Digital Archives: Semantic Search and Retrieval

Dimitris Spiliotopoulos[1], Efstratios Tzoannos[1], Cosmin Cabulea[2], and Dominik Frey[3]

[1] Innovation Lab, Athens Technology Centre, Greece
{d.spiliotopoulos,e.tzoannos}@atc.gr
[2] New Media, Innovation, Deutsche Welle, Bonn, Germany
cosmin.cabulea@dw.de
[3] Documentation and Archives, Suedwestrundfunk, Baden-Baden, Germany
dominik.frey@swr.de

Abstract. Social media, in the recent years, has become the main source of information regarding society's feedback on events that shape the everyday life. The social web is where journalists look to find how people respond to the news they read but is also the place where politicians and political analysts would look to find how societies feel about political decisions, politicians, events and policies that are announced. This work reports on the design and evaluation of a search and retrieval interface for socially enriched web archives. The considerations on the end user requirements regarding the social content are presented as well as the approach on the design and testing using a large collection of web documents.

Keywords: digital archives, social networks, user experience, big data.

1 Introduction

News, events and views are part of the everyday people's interaction. In the times of social media, people communicate their thoughts and sentiments over the events and views that are presented to them via social networks [1, 2]. Their views are opinionated and can be viewed and responded to by the rest of the community. Organizations that are involved in the process of collecting and processing the people's views, such as Broadcasters and Political Analyst Groups, try to harvest as much of the content as possible on an everyday basis. Instead of blindly searching for possible responses to news one by one, they use specialized tools to collect, analyze, group, aggregate, fuse and deliver the people's opinions that are relevant to their analyses [3]. Broadcasters use the social network information for two purposes. The first is for the classification of importance of events or entities (persons, locations, etc.) reported in the news. The second is to further refine their search on opinions for specific entities or events that have exhibited some kind of importance, for example opinions about a certain person that were very diverse or polarized.

The target of this work is an important problem of the recent years: big data and the approach that has to be adopted by the HCI researcher in order to create an interactive system for users to appreciate and reason the results of the big data

A. Holzinger and G. Pasi (Eds.): HCI-KDD 2013, LNCS 7947, pp. 173–182, 2013.

analysis. It presents the exploratory usability studies, in which user groups and usability requirements are identified and the follow up report on the archivist and end user findings.

This work reports on the design and evaluation of a user interface for search and retrieval of archived web documents. We are particularly interested in the user experience design and testing involved with the aim of building a final user interface prototype that will enable both archivists and end users to search into collections of archived content and retrieve, social web information such as opinions, sentiments, key peoples' names, and so on.

The next session provides the related work followed by a discussion on the user requirements for the Search and Retrieval Application interface (SARA) and the core functionalities. Then, the user experience methodology is described along with preliminary results and challenges faced. Next, the design considerations along with certain technical approaches are presented. Finally, the user interface prototype is presented along with the analyzed results on the archivist and end user activity.

2 Related Work

Web archiving is about striving to preserve a complete and descriptive snapshot of the available web data for the future. In the always-changing Web, a dynamic approach of content selection and appraisal is important to ensure that the web data that will be archived are of high quality, present a complete description of the selected area of context and that this description is persistently retained in the resulting archives.

For content and language analytics, social networks are a major part of the Semantic Web [4]. Social network information is the focus of major research because of the vast variety of content authors and the potential of the analyses that can be performed [5]. The crawling process collects the data according to parameters set by the archivist. The data quality at this point can be measured using specific crawling strategies [6]. The collected content is analyzed and annotated with semantic meta-information. The semantic data are archived as meta-data for the web data.

Identified *entities* are the most common result of linguistic analysis and the task of searching for entities in social networks involves the recognition of entities and relations [7]. The *news* domain is a large area of application for sentiment and includes many sources such as news web sites, blogs, RSS feeds and social media [8, 9] Entities are the most important element for creating training sets for sentiment analysis [10]. They are used to describe the ontologies needed for sentiment analysis for texts in both generic and specialized domains such as the arts [11]. Moreover, entities and relations can be modeled for ontological approaches beyond traditional polarity sentiment, for example for modeling emotion recognition by relating entities with human emotional states such as arousal and pleasantness [12].

