Relation Visualization for Semantically Enriched Web Content

Dimitris Spiliotopoulos[1], Ruben Bouwmeester[2], Dimitris Proios[3]

[1] Innovation Lab, Athens Technology Centre, Greece
d.spiliotopoulos@atc.gr
[2] New Media, Innovation, Deutsche Welle, Bonn, Germany
ruben.bouwmeester@dw.de
[3] Dept. Information and Telematics, Harokopio University of Athens, Greece
dimi.proios@gmail.com

Abstract. Many methods for web archiving include semantic analysis of the archived content. Texts from social media provide a rich feature set of analyzed data, such as topics, opinions, events, and more. One of the main challenges is the successful utilization of the semantic and raw data for an optimized search and retrieval approach. This work reports on the user needs for semantic relation visualization and the experimental approaches for creating visualizations that answer to specific user queries, based on the user interaction monitored by a dedicated semantic search tool.

Keywords: Semantic analysis / Twitter data / visualization

1 Introduction

Social media is a source of abundant user-created information that exhibits certain characteristics. The content is diverse, multimodal, opinionated, and can be classified according to popularity, influence, and other social factors. Current applications try to harvest social media content and present aggregated results to users. For Twitter data, there exist many applications that analyze and visualize sets of data according to user location information, Twitter post language, hashtag mentions, user mentions, post volume over time, keyword tag clouds, top X influencers, sources, URLs. The data that are used are meta-information already provided by the Twitter API and can be used for a helpful quick overview of statistics from data samples.

Recent approaches, mainly in web archiving, include social media content collection and analysis. Some of them, such as the ARCOMEM[1] approach, include a thorough post-level analysis on the semantic level of the content, identifying persons, locations, organizations, opinions expressed on them, topics discussed and events identified, clusters of semantically related entities, hashtags and users, and so on [1,2]. Applications may use the semantic meta-information to answer user queries such as “return all positive opinions on President Obama” that can also, with the use

[1] ARCHive COmmunities MEMories, www.arcomem.eu

of faceting, be refined to “return all positive opinions on Obama from European politicians on the economic crisis”. Such queries are concrete and clearly formulated to retrieve web and social web data that match one or more parameters, either from raw data or semantically analyzed meta-data. Results can be visualized in the form of bubble graphs and timelines using frameworks such as D3 and present the current common practices [3].

Semantic analysis is a key feature in web archiving. In the cases where social media content is also archived and analyzed, the abundance and complexity of the semantic information provides a key advantage for retrieval tasks as well as high complexity for user interaction. Semantic-centric approaches to data visualization may provide a direct mapping between user and data semantics [4]. That can be used to guide navigation over large collections of documents through semantic visualization [5]. However, one of the most important issues is the unobtrusive integration with the user interaction [6].

Semantic analysis has also enabled the design of semantic search and retrieval methods and applications that utilize the semantic information to a great extent [7]. Such approaches provide results that combine raw data (such as hashtags and users), semantic data (such as opinions and topics) and statistical figures on combinations of the above. This work examines two case studies. The first uses the #BostonMarathon #Bombings of 2013 Twitter data and examines how specific data may stand out during visual exploration. The second uses the US Election 2012 web archives and the ways of building semantically related content visualizations that can be used to display search results that correspond to semantic queries such as “show me everything on Obama and Romney” and follow up.

This section provided the introduction, motivation and related work while the next sections discuss the social media and user interaction driven visualization problem, experimentation, implication and methodology for search and retrieval from big data archives. Finally, conclusions and further work are presented.

3 Twitter Data Experimentation

The first part of the work was aimed at establishing a base of user understanding on the importance of Twitter specific features, such as hashtags, with topics and sentiment. The generic problem that is faced in this domain is that Twitter posts are by default too small to extract topic information. As such, it is impossible to align Twitter data with topics identified in the web documents.

