Metadata Enriched Visualization of Keywords in Context

Daniel Fischl
MODUL University Vienna
Am Kahlenberg 1, 1190 Vienna
daniel.fischl@modul.ac.at
+43 (1) 3203555 534

Arno Scharl
MODUL University Vienna
Am Kahlenberg 1, 1190 Vienna
arno.scharl@modul.ac.at
+43 (1) 3203555 500

ABSTRACT
This paper presents an interactive, synchronized and metadata enriched implementation of the Word Tree metaphor, which is an interactive visualization technique to show Keywords-in-Context (KWIC). Embedded into a Web intelligence platform focusing on climate change coverage, it provides users with a tool to better understand the usage of terms in large document collections. One of the novelties is the implementation of filters for the Word Tree, which shifts the focus of attention directly onto significant phrases, instead of punctuation or fill-words inherent to natural language usage.

Author Keywords
Algorithms; design; human factors; keywords in context

ACM Classification Keywords
D.5.2 Graphical user interfaces (GUI); H.3.3: Information filtering; I.3.6: Interaction techniques

INTRODUCTION
If an analyst searches through the Internet for instances of a bank, product or any other items of interest, he or she will be confronted with a great amount of unstructured and unordered textual information to review. To help users achieve a better understanding of how terms or phrases are being used in articles and posts in different media (to which we will subsequently refer to as "documents"), and how they are perceived by these media, we adopt and extend the functionality of the well-known Word Tree metaphor [5].

Our motivation, and subsequent use case for this work, is to better visualize the context of user-defined search terms inside the *Media Watch on Climate Change* [3] as shown in Figure 1, where the Word Tree is shown in the center part. Publicly available at www.ecoreserach.net/climate, the system collects documents from various news channels, social media platforms and the Web sites of NGOs and large corporations. It allows users to query the document collection and uses multiple coordinated views [2] to display the results including various types of metadata – e.g., associated keywords, polarity (positive and negative sentiment), geographic location, etc.

EICS'14, Jun 17-20 2014, Rome, Italy
ACM 978-1-4503-2725-1/14/06.
http://dx.doi.org/10.1145/2607023.2611451

WORD TREE VISUALIZATION
The Word Tree technique is a visual tool to show the different contexts in which certain terms appear. Its graph-based display facilitates the rapid exploration of search results and conveys a better understanding of how language is being used surrounding a certain topic. To generate the display, the system processes the list of concordances of the focus term and presents them in a structured manner. It complements other visualizations such as tag clouds and keyword graphs [3], which give a good overview of the main keywords, but do not reflect their usage context within specific sentences.

Unlike the original Word Tree [5], this work adopts a symmetrical approach [1] to directly visualize how the root of the tree, the search term, is embedded in the context. This allows representing the full sentence structures rather than fragments, which might leave out valuable information. The left part of the tree displays all sentence parts that occur before the search term (prefix tree), while the right part displays 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. The disadvantage of this symmetrical representation in comparison to the original version is that a link must be provided to show which sentence parts belong together. This is handled via mouse-over (see next Section).

Visual cues include different font sizes to indicate the frequency of phrases, and connecting lines to highlight typical sentence structures.

INTERACTIVE FEATURES
To explore the displayed information, the module provides several possibilities through interaction (we use the term "node" to refer to a specific word or phrase occurring multiple times, which is displayed between connecting lines):

- **Hovering** over a node highlights all connected sentences – only a single (complete) sentence in the case of leaf nodes, or all sentences containing the phrase from the root to the hovered branch in the case of intermediate nodes.