Especially for the Social Web, from the moment that social networks such as Twitter provided an API for collecting information, sentiment analysis can be performed in a multitude of ways. Saif et al. have used features like semantic concept to predict sentiment on Twitter data sets [13] while other works present ways of

analyzing sentiment using hashtags as feature annotation [14]. Applying sentiment analysis on collected texts requires that the text be processed to ensure that is valid and clean such as all data are for the intended language, the non-textual segments from the collection process are removed and so on. Petz et al. presented a process model for preprocessing such corpora [15].

The design and development of dedicated search and retrieval interfaces for accessing archived content is a dedicated task in the area of web archiving [16]. Hearst et al. presented the requirements for a successful search interface where they stressed the importance of successful faceting and browsing of the content using metadata [17].

Semantic search [18] can be performed over data that include semantic metadata and involves the use of complex semantic queries. Methods have been proposed to simplify the complexity of the semantic queries by translating keywords to formal queries [19]. However, from the end user point, the complexity of the semantic queries should be left out of the user interaction flow by providing the means to make semantic search simple of all users as a step towards maximizing usability [20]. Such simplicity can be achieved by providing ways to faceting through metadata after a simple initial search, allowing the user to dive deeper into the semantic context rather instead of requiring complex semantic queries to be entered at the initial search [21]. Including both browsing and refining qualities on the faceted metadata can further optimize faceting [22].

Usability is a major factor for measuring success for semantic web applications [23] while success is measured taking in to account the human factor, the user [24]. Understanding the user behavior behind the search workflow is paramount to designing a successful search interface [25]. Studying the intentions of the users and their expectations of the search and retrieval process can be the basis of a successful user centered design [26]. Works on usability testing suggest that a search interface should empower the users during the whole information retrieval process [27].

There are many usability evaluation methods that can be deployed for evaluating web interfaces. Evaluating usability entails both usability inspection and usability testing methods to be applied and a successful usable design should be a product of iterative usability evaluation that involves the users in all stages of the development lifecycle [28].

3 The Rationale behind the SARA User Interface

SARA is an integral part of the ARCOMEM project [29], the purpose of which is to leverage the wisdom of the people for web content selection and appraisal for digital preservation. The typical system process for the collection and analysis of the web and social web data is as follows:

(a) The crawling process collects web pages [30] and social web sources [31], based on initial seed lists and keywords that collectively describe a domain (e.g. EU economic crisis).

2. During the online phase the system analyzes the collected data and produce information regarding the sources, dates, reputation statistics, etc. The data are stored in an appropriate document store [32].
3. The semantic analysis, the offline phase, analyses the web documents and the social data and produces the semantic information. The semantic data are stored in a Resource Description Framework (RDF) structure in order to be easily searchable by semantic queries.
4. A high-level analysis is performed on the socially-derived data from which statistics like opinion mining, cultural analysis, entity analysis, are obtained.

Users that comment on news, events and entities are part of the data that are collected and analyzed in order to select the most important opinions, views, key players and roles that are, in effect, the targets of interest of the communities. Entities, such as people and locations, are identified and their importance as well as the user opinions on them is examined. Related entities are also discovered and further analyzed. Information from the social networks is linked to other web sources in order to provide a complete picture of the events that shaped the opinions of the readers.
The aim of the SARA web interface is to provide the means for the archivists and journalists to semantically search the vast amount of data (raw, semantic, social, analytics), retrieve and visualize the content so that all semantic links between the user search and the data are retained. Data include:

- Web resources (text, images, videos)
- Semantic information (sentiments, opinions)
- Entities (people, locations, events, etc.)
- Social network sources (statistics, user name, location, activity, etc.)

