

Visualizing Contextual Information in Aggregated Web Content Repositories

Arno Scharl, Ruslan Kamolov, Daniel Fischl, Walter Rafelsberger, Alistair Jones

Department of New Media Technology

MODUL University Vienna

Vienna, Austria

scharl@modul.ac.at

Abstract—Understanding stakeholder perceptions and the impact of campaigns are key insights for communication experts and policy makers. A structured analysis of Web content can help answer these questions, particularly if this analysis involves the ability to extract, disambiguate and visualize contextual information. After summarizing methods used for acquiring and annotating Web content repositories, we present visualization techniques to explore the lexical, geospatial and relational context of entities in these repositories. The examples stem from the *Media Watch on Climate Change*, a publicly available Web portal that aggregates environmental resources from various online sources.

Keywords—Web intelligence; context; visual analytics; word tree; named entity detection; relation extraction; climate change.

I. INTRODUCTION

Media analytics solutions have been developed for various domains including sports [6], politics [2; 9] and climate change [4; 8], often focusing on specific aspects such as (sub-)event detection [1], content classification [4] and the automated annotation of video broadcasts [2]. Such media analytics systems face two major challenges:

- compile and annotate large document collections from online sources that are heterogeneous in terms of authorship, formatting, style (e.g., news article versus tweets), and update frequency;
- provide an interactive dashboard to select the most relevant subsets of the information space, and to analyze and visualize the extracted information.

Contextual information, especially when properly disambiguated, plays a vital part in addressing both challenges. It improves several steps in the processing pipelines of media analytics platforms – targeted content acquisition via focused crawling [5], for example, or more accurate knowledge extracting algorithms tailored to the specifics of user-generated content [10] – especially when trying to understand the role of affective knowledge in the decision-making process [3].

II. MEDIA WATCH ON CLIMATE CHANGE

The *Media Watch on Climate Change* (MWCC) is a content aggregation and online collaboration platform publicly available at www.ecoresearch.net/climate [4; 8]. Using the Web intelligence and media analytics platform of webLyzard (www.weblyzard.com), it compiles large archives of Web

content from multiple online sources, and provides a variety of knowledge co-creation and visualization tools [8]. MWCC also serves as the knowledge repository for DecarboNet, a three-year research project funded by the European Commission via the *7th Framework Program* (www.decarbonet.eu).

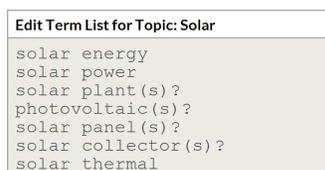
MWCC integrates multilingual content from English, French and German online sources: social media including *Twitter*, *Facebook*, *Google+* and *YouTube*, and the Web sites of news channels, Fortune 1000 companies, municipalities, and environmental NGOs. Automated document enrichment services then transform the gathered information into a contextualized information space spanning geospatial, temporal and social dimensions.

Analyzing this information space sheds light on stakeholder perceptions, reveals flows of relevant information, and provides indicators for assessing the impact of large environmental campaigns such as the *Earth Hour* [11].

III. WEB CRAWLING AND PREPROCESSING

To process and enrich data from unstructured, structured and social evidence sources, MWCC pursues a focused crawling strategy. Managing the abundant quantity and dynamic nature of news and social media content requires efficient pre-processing to remove irrelevant content at an early stage of the processing pipeline. This filtering reduces the number of documents to be processed by computationally expensive information extraction algorithms.

MWCC relies on a domain specificity measure based on a combination of blacklists and whitelists to assess the relevance of gathered documents in the context of climate change and related environmental issues (see Figure 1).



```
Edit Term List for Topic: Solar
solar energy
solar power
solar plant(s)?
photovoltaic(s)?
solar panel(s)?
solar collector(s)?
solar thermal
```

Figure 1. List of regular expressions to define the topic “solar energy”

IV. EXTRACTING FACTUAL AND AFFECTIVE KNOWLEDGE

Which organizations tend to have a negative reputation among social media users? Who are the most visible climate change activists, and what are mainstream media associating with their recent public appearances?

