# Topic Wizard – Interactive Visual Tool for Defining and Disambiguating Topics via Regular Expressions

**Arno Scharl, Daniel Fischl**
MODUL University Vienna, Department of New Media Technology
Am Kahlenberg 1, A-1190 Vienna, Austria
{arno.scharl, daniel.fischl}@modul.ac.at

## ABSTRACT

The *Topic Wizard* presented in this paper is a new interactive tool that supports the definition and revision of topics in form of regular expressions, combining a dialog for prefix and suffix extension with a word tree-based representation of phrases for restricting the query to specific expressions. The examples stem from the *Media Watch on Climate Change*, a public Web portal based on the webLyzard Web intelligence platform that aggregates environmental stakeholder communication from multiple online sources.

## Author Keywords

Topic definition; term disambiguation; word tree; editor; regular expressions.

## ACM Classification Keywords

H.3.3 [Information Search and Retrieval]: Information Filtering; H.5.2 [Information Interfaces and Presentation]: User Interfaces and Interaction styles; H.5.3 [Group and Organization Interfaces]: Web-based Interaction.

## INTRODUCTION

The *Media Watch on Climate Change* (MWCC) is a Web intelligence and online collaboration platform available at www.ecoresearch.net/climate. It compiles large archives of digital content from multiple sources, and provides a variety of knowledge co-creation and visualization services [3].

MWCC integrates multilingual content from English, French and German source: social media including *Twitter, Facebook, Google+* and *YouTube,* and the Web sites of news organizations, companies, municipalities, and environmental NGOs. To classify and annotate this content, MWCC utilizes the *webLyzard* media monitoring and Web intelligence platform (www.weblyzard.com). The platforms' data acquisition and information extraction services have been optimized for Web-scale applications in terms of throughput and scalability. The result is a comprehensive information space spanning geospatial, temporal and thematic dimensions. A visual dashboard lets users navigate this information space, and structure the search results along these multiple dimensions.

The metadata enriched visualization of keywords in context ("word tree") shown in Figure 1, for example, sheds light on typical usage patterns of a term within a larger corpus [1]. Within the MWCC portal, the color-coding reflects normalized document sentiment [6], ranging from green (positive) to grey (neutral) and red (negative).

This paper presents a novel application of the graph rendering capabilities underlying the *word tree*. Instead of showing the lexical context of query results [4], the *Topic Wizard* uses the visual metaphor of the *word tree* to guide users in their task of defining and disambiguating topics, and to generate the corresponding regular expressions.

The *Topic Wizard* thereby helps to infer the informational needs of a user, which is a primary goal of any search system. Creating accurate regular expressions that result in high recall and precision usually requires significant manual effort and expert knowledge, and there is ongoing research effort in partially automating the regular expression learning pipeline [7].

The remainder of this paper presents the word tree visualization, outlines its integration into the topic definition workflow, describes the main interface components, gives a number of specific examples from the MWCC, and concludes with a summary and outlook.

## WORD TREE VISUALIZATION

MWCC can list search results on a document level, or provide matching quotes in the form of a *concordance list*. Users can sort the concordances by 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.