The above data had already been processed by the analysis modules of the ARCOMEM system so that more information about the semantic relations and the social web analytics has been stored. A non-definitive list of such data for a specific set of search parameters is:

- List of most relevant social media posts for one or more entities
- List of most diverse social media posts
- Topic detection
- The most influential users from social media
- Entity evolution information

The images and videos are content types that are not processed semantically but directly to provide indication of duplication of documents (i.e. news articles that contain the same video or picture may also be duplicates) or information on the evolution of entities over time. An example for the latter are pictures depicting the same entity at different points in time that provide explicit verification that the entity description has evolved, e.g. Cardinal Francis was formerly Cardinal J.M. Bergoglio.

4 User Experience Considerations

The user requirements collection was initially based on the expected functionalities by the two main user groups, Broadcasters and Parliament Archivist. The initial list of requirements that was based on the conceptual design of such web interface was very long, because of all the possible results expected from the analyses of the web data. The target users were overwhelmed by both the broad potential uses of the analyzed social information but also from the, at that time, unknown usefulness of each bit of that vast amount of data. That realization has made obvious the fact that the core advantage of the ARCOMEM approach, that is the content diversity and social web data potential, was also the main problem to solve regarding the actual design of the user interface. The provision, type and quality of the analyzed data would have to drive the user interface design and interaction process. In order to tackle this, it was decided that the best approach would have to be a combined focus group discussion on the user-system interaction and a heuristic approach during the requirements gathering. In our case, the classic low-fidelity prototyping would have had minimal, if any, success, since the user interaction would be driven by the actual content.

During the focus group discussions the archivists, which are experts in search and retrieval interfaces, were presented with possible approaches using examples of real web interfaces that are used for archived documents, like the Europeana portal [33]. A quick breakthrough came when it was realized that the social media content itself as well as the analyzed semantic information from the social media was the most controversial part of the user interaction. The users were very interested in using the semantic data for their search but were not fully aware of why that information was there and where it was derived from. It was also obvious that different levels of importance could be assigned to the types of semantic data.

The user perception on the importance of the types of semantic information for search and retrieval of web documents was an open research question as well as a pre-requisite for the user experience design of the SARA interface. A series of experiments were run on first time users using an early demo interface populated with semantic information [34]. The populated data were carefully selected and several delivery options in the user interface were explored (in text, separate lists, tag clouds, facets, etc.). The think aloud approach was used to get the subjective user feedback but also a simplistic analysis of the user path selections was done by logging the user clicks and time. The above approach has led to an initial set of functional and non-functional requirements that were used to create the high-quality interface prototype using a small subset of content data for the formative evaluation.

5 Design and Technical Challenges

The design of the interface prototype was based on the archivist and journalist end-user feedback. The expected type of information from the retrieval process by both main user groups was formalized. The obligatory entities and opinions were first on the list but so was marking and ranking of the most reputable sources. The journalists

reported that events, both standalone (if unprocessed) or aggregated (if such process was available) were at the top of their priority. It was also asked that Twitter users, blog posters and other entities that report events should be analyzed for their reputation status, location and any other values that would assign them a "trustworthy source" for the reported event. This requirement bears similarities to an earlier study where journalists placed importance to the type of user that reports in the social media or the web, assigning the "eyewitness" versus "non-eyewitness" and the "journalist/blogger" versus "ordinary individual" values [35]. Apart from the above, the users asked for lists of relevant Twitter posts for each web resource. A web resource can be a web page, blog post, wiki page, and so on, but not only. Web resources are comprised of several web objects such as Twitter posts, blog comments, and other pieces of information that bear direct semantic relation to the bound entities and events.