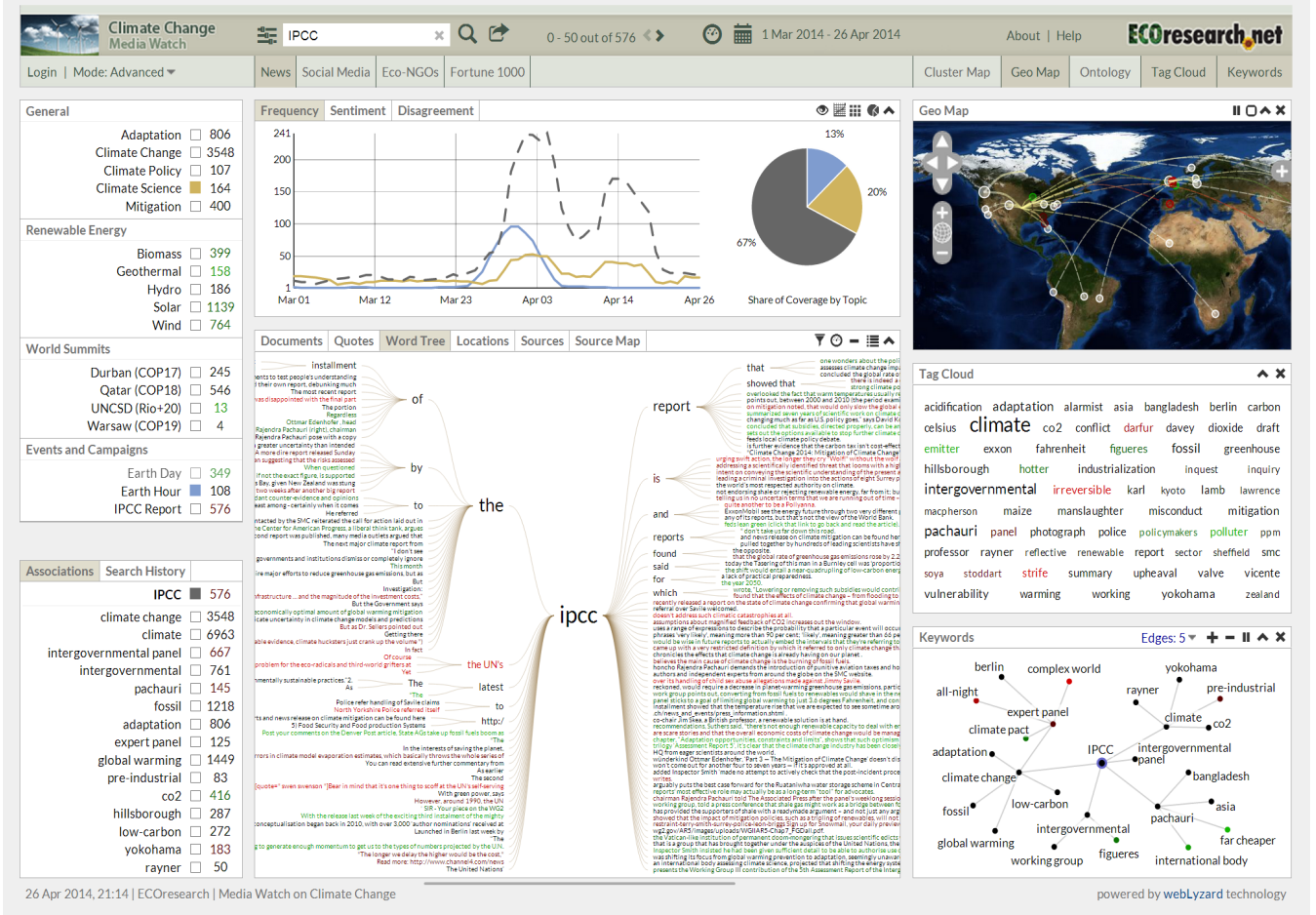

Figure 1: Screenshot of the Media Watch on Climate Change (www.ecoresearch.net/climate) with a Word Tree that shows Anglo-American news media coverage on the IPCC Assessment Report between March and April 2014

- **Single clicking** on the *root node* (= terms matching the search query) displays alternative root terms, which can be used to create variations of the tree based on the same set of search results. This is useful when searching for multiple terms via complex queries or predefined topics.
- **Single clicking** on *any of the other words* reprocesses the shown data to create a new tree (the phrase from the previous root to the clicked word becomes the new root). This drill down operation limits the amount of information shown; e.g., to explore sub-branches of the tree containing specific phrases.
- **Double clicking** on any of the shown word triggers a new full-text search.

Filtering

Word Trees were initially conceptualized as tools for the lexical analysis of texts. One of the main goals of this work, however, is to help users gain quick insights into a collection of documents. This is done by providing them with a tool to see how, and in combination with which words, the search term in the center of the tree is used in different contexts. Basically the original Word Tree is, by means of grouping, able to provide a structured view of the underlying text. Nevertheless, in order to understand the general message of a collection of texts, some characters and fill-words inherent to natural language use are not necessarily required. Since the original version treats each of those as a token when comparing phrases, a grouping which includes these might lessen the significance of the tree's overall structure. Therefore, we have implemented several possible filters, which adapt to the underlying data. A comparison of the different filter types is depicted in Figure 2, where each of the trees is based on the same data. Figure 2(a) shows the outcome of the original Word Tree algorithm. In all subfigures, the newly gained groupings compared to the original version have been highlighted. For reasons of simplicity we only show the suffix-part of the tree. The exact differences will be further explained in the following subsections.

Filter punctuation

As seen in Figure 2(a), building the tree using the original method would lead to a grouping by comma, since several sentences contain the phrase "energy,". Our version of filtering by punctuation removes such groupings, which al-

lows the algorithm to find groupings which have previously been obscured because of punctuation. An example of a newly found grouping can be seen in Figure 2(b), where new groupings based on the words "like" and "is" have been found.

Filter stop-words

Based on the punctuation filter, we also provide a more restrictive filter, which filters not only punctuation marks, but also prevents groupings by a pre-defined list of stop-words.

