Extracting Knowledge from the Web and Social Media for Progress Monitoring in Public Outreach and Science Communication

Arno Scharl
MODUL University Vienna
Department of New Media Technology
Am Kahlenberg 1, 1190 Vienna, Austria
+43 (1) 3203555 500
scharl@modul.ac.at

David D. Herring
National Oceanic and Atmospheric Administration
NOAA Climate Program Office, 1315 East-West
Highway, Silver Spring, MD 20910-3282, USA
+1 (301) 7341207
david.herring@noaa.gov

ABSTRACT

Given the intense attention that environmental topics such as climate change attract in news and social media coverage, key questions for large science agencies such as the *National Oceanic and Atmospheric Administration* (NOAA) are how different stakeholders perceive the observable threats and policy options, how public media react to new scientific insights, and how journalists present climate science knowledge to the public. This paper investigates the potential of semantic technologies to address these questions. It introduces the NOAA Media Watch and presents a detailed case study of how the metrics and visualizations of the webLyzard Web intelligence platform are used to track information flows across online media channels. Building upon this platform, we present a novel framework to measure the impact of science communication and public outreach campaigns – through a combination of quantitative and visual methods that go beyond sentiment analysis and related opinion mining approaches.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Linguistic processing; H.3.3 [**Information Search and Retrieval**]: Information Filtering; I.7.5 [**Document Capture**]: Document analysis.

General Terms

Measurement, Documentation, Performance, Human Factors.

Keywords

Web intelligence, visual analytics, semantic technologies, online media monitoring, science communication.

1. INTRODUCTION

People's health, security and economic well-being are closely linked to weather and climate. Every day communities around the world grapple with environmental challenges due to extreme weather and changing climate conditions. Policy leaders, businesses, resource managers, and citizens are asking for information to help them address climate-related risks and opportunities.

WebMedia'13, November 5–8, 2013, Salvador, Brazil.

ACM 978-1-4503-2559-2/13/11…$15.00.
http://dx.doi.org/10.1145/2526188.2526219

Web-based communication and engagement channels such as online news and social media play important roles in the information-gathering process. Leading science organizations such as the *National Oceanic and Atmospheric Administration* (NOAA) benefit from a thorough understanding of who uses these media channels, and how. NOAA is mandated by the United States Congress to advance scientific understanding of how the Earth's climate system works, and to share the resulting knowledge with the public. NOAA concentrates on climate-related challenges facing society today in the following areas: (i) reducing vulnerability and improving resilience to extreme weather and climate events; (ii) preparing for drought and water resource challenges; (iii) protecting coastal communities and coastal infrastructure from inundation due to sea level rise and storm surges; (iv) identifying and managing threats to marine ecosystems; and (v) developing strategies for mitigating and adapting to climate-related changes.

The NOAA Climate.gov Web portal provides climate-related information in a way that is easy to access and use, helping people to understand the state of the climate system and its impacts on their lives and livelihoods – both now and in the future so that they can make more informed decisions. To engage with stakeholders and disseminate scientific results, the NOAA Climate Program Office employs a three-pronged strategy:

- Publishing data and information via www.Climate.gov, an interactive Web portal that is designed to reach large numbers of people across four public segments: policy leaders and decision makers, scientists and data users, educators, and the climate-interested public.
- Directly engaging with small numbers of our target groups in interactive events designed to build relationships and to foster deeper, richer exchanges of information.
- Encouraging our partners and news media to syndicate, host, and republish our content in their websites and broadcasts.

Like all U.S. federal agencies, NOAA must demonstrate measurable progress toward its goals. Because there is no one-size-fits-all approach, the Climate Program Office uses a combination of evaluation approaches: (i) track visitors to Climate.gov and review how these visitors rate our online content; (ii) conduct focus groups and surveys designed to measure our Quality of Relationship with each target public; and (iii) use the NOAA Media Watch, a Web intelligence tool to monitor how others host, report on, and discuss published climate science information.

Today, Web-based media channels and semantic technologies add an important dimension to science communicators' ability to assess and evaluate online communications. For exploring this di-

mension, the webLyzard platform (www.weblyzard.com) has become an essential part of NOAA's evaluation strategy. Several years ago we began using the system to track online communications efforts that otherwise could not be measured – e.g., learning about sudden spikes or dips in news or social media coverage on NOAA's climate research and data products, or climate-related topics and relevant societal challenge areas. The NOAA Climate Program Office also wants to know how the number of articles in a given period compares to the long-term average for a given topic, especially during spike events [5; 8], and whether overall sentiment in online articles is positive or negative [11].

