The Iterative Design and Evaluation Approach for a Socially-aware Search and Retrieval Application for Digital Archiving

Dimitris Spiliotopoulos
Innovation Lab
Athens Technology Centre
Athens, Greece
d.spiliotopoulos@atc.gr

Ruben Bouwmeester
New Media, Innovation
Deutsche Welle
Bonn, Germany
ruben.bouwmeester@dw.de

Dominik Frey
Documentation and Archives
Suedwestrundfunk
Baden-Baden, Germany
dominik.frey@swr.de

Georgios Kouroupetroglou, Pepi Stavropoulou
Department of Informatics and Telecommunications
University of Athens
Athens, Greece
{koupe; pepis}@di.uoa.gr

***Abstract*—Designing user interfaces involves several iterations for usability design and evaluation as well as incremental functionality integration and testing. This paper reports on the methodological approach for the design and implementation of an application that is used for search and retrieval of socially-aware digital content. It presents the archivist view of professional media organizations and the specific requirements for successful retrieval of content. The content derived from the social media analysis is enormous and appropriate actions need to be taken to avoid irrelevant and/or repeated social information in the displayed results as well as over-information. The archivist feedback reveals the way humans address the social information as presented in the form of metadata along with the archived raw content and how this drives the design of a dedicated search and retrieval application.**

Keywords-search and retrieval user interfaces; social network information; archiving; preservation; user interface design; usability

I. INTRODUCTION

Social Media provide a vast amount of information identifying stories, events, entities that play the crucial role of shaping the community from continuous user involvement [2,3,4]. This work reports on the design and development of a socially-aware search and retrieval application that has the main goal of enabling archivists and researchers to retrieve archived web content based on semantic information derived from social media. Social media categories (SMC) cover almost all existing social media networks that information can be harvested from [1]. This information can be categorized according to type (multimodal: text, video, sound, picture) and role players (agents, users, opinion leaders) and actual semantic meta-data (entities, opinions on entities, etc.). The aim is to design a user interface that uses an engine for semantic analysis taking into account several modalities (plain text, sound, image, video), social media crawling, contextual search fusion and archivist usage requirements. The interface is the means to accessing the vast amount of information that can be archived. This work is ongoing and the findings of the pilot user testing and evaluation provide indications on how the semantic analysis of the social media information can be integrated to the design methodologies for user interfaces resulting in maximization of user experience in terms of social information involvement. The following sections describe the motivation and related work, the initial design considerations and the user evaluation that resulted in updated requirements for the next iteration phase of the design.

II. MOTIVATION

There are several approaches and applications in the recent years for digital archiving. However, these approaches mostly specialize in the archiving process itself without taking into account the actual impact of the archived information to the world. The need for leveraging the wisdom of the crowd for selecting the optimal content for future preservation has been stressed by digital libraries and the broadcasting world [5]. Concrete requirements for reporting the peoples' opinions on important events and associating them with candidate content pages for have been laid out. These stress the important role of the social media for such task. Harvesting the opinions of the people and enabling the retrieval of such social information for the future generations will provide a unique eye in the history and preservation of events. It also validates the selection of the content appropriate for archiving based on the impact of the recorded information that was reported in it. A way of gathering and analyzing such information from the web and building innovative ways to retrieve it was, therefore, a crucial requirement for the above process.

Earlier works treat archives independently of any externally provided social information but allow for manual

 ISBN: 978-1-61208-250-9

or semi-automatic annotation of the relational tags between items [6]. Such applications are in effect annotation interfaces rather than search and retrieval. And for such tasks there were formal models used for the annotation procedure [7]. The approach discussed in this paper goes beyond simple consideration of either an interface for search and retrieval of archived documents or social network dedicated interface. It is a dedicated interface for archived documents that are heavily enriched with social web metadata as analyzed and ingested data derived from social media.

The task before us was to design an innovative interface that would use the semantic information to satisfy complex queries as well as provide semantically enriched views of the archived document information. In order to do that, the user feedback was collected to refine the design considerations.

III. INITIAL DESIGN CONSIDERATIONS

The requirements for the search and retrieval of semantically analysed archived content were considered on the domain, content use and interaction levels. As it can be seen, the social web content integration and the raw data and their relations were a key to this exploration.

A. Domain-specific Requirements

Domain-wise there are certain needs for specific use within organizations. Media organizations have journalists that retrieve selected data on events, topics and entities for articles or documentaries that are under production. Those data need to be as complete as possible and provide a variety of sources and other related articles.