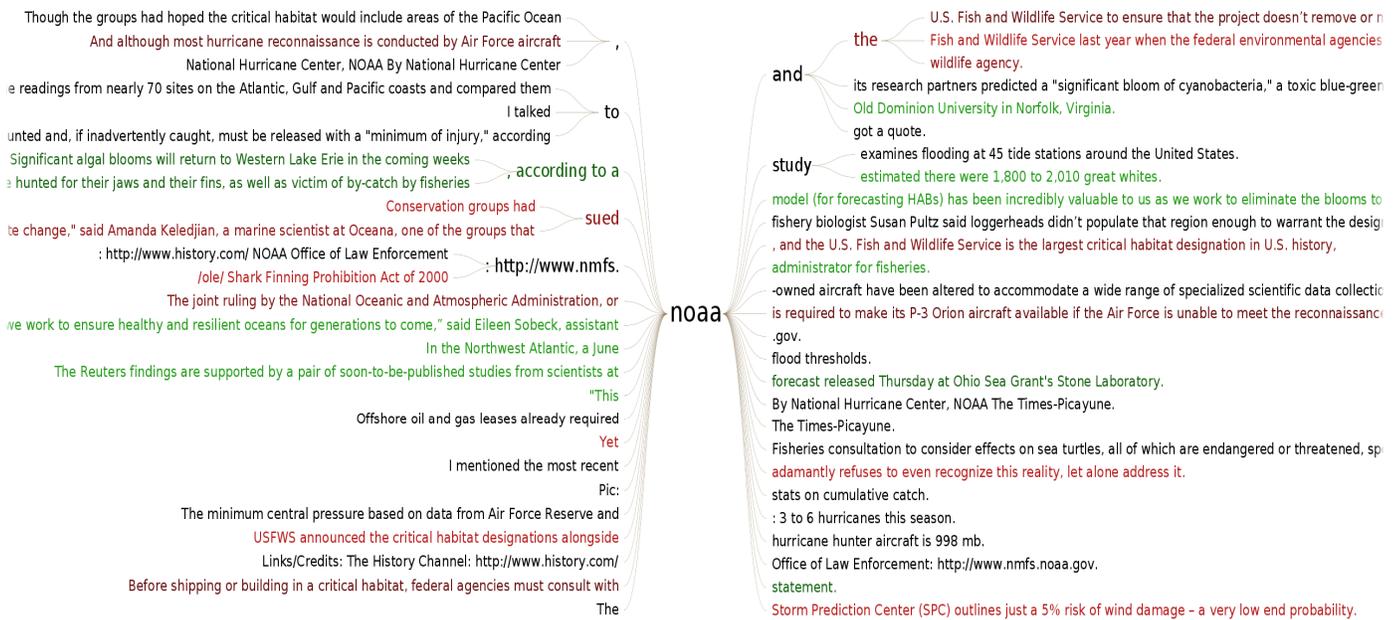


Figure 2. Word tree representation for the search term “NOAA” in Anglo-American news media coverage (Q2/2014)

For properly answering such communication questions, MWCC utilizes *Recognize* [15] – a named entity recognition and resolution component that draws upon structured external knowledge repositories such as *DBpedia.org*, *Freebase.com* and *GeoNames.org* to identify and disambiguate named entities (organizations, persons and locations), assigning confidence values to align them with the items contained in the external knowledge repositories.

The result is a continuously evolving knowledge repository that helps to better understand networks and the dynamic relations [1] among their actors.

The platform provides a seamless integration of factual (concepts, instances, relations) and affective (beliefs, opinions, arguments, etc.) knowledge:

- *Factual Knowledge*. The *Recognize* named entity recognition and resolution component not only identifies and classifies entities, but also grounds them to external knowledge bases or corporate databases.
- *Affective Knowledge* includes sentiment and other emotions expressed in a document, which are captured and evaluated by opinion mining algorithms [13; 14].

V. VISUALIZING CONTEXTUAL INFORMATION

The MWCC visual dashboard [8] reveals popular issues that are being discussed in conjunction with a given topic. This section describes three new visualization components to reveal contextual information in such online discussions – a *word tree* for lexical context, a *map projection* for geospatial context, and an *entity tracker* for relational context across organizations, persons and locations.

The color coding of the diagrams reflects normalized document sentiment, ranging from green (positive) to grey (neutral) and red (negative).

A. Lexical Context

Once a user has entered a search query, MWCC ranks the matching results by relevance, date, or geographic location. Sentiment information is available in a separate column. Clicking on a quote extends the entry; a second click activates the full text mode. When the full text of a document is shown, the header of the page includes document keywords and the source URL, while the footer summarizes the document’s other annotations including source category, source location, target location, sentiment, and relevance.

