledge base to detect the degree of relevance of a page to the crawl specification, to rank the extracted links and to update the priority queue of the crawler accordingly. The crawler processes the priority queue and hands over new pages to the online analysis.

5. DEMONSTRATION

The ARCOMEM project is now at the end of the first year. Since the first integrated prototype is planned for mid of the 2nd year, we are only able to demonstrate some key system components. These components consist of entity extraction and consolidation, and event detection. The output of these components concerns part of the actual content of the extraction processes in the offline phase, and give as such a good first impression of the nature of information archivists will be enabled to work with in their curation effort.

These components will be run on a set of web pages in both English and German, and the results will be presented in the form of:

- A summary detailing the results of the analysis;
- RDF ontological resources containing extracted entities and events;
- Source text annotated with ARCOMEM concepts. For this purpose we will use the graphic user interface of the GATE system [6], which unpins most extraction efforts in ARCOMEM. Figure 3 illustrates the browser functionality of the

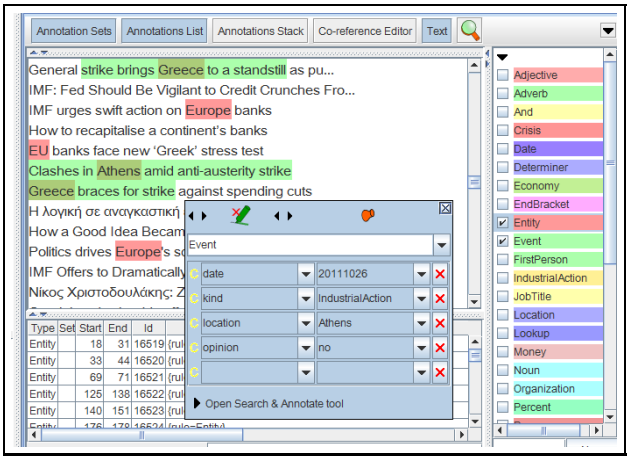

Figure 3: Screen shot from the GATE GUI

GATE interface. The colours in the text pane on the left hand pane correspond to ticked concepts from the right hand pane. The pop up window shows the annotations associated with the recognized event expressed by the text span "Clashes in Athens amid anti-austerity strike".

6. RELATED WORK

Since 1996, several projects have pursued Web archiving (e.g. [1, 15]). The Heritrix crawler [15], 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 by the archiving community to achieve a better completeness of capture and to reduce temporal coherence of crawls. These two requirements come from the fact that, for web archiving, crawlers are used to build collections and not only to index [11]. These issues were addressed in the European project LiWA (Living Web Archives) .

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. The filtering of URLs is necessary to avoid unrelated content in the archive. For content that is partly relevant, URLs need to be prioritized to focus the crawler tasks to crawl in order of relevancy. A number of strategies and therefore URL ordering metrics exist for this, such as breadth-first, back link count and PageRank. PageRank and breadth-first are good strategies to crawl "important" content on the web [5, 2], but since these generic approaches do not cover specific information needs, focused or topical crawls have been developed [4, 14]. However, these approaches have only a vague notion of topicality and do not address event-based crawling.

7. CONCLUSIONS & FUTURE WORK

In this paper we have presented the approach we follow in the ARCOMEM project to build Web archives as community memories that revolve around events and the entities related to them. The content selection and appraisal task will be supported by information from the Social Web. In the demonstration we will show the entity extraction and consolidation, and event detection component - a core part of the upcoming system. For mid 2012 it is planned to release a first integrated version of the ARCOMEM system.

8. REFERENCES

[1] A. Arvidson and F. Lettenström. The Kulturarw Project - The Swedish Royal Web Archive. *Electronic library*, 16(2), 1998.

[2] R. Baeza-Yates, C. Castillo, M. Marin, and A. Rodriguez. Crawling a country: better strategies than breadth-first for web page ordering. In *Special interest tracks and posters of the 14th international conference on World Wide Web*, WWW '05, pages 864–872, New York, 2005. ACM.

[3] Blue Ribbon Task Force on Sustainable Digital Preservation and Access. Sustainable economics for a digital planet, ensuring long-term access to digital information, 2010.

[4] S. Chakrabarti, M. V. D. Berg, and B. Dom. Focused crawling: a new approach to topic-specific web resource discovery. In *Computer Networks*, pages 1623–1640, 1999.

[5] J. Cho, H. Garcia-Molina, and L. Page. Efficient crawling through url ordering. In *Proceedings of the seventh international conference on World Wide Web 7*, WWW7, pages 161–172, Amsterdam, The Netherlands, The Netherlands, 1998. Elsevier Science Publishers B. V.

[6] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proc. of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, 2002.

[7] P. G. Enser, C. J. Sandom, J. S. Hare, and P. H. Lewis. Facing the reality of semantic image retrieval. *Journal of Documentation*, 63(4):465 – 481, 2007.

[8] A. Goyal, B.-W. On, F. Bonchi, and L. V. S. Lakshmanan. Gurumine: A pattern mining system for discovering leaders and tribes. In *Proceedings of the 2009 IEEE International Conference on Data Engineering*, pages 1471–1474, Washington, DC, USA, 2009. IEEE Computer Society.

[9] S.-M. Kim and E. Hovy. Automatic detection of opinion bearing words and sentences. In *Companion Volume to the Proceedings of IJCNLP-05, the Second International Joint Conference on Natural Language Processing*, pages 61–66, Jeju Island, KR, 2005.

[10] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *Proceedings of the 12th international conference on World Wide Web*, WWW '03, pages 568–576, New York, NY, USA, 2003. ACM.

[11] J. Masanès. *Web archiving*. Springer, 2006.

[12] D. Maynard, Y. Li, and W. Peters. NLP Techniques for Term Extraction and Ontology Population. In P. Buitelaar and P. Cimiano, editors, *Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text*. IOS Press, 2008.

[13] D. McClosky, M. Surdeanu, and C. D. Manning. Event extraction as dependency parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1626–1635, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[14] F. Menczer, G. Pant, and P. Srinivasan. Topical web crawlers: Evaluating adaptive algorithms. *ACM Trans. Internet Technol.*, 4:378–419, Nov. 2004.

[15] G. Mohr, M. Kimpton, M. Stack, and I. Ranitovic. Introduction to Heritrix, an archival quality web crawler. In *4th International Web Archiving Workshop (IWAW04)*, 2004.