Figure 2(c) shows the outcome of this approach, leading to a tree where additional groupings have been found, like "sustainable", "development" and "bills". Note that the cluster "is" has been removed, since it is considered a stop-word and therefore a grouping by it has been prevented.

Adaptive filtering

The results of an adaptive filter are presented in Figure 2(d). It is based on the previous two filters, but strips the text of punctuation or stop-words, only if a better grouping can be found without them. Otherwise, it maintains the original grouping.

Figure 2(d) shows that the same groupings were found as in the stop-word filtering process except for two groupings by "," and "is" – which are still smaller than the same groupings in Figure 2(a). This is because no superior solution could be found for the affected phrases.

Hierarchical Layers

To allow the users to focus on often occurring phrases, buttons to add and remove nodes have been provided. These buttons extend or trim the tree by either adding a hidden hierarchical layer of branches, or by removing the current set of leaf nodes. This allows users to hide single sentences which occur only once in the document collection and focus on the main tree structure by displaying only those phrases that have occurred multiple times. This functionality, in combination with options to reduce the result set and drill-down into sub-branches allow users to focus on relevant data and sub-branches to gain further insights. An example of such a trimmed tree, where the first layer has been hidden, is shown in Figure 3.

Sentiment Display

The *Media Watch on Climate Change* computes a sentiment value for each sentence [4].

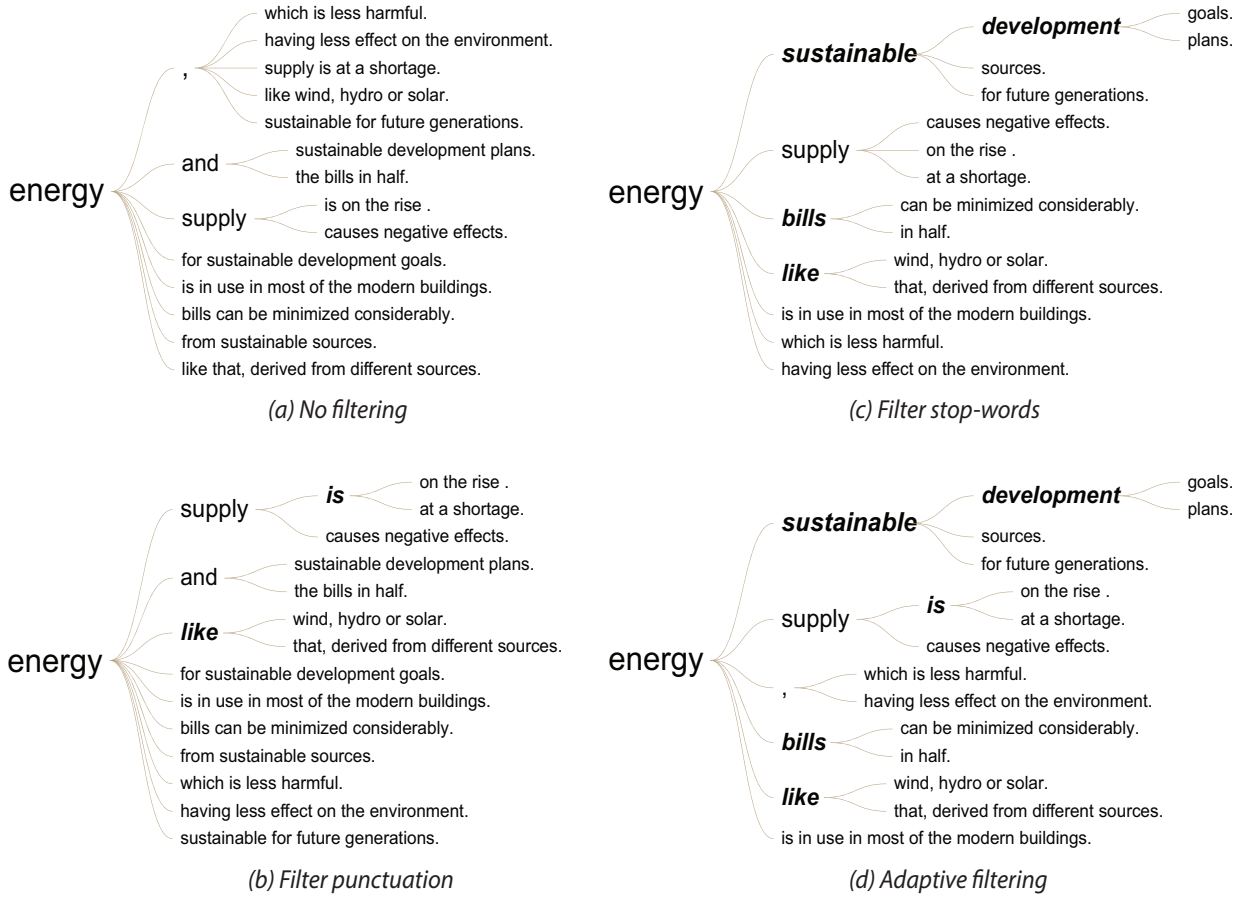

Figure 2: Comparison of different filtering techniques: (a) shows the original Word Tree structure, (b) filters groupings by punctuation (c) filters groupings by stop-words and (d) adapts the filtering by filtering only in the case of more favorable groupings