Alternatively, the system lists matching quotes as a *concordance list*. Users can sort the concordances by their source, date of publication, and sentiment on either a document or sentence level.

The *word tree* module presents the concordance list in a visual and more intuitive manner, summarizing the different contexts in which certain entities or topics are being discussed. Its graph-based display facilitates the rapid exploration of search results and conveys a better understanding of how language is being used surrounding a topic of interest.

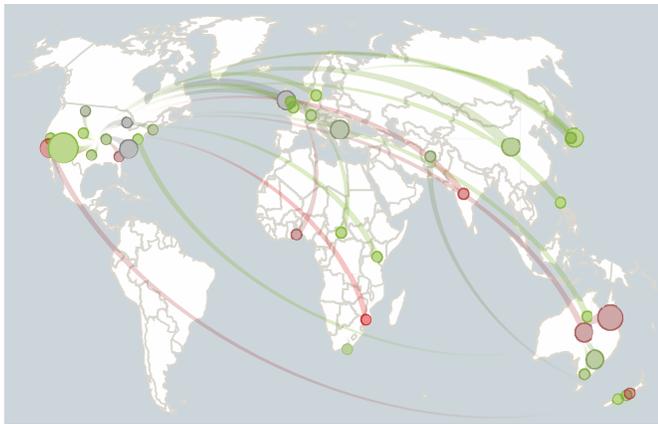
Based on the popular keyword-in-context technique [12], our specific implementation of the *word tree* metaphor adopts a symmetrical approach [7]. The root of the tree is the search term. The left part of the tree displays all sentence parts that occur before the search term (prefix tree), and the right part those that follow the search term (suffix tree). These branches to the left and to the right help users to spot repetition in contextual phrases that precede or follow the search term.

Mouse-over highlights connected elements, allowing users to reconstruct entire sentences. Visual cues include different font sizes to indicate the frequency of phrases, and connecting lines to highlight typical sentence structures.

Figure 2 shows how the tree-like structure is built after searching for the term “NOAA”, which is the acronym of the *National Oceanic and Atmospheric Administration*, and grouping identical phrases containing the term into nodes (e.g., “NOAA study”). This grouping together of equal phrases into a connected tree structure sheds light on word usage within the selected source(s) in a given time interval.

B. Geospatial Context

The results of searches within the MWCC portal are also projected onto a geographic map that shows the regional distribution of Web coverage – e.g., references to locations co-occurring with the term “solar energy” as shown in Figure 3. The position of circles is determined by the geographic coordinates of these references, their size is proportional to the number of documents referring to a specific position.



Location	Count	Latitude	Longitude	Sentiment
California california san francisco chronicle solarcity	86	37.3	-119.8	+0.3
Denver proctor cathy proctor cathy	51	39.7	-105.0	+0.3
Washington american coalition asthma avalanche	43	47.5	-120.5	+0.2
New York solarcity rive musk	37	43.0	-75.5	+0.3
People's Republic of China solar solarcity commerce	36	35.0	105.0	+0.1
Arizona arizona public service co corporation commission	32	34.5	-111.5	+0.3
Idaho solar roadways roadways brusaw	24	44.5	-114.3	+0.3
North Carolina solar close proximity apple	24	35.5	-80.0	+0.5
Colorado xcel xcel energy cathy	21	39.0	-105.5	+0.5
Europe detailed analysis decc britain	21	48.7	9.1	+0.2
United States pbn clean energy address	20	38.9	-77.0	+0.5
Republic of India india rajasthan solar	19	20.0	77.0	+0.2
Florida hacking alibaba chinese	18	28.8	-82.5	+0.1
Hawaii pbn shimogawa hawaii public	15	20.8	-156.5	+0.6

Figure 3. Geographic map and list of locations that co-occur with the term “solar energy” in Anglo-American news media coverage (06-07/2014)

When rendering documents in their geospatial context, the system distinguishes between source and target information – i.e., the authors’ locations versus the primary locations referenced in the documents, which is determined by applying the above mentioned *Recognize* component to a geo-tagging process (the table underneath the map shows a list of the identified geographic entities, sorted by decreasing co-occurrence frequency with “solar energy”).