In the search and retrieval tasks, accuracy is of paramount importance, so specific solutions must be applied in order to associate Twitter data with topics, entities and sentiment values. One way would be to group similar Twitter posts and treat them together as a web document. The hashtag information can be used in this set up. However, this could result in large numbers of posts under each hashtag, manifesting itself in the visualization of the search results, since it would be impossible to return all data from one or more hashtags. To add to that, the nature itself of the Twitter data, governs the expectations of the users. In that respect, the users expect to be able to retrieve keywords/catchwords, opinions over time frames of events, events that the posts refer to, influence information, in effect a reply to the question of what have people been talking about regarding an event.

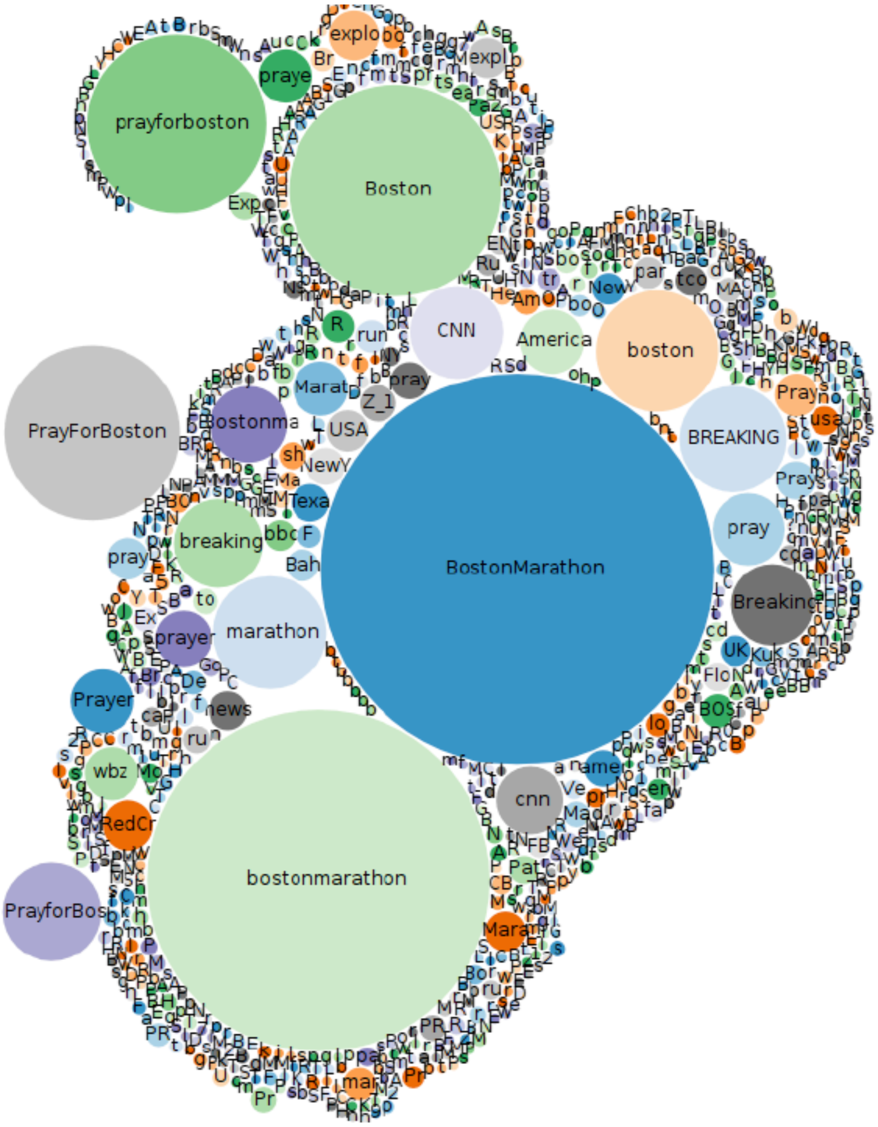

Fig. 1. General Hashtags Bubble Graph

The dataset used for this exercise contained 5000 tweets related to the Marathon event and the bombings that took place in Boston on 15/4/2013. The Marathon started at 09:00 and the explosions occurred at 14:59. The tweets were collected in the timeframe between 14/4/2013 at 21:00 and 15/04/2013 at 19:46. Each of these tweets was annotated for polarity (positive, negative, neutral) according to its content. Some examples of the tweets are:

a. Excited for the #bostonmarathon tomorrow #makeitcount #findgreatness
b. Best of luck to all those running #bostonmarathon today! Have FUN and enjoy!!
c. #bostonmarathon explosions! Horrifying site :(let's pray for the may affected
d. #PrayersforBoston our prayers go out to Boston this afternoon.