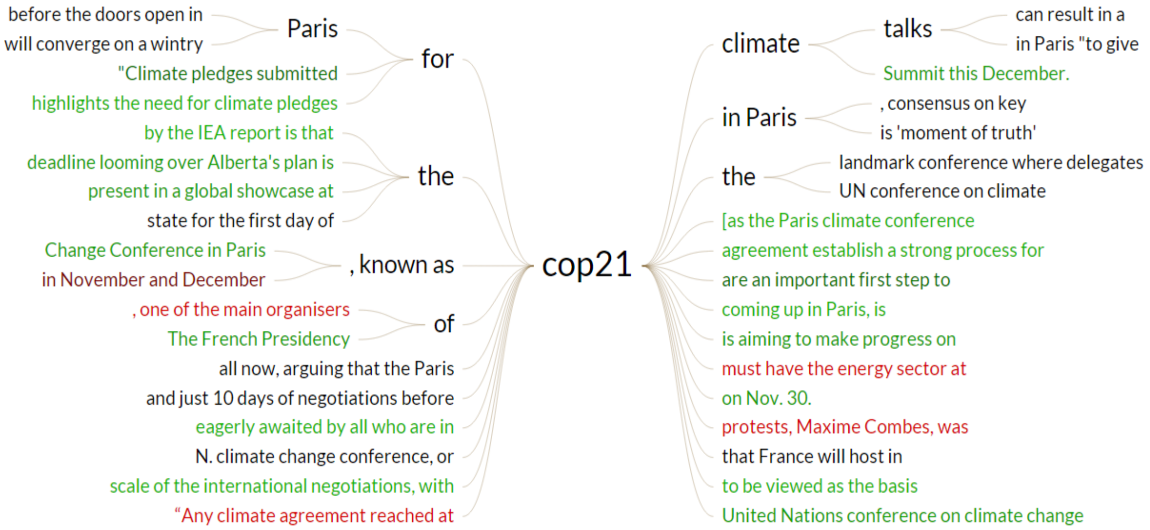

before the doors open in
will converge on a wintry
Paris
"Climate pledges submitted
highlights the need for climate pledges
for
by the IEA report is that
deadline looming over Alberta's plan is
present in a global showcase at
state for the first day of
the
Change Conference in Paris
in November and December
, known as
, one of the main organisers
The French Presidency
of
all now, arguing that the Paris
and just 10 days of negotiations before
eagerly awaited by all who are in
N. climate change conference, or
scale of the international negotiations, with
"Any climate agreement reached at
cop21
climate
talks
can result in a
in Paris "to give
Summit this December.
in Paris
, consensus on key
is 'moment of truth'
the
landmark conference where delegates
UN conference on climate
[as the Paris climate conference
agreement establish a strong process for
are an important first step to
coming up in Paris, is
is aiming to make progress on
must have the energy sector at
on Nov. 30.
protests, Maxime Combes, was
that France will host in
to be viewed as the basis
United Nations conference on climate change

**Figure 1.** Word tree for the query "COP21" in news media coverage between May and June 2015

Based on the popular keyword-in-context technique [5], the MWCC implementation of the *word tree* metaphor adopts a symmetrical approach [2]. 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 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 "COP21" (acronym of the *United Nations Climate Change Conference* to be held in Paris in December 2015; www.cop21paris.org), and grouping identical phrases containing the term into nodes (e.g., "COP21 climate talks"). 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.

## TOPIC DEFINITION WORKFLOW

Accurately defining a topic in Web intelligence and semantic search applications such as MWCC goes beyond simply searching for a particular term. For ad-hoc queries, simple text fields typically suffice. For defining and disambiguating topics, however, the number of terms to include is often larger. To properly describe abstract concepts like *climate change* or popular but lexically ambiguous brand names such as *Amazon*, *Apple*, and *Three*, one needs to consider synonyms, singular and plural versions of a term, grammatical variations, lists of related products and services, etc.

To enumerate all the different forms a base term might appear in, a variety of prefixes and suffixes need to be specified. In addition, users might want to include (whitelist) or exclude (blacklist) specific term combinations. To formally describe such multiple appearances, regular expressions are very useful. However, defining a "match-all" regular expression such as ".*solar.*" that includes all possible prefixes and suffixes might not always be the desired behavior:

- users might want to disambiguate a term and disregard terms not related to their desired topic – e.g. "solaris" and "solarium" when aiming to analyze recent Web coverage on solar energy.
- searching with such "match-all" regular expressions will result in reduced response times and negatively impact the "snappiness" of the application.

To prevent this, users manually have to define regular expressions with prefixes and suffixes that surround a specific base term, to cover the desired occurrence forms of their concept. This definition process can dwindle into a tedious task of pondering over a wide range of possible combinations, and will most likely result in incomplete definitions since it is unlikely that a user would be able to think of every possible combination while defining the topic.

### Interaction Design

To assist users with the enumeration of occurrence forms, the wizard collects a repository of domain-independent text documents from online sources (e.g. English-speaking news media channels, or a generic set of social media postings). To maximize query responsiveness, all content elements are indexed using Elasticsearch (www.elastic.co). The underlying content repository is updated on a regular basis to keep the dataset current and consider new terms and expressions as they emerge in the online dialog.

Instead of having to specify all possible inflections and term combinations for a given base term, the generic index is queried to present users with a scrollable list of prefixes and suffixes. This list can be used to select prefixes and suffixes to be included in the regular expression.

This process can be seen as a measure to improve the regular expression's *recall* and *precision*, since users get to expand their topic definition based on a screened and continuously updated selection of the most common word forms occurring in current media coverage.

In many cases, even an accurate definition of a single term including all the relevant prefixes and suffixes might not suffice to define a topic - e.g., when looking for a specific set of multi-word expressions, or when disambiguating a base term with multiple meanings. For this purpose, the *Topic Wizard* allows users to restrict the regular expression to exact phrases. Taking into account all selected pre- and suffixes, the background queries provide all phrases surrounding the defined set of terms, grouping them together in word tree-based fashion. To not overload the user with unnecessary information, only the most frequent terms directly following the base terms are displayed, with the possibility to expand the subtrees on demand if longer phrases need to be specified.