The semantic information was chosen to be indexed using the named entities as the basic starting point. Each entity may have other entities associated with it, opinions, participate in an event, and so on. This approach provides the advantage that a complete list of attributes is always available for each entity. But it also requires several hops through the RDF storage via SPARQL queries in order to collect it. Moreover, SPARQL is slow and inefficient for such queries and does not support full text search. RDF storage search does not offer functionalities like faceting, hit highlighting, lemmatization, stemming, etc. In order to allow for fast indexing, it was decided that a full text search engine should be deployed.

In order to minimize response times, it was decided to fully populate the Solr index offline. For this purpose, it was needed to migrate most of the existing information from the RDF triple store to the Full Text Search Engine beforehand. The best practice for this was to convert the RDF triple store into a set of flattened documents that are compatible with the Solr schema. These flattened documents had to be populated one by one through an indexing process with values retrieved from the RDF triple store.

For this task, a custom indexing module in Java and Groovy was implemented. Parts of Groovy dynamic programming language were added to create a kind of domain specific language. This small domain specific language (a set of classes) offers to an external user the possibility to easily change indexing rules on the fly.

There are several approaches for the indexing procedure. Nevertheless, the most appropriate due to the large volume of the knowledge base was to apply an incremental formulation of Solr documents and index them one by one. This practice, though more time consuming, has minimal requirements for memory in both indexer and Solr Handlers while indexing. Moreover, it can be stopped anytime during indexing and resumed afterwards, without affecting the whole process.

Each Solr document is representing an RDF web resource instance. It starts by retrieving a set of web resource-ids Iterator from the RDF triple store. Then, the process iterates through the ids of the web resources, and, for each id, all the required fields referring to this web resource instance are retrieved by SPARQL queries. Once the Solr document is fully populated it is inserted into Solr. The type of fields of the Solr schema has been appropriately configured for the purposes of this case.

6 The SARA Prototype User Interface

The interface prototype was releases and a small scale domain on the Greek economic crisis was indexed. This domain was analyzed and tested manually in order to produce a large set of semantic data that are free of statistical errors. The obvious purpose for this was that the user feedback should be unaffected by potential error from the semantic analysis.

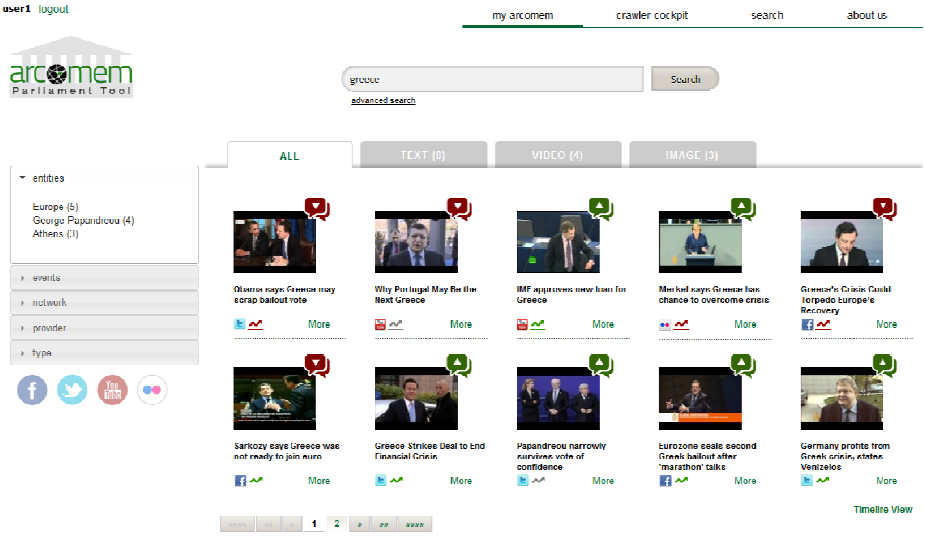

Fig. 1. Results of semantic search within a political domain

Figure 1 depicts a typical search results page. Social media derived data are used for search refinement and for the document opinion and trending visual preview. The entity terms that were extracted from the social media are used both as search parameters as well as facets.