With the above considerations, two types of data graph visualizations were considered: the hashtag graph and the sentiment chart. Since topic detection for single tweets is not recommended, hashtags could be used for the top-level visualization. The main tool that was used to create the various implemented graphs is JavaScript and more specifically the d3.js JavaScript library [3]. D3 is used for manipulating database documents using HTML, CSS and SVG components displaying JSON-formatted data.

As seen in figure 1, each topic may correspond to a set of twitter hashtags whose names may exhibit similarities as close as very minor lexical or stylistic transformations (e.g. "BostonMarathon", "bostonmarathon"). These hashtags are detected using simple heuristics or Levenshtein distance [8] and are consolidated as shown in figure 2.

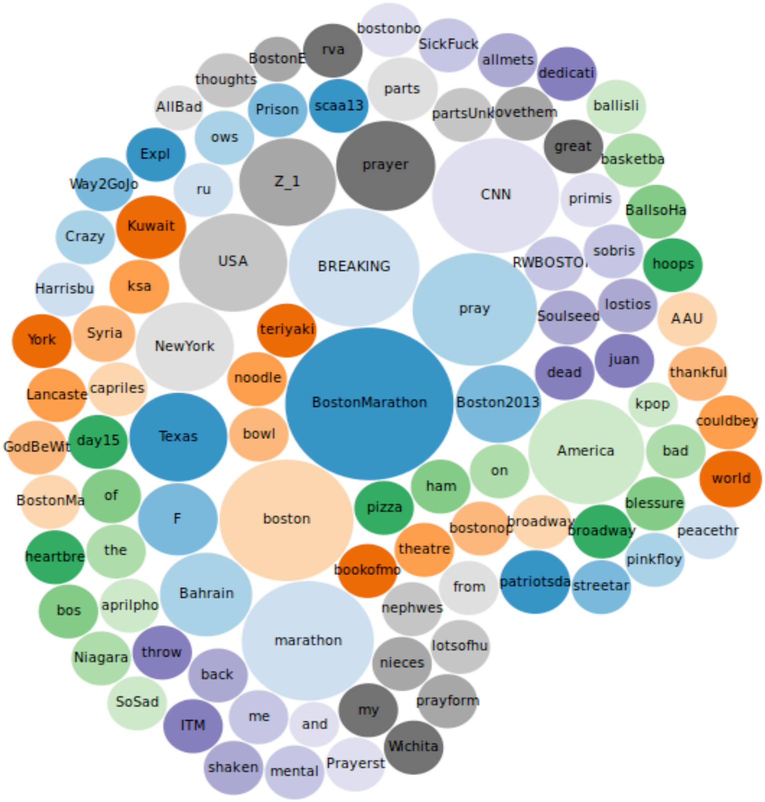

Fig. 2. Bubble Chart of consolidated hashtags using logarithmic scale for radial size

Sentiment charts are also one of the standard requirements. Clicking on the hashtag of interest can access them. They present the frequency of the positive and negative tweets over time. As shown in figure 3 below, the number of tweets for the Boston Marathon was relatively small in the beginning, however, after the bomb explosion it was rapidly increased. As was also shown, the negative tweets dominated over the positive ones as time was passing, since more people expressed their sadness or anger about the event. The fact that many positive tweets are detected (as shown in the graph) is mainly due to many people expressing hopes and wishes (e.g. “Best wishes to those at #BostonMarathon”, “I hope everyone is ok”).