B. High-level Requirements

In the broadcasting world, collections of information targeting a common set of topics and events are called “campaigns”. Examples of campaigns span from generalized, large size, long-time types like “The EU financial crisis” to the more focused, short-lived types like “EU Summit 2012” or “Presidential Primaries 2012”. The former may span several years, the latter only a few months. Taking the above into consideration, a main functionality for our initial design approach is continuity of data collection and archiving. The scheduled crawled sets and analysis should span the entire campaign both in terms of time as well as in terms of social network content provision. Activity should be monitored and data collected accordingly. The sub-campaigns should correlate to the main campaign that they are supposed to belong to. The users should be able to retrieve information from all relevant archives and continuity of the semantic analysis should be preserved throughout the campaign search and retrieval process.

The database plays an important role in the architecture of such application, as it provides the storing, indexing and retrieving functionality for all the data collected and utilized by the archivists for the refinement process and searched upon by the generic users. As such, it is expected to support various types of data, handle updates and accommodate many types of queries.

The required functionality of the database is defined by its interaction with the search and retrieval user interface.

The database is expected to store: a) the original web content fetched by the crawler during for specific campaigns, b) metadata, meaning the extra attributes that are derived from the application aware crawling (e.g. author information, #retweets for a tweet etc.). The database should be able to answer quantitative as well as qualitative related queries posed by the archivists to further guide the filtering and retrieval processes. All end-users must be able to navigate through the stored content by posing queries that concern either the annotated meta-data or the semantic information derived from the analysis of the social content.

C. Non-functional requirements

Usability testing in our case required the study of users preferences towards the social media derived information. Semantic analysis results have enriched the archived content with semantic descriptions and tags that include entities (persons, locations, dates, etc.), events, topics, opinions on them, trending, cultural dynamics and more. Each web document may include several pieces of semantic information that may or may not be useful to the archivist or end user that would search and retrieve. Before proceeding with mockup design, experimentation has taken place in order to derive knowledge on the user perspective. Results from experimentation with fusing and visualizing social content with semantically driven context-sensitive information were derived [8]. Based on them, the mockups have been designed so that the socially aware semantic information would fully complement the search results. The users have indicated that opinions from social networks were considered fundamental to their understanding of the impact of searched web content. Furthermore, the semantic meta-data were integrated to the design of the interaction flow, so that users may facet and filter their search by using several levels of semantic content as well as prominent traditional information such as social network source information.

IV. THE PROTOTYPE

Iterative design process from the start to the implementation of the first prototype was followed from the initial design requirements gathering (made via wire-framing) to the mockup user environment, to the online prototype. The mockups were constructed based on the three types of information from the functional specifications (Search filtering, Core content, Social content) and the non-functional specifications described above. The initial experimentation was on the observation of the user perception of processed content (filters, tag clouds, paths), direct content (item descriptions, authors, dates, type of modality) and social content (opinions, trends, semantics, entities, events).

Based on the above, low fidelity mockups were created for the web retrieval interface. Those were subsequently evaluated for the core functionalities (filtering of results, follow-up search, result visualization) as well as usability (user approach to semantic search, information load, user effort, acceptance). The initial evaluation resulted in the set of specifications laid out by professional users (archivists,

 ISBN: 978-1-61208-250-9

broadcaster researchers). This specification was used for the first version of the application prototype.

The prototype itself used real data from the #greekfinancialcrisis (Twitter) as the main social web source as well as crawled web pages using as parameters the entities that were identified from the #greekfinancialcrisis text analysis. The raw content and the semantic metadata were indexed in Solr. Free text search as well as query enabled semantic search comprised the search functionality. The returned results were web resources that matched the search string. The users were able to view the search results in the main content frame. The results had distinct indicators on the opinion and trending values for each web resource. Dynamically created facets were provided for refinement of the results. Each result was, in fact, web page from the raw content archives. The content that could be viewed were not the original web pages themselves but descriptive data taken from them, such as title, initial description text, source, author. That view was enriched with lists of named entities (people, organizations, locations, etc.), events, and opinions derived from the text analysis. Furthermore, data from the most relevant tweets were provided where the users could use to see the following:

- Timeline showing the opinions of the Twitter authors on the entities contained in the web page
- Lists of positive and negative tweets on the entities of the web resource