2. WEB INTELLIGENCE

The NOAA Media Watch allows us to determine trends and semantic associations with just a few clicks. Its interactive information exploration and retrieval interface sheds light on stakeholder perceptions, reveals flows of relevant information between these stakeholders, and provides success metrics for assessing the effectiveness of awareness and public outreach campaigns.

Figure 1 shows an analysis of online media coverage related to the term "coast" and a range of associated topics, including: coastline, coastal inundation, coastal flooding, coastal communities, coastal infrastructure, and coast erosion. The trend chart depicts the weekly frequency of news reports on the topic "coast" from May 1 to Nov 7, 2012 (the Atlantic hurricane season).

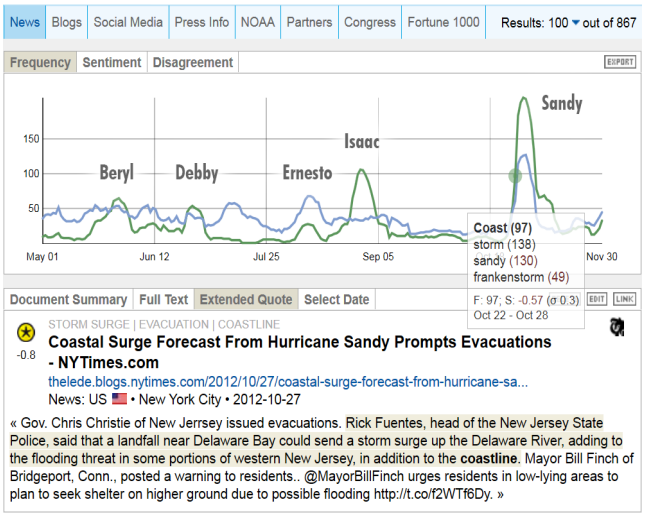

Figure 1. Media monitoring of the Atlantic hurricane season 2012 using the NOAA Media Watch (search term: "coast")

On average, between 50 and 75 news articles per week referred to coast-related terms. We saw six spikes during that period in which the number of reports climbed significantly higher than average due to tropical storms or hurricanes approaching the U.S. coast – Tropical Storm Beryl received media attention in late May (207 articles), TS Debby in late June (221 articles), TS Ernesto in early August (148), TS Isaac in late August (376), Hurricane Leslie in mid-September (145), and Hurricane Sandy in late October (659).

Figure 2 presents the complete dashboard [7] of the NOAA Media Watch. It provides rapid synchronization of multiple coordinated views [6] including tag clouds, geographic maps, keyword and ontology graphs, as well as document clusters in two- and three-dimensional information landscapes [9]. These visualizations help users to understand the context of the extracted knowledge while navigating the knowledge repository. The shown query on "climate science" for the first quarter of 2013, for example, reveals that discussions about extreme weather events including droughts in North America and the unprecedented heat wave in Australia remain dominant topics in Anglo-American news media coverage.

When extreme events like Hurricane Sandy happen, people often turn to NOAA and ask: "Was this due to global climate change?" Such questions often cannot be answered by a simple yes or no. Extreme events happen with or without climate change. Yet today, such events happen in the context of a warming world. We know that a warmer atmosphere has a greater capacity to hold water vapor, which means heavy rain events are likely to grow more extreme. We also know that sea level along New York's coast is about 30 cm (1 foot) higher today than it was a century ago, which means hurricane Sandy's surge of seawater was made more severe. But these points are nuances, and nuances like these often get overlooked in today's fast-paced lifestyles in which we demand instant answers packaged into easily digestible sound bites. Climate science almost never lends itself to sound bites, which increases the challenge to science agencies like NOAA when communicating with non-scientist publics [2].

Being able to monitor the number and nature of news reports and public dialogs in the blogosphere is very important to assess whether authors and their readers understand the nuances of climate science, such as the differences between natural climate variability and human-induced climate change. In case of misleading or wrong conclusions, a real-time analysis of online media can help agencies like NOAA to diagnose the problem and take corrective action – e.g. using semantic technologies to improve our understanding of when and how people are "spinning" information about climate science in ways that are not accurate.