It is important to keep in mind that selecting prefixes and suffixes will result in an *extended* result set, while restricting the topic to certain phrases will *limit* the result set.

### Main Interface Components

The user interface of the *Topic Wizard* consists of a mode selector, the central term list, visual tree structures to the left and right, and additional options in the footer menu:

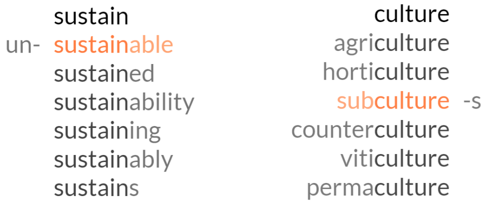
- *Mode Selection.* Located in the lower part of the window, users can switch between the suffix and prefix modes, or activate the word tree module to define phrase restrictions.
- *Term List.* Depending on the selected mode, the area in the center of the window shows either a scrollable list of the base term plus the identified prefixes or suffixes, or the currently selected pre-/suffix combinations in the phrase restriction mode.
- *Phrase Tree.* Left and right of the center area, a tree structure shows the identified phrases surrounding the term(s) - initially collapsed and activated upon switching to the phrase restriction mode.
- *Additional Options*. The footer menu provides (i) sorting criteria next to the mode selector to list query results alphabetically, or by frequency of occurrence in the generic corpus; (ii) stop word filtering in phrase restriction mode to filter for common stop words; and (iii) a result counter in the lower right corner that indicates the number of matches in the selected data source(s) and date range.

Using the components listed above, c*licking* on a *term* adds the highlighted prefixes or suffixes. Hovering over a *prefix term* will then display all suffixes that the prefixed term appears with, and vice versa for suffix terms.

Hovering over a *phrase* will display a list of base terms preceding or following this phrase; if available (depending on the local tree structure), an icon to expand or collapse a subtree is visible, which can be used to show or hide the different variations of how the phrase can continue. Clicking on a sub-tree restricts the query to the selected *phrase*.

### Example Query

The following example illustrates the topic definition process. Starting from a single base term such as "sustain", the visualization supports the identification and addition of common prefixes (**un**sustain) or suffixes (sustain**ability**, sustain**able**, etc.) to enrich the query and increase its recall.



sustain
un- sustainable
sustained
sustainability
sustaining
sustainably
sustains

culture
agriculture
horticulture
subculture -s
counterculture
viticulture
permaculture

**Figure 2.** Examples for the prefix and suffix selector for the two base terms "sustain" (left) and "culture" (right)

Users who are interested in a more narrow specification can switch to the *"Restrict to Phrases"* mode. In the screenshot of Figure 3, for example, the phrase "sustainable energy" is activated, and extensions such as "generation", "market" and "policy" shown. This active selection would result in the following regular expression:

```
sustain(ability|able)? (energy policy)
```

Stored as a topic definition, this regular expression would return all documents containing the base term "sustain" (plus the selected prefix and suffix variations), followed directly by the term "energy policy".

### CONCLUSION AND OUTLOOK

This paper has introduced *Topic Wizard*, a novel tool to aid non-expert users in the generation of regular expressions, in order to define input filters and topics for the *Media Watch on Climate Change* and other applications based on the webLyzard Web intelligence platform.

The *Topic Wizard* allows users to iteratively describe a desired result, evolving from a descriptive concept term through specification and disambiguation. The flexibility of this approach results in a wide area of application – not only for specifying search queries, but also for developing blacklists and whitelists effectively.

Future work will (i) extend the prefix and suffix selector with the ability to identify and select *infixes*, (ii) embed a lookup table to correct typos and account for both US and British spelling, and (iii) provide a recommender system to identify *synonyms* and integrate them into the resulting search query automatically.