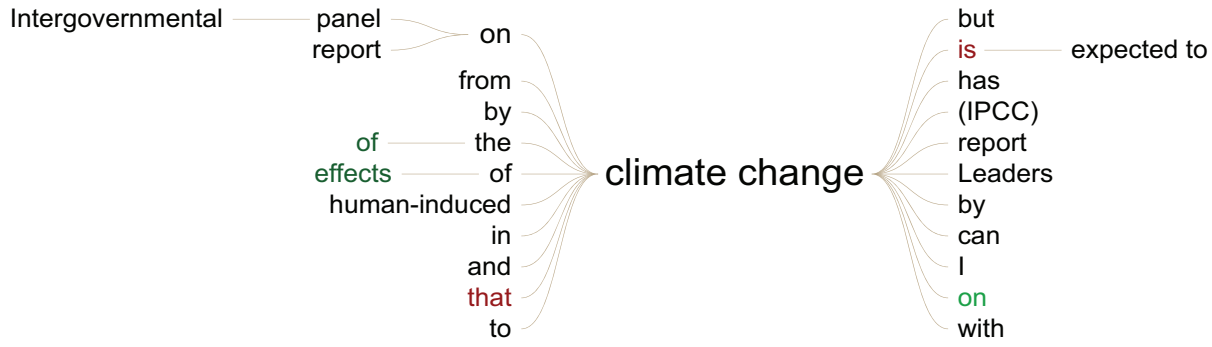

Figure 3: Example of the word tree with the root term "climate change" with the first layer (all leaf nodes) hidden and sentiment coloring active.

The system determines the ratio of positive and negative terms found in a document (based on a sentiment lexicon, an enumerative list of sentiment terms with indicators of their sentiment charges). It then uses this value as an indicator of overall polarity. The accuracy of this metric is further improved by considering linguistic features such as negations and intensifiers. The user has the option to enable color coding based on the distribution of these sentiment values. Each sentence is then colored in the range from red (negative) to black (neutral) to green (positive).

Intermediate nodes receive the average color of all their connected child nodes – this might result in neutral colors in the case of controversial phrases with a roughly similar number of positive and negative mentions.

CONCLUSION AND OUTLOOK

This paper presented the integration of the Word Tree concept into a news and social media aggregator focusing on the environmental domain, the *Media Watch on Climate Change* (www.ecoresearch.net/climate).

Adapted to the specific requirements of a multiple coordinated view interface, this implementation of the Word Tree metaphor offers a range of interaction possibilities to enable an effective and quick manipulation of the retrieved content, including filtering techniques to eliminate irrelevant terms and focus on the most significant information.

The display of additional metadata such as the color-coding of sentiment information helps users to quickly understanding the relevance of the various topics. All the required filtering is done in real time during the generation process, based on the retrieved set of search results. This one-pass generation eliminates the need for additional post- or pre-processing steps.

Future work will continue to explore Word Tree usage when exploring large document collections of the *Media Watch on Climate Change*, as well as other domain-specific content repositories gathered through the *webLyzard Web intelligence platform* (www.weblyzard.com).

We will investigate the possibility of stripping the Word Tree of contextual information step-by-step until it transforms into a keyword-graph-like structure, where only the most important keywords remain, without additional context information. This could potentially facilitate the perception of information in the underlying data.

ACKNOWLEDGEMENT

The presented work was supported by DecarboNet.eu, which receives funding from the European Union's 7th Framework Program for research, technology development and demonstration (GA No. 610829), as well as uComp.eu, funded by the Austrian Science Fund through the European CHIST-ERA program line (GA No. I 1097-N23).

REFERENCES

[1] Culy, C. and Lyding, V. 2010. Double Tree: An Advanced KWIC Visualization for Expert Users. In *Proceedings of the 2010 14th Int'l Conference Information Visualisation* (IV '10). IEEE Press, Washington, DC, USA, 98-103.

[2] 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.

[3] Scharl, A., Hubmann-Haidvogel, A., Sabou, M., Weichselbraun, A. and Lang, H.-P. (2013). "From Web Intelligence to Knowledge Co-Creation – A Platform to Analyze and Support Stakeholder Communication", *IEEE Internet Computing*, 17(5): 21-29.

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

[5] 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.