User Perception Knowledge for Socially-Aware Web Document Accessibility

Dimitris Spiliotopoulos[1], Pepi Stavropoulou[2], Georgios Kouroupetroglou[2], and Dimitrios Tsonos[2]

[1] Innovation Lab, Athens Technology Centre, Greece
d.spiliotopoulos@atc.gr
[2] National and Kapodistrian University of Athens, Department of Informatics and Telecommunications, Greece
{pepis,koupe,dtsonos}@di.uoa.gr

Abstract. Social Media provide a vast amount of information identifying stories, events, entities that play the crucial role of shaping the community in an everyday heavy user involvement. This work involves the study of social media information in terms of type (multimodal: text, video, sound, picture) and role players (agents, users, opinion leaders) and the potential of using that information for the design of accessible, usable preservation strategies. The challenge was to analyze the social web and present ways of preserving the web documents with social content in such way as to make them accessible for the future. The web documents should preserve accessible data and stored in such way as to enable intelligent retrieval.

Keywords: social media, meta-information analysis, user-driven, document accessibility.

1 Introduction

Web archiving approaches have recently included social media content collection procedures in order to allow greater coverage of the web content to be preserved. User input in social media had been vastly rising in the recent years, transforming them into major hubs of web information. This trend continues to rise and the social media are becoming more and more pervasive in all areas of life [5]. The major social media (Twitter, Facebook, and others) are the main means of communication for many people and information is constantly updated. The ephemeral nature of this information is a factor that has been very important for the analysis of social events and opinions that are expressed in the social media. People's opinions, actions, and even news network reporting are shaped by the very nature of the social web, where the social media information creation and communication now characterizes the information society.

This work described here is part of the research on social media analysis for web document archiving for the ARCOMEM project [3]. One of the major challenges of the web preservation process is the successful inclusion of social media information

C. Stephanidis and M. Antona (Eds.): UAHCI/HCII 2013, Part II, LNCS 8010, pp. 710–717, 2013.

[1]. Consequently, the social information must be handled cautiously in order to retain the social and semantic value that it bears throughout the whole archival process. This process includes online collection and analysis of the web content, linguistic analysis of the textual information and storing of the resulting meta-information to a semantic structure. From that point onwards, that information will need to be searched and retrieved in such way so that it actually makes sense for the user.

This work builds upon the results of the earlier experimentation on the user understanding of semantic information such as opinion, trending, and social network source identification [4]. The purpose is to try to associate the reader perception of the social information with the actual information types that were identified as main means of information provision from the social media texts on actual analyzed web documents. The results of this study can be the basis of the experience that can be applied to user interface design.

This document introduced the reader to the generic idea of the web preservation considerations regarding the social web. The following sections describe the motivation behind this work, the experimentation and results on the user perception of socially derived information. Finally, a discussion on how the results on the user perception knowledge can be used for the archived document information provision and visualization user interface design is provided.

2 Motivation

The motivation behind this work is that the design of any means of handling social information, such as the search and retrieval process of archived content, should include meaningful ways of utilizing that information as part of a design-for-all approach. The analysis of web documents on the semantic level based on social information leads to large amounts of information that can be used as is or combined in order to provide the text it refers to with high level data. That data should be preserved in order to be accessible; otherwise the source information connection with the text will be lost.

The analysis of the ARCOMEM system provides analytical data that can be used to deduct semantic entities/events/topics, sentiment and opinions, social and demographic information, user trust and loyalty, thematic context, timeline and impact, social multimedia content, contextual information (actors, influences) as well as generic social media specific data (e.g. social dynamics for Twitter). The social media-related sources include Blogs, Microblogs, Wikis, Social Networks, Video Networks, Photo Networks, Music/Audio Networks, Discussion groups, Social Bookmarks.

The content of web documents is archived in the form of *web resources*. Web resources may contain one or more web objects. Web objects may be text documents such as paragraphs or snippets of text from the web page, Twitter ports, blog comments, and wiki comments. For this phase of the work, from social networks, only Twitter posts were treated as web documents since they provide text-only posts that are heavily opinionated.

