# Using Events for Content Appraisal and Selection in Web Archives*

Thomas Risse[1], Stefan Dietze[1], Diana Maynard[2], and Nina Tahmasebi[1]

[1] L3S Research Center, Hanover, Germany
{risse|dietze|tahmasebi}@L3S.de

[2] University of Sheffield, Sheffield, UK,
{diana}@dcs.shef.ac.uk

**Abstract.** With the rapidly growing volume of resources on the Web, Web archiving becomes an important challenge. In addition, the notion of community memories extends traditional Web archives with related data from a variety of sources on the Social Web. Community memories take an entity-centric view to organise Web content according to the events and the entities related to them, such as persons, organisations and locations. To this end, the main challenge is to extract, detect and correlate events and related information from a vast number of heterogeneous Web resources where the nature and quality of the content may vary heavily. In this paper we present the approach of the ARCOMEM project which is based on an iterative cycle consisting of (1) targeted archiving/crawling of Web objects, (2) entity and event extraction and detection, and (3) refinement of crawling strategy.

**Keywords:** Event Detection, Web Archiving

## 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. However, in addition to the "common" challenges of digital preservation, such as media decay, technological obsolescence, authenticity and integrity issues, Web preservation has to deal with the sheer size and ever-increasing growth rate of Web data. Hence, selection of content sources becomes a crucial task for archival organizations. 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, organisations and locations. These may be unique events such as the first landing on the moon or a natural disaster, or regularly occuring events such as elections or TV serials.

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

In this paper, we provide an overview of the approach we follow in the European project ARCOMEM[3]. The overall aim is to create incrementally enriched Web archives which allow access to all sorts of Web content in a structured and semantically meaningful way. In addition to topic-centred preservation approaches, we are exploring event- and entity-centred processes for content appraisal and acquisition as well as rich preservation. By considering a wide range of content, a more diverse archive is created, taking into account a variety of dimensions including perspectives taken, sentiments, images used, and information sources.

To build a community archive from Web content, a *web crawler* needs to be guided in an intelligent way based on the events and entities derived from previous crawl campaigns so that pages are crawled and archived if they relate to a specified event or entity. While at the beginning of any crawl campaign the amount of information is very limited, the crawler needs to learn about the event incrementally, while at the same time it has to decide about following links. Therefore, our approach is based on an iterative cycle consisting of the following steps:

1. Targeted archiving/crawling of Web objects;
2. Entity and event extraction and detection;
3. Refinement of crawling strategy.

To this end, the main challenges are related to *entity and event extraction*, *detection* and *correlation* of events and related information in a vast variety of heterogeneous Web resources. Please note, while *extraction* covers the identification and structured representation of knowledge about events and entities from previously unstructured material from scratch, *detection* refers to the detection of previously extracted events and entities. Therefore, in contrast to the extraction step, detection takes advantage of existing structured data about events and entities. Obviously, both extraction as well as detection face issues arising from the diversity of the nature and quality of Web content, in particular when considering *social media* and *user-generated content*, where further issues are posed by informal use of language.

In the following section, we give an overview of related work and introduce the ARCOMEM approach and architecture in Section 3. Section 4 provides an overview on the event detection mechanisms deployed by ARCOMEM while we discuss some key challenges in Section 5 followed by the conclusion in Section 6.

## 2 Related Work

Since 1996, several projects have pursued Web archiving (e.g., [AL98,ACMS02]). The Heritrix crawler [MKSR04], jointly developed by several Scandinavian national libraries and the Internet Archive through the International Internet

[3] ARCOMEM - From Collect-All Archives to Community Memories, http://www.arcomem.eu/

Preservation Consortium (IIPC)[4], 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 [Mas06]. These issues were addressed in the European project LiWA (Living Web Archives)[5].

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 [CGMP98,BYCMR05], but since these generic approaches do not cover specific information needs, focused or topical crawls have been developed [CBD99] [AAGY01,MPS04]. However, these approaches have only a vague notion of topicality and do not address event-based crawling.

Entity and event recognition are two of the major tasks within Information Extraction, and have been successfully applied in research areas such as ontology generation, bioinformatics, news aggregation, business intelligence and text classification. Recognizing events in these fields is generally carried out by means of pre-defined sets of relations, possibly structured into an ontology, which makes such tasks domain dependent, but feasible. Entity extraction in this case comprises both named entity recognition [CMBT02] and term recognition [BS09,MLP08].