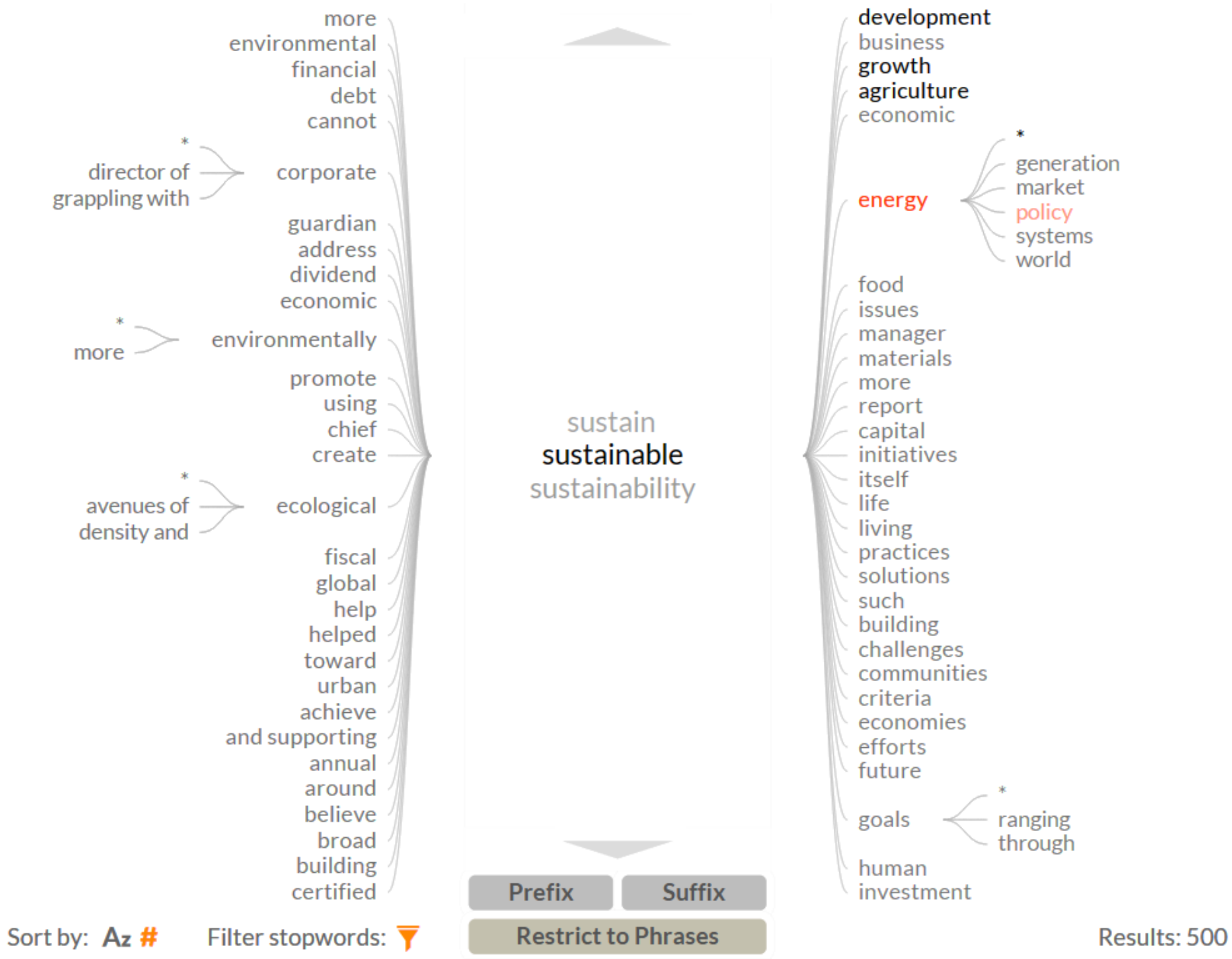


**Figure 3.** Screenshot of the Topic Wizard for the base term "sustain", including word tree-based phrase restriction components

**Acknowledgement.** This research has been conducted as part of the DecarboNet project (www.decarbonet.eu), which has received funding by the European Union's 7th Framework Program for Research, Technology Development and Demonstration under the Grant Agreement No. 610829.

## REFERENCES

1. Fischl, D. and Scharl, A. (2014). Metadata Enriched Visualization of Keywords in Context. *6th ACM SIGCHI Symposium on Engineering Interactive Computing Systems (EICS-2014)*. Italy, Rome: ACM Press: 193-196.
2. Muralidharan, A., Hearst, M.A. and Fan, C. (2013). WordSeer: A Knowledge Synthesis Environment for Textual Data. *22nd ACM International Conference on Information and Knowledge Management (CIKM-2013)*. San Francisco, USA: ACM: 2533-2536.
3. 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 Compting,* 17(5): 21-29.
4. Scharl, A., Kamolov, R., et al. (2014). Visualizing Contextual Information in Aggregated Web Content Repositories. *9th Latin American Web Congress (LA-WEB 2014)*. J.M. Almeida et al. Ouro Preto, Brazil: IEEE Press: 114-118.
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.
6. Weichselbraun, A., Gindl, S. and Scharl, A. (2013). "Extracting and Grounding Contextualized Sentiment Lexicons", *IEEE Intelligent Systems,* 28(2): 39-46.
7. Yang, H., Pupons-Wickham, D., et al. (2013). I Can Do Text Analytics!: Designing Development Tools for Novice Developers. *SIGCHI Conference on Human Factors in Computing Systems (CHI -2013)*. New York, USA: ACM Press: 1599-1608.