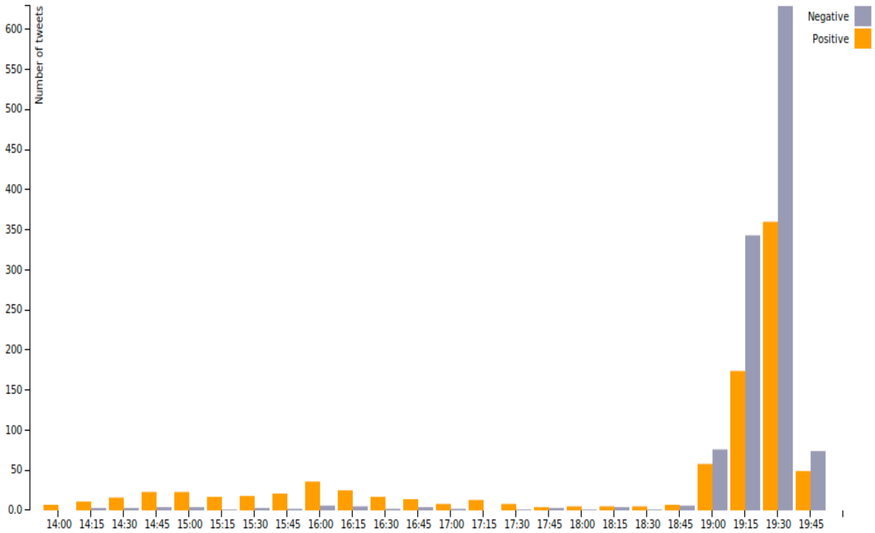

Fig. 3. The distribution of positive and negative tweets per hour

4 Semantic Relation Visualization

For the purpose of this task the Search and Retrieval Application (SARA) was used to measure user interaction [4]. As part of the heuristic evaluation, the users were asked to search using both a scenario that was designed to cover all functionalities and features of the SARA interface and as free search using any means necessary to complete their tasks. Their interaction was automatically logged and their intentions were recorded using the think aloud method.

The archived web documents were indexed in Solr [9] as semantic data including named entities, opinions, events, topics, as well as post-processed entries such as clustered entities. One of the major findings of the users was the fact that, after the initial search by one or more entities of interest, filtering by topic is the optimal way of retrieving the required information. However, only the nearest topics are suggested since that process is search-specific. Follow up queries on topics and information contained in the topics has to be explored. This can be remedied by visualizing the map of the detected topics and, even better, as a follow-up requirement, provide a high level visual of all topics for a whole set of collected resources. Figure 4 shows the top level visual regarding topics. The distance of the topics between them is provided by the layout itself while the number of contained web resources by size of each topic bubble.