The identification of relations between entities in text is generally performed by means of heuristic, rule-based applications using background knowledge from instantiated ontologies and lexico-syntactic patterns to establish links between textual entities and their ontological provenance [MFP09a], or a combination of statistical and linguistic techniques [MPB08]. Tools such as Espresso [PP06] and Text2Onto [CLS05] make use of predefined or automatically extracted text patterns in order to structure the domain in terms of classes and relations. Furthermore, shallow parsing techniques such as semantic role labelling [Gil02] characterise the relationship between predicates (relations) and their arguments (entities) on a semantic level by means of roles such as agent and patient. On the other hand, unsupervised machine learning techniques such as TextRunner[BE08] and Powerset[6] scale to the extraction of facts from hundreds of millions of web pages, but they use only very shallow linguistic analysis and may not be so accurate.

[4] http://netpreserve.org/
[5] http://wiki.liwa-project.eu/
[6] http://www.powerset.com/

While PowerSet, for example, uses advanced parsing and some NLP techniques, it does not understand word and phrase meanings in context. In this work, we position our event extraction approach somewhere between the very constrained template-filling approach used in MUC, and the open domain approach of finding new relations over the whole web, used by systems such as TextRunner and Powerset.

In addition, for representation of events and entities we consider Semantic Web and Linked Data-based approaches, as it is one of the fundamental aims to expose the generated knowledge in an interoperable and reusable way. We consider in particular Linked Open Descriptions of Events, LODE [STH09], Event-Model-F [ASS09] and the Event Ontology[7]. While LODE and the Event Ontology follow a similar approach to and provide rather light-weight RDF schemas for event description, the Event-Model-F is a more formal OWL ontology which applies the DOLCE Descriptions and Situations pattern by using DOLCE+DnS Ultralight (DUL)[8] as upper level ontology.

## 3 Approach and Architecture

### 3.1 Overall Approach

The goal for the ARCOMEM system 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 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.

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, the context and the 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*.

[7] http://motools.sourceforge.net/event/event.html
[8] http://www.loa-cnr.it/ontologies/DUL.owl

### 3.2 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 1.



**Fig. 1.** Architecture for the Content Selection

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"], [Person: "Coldplay"] , [Location: "Nürburgring"]. The idea behind the following process is that the archivist is not able to give a full crawl specification as he or she cannot be fully aware of how the events, topics, etc. he or she is 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 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 2-3 secs. 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.

## 4 Event Extraction

The event extraction method we propose involves the recognition of entities and the relations between them in order to find domain-specific events and situations. As discussed in Section 2, in a (semi-)closed domain, this approach is preferable to an open IE-based approach which holds no preconceptions about the kinds of entities and relations possible. Building on the work of [MYKK05], we combine a number of different techniques, using two parallel strategies for event detection. The **top-down approach**, similar to a template-based IE approach as used in the Message Understanding Conferences [CHL93], consists of identifying a number of important events, based on analysis of the user needs and manual inspection of the corpora. Here, the slots are known in advance and the values are entities extracted from the text. In our Rock am Ring use case, the following example depicts a band perfmance event:

```
band:Kings of Leon  relation: performed  date: 3 June 2011
```

The technique consists of pre-defining a set of templates for the various relations, and then using a rule-based approach based on GATE [CMBT02] to identify the relevant slot values. First, we perform linguistic pre-processing (tokenisation, sentence splitting, POS tagging, morphological analysis, and verb and noun phrase chunking), followed by entity extraction, which includes both named entities and terms: for this we make use of slightly modified versions of ANNIE [CMBT02] and TermRaider[9] respectively. The third stage involves a semantic approach to finding the verbal expressions which represent the relations. We automatically create sets of verbs representing each relation, using information from WordNet and VerbNet to group verbs into semantic categories: for example, the relation "perform" might be represented by any morphosyntactic variant of the verbs "perform", "play", "sing", "appear" etc. We then develop hand-crafted rules to match sentences containing the relevant entities and verbs: for example, a rule to match the "performance" event described above should contain an entity representing a band name as the subject of a "perform" verb, and optionally a date and/or time within the sentence.

This kind of rule-based approach tends to be very accurate, achieving relatively high levels of precision (depending on how specific the rules are), but can suffer from low recall. On the other hand, a **bottom-up** technique involving open-domain IE can find previously unknown events and does not limit us to a fixed set of relations. This can be vital for discovering new information. By

[9] http://gate.ac.uk/projects/neon/termraider.html

combining the high precision of the top-down method with the high recall of the bottom-up method, we can get the best of both worlds if done correctly.