C. Relational Context

To identify opinion leaders and reveal key factors influencing social conversations about a topic, the webLyzard platform detects not only locations, but also other named entities such as persons and organizations that have an impact on news and social media coverage. To develop a deeper understanding of this process, analysts must not only understand how these entities influence topics of interest, but also unravel the interconnected relations among the entities themselves. How did a public appearance of the CEO impact a company’s perceived relation to main competitor, for example, and what are journalists associating with the competitor’s latest announcement?

To help answer such questions, the *Entity Map* shown in Figure 4 visualizes (i) relations among named entities in the analyzed corpus, and (ii) co-occurrence patterns between these entities and user-defined search terms. In the case of MWCC coverage from April to July 2014, the list of referenced entities includes politicians such as U.S. President *Barack Obama* and the Australian Prime Minister *Tony Abbott*, organizations such as the *Green Energy Collective*, and various locations including *Washington DC* and the *State of California*. The Entity Map component combines a line chart with a radial imposition, and a radial convergence diagram:

- **Radial Convergence Diagram.** Located in the center of the graph, the radial convergence diagram displays relations among different entities using ribbons. Entity names are displayed along a circle – their font size indicates the number of documents that mention the entity, their color ranges from red to green depending on the average sentiment (in line with the sentiment color coding of the word tree and the geographic map). The thickness of an arc represents the number of co-occurrences between an entity pair. On mouse-over, the opacity of arcs that connect the selected entity to other entities is increased. A slider element in the lower left corner controls the level of detail in the radial convergence diagram – i.e. it determines the threshold for showing relations among entities. The second slider element in the lower left corner adjusts the number of entities to be shown, between a minimum of three and a maximum of 50 entities.
- **Line Chart.** Surrounding the radial convergence diagram in the center, the data points in the line chart show the number of co-occurrences between an entity and the selected topics (using the same color-coding as the trend chart). To increase the readability of the display and facilitate comparisons across topics, the line chart uses a logarithmic scale.