To embed the collected knowledge into existing workflows, the NOAA Media Watch supports a range of export formats, including RSS, HTML, and PDF for textual data, and CSV for time series date. This is in line with calls for a *Semantic Social Web*, in which data isn't locked away within data silos but can be easily integrated and exchanged between applications [3].

A public showcase of the semantic technologies presented in this paper in the form of a Web content aggregator about climate change and related environmental issues [7] is available at www.ecoresearch.net/climate; its dashboard resembles the NOAA Media Watch, but provides fewer analytical functions and uses different Web media sources and input filter criteria.

3. ONLINE SUCCESS MEASURES

The potential of semantic technologies for progress monitoring in science communication is not only reflected in the interactive visualizations of the webLyzard dashboard, but also in its analytical services and success metrics. The emphasis in communication assessments needs to shift from measures of output to measures of *outcome*. The majority of Web intelligence and social media monitoring tools provide frequency statistics and language characteristics such as positive and negative sentiment [10; 11]. While sentiment is an important and insightful indicator, even if measured accurately it fails to address some of the most fundamental questions of decision makers. The ongoing research presented in this paper will provide implicit observations (as compared to explicit data collection methods such as questionnaires) and tailored success measures (as compared to generic metrics from natural language processing such a sentiment) to measure the attitudes and preferences of social media users.