The search results may be sorted by modality and social network source. Each search result contains the aggregated opinion of the web resource textual data (top right) and the social network source and opinion trending (bottom left). The user may perform semantic or full text search and retrieve all web resources that contain the list of search terms. The faceting is used to drill into the search parameters.

Figure 2 shows the web resource page that lists, apart from the raw content:

- Entities and events extracted from the document text
- Related events and latest social posts that are related to the search results (bottom left),
- Tag cloud that is dynamically comprised of entities and events retrieved from the search results (top left) and can be used for quick follow-up search,
- Twitter posts ranked by opinion polarity (far bottom),
- The most influential users relevant to the web resource content (bottom right),
- Timeline depicting the opinion over time for the associated entities of the resource.

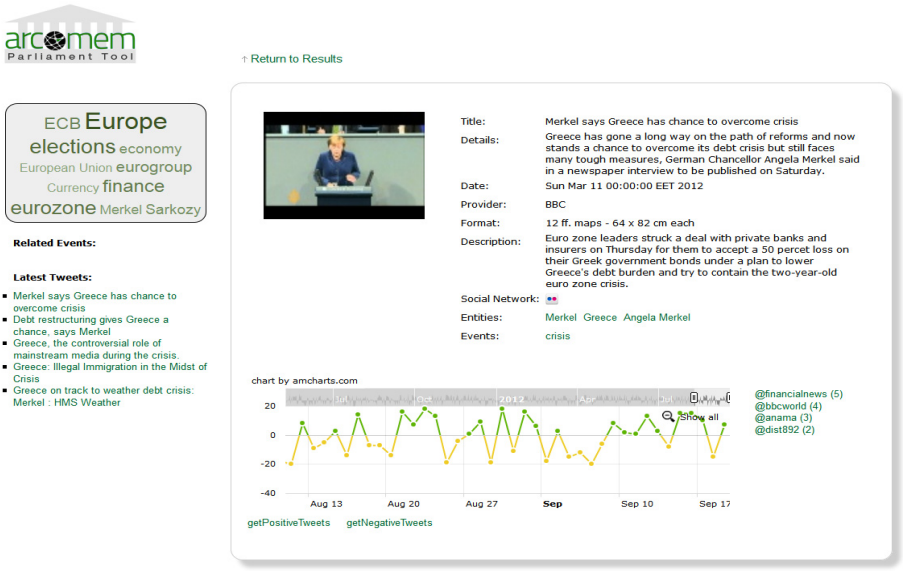

Fig. 2. A web resource page, enriched with social information

From the logs of the user interface usage, there are indications that two main user types are identified: the archivist and the end-user. This is a significant change from the initial assumption that user types such as a journalist and a political analyst end user are different types. The archivist proceeds in a well structured manner and goes through several items of the same hierarchical level in order to fulfill a sense of completeness for the archived data. The end-user, on the other hand, employs a more aggressive approach that drills down into the content, entities and opinions, in order to collect several examples that establish and validate a news story.

Archivist. Views the content and evaluate availability and completeness of the selected web documents for digital preservation. Expert user, highly trained to locate missing groups of information, web resources and semantic.

End-User. The generic user type. It includes researchers in news reporting, such as broadcasters and journalists, as well as researchers in other fields that are interested in the social web information. The latter may be policy makers, political/parliamentarian assistants, students of social sciences, law, and so on.

7 Conclusion and Further Work

This work reported on the user experience design considerations and findings for the design of a search and retrieval web interface for socially-aware web preservation. The next iteration of the design will involve the end users more actively. Transcribed scenarios will guide the users to perform certain actions in order to evaluate their interaction with the system. Free form search and retrieval will also be monitored in order to assess the efficiency of the current design and identify factors that may lead

to usability performance optimizations by letting the users exploit the full potential of the social content. User training sessions are expected to provide new insight into how archivists and journalist end users extensively access the interface to its full potential. The expected results could be very revealing as to the nature of the archivist search behavior, eventually leading to a more refined user experience for both user types.