The bottom-up approach we adopt is rather different from the machine learning approach adopted by e.g. [BE08], in that we still specify hand-coded rules. However, these rules are flexible and under-specified, making use of linguistic structure and semantic relations from WordNet [ME90] rather than pre-specifying exact relations. We use the Noun Phrase and Verb Phrase chunker from GATE to identify certain linguistic patterns contextualising verb phrases, and then cluster these verbs into semantically related categories to find new relations. The participants in the relations can also be semantically clustered around similar relation types, such that an iterative development cycle can be produced. We also combine rules for ontology learning developed in SPRAT [MFP09b] which can be used to find patterns denoting relations between entities, such as hyponyms and properties. Preliminary experiments with news texts in English have found relations such as the following:

```
Mr Woerfel        represented    Daimler Benz-Aerospace
Gen Musharraf     reshuffled     two pro-Taliban generals
Gen Musharraf     appointed      Lt Gen Mohammed Yousuf
Mr Daoudi         was arrested   in Leicester
```

We do not only restrict ourselves to verbal relations, but also look for nominalisations. For example, "the arrest of Mr Daoudi in Leicester" is semantically equivalent to "Mr Daoudi was arrested in Leicester".

The work on event detection is still very much in progress, and it is clear that there are many difficult issues to solve. We do not use full parsing because it is very slow and because it does not work so well on social media where English is often not written correctly in full sentences. Related work on opinion mining from tweets [MF11] has proved that shallow linguistic techniques are, however, promising for extracting knowledge from this kind of noisy data, using backoff strategies and fuzzy matching where necessary.

## 5 Challenges

For the long-term availability and usage of Web content, it is important to preserve not only the content itself but also its context and interactions from relevant Web destinations. Relevant Web destinations include those that the content providers own (the main portal, channel portals or programme portals), those that they partner with (e.g. joint broadcaster portals), social media services or platforms, and both professional and user blogs/websites. This type of content is varied and ranges between e.g., general content, commenting, rating, ranking and forwarding and contains structured data as well as unstructured free text.

To this end, it is a challenge to manage and correlate content from these information sources, differing in quality, form (e.g. both audiovisual and textual material) and structure. In order to achieve a focused crawl it is necessary to identify semantically related objects, e.g. ones which discuss the same events or

entities. However, the preservation and identification of correlations within such a diverse variety of Web sources poses a number of key challenges:

1. Extraction of events and entities from heterogeneous and unstructured content;
2. Detection of events and entities in heterogeneous and unstructured content;
3. Targeted Web crawling.

*Entity and event extraction* for information from unstructured and heterogeneous Web data is one of the key challenges. This involves the use of natural language processing (NLP) techniques to extract events and entities from unstructured and heterogeneous text, and video analysis techniques to deal with audiovisual material. Although extraction is performed in the offline phase (see Fig. 1), there are still time requirements. Because the newly extracted entities and events are used in the online phase to focus the crawl, the extraction must be reasonably fast. To keep the crawl from getting diffuse the extraction must also provide accurate results which provides an additional challenge.

In contrast to the extraction, the *detection of events and entities* needs to exploit the data captured in the knowledge base in order to automatically detect events and entities. The difference between extraction and detection is that extraction populates the knowledge base with new information in an offline phase, while detection identifies previously extracted events (matching the stored set) from newly crawled Web content. Both NLP and video processing techniques need to be exploited here too but with much less time for analysis providing more shallow results. Because the detection occurs in the online phase (see Fig. 1) and is in close interaction with the crawler, a key challenge is to perform the detection in a very short time frame and with limited time for deep, linguistic analysis.

Finally, the results of both processing phases in Section 3.2 are used for *targeted Web crawling*. This allows the crawling strategy to be gradually refined, based on the outcomes of the previous crawling, extraction and detection activities. It is a challenge to make appropriate use of these outcomes to create focussed archives.

## 6 Conclusions

In this paper we 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 such as persons, organisations and locations. The need to decide during the crawl process with a limited amount of information raises a number of issues. The division of online and offline processing allows to separate the complex event and entity extraction from the necessarily fast detection of them at crawl time. Furthermore, it allows learning more about the events and topics the archivist is interested in. However, the typically limited set of reference pages and the limited time to detect events in during crawling are open issue to be addressed in the future. Moreover, the whole approach needs

to be evaluated in real world scenarios e.g. for crawling election related pages or the upcoming Olympic Games.