V. REFINING THE GOALS

The prototype was used for the user evaluation which, itself, had the main purpose of monitoring the usage in order to construct and validate updated user requirements

A. *Application specifications*

A screenshot of the prototype is shown in Figure 1. The points of interest are pointed out by the highlighted numbers and included the following:

- Single semantic search (Fig.1 point 3)
- Advanced search
- Refine search
- Dynamic filtering via faceting (Fig.1 point 4)
- Sorting (by modality, by source SMC) (Fig.1 point 5, 6)
- Item viewing properties and derived functionalities:
  - Modality views (all, text, image, video, sound)
  - Social Media related information per item (current opinion, trending, latest tweets, source SMC)
  - Text/image analysis related information (entities, events)
  - Tag cloud support for major entities/events for quick refinement
  - SMC and text/image analysis items with linked specific functions (preliminary functions include search refinement, new search)

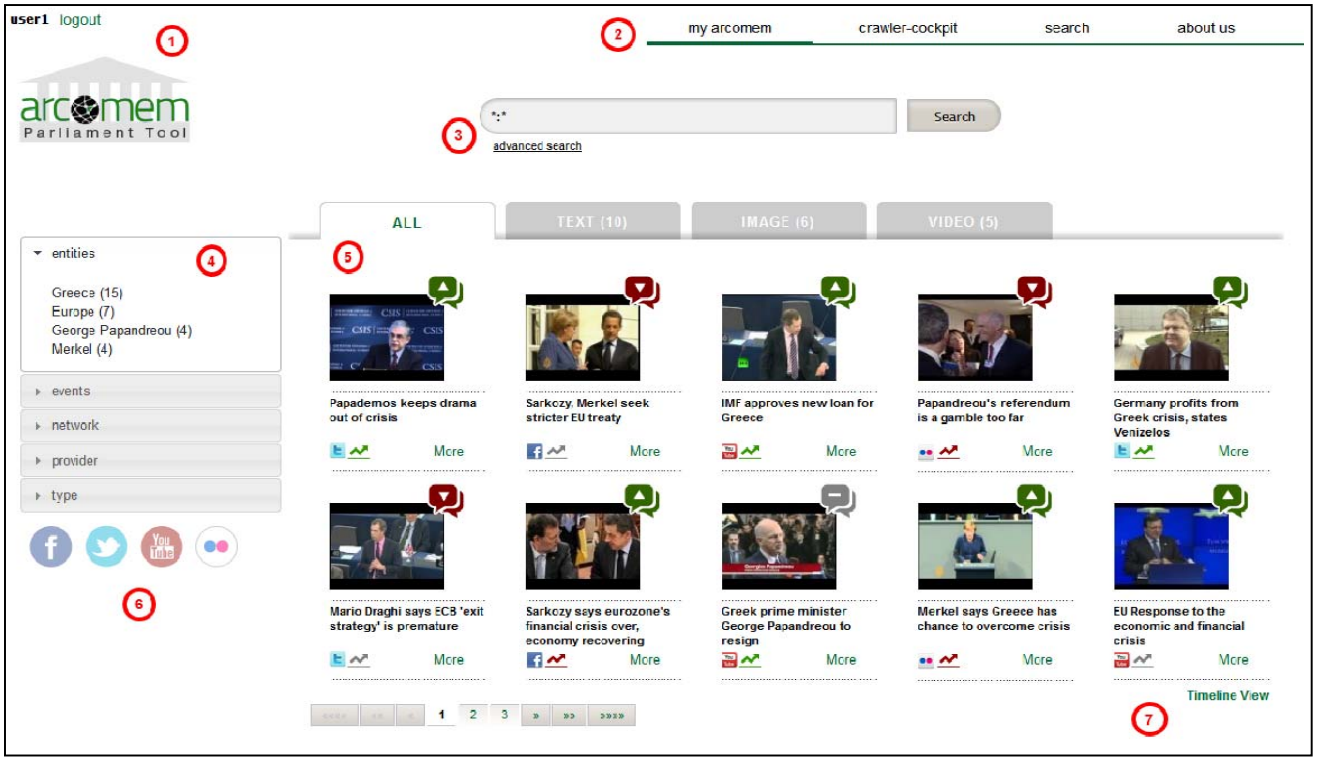

Figure 1. Search results for semantic search within a political domain.