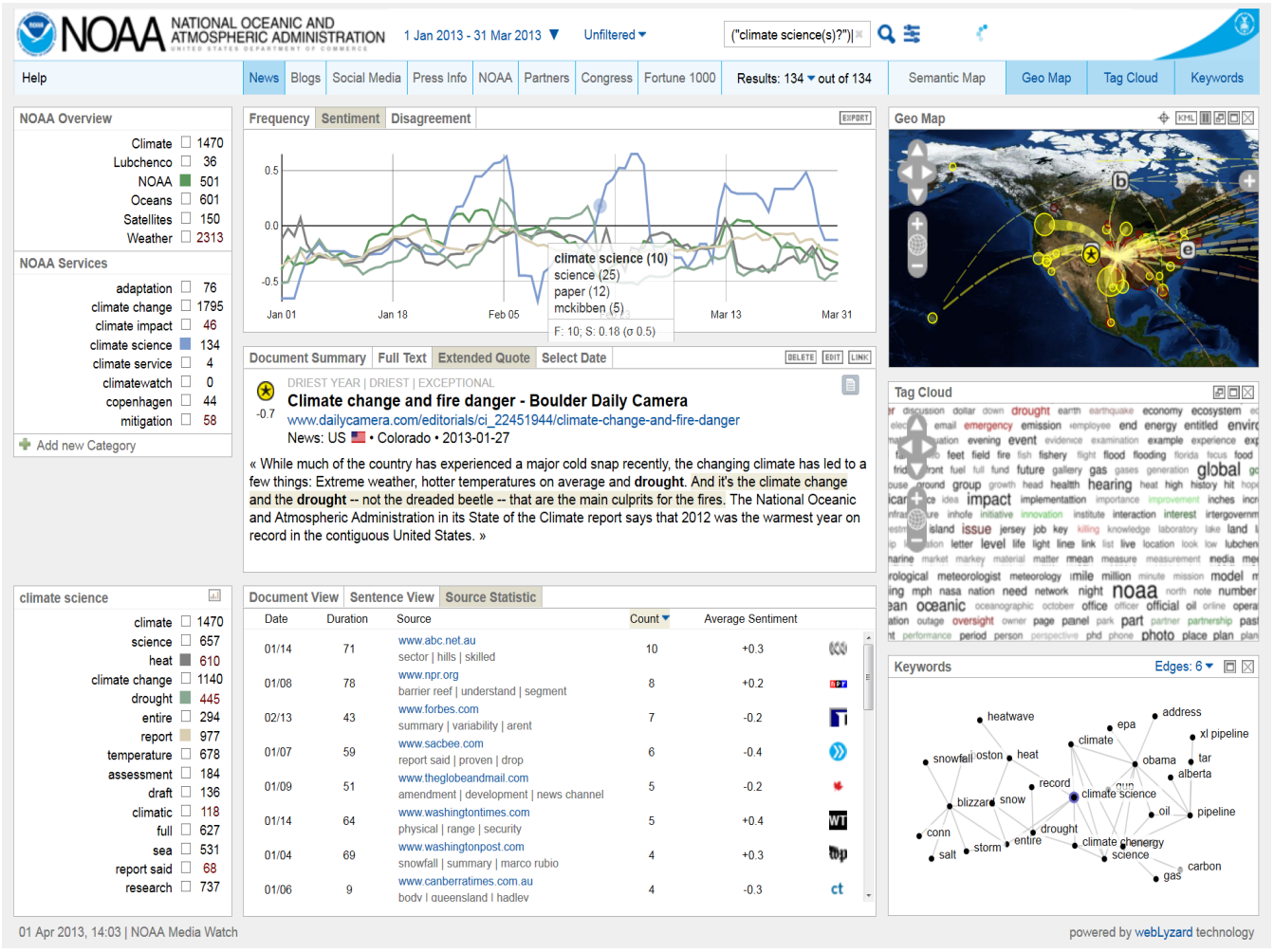

Figure 2. Visual analytics dashboard of the NOAA Media Watch (search term: "climate science")

3.1 Quantitative Analysis

From a commercial and exploitation perspective, such a dynamic assessment beyond sentiment allows real-time insights into the success of marketing and public outreach activities. Measuring attention and sentiment is descriptive in nature.

What we intend to measure, by contrast, is whether communication targets have been reached – i.e., whether the chosen communication strategy has an impact on observable patterns in online coverage, and how consistently a message is being conveyed.

The new metric shows whether an organization or its products and services were associated with desired topics considered important, whether they were in line with corporate communication goals, and whether undesired topics and media coverage were avoided successfully. This analysis goes far beyond language characteristics such as positive and negative sentiment.

In the case of the NOAA Climate Program Office, for example, a desired association with "*climate change*" contributes positively to the success metric, although the term typically carries a negative sentiment. The new metric is adaptive and part of an iterative feedback cycle, customized to an organization's evolving communications and dissemination goals. To specify these goals, analysts have the opportunity to create lists of desired and undesired topics, and update them in line with changing priorities.

3.2 Visual Analysis

This section describes how measures of online success can be represented in visual form, including an interactive stacked bar chart to be implemented using the *Data-Driven Documents* (D3) JavaScript library [1], as well as information landscapes computed through a combination of hierarchical cluster analysis [4] and force-directed placement algorithms [9].

The stacked bar representing the metric will consider the number of desired vs. undesired associations, as well as the number of positive vs. negative references (measuring the popularity of a brand, organization, person, or topic). While the algorithmic part and work on the computation of the metric has been completed at the time of writing, the authors expect the visual chart representation to be completed in the third quarter of 2013. The representation will be synchronized with the dashboard shown in Figure 2. Tooltips will display additional metadata; e.g., topics and opinion leaders responsible for observable changes. The adaptive calculation enables analysts to assign weights to the individual components of the success metric and thereby configure the computation according to their perceived importance for progress monitoring.

Dynamic topography information landscapes [9] are a powerful way to show the positioning of an organization vis-à-vis desired and undesired terms. NOAA Media Watch visualizes longitudinal

changes in large document repositories in near real time. Based on a dynamic clustering of news and social media documents, the resulting landscape resembles a geographic map. Hills represent document clusters around a common topic, while valleys or oceans indicate low document density in sparsely populated areas of the information space. The height of a hill indicates the size of the corresponding topical cluster, while its compactness corresponds to the cluster's cohesion – i.e., the similarity of articles based on a vector space representation. Hills (clusters) are labeled with dominant terms and phrases from the underlying documents to provide a high-level navigational aid for the users.

Figure 3 exemplifies this type of representation based on rendering approximately 5,000 documents taken from environmental blogs. By mapping search results onto the information landscape, communication analysts can track which terms are closely related with an organization, and whether a campaign had an impact on the relative position.

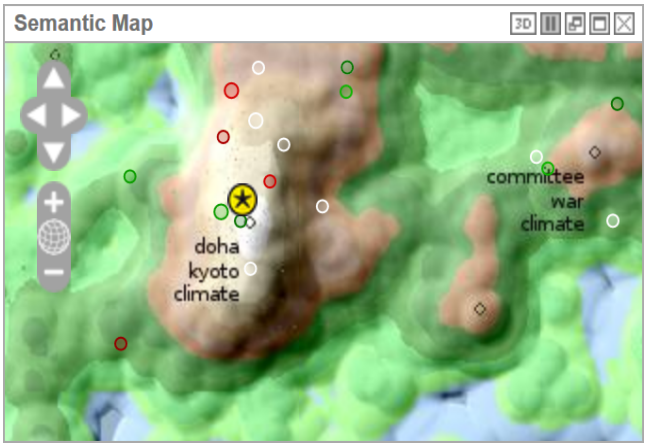

Figure 3. Dynamic topography information landscape on news media coverage about "climate change"