## References


[AAGY01] Charu C. Aggarwal, Fatima Al-Garawi, and Philip S. Yu. Intelligent crawling on the world wide web with arbitrary predicates. In *Proceedings of the 10th international conference on World Wide Web*, WWW '01, pages 96–105, New York, NY, USA, 2001. ACM.

[ACMS02] Serge Abiteboul, Gregory Cobena, Julien Masanes, and Gerald Sedrati. A First Experience in Archiving the French Web. In *Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL '02, pages 1–15, London, UK, 2002. Springer-Verlag.

[AL98] Allan Arvidson and Frans Lettenström. The Kulturarw Project - The Swedish Royal Web Archive. *Electronic library*, 16(2), 1998.

[ASS09] C. Saathoff A. Scherp, T. Franz and S. Staab. F-a model of events based on the foundational ontology dolce+dns ultralight. In *International Conference on Knowledge Capturing (K-CAP)*, 2009.

[BE08] M. Banko and O. Etzioni. The tradeoffs between open and traditional relation extraction. In *Proceedings of ACL-08*, 2008.

[BS09] C. Buckley and G. Salton. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 2009.

[BYCMR05] Ricardo Baeza-Yates, Carlos Castillo, Mauricio Marin, and Andrea 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, NY, USA, 2005. ACM.

[CBD99] Soumen Chakrabarti, Martin Van Den Berg, and Byron Dom. Focused crawling: a new approach to topic-specific web resource discovery. In *Computer Networks*, pages 1623–1640, 1999.

[CGMP98] Junghoo Cho, Hector Garcia-Molina, and Lawrence 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.

[CHL93] N. Chinchor, L. Hirschman, and D.D. Lewis. Evaluating Message Understanding Systems: An Analysis of the Third Message Understanding Conference (MUC-3). *Computational Linguistics*, 19(3):409–449, 1993.

[CLS05] P. Cimiano, G. Ladwig, and S.Staab. Gimme' The Context: Context-driven automatic semantic annotation with C-PANKOW. In *Proceedings of the 14th World Wide Web Conference*, 2005.

[CMBT02] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, 2002.

[Mas06] Julien Masanès. *Web archiving*. Springer, 2006.

[ME90] G. A. Miller (Ed.). WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–312, 1990.

[MF11] D. Maynard and A. Funk. Automatic detection of political opinions in tweets. In *Proceedings of MSM 2011: Making Sense of Microposts Workshop at 8th Extended Semantic Web Conference*, Heraklion, Greece, May 2011.

[MFP09a] D. Maynard, A. Funk, and W. Peters. NLP-based support for ontology lifecycle development. In *CK 2009 – ISWC Workshop on Workshop on Collaborative Construction, Management and Linking of Structured Knowledge*, Washington, USA, October 2009.

[MFP09b] D. Maynard, A. Funk, and W. Peters. SPRAT: a tool for automatic semantic pattern-based ontology population. In *International Conference for Digital Libraries and the Semantic Web*, Trento, Italy, September 2009.

[MKSR04] 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.

[MLP08] Diana Maynard, Yaoyong Li, and Wim 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.

[MPB08] A. Moschitti, D. Pighin, and R. Basili. Tree kernels for semantic role labeling. *Computational Linguistics*, 34(2):193–224, 2008.

[MPS04] Filippo Menczer, Gautam Pant, and Padmini Srinivasan. Topical web crawlers: Evaluating adaptive algorithms. *ACM Trans. Internet Technol.*, 4:378–419, November 2004.

[MYKK05] D. Maynard, M. Yankova, A. Kourakis, and A. Kokossis. Ontology-based information extraction for market monitoring and technology watch. In *ESWC Workshop "End User Aspects of the Semantic Web")*, Heraklion, Crete, 2005.

[PP06] P. Pantel and M. Pennacchioni. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of Conference on Computational Linguistics / Association for Computational Linguistics (COLING/ACL-06)*, pages 113–120, Sydney, Australia, 2006.

[STH09] Ryan Shaw, Raphaël Troncy, and Lynda Hardman. Lode: Linking open descriptions of events. In Asunción Gómez-Pérez, Yong Yu, and Ying Ding, editors, *ASWC*, volume 5926 of *Lecture Notes in Computer Science*, pages 153–167. Springer, 2009.