Acknowledgements. 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. Schefbeck, G., Spiliotopoulos, D., Risse, T.: The Recent Challenge in Web Archiving: Archiving the Social Web. In: Int. Council on Archives Congress, Brisbane, Australia, August 20-24 (2012)
2. Anderson, R.E.: Social impacts of computing: Codes of professional ethics. Social Science Computing Review 10(2), 453–469 (1992)
3. Golberg, J., Wasser, M.: SocialBrowsing: Integrating social networks and web browsing. In: Proc. CHI 2007, San Jose, USA, April 28–May 3 (2007)
4. Musial, K., Kazienko, P.: Social Networks on the Internet. World Wide Web 16, 31–72 (2013)
5. Torre, L.: Adaptive systems in the era of the semantic and social web, a survey. User Model. User-Adapt. Interact. 19(5), 433–486 (2009)
6. Denev, D., Mazeika, A., Spaniol, M., Weikum, G.: The SHARC framework for data quality in Web archiving. VLDB 20(2), 183–207 (2011)
7. You, G., Park, J., Huang, S., Nie, Z., Wen, J.-R.: SocialSearch+: enriching social network with web evidences. World Wide Web (2013)
8. Godbole, N., Srinivasaiah, M., Skiena, S.: Large-scale sentiment analysis for news and blogs. In: Proceedings of the International Conference in Weblogs and Social Media (2007)
9. Ruiz-Martinez, J.M., Valencia-Garcia, R., Garcia-Sanchez, F.: Semantic-Based Sentiment analysis in financial news. In: Proc. 1st Int. Workshop on Finance and Economics on the Semantic Web (FEOSW 2012) in conjunction with 9th Extended Semantic Web Conference (ESWC 2012), Heraklion, Greece, May 27-28 (2012)
10. Kumar, A., Sebastian, T.M.: Sentiment Analysis: A Perspective on its Past, Present and Future. IJISA 4(10), 1–14 (2012)
11. Baldoni, M., Baroglio, C., Patti, V., Rena, P.: From tags to emotions: Ontology-driven sentiment analysis in the social semantic web. Intelligenza Artificiale 6(1), 41–54 (2012)
12. Zhang, X., Hu, B., Chen, J., Moore, P.: Ontology-based context modeling for emotion recognition in an intelligent web. World Wide Web (2013)
13. Saif, H., He, Y., Alani, H.: Semantic Sentiment analysis of Twitter. In: Cudré-Mauroux, P., et al. (eds.) ISWC 2012, Part I. LNCS, vol. 7649, pp. 508–524. Springer, Heidelberg (2012)
14. Mukherjee, S., Malu, A., Balamuralli, A.R., Bhattacharyya, P.: TwiSent: A Multistage System for Analyzing Sentiment in Twitter. In: Proc. 21st ACM Int. Conf. on Information and Knowledge Management (CIKM 2012), Maui, USA, October 29–November 02, pp. 2531–2534 (2012)
15. Petz, G., Karpowicz, M., Fürschuß, H., Auinger, A., Winkler, S.M., Schaller, S., Holzinger, A.: On text preprocessing for opinion mining outside of laboratory environments. In: Huang, R., Ghorbani, A.A., Pasi, G., Yamaguchi, T., Yen, N.Y., Jin, B. (eds.) AMT 2012. LNCS, vol. 7669, pp. 618–629. Springer, Heidelberg (2012)
16. Lesk, M.: Understanding Digital Libraries. Morgan Kaufmann (2004)