- Generic viewing properties:
  - Recent searches

- Recent campaigns
- Latest news (promotional, user-specific)

Also, as highlighted in Figure 1, main placeholders and associated functions or visual elements we clarified:

- User login and application logo
- Top level menu reserved for high level functions
- Search box (also toggling single and advanced search)
- Dynamic filtering via facets
- Main results page with modality sorting tabs
- SMC filtering (results are filtered by selecting Social Media icons - toggling)
- Timeline view enablement button

The item results page (Fig. 2) provides all retrieved information about an archived web resource. The standard information includes:

- Title
- Details
- Date
- Provider
- Format
- Description

The semantic information includes and is presented to the user as:

- the identified associated entities
- the identified events and topics
- a tag cloud of entities derived from items that include the entities/events of currently web resource.
- The related events of the current identified event (if any)
- the latest tweets that contributed the major entities and opinions for the current web resource
- the major twitter accounts that contributed to the above
- the timeline that shows the positive-negative social network input regarding the associated entities (currently only for Twitter for the first prototype)
- the lists of positive and negative Tweets for the above

B. *Evaluation*

The early prototype that was build based on the previous feedback and specification has been evaluated by the professional end-users in order to refine the goals of the application based on the archivist/broadcaster workflow. So, in that respect, usability was measured qualitatively rather than quantitatively and was guided by specific search and retrieval scenarios.