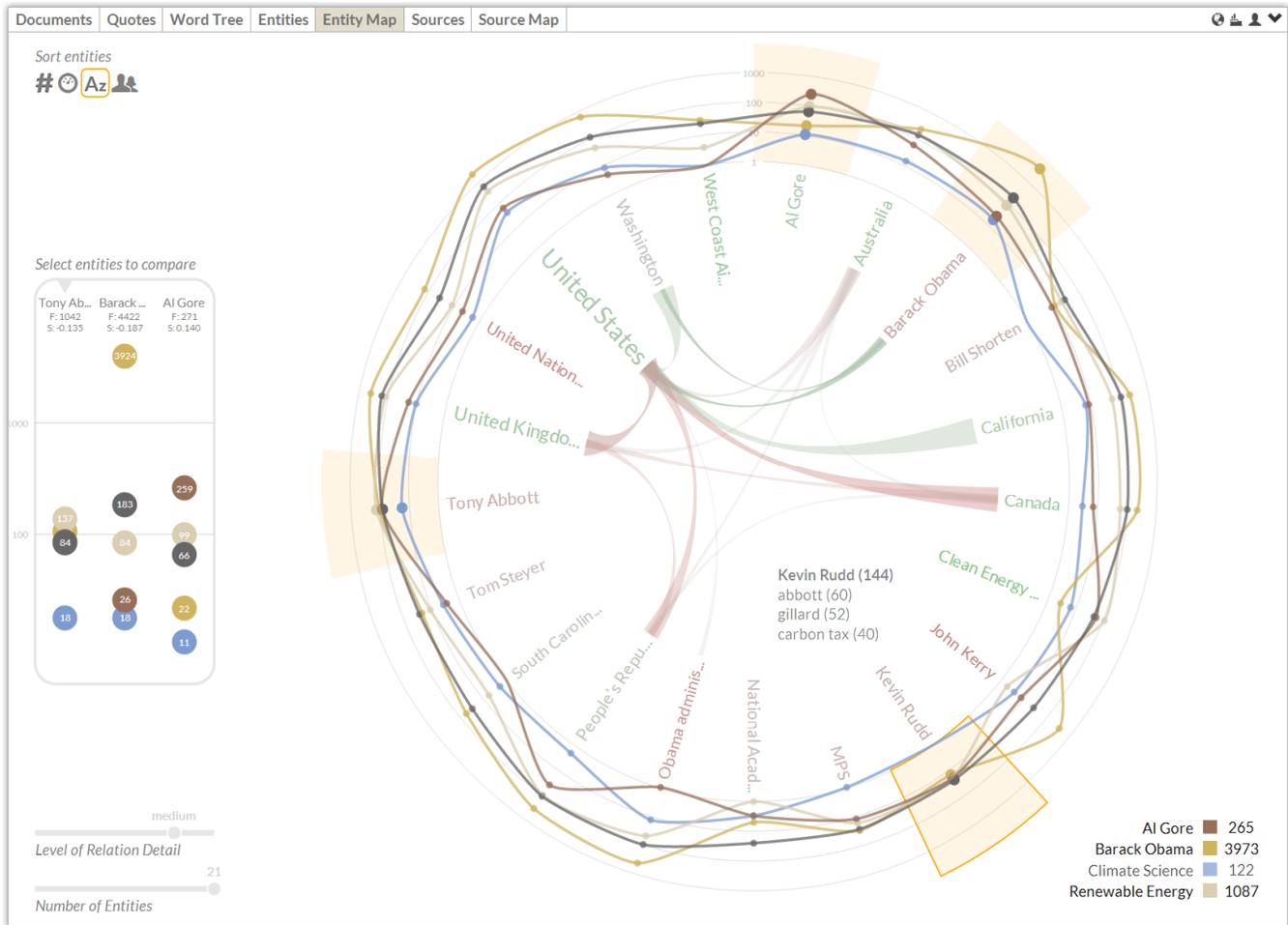


Figure 4. Entity Map showing the co-occurrence frequency of the search terms “Al Gore”, “Barack Obama”, “Climate Science” and “Renewable Energy” with identified named entities (organizations, persons and locations), as well as the strength of the relations among these entities

Three small icons in the upper right corner control which entity types are displayed – persons, organizations and locations (of which at least one needs to be active). Clicking on an icon in the upper left corner causes the entities to be rearranged by (i) entity type, (ii) name, (iii) the number of documents which contain an instance of the entity’s name in descending order, or (iv) the average sentiment of the documents containing the entity, from positive to negative.

Both the line chart and the radial convergence diagram are being updated by means of smooth, animated transitions. Hovering over an entity highlights the corresponding sector, shows a tooltip with the top three keywords associated with the chosen entity, and highlights the arcs in the radial convergence diagram. Hovering over one of the search terms in the list on the left removes all lines in the chart except for the one corresponding to the selected search term; this allows for a cleaner view of a single search term.

Additional interactions support more detailed comparisons. Clicking on an entity causes supplemental information to be displayed in a sidebar, which includes the data points with the co-occurrence values and the entity information – i.e., name, document count (d) and average sentiment (s).

The logarithmic scale of the sidebar adjusts automatically to accommodate the range of data values. The box contains the three most recently selected entities, which remain highlighted in the graph.

VI. SUMMARY AND CONCLUSION

The visualizations presented in this paper allow users to interactively explore the lexical, geospatial and relational context of Web documents. The underlying data stems from the *Media Watch on Climate Change*, a media analytics portal available at www.ecoresearch.net/climate. The system is currently being extended into a collective awareness platform through the *DecarboNet* project (www.decarnonet.eu). The context of Web coverage is important when aiming to investigate and better understand the various processes that lead to collective awareness, since it impacts opinions and decision making on both individual and collective levels.

The tools presented in this paper help to understand the context of Web coverage by establishing connections between named entities (persons, organizations, and locations), based on references to these entities in aggregated content from English, French and German news channels, and from

social media platforms such as Twitter, Facebook, Google+ and YouTube. The *Media Watch on Climate Change* utilizes the *Recognize* component (www.weblyzard.com/recognize) to identify and resolve named entities. *Recognize* draws upon structured external knowledge repositories such as *DBpedia.org*, *Freebase.com* and *GeoNames.org* to disambiguate these entities via confidence values that align entities with the items of the knowledge repositories.

Extracting and visualizing contextual information transforms unstructured collections of crawled Web content into structured repositories of actionable knowledge. Thereby, the presented techniques to reveal context in Web coverage provide value for a wide range of organizations including enterprises, non-government entities, news media outlets, science agencies, and policy makers. Uncovering patterns and trends in Web coverage can help these organizations to adopt better strategies for engaging audiences, guide their communication and public outreach campaigns, and increase the effectiveness of their decision making processes.