17. Hearst, M., English, J., Sinha, R., Swearingen, K., Yee, P.: Finding the Flow in Web Site Search. Communications of the ACM 45(9), 42–49 (2002)
18. Guha, R., McCool, R., Miller, E.: Semantic Search. In: Proc. 12th Int. Conf. on World Wide Web, pp. 700–709 (2003)
19. Wang, H., Zhang, K., Liu, Q., Tran, T., Yu, Y.: Q2Semantic: A lightweight keyword interface to semantic search. In: Bechhofer, S., Hauswirth, M., Hoffmann, J., Koubarakis, M. (eds.) ESWC 2008. LNCS, vol. 5021, pp. 584–598. Springer, Heidelberg (2008)
20. Lei, Y., Uren, V.S., Motta, E.: SemSearch: A Search Engine for the Semantic Web. In: Staab, S., Svátek, V. (eds.) EKAW 2006. LNCS (LNAI), vol. 4248, pp. 238–245. Springer, Heidelberg (2006)
21. Makela, E., Hyvonen, E., Sidoroff, T.: View-Based User Interfaces for Information Retrieval on the Semantic Web. In: Proc. ISWC-2005 Workshop on End User Semantic Web Interaction (November 2005)
22. Ziang, J., Marchionini, G.: Evaluation and Evolution of a Browse and Search Interface: Relation Browser++. In: Proc. Conf. on Digital Government Research, pp. 179–188 (2005)
23. Nedbal, D., Auinger, A., Hochmeier, A., Holzinger, A.: A Systematic Success Factor Analysis in the Context of Enterprise 2.0: Results of an Exploratory Analysis Comprising Digital Immigrants and Digital Natives. In: Huemer, C., Lops, P. (eds.) EC-Web 2012. LNBIP, vol. 123, pp. 163–175. Springer, Heidelberg (2012)
24. Calero Valdez, A., Schaar, A.K., Ziefle, M., Holzinger, A., Jeschke, S., Brecher, C.: Using mixed node publication network graphs for analyzing success in interdisciplinary teams. In: Huang, R., Ghorbani, A.A., Pasi, G., Yamaguchi, T., Yen, N.Y., Jin, B. (eds.) AMT 2012. LNCS, vol. 7669, pp. 606–617. Springer, Heidelberg (2012)
25. Teevan, J., Alvarado, C., Ackerman, M.S., Karger, D.: The perfect search engine is not enough: a study of orienteering behavior in directed search. In: Proc. SIGCHI Conf. on Human Factors in Computing Systems (CHI 2004), pp. 415–422 (2004)
26. Chen, Z., Lin, F., Liu, H., Liu, Y., Wenyin, L., Ma, W.: User Intention Modeling in Web Applications Using Data Mining. World Wide Web: Internet and Web Information Systems 5, 181–191 (2002)
27. Taksa, I., Spink, A.H., Goldberg, R.R.: A task-oriented approach to search engine usability studies. Journal of Software 3(1), 63–73 (2008)
28. Holzinger, A.: Usability engineering methods for software developers. Communications of the ACM 48, 71–74 (2005)
29. ARCOMEM: Archive Communities Memories. FP7-ICT-270239, http://www.arcomem.eu
30. Faheem, M.: Intelligent crawling of Web applications for Web archiving. In: Proc. PhD Symposium of WWW, Lyon, France (April 2012)
31. Gouriten, G., Senellart, P.: API Blender: A Uniform Interface to Social Platform APIs. In: Proc. Developer Track of WWW, Lyon, France (April 2012)
32. WARC File Format specifications, http://archive-access.sourceforge.net/warc/
33. Europeana Cultural Collections Archive Portal, http://www.europeana.eu/portal/
34. Spiliotopoulos, D., Tzoannos, E., Stavropoulou, P., Kouroupetroglou, G., Pino, A.: Designing user interfaces for social media driven digital preservation and information retrieval. In: Miesenberger, K., Karshmer, A., Penaz, P., Zagler, W. (eds.) ICCHP 2012, Part I. LNCS, vol. 7382, pp. 581–584. Springer, Heidelberg (2012)
35. Diakopoulos, N., De Choudhury, M., Naaman, M.: Finding and assessing social media information sources in the context of journalism. In: Proc. 2012 ACM Annual Conf. Human Factors in Computing Systems (CHI-2012) (2012)