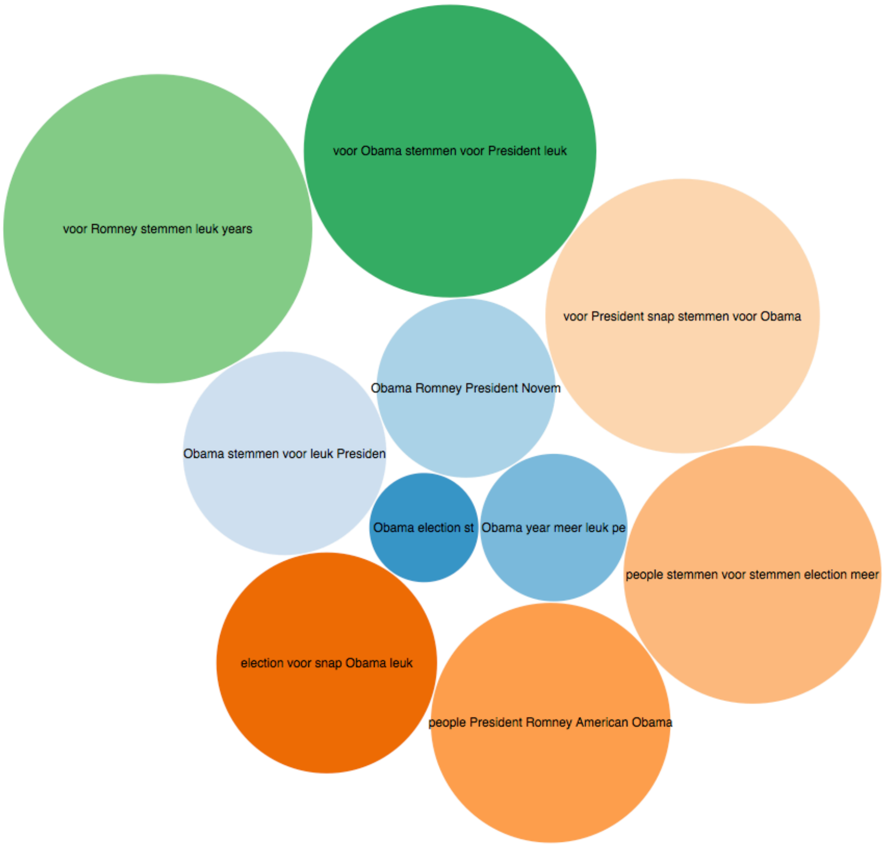

Fig. 4. Topic influence display

The users may click on the topics to explore them. By doing so, the major associated entities of the chosen topic can be viewed as well as indicators on detected opinions on each, as shown in figure 5. That, in effect, projects the overview of the entities of a topic that span all the web resources of that topic. At that point an entity may be clicked or the topic in order to zoom out to the figure 4 overview. By choosing an entity, all the topics that contain that entity as a major entry can be viewed as in figure 6. In effect, the users may always go forward in their search by either “zooming out” to topic view or “zooming in” to entity/opinion level.