The procedure for collecting and archiving web content is as follows: the archivist sets up a crawling campaign, then the web content is collected and analyzed and

finally the content is stored in an appropriate document store such as WARC format [2]. The web documents are collected through crawling processes based on seed lists and keywords for web pages [6] and social web content [7]. Then, both the web and social web contents are analyzed in order to extract named entities such as people and organizations, events, opinions, cultural analytics, and so on [8, 9]. The semantic information is stored in an RDF format in order to be accessible by semantic search queries.

After the semantic analysis, there is an abundance of semantic meta-information available to be associated with the web documents. There are many possibilities that may drive that association and, from the user interface designer point of view, all depend on the usefulness and semantic bond between the raw content and the socially-derived semantic data. Those possibilities needed to be explored in order to successfully build an interface for search and retrieval of the archived content. A key factor for this approach was that the actual feedback from the users would be collected through a proper prototype interface. The users would be able to actually search and evaluate the results of the search on the semantic level. The social media derived information would be evaluated on the basis of importance and impact to the user refinement process.

3 Experiment Set Up

Ten users with experience in user interface evaluation were involved. The first task of the experiment was to evaluate the user perception of usefulness of certain pieces of information on actual web documents. Ten web documents were presented to users in random order. In our case, the web documents were comprised of webpages, blog posts and Twitter posts. Those web documents contained the social information (entities, opinions, events) derived from the search and retrieval application. For this task the actual documents were retrieved from the application and manually annotated with the social information so that the meta-information was fully visualized. The users were asked to click on items of social/semantic value and their actions were logged.

The second part experiment was performed by using the actual search and retrieval prototype interface in order to retrieve documents using semantic queries. The users were asked to evaluate the non-textual information (social trending graph, user location visuals) as to their perceived link in the document. The idea was to see how document specific information or even search specific information can be perceived by the users as important to the results. That would provide enough initial data for the designer to select and prioritize the work that would result in an accessible design of the – at the moment – visual only information.

Certain results may be either presented by graphical means, like interactive timelines or visual statistics, or by lists of data or aggregated data. Figure 1 depicts the visual representation of the location distribution of the Twitter posters for the collected web resources. This information is what the user may consult in order to

create the baseline for the results and to be able to make more educated judgments on the actual social results such as opinions, trends, etc.

Figure 2 shows a timeline presenting the opinions expressed from the Twitter posts referring to the specific search results. The opinion timeline shows the opinion about a European politician and that statement could be the basic information presented to the users by a modality like speech by an accessible interface. But, the important parameter regarding the social media user location information from Figure 1 would have to be included in order to transfer the full context of the timeline in Figure 2.

The visualization from Figure 1 provides the users the basis for all the visualized analytics that will be presented to them that are derived from the social data.

4 Results and Discussion

The above procedure was designed to evaluate the semantic information of the archived content and to provide feedback as to the expected usefulness as a total approach for the accessible design. The participants of the first part of experiment clicked over the semantic items and exposed opinions about entities and opinions as well as authors. The duration of each task was quite short, less than a minute for each of the 10 web documents. The major finding was revealed from the logging of their behavior. While they were interested in the opinions, they collected both positive and negative opinions at the same rate and almost in parallel. Their questionnaire response mentioned that they wanted to form a complete idea about the terms mentioned in the text before they would proceed to read the actual opinions. After they collected that quantitative information, they proceeded to the actual qualitative data by sampling the actual Twitter posts that contained the opinionated text. There were certain points when the participants selected to view all opinions from certain social media authors. Those opinions were mostly targeting terms that belonged to the same domain (e.g. politicians) but were not part of the data contained in the current web document. The participants later reported that they were intrigued by the choice of words by certain authors so they looked for other posts by the same authors. These authors were classified as *influencers*.

The second part of the experiment validated the findings of the first part. Moreover, specific feedback regarding the visualized data was collected. The users were asked to search for a specific entity and view the top five web resources that from the search results. They were asked to evaluate the visualized information and report their findings. The visualized data included the following: Twitter user geolocation (for all web documents), social media source distribution (for all web documents), opinion timeline (per web document), trending (per web document).