ACKNOWLEDGEMENT

The research presented in this paper has been conducted as part of several European research projects: DecarboNet (www.decarbonet.eu) and PHEME (www.pHEME.eu), which have received funding by the European Union's 7th Framework Program for research, technology development and demonstration under the Grant Agreements No. 610829 and 611233, respectively; as well as uComp (www.ucomp.eu) with funding support of EPSRC EP/K017896/1, FWF 1097-N23, and ANR-12-CHRI-0003-03, in the framework of the CHIST-ERA ERA-NET program line.

REFERENCES

- [1] Adams, B., Phung, D. and Venkatesh, S. (2011). Eventscape: Visualizing Events Over Times with Emotive Facets. 19th ACM International Conference on Multimedia (MM-2011). Scottsdale, USA: 1477-1480.
- [2] Diakopoulos, N., Naaman, M. and Kivran-Swaine, F. (2010). Diamonds in the Rough: Social Media Visual Analytics for Journalistic Inquiry. IEEE Symposium on Visual Analytics Science and Technology (VAST-2010). Salt Lake City, USA: IEEE: 115-122
- [3] Hoang, T.-A., Cohen, W.W., et al. (2013). Politics, Sharing and Emotion in Microblogs. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Niagara Falls, Canada: ACM Press: 282-289.
- [4] 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.
- [5] Mangaravite, V., Assis, G.T.d. and Ferreira, A.A. (2012). Improving the Efficiency of a Genre-aware Approach to Focused Crawling Based on Link Context. Eighth Latin American Web Congress (LA-WEB 2012). Cartagena de Indias, Colombia: IEEE CPS: 17-23.
- [6] Marcus, A., Bernstein, M.S., et al. (2011). Twitinfo: Aggregating and Visualizing Microblogs for Event Exploration. 2011 Annual Conference on Human Factors in Computing Systems (CHI-11). Vancouver, Canada: ACM: 227-236.
- [7] Muralidharan, A., Hearst, M.A. and Fan, C. (2013). WordSeer: A Knowledge Synthesis Environment for Textual Data. 22nd ACM International Conference Information and Knowledge Management (CIKM-2013). San Francisco, USA: ACM: 2533-2536.
- [8] Scharl, A., Hubmann-Haidvogel, A., et al. (2013). "From Web Intelligence to Knowledge Co-Creation – A Platform to Analyze and Support Stakeholder Communication", *IEEE Internet Computing*, 17(5): 21-29.
- [9] Shamma, D.A., Kennedy, L. and Churchill, E.F. (2010). Tweetgeist: Can the Twitter Timeline Reveal the Structure of Broadcast Events? ACM Conference on Computer Supported Cooperative Work (CSCW-2010). Savannah, USA.
- [10] Sipos, R., Ghosh, A. and Joachims, T. (2014). Was this Review Helpful to You?: It Depends! Context and Voting Patterns in Online Content. 23rd International World Wide Web Conference (WWW-2014). Seoul, Korea: World Wide Web Consortium: 337-347.
- [11] Sison, M.D. (2013). "Creative Strategic Communications: A Case Study of Earth Hour", *International Journal of Strategic Communication*, 7(4): 227-240.
- [12] Wattenberg, M. and Viégas, F.B. (2008). "The Word Tree, an Interactive Visual Concordance", *IEEE Transactions on Visualization and Computer Graphics*, 14(6): 1221-1228.
- [13] Weichselbraun, A., Gindl, S. and Scharl, A. (2013). "Extracting and Grounding Contextualized Sentiment Lexicons", *IEEE Intelligent Systems*, 28(2): 39-46.
- [14] Weichselbraun, A., Gindl, S. and Scharl, A. (2014). "Enriching Semantic Knowledge Bases for Opinion Mining in Big Data Applications", *Knowledge-Based Systems: Forthcoming* (Accepted 26 Apr 2014).
- [15] Weichselbraun, A., Streiff, D. and Scharl, A. (2014). Linked Enterprise Data for Fine Grained Named Entity Linking and Web Intelligence. 4th International Conference on Web Intelligence, Mining and Semantics (WIMS-2014). Thessaloniki, Greece: ACM Press.