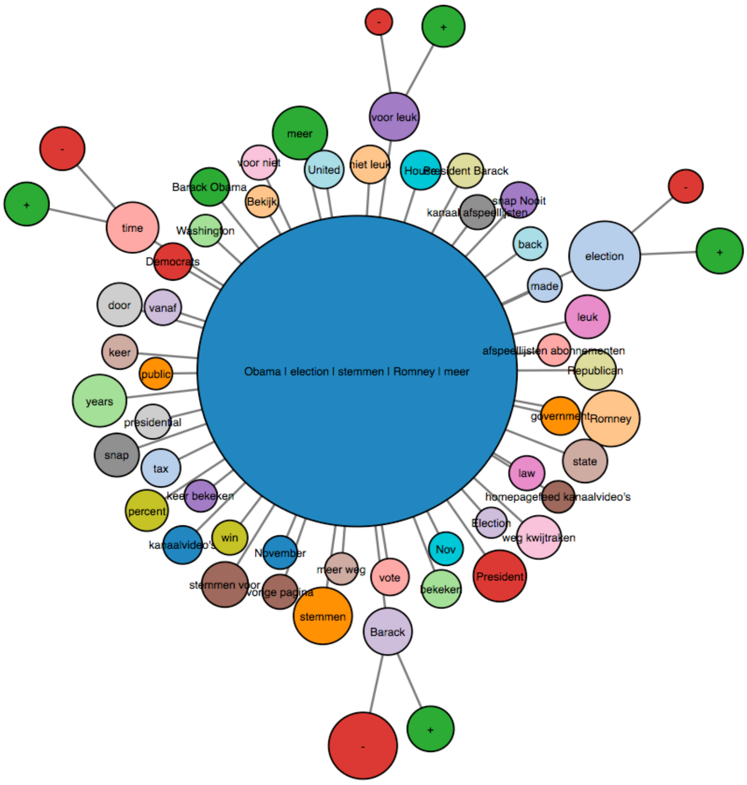

Fig. 5. Topic, entities and opinions

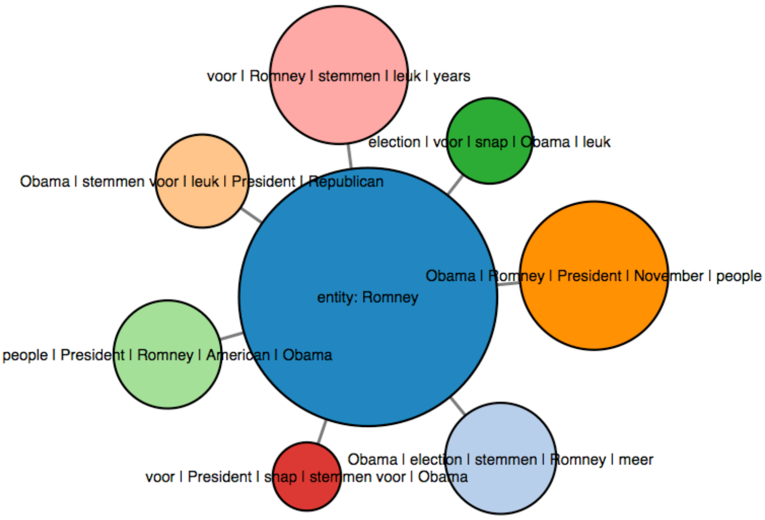
voor | Romney | stemmen | leuk | years
election | voor | snap | Obama | leuk
Obama | stemmen voor | leuk | President | Republican
entity: Romney
Obama | Romney | President | November | people
people | President | Romney | American | Obama
Obama | election | stemmen | Romney | meer
voor | President | snap | stemmen voor | Obama

Fig. 6. Entity and relevant topics

5 Evaluation

The US Election 2012 dataset was used for the evaluation. The interactive visualizations were made available through the SARA web interface from the start of the search process. The purpose of the usability evaluation, at that point, was to measure user engagement with the newly introduced visualizations, changes in the key performance indicators regarding the search and retrieval core tasks, verification of the findings of the heuristic evaluation mentioned in section 4, and evaluation of impact and acceptance by the end users on the interaction level.

The typical scenarios for search and retrieval used in earlier iterations of usability evaluations were slightly adapted to include optional use of the topic-based visualizations. The hypotheses for the above were:

i. Taking into account the clear preference that users have shown to the introduction of topics in the ARCOMEM data and respective semantic search functionality, it was expected that the end users would actively show clear preference to using the interactive visualization process. The drawback is that using the visualization exclusively would yield results for certain typical search expectations, but it was not designed to single-handedly replace the traditional features of such application. For example, the user cannot filter or facet the results, by other entities, source network, nor rank the results according to opinions. Furthermore, many users may search by topic or by entity, entity being the norm for non-experts or non-archivists.

ii. There was already a significant improvement in accuracy and minimization of exploratory or corrective backtracking with the introduction of topics. Given the fact that, by design, the visualizations provide an accurate overall view on topics, it was expected that the indicative times to reach the same state would be even shorter.

iii. The visualizations were fine-tuned on the interaction level, based on the heuristic evaluation that was also used to establish the baselines for visual interaction. The findings were expected to verify the previous conclusions, however, this time around it was the end users of all levels that participated.

iv. The impact and acceptance was expected to be higher than the text-only interaction that was evaluated at earlier stages. Items and actions of interest were flagged for further investigation such as the topics/entities/opinions relation view (figure 5) regarding over-information and understandability.

The feedback from the users indicated that the topic-centered approach was welcomed and accepted by most users, however, an additional request was to provide a higher-level relation of entities and topics at the start of the search. This was expected, since it is common, especially for targeted search, to use entities for searching archives and move to more abstract level, such as topics, later if further search is required. A simple response to that was to include an entity word cloud (figure 7) at the same time, connecting the major entities of the web documents to other entities (i.e. clustering) or topics by colour.

The speed and accuracy were also improved since the visualizations allowed the users to engage in topics and described entities is just two clicks while, previously, topics were made available only during the search, whether at web resource level (lowest) or as prediction based on the interaction. The users now felt that they had more choice, especially true for the experts.