The participants went over all the information and were asked to fill in the questionnaire on how they perceived the above information. They reported that the combination of the visualized data provided the full results they were looking for. The opinions they read and the entities they retrieved and associated with those opinions made sense only in the context of the actual archived social data. The influencers were the sources of the most interesting opinions while they were themselves the

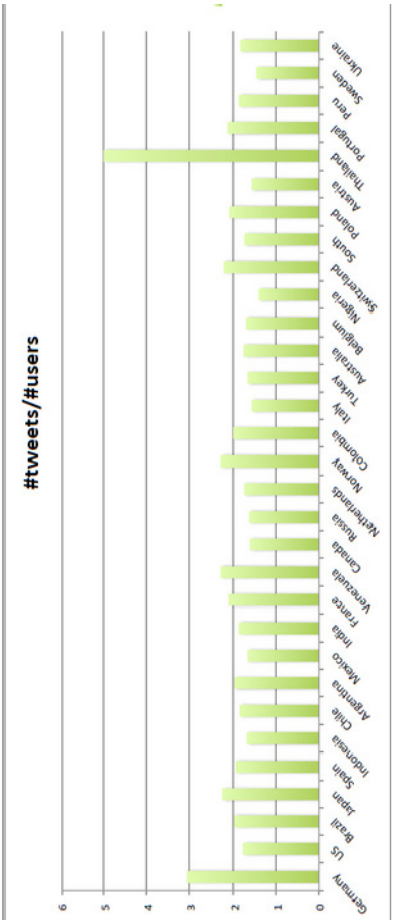

Fig. 1. Social media analysis visualizations

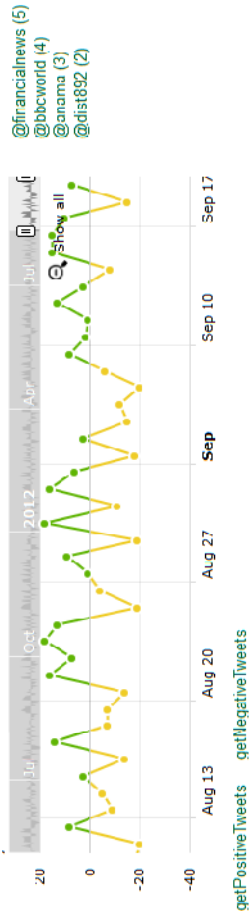

Fig. 2. Timeline: Peoples' opinion over time

product or content of a specific snapshot of a group of opinion leaders from a specific space in time.

The above analysis resulted in a formulation of the end-user perception knowledge of the socially derived information from web document. Human-computer interface designers go to great lengths in order to design and construct proper visualization models that would provide the users with the complete picture of the intended information through the user interface. It is, therefore, prudent that the model itself serves as the template for the provision of the necessary data for making the results of such analysis accessible.

5 Conclusion

This work had the task to measure certain parameters regarding the user perception of analyzed social data that are preserved in web archives from the social web. The results have provided insight on the use of social media derived semantic information on web documents that were retrieved from archives. The optimal use of that information will be the target for the optimization of the design of the search and retrieval interface that would eventually present the information to archivists or ordinary end-users along with the opinions, important entities, and other socially derived data. Additionally, the designer can use the results in order to make the analyzed data accessible through the interface.

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. International Council on Archives Congress, Brisbane, Australia (2012)
2. WARC File Format specifications, `http://archive-access.sourceforge.net/warc/` (retrieved: February 2013)
3. ARCOMEM: Archive Communities Memories, FP7-ICT-270239, `http://www.arcomem.eu` (retrieved: February 2013)
4. 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)
5. Brussee, R., Hekman, E.: Social Media are Highly Accessible Media. Paper presented at the WWW/Internet 2009, Rome, Italy (2009)
6. Faheem, M.: Intelligent crawling of Web applications for Web archiving. In: Proc. of the 21st Int. Conference on World Wide Web, Lyon, France, pp. 127–132. ACM (2012)
7. Gouriten, G., Senellart, P.: API Blender: A Uniform Interface to Social Platform APIs. In: Proc. Developer Track of WWW, Lyon, France (2012)

8. Maynard, D., Funk, A.: Automatic detection of political opinions in tweets. In: García-Castro, R., Fensel, D., Antoniou, G. (eds.) ESWC 2011. LNCS, vol. 7117, pp. 88–99. Springer, Heidelberg (2012)
9. Risse, T., Dietze, S., Maynard, D., Tahmasebi, N., Peters, W.: Using Events for Content Appraisal and Selection in Web Archives. In: Proc. Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011), in Conjunction with The 10th Int. Semantic Web Conference 2011 (ISWC 2011), Bonn, Germany (2011)