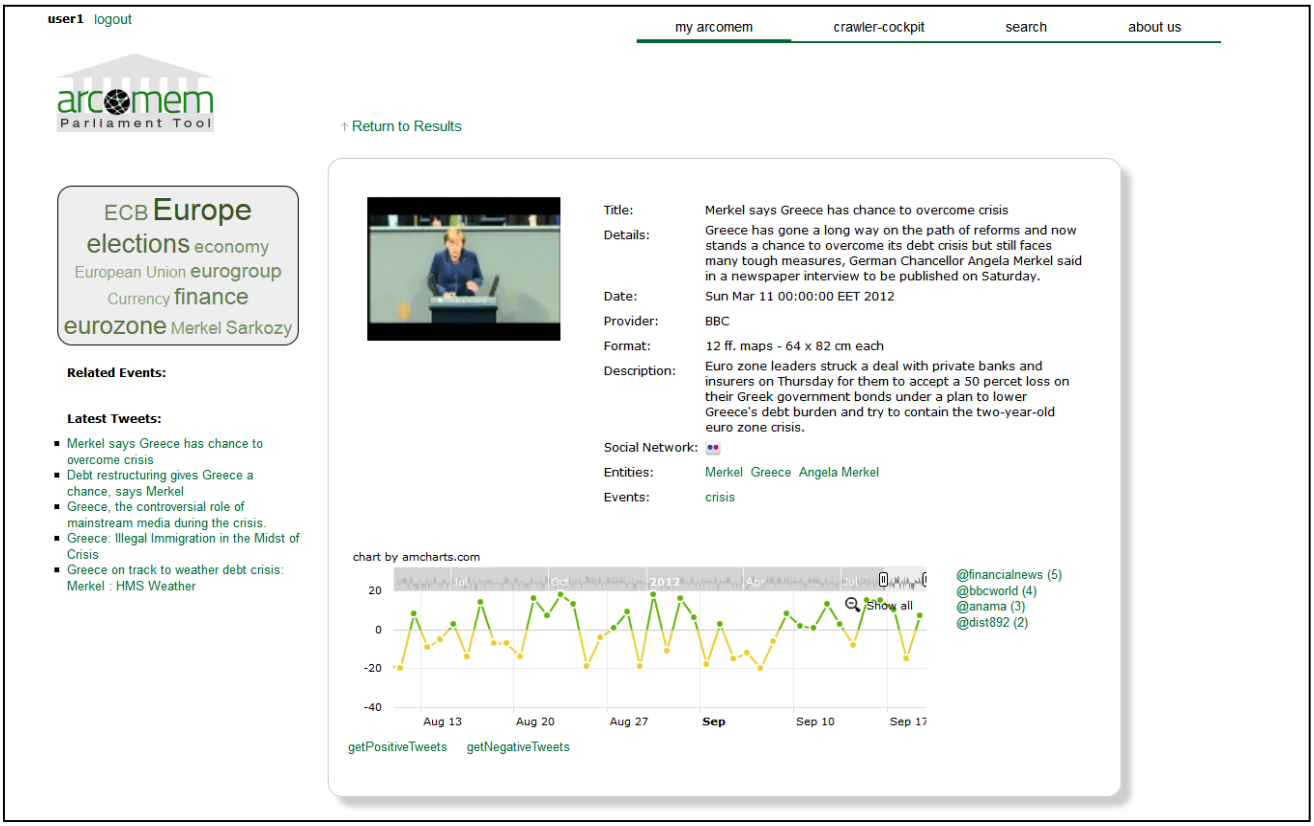

Figure 2. Web object – an archived web page enhanced with semantic information

Evaluation scenarios were created and deployed for novice users. The expert archivists performed free search actions aiming for in-depth social information research on specific content that they knew already. The scope of the evaluation was to enable validation of present functionalities as well as requests for extensions. Another task was to rank the types of semantic information from highly relevant to irrelevant for as many queries as possible.

Those data are fed to the facet ranking algorithm, so that the dynamic facet selection options are optimized. Table 1 provides an overview of the elements of the prototype testing that were evaluated.

TABLE I. EVALUATION OF THE PROTOTYPE APPLICATION

Actions	**Evaluation consolidated overview**		
	Expected results	***Validation***	***Extension***
query	Single and advanced query	Yes	Yes
filter	Facetting/filtering search	Yes	No
single page	Viewing a web page	No	Yes

This evaluation provided an updated set of specifications from the point of usability in terms of system interaction but unavoidably that would be dependent on actual content. The new specification requirements include:

- Ability to search for campaigns (top page functionality) based on campaign tags or description.
- Latest/popular campaigns
- Tag cloud that displays campaign keywords.
- Campaign overview (characteristic entities, events, SMCs, opinions)
- Campaign information (ID, logo, title).
- Information on volume of social/semantic data
- Types of result presentation (list, timeline, map)
- Connection to Wayback machine.

The next iteration will involve the advanced prototype with vast amount of semantic data attached to the archived pages that should be tested both functionally and for usability.

VI. CONCLUSION

This paper reported on the design and testing issues for an application for socially-aware web archiving. Important findings on how archivists and researchers view the social information and how that is integrated to their research workflow have been identified. So far, the social/semantic information is used to guide the search and retrieval process by a large margin. More than 80% of the content used comes from social networks. That shows how important that information is for web archiving and how that information provision can be optimized from a dedicated application interface.

ACKNOWLEDGMENT

The work described here was partially supported by the EU ICT research project ARCOMEM: Archive Communities Memories, www.arcomem.eu, FP7-ICT-270239.

REFERENCES

[1] ARCOMEM: Archive Communities Memories, www.arcomem.eu, FP7-ICT-270239 [retrieved: February, 2013].

[2] K. Church, J. Neumann, M. Cherubini, and N. Oliver, “SocialSearchBrowser: A novel mobile search nd information discover tool”, Proc. 14th Int. Conf. Intelligent User Interfaces, ACM, 2010, pp. 101-110.

[3] J. Golberg and M. Wasser, “SocialBrowsing: Integrating social networks and web browsing”, Proc. ACM CHI2007 Conf. Human Factors in Computing Systems, April 28 – May 3, 2007, San Jose, USA. pp. 2382-2386.

[4] M. De Choudhury, S. Counts, and M. Czerwinski, “Identifying Relevant Social Media Content: Leveraging Information Diversity and User Cognition”, Microsoft Research, ACM, 2010.

[5] G. Schefbeck, D. Spiliotopoulos, and T. Risse, “The Recent Challenge in Web Archiving: Archiving the Social Web”, International Council on Archives Congress, 20-24 August 2012, Brisbane, Australia.

[6] M. Agosti and N. Ferro, “A formal model of annotations of digital content”, ACM Transactions on Information Systems (TOIS), 26(1), pp. 3:1–3:57.

[7] C. Kohlschütter, F. Abel, and D. Skoutas, “A Novel Interface for Exploring, Annotating and Organizing News and Blogs Articles”, Proc. 3rd Annual Workshop on Search in Social Media (SSM 2010), co-located with the ACM WSDM 2010 Conference, 3rd February, 2010, New York City, USA.

[8] D. Spiliotopoulos, E. Tzoannos, P. Stavropoulou, G. Kouroupetroglou, and A. Pino, “Designing user interfaces for social media driven digital preservation and information retrieval”, Proc. 13th Int. Conf. on Computers Helping People with Special Needs, 11-13 July, 2012, Linz, Austria, pp. 581-584, doi: 10.1007/978-3-642-31522-0_87

 ISBN: 978-1-61208-250-9