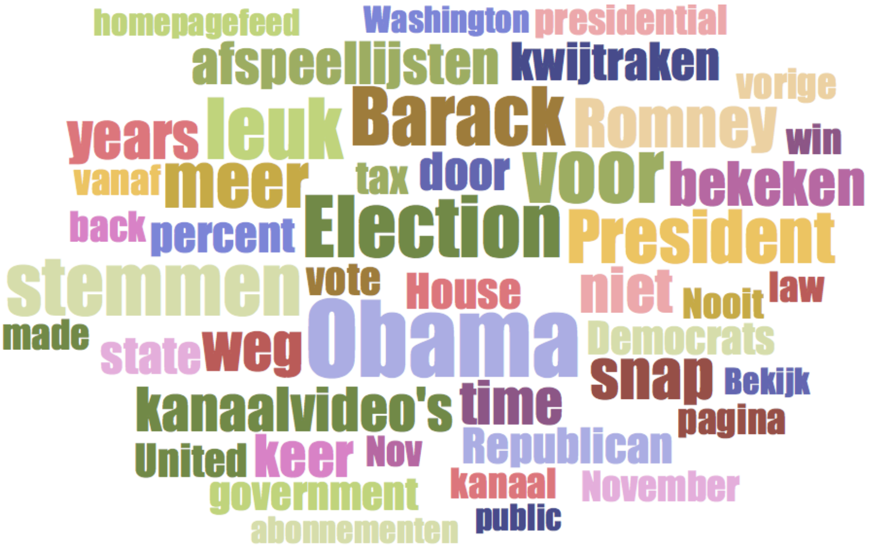

Fig. 7. Entity word cloud, clustering is indicated by matched colour

For the topic/entity/opinion view, some users had difficulty scanning the graph. There is clear feedback that the type of entity should be shown, preferably by color while only the most opinion-heavy entities should be visualized, resulting in less clutter. On the interaction level, the process was deemed highly acceptable. It was also mentioned by a large percentage of the participants that the social media results should be also visualized as part of a comparative view with the traditional web results or aggregated.

6 Conclusion and Further Work

The idea behind this work was not to build an ontology visualization toolset [10] but provide an interaction-aware interactive visual way for optimised semantic search and retrieval. Users perceive the potential of interactive visualized content by making semantic inferences [11] and the main task was to meet those efficiently on the interaction level. The described approach does not require a query formulation by the user at the start of the search and retrieval process but rather automatically generates the queries needed during the interaction.

Further work includes experimentation on a combined semantic relation visualization of social media meta-information, such as hashtags, with higher level semantic descriptions such as topics. That is expected to result in realizations of summative results, like topic-ver-time, topic influence over hashtags, and so on. It is also expected that nested levels of topics or other interemediate semantic levels may be expressed, especially for very large collections and broader domains.

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. Risse, T., Peters, W., Senellart, P.: The ARCOMEM Approach for Social and Semantic Driven Web Archiving, In: Proc. 1st Int. Workshop on Archiving Community Memories, Lisbon, Portugal (2013)
2. Schefbeck, G., Spiliotopoulos, D., Risse, T.: The Recent Challenge in Web Archiving: Archiving the Social Web. In: Int. Council on Archives Congress, Brisbane, Australia (2012)
3 Data-Driven Documents, http://d3js.org/
4. Chen, C., & Carr, L.: A semantic-centric approach to information visualization. In: IEEE Int. Conf. on Information Visualization (1999)
5. Kboubi, F., Chaibi, A. H., BenAhmed, M.: Semantic visualization and navigation in textual corpus. Int. J. Inf. Sciences and Techniques (IJIST) Vol.2, No.1 (2012)
6. Voigt, M., Pietschmann, S., Meißner, K. Towards a semantics-based, end-user-centered information visualization process. In: 3rd Int. Workshop on Semantic Models for Adaptive Interactive Systems (2012)
7. Spiliotopoulos, D, Tzoannos, E, Cabulea, C., Frey, D.: Digital Archives: Semantic Search and Retrieval, In: Holzinger, A., Pasi, G. (eds.) HCI-KDD 2013, LNCS 7947, pp. 173-182 (2013)
8. Levenshtein, V.I. Binary codes capable of correcting deletions, insertions, and reversals. Technical Report 8 (1966)
9 Apache Solr, http://lucene.apache.org/solr/
10.Temitayo, F., Stephen, O., Abimbola, A.: RROVT: A Proposed Visualization Tool for Semantic Web Technologies. J. Information Engineering and Applications, 2(3), 7-25 (2012)
11.Keller, R.M., Hall, D.R.: Developing Visualization Techniques for Semantics-based Information Networks. In: ACM Workshop on Visualization in Knowledge Engineering, 2nd Int. Conf. on Knowledge Capture (2003)