4. CONCLUSION AND OUTLOOK

While Web-based media channels have provided us with quick access to exponentially greater amounts of information, the public's ability to discern credible from non-credible sources appears to be declining. We need to better understand human perceptions and biases, and challenges associated with clear and effective communication. This is particularly true for science agencies like NOAA that are mandated to share scientific data with the public in ways that benefit society and enhance decision making. This requires the effective use of new media to listen to our audiences, directly engage with stakeholders, create shared meaning, and identify and address myths and misconceptions. If we are to succeed in making in-roads into the myriad societal challenges presented by climate change then it is clear that semantic technologies and media monitoring platforms such as the NOAA Media Watch have a vitally important role to play.

Future research will incorporate additional features into the computation of the hybrid success metric, and transform the bitmap-based information landscape component to a dynamic Scalable Vector Graphics (SVG) representation, supporting on-the-fly computations instead of weekly updates. This will enable us to extend the synchronization mechanism and highlight areas and peaks in the landscape that are related to the current search and the content of the displayed subset of Web documents.

Acknowledgement

Key components of the presented system were developed within the DIVINE (www.weblyzard.com/divine) research project, funded by *FIT-IT Semantic Systems* of the Austrian Research Promotion Agency (www.ffg.at) and the Austrian Federal Ministry for Transport, Innovation and Technology (www.bmvit.gv.at). The data acquisition component has recently been extended within the uComp research project (www.ucomp.eu), funded by the Austrian Science Fund (www.fwf.ac.at) through the European CHIST-ERA program (www.chistera.eu).

5. REFERENCES

[1] Bostock, M., Ogievetsky, V. and Heer, J. (2011). "D3: Data-Driven Documents", *IEEE Transactions on Visualization and Computer Graphics,* 17(12): 2301-2309.

[2] Bowman, T. (2008). Summary Report: A Meeting to Assess Public Attitudes about Climate Change. Silver Springs: National Oceanic and Atmospheric Administration (NOAA), George Mason University Center for Climate Change Communications.

[3] Breslin, J.G. and Decker, S. (2007). "The Future of Social Networks on the Internet: The Need for Semantics", *IEEE Internet Computing,* 11(6): 86-90.

[4] Farahat, A.K. and Kamel, M.S. (2009). Document Clustering Using Semantic Kernels Based on Term-Term Correlations. *IEEE International Conference on Data Mining (ICDM-2009), Second International Workshop on Semantic Aspects in Data Mining*. Miami, United States: 459-464.

[5] Gruhl, D., Guha, R., Liben-Nowell, D. and Tomkins, A. (2004). "Information Diffusion Through Blogspace", *13th International World Wide Web Conference*. New York, USA: ACM Press. 491-501.

[6] Hubmann-Haidvogel, A., Scharl, A. and Weichselbraun, A. (2009). "Multiple Coordinated Views for Searching and Navigating Web Content Repositories", *Information Sciences,* 179(12): 1813-1821.

[7] Scharl, A., Hubmann-Haidvogel, A., et al. (2013). Media Watch on Climate Change – Visual Analytics for Aggregating and Managing Environmental Knowledge from Online Sources. *46th Hawaii International Conference on Systems Sciences (HICSS-46)*. R.H. Sprague. Maui, USA: IEEE Press: 955-964.

[8] Scharl, A., Weichselbraun, A. and Liu, W. (2007). "Tracking and Modelling Information Diffusion across Interactive Online Media", *International Journal of Metadata, Semantics and Ontologies,* 2(2): 136-145.

[9] Syed, K.A.A., Kröll, M., et al. (2012). "Incremental and Scalable Computation of Dynamic Topography Information Landscapes", *Journal of Multimedia Processing and Technologies* 3(1): 49-65.

[10] Weichselbraun, A., Gindl, S. and Scharl, A. (2010). "A Context-Dependent Supervised Learning Approach to Sentiment Detection in Large Textual Databases", *Journal of Information and Data Management,* 1(3): 329-342.

[11] Weichselbraun, A., Gindl, S. and Scharl, A. (2013). "Extracting and Grounding Contextualized Sentiment Lexicons", *IEEE Intelligent Systems,* 28(2): 39-46.