LREC 2018 Workshop

The 3rd Workshop on Visualization as Added Value in the Development, Use and Evaluation of Language Resources (VisLR III)

PROCEEDINGS

Edited by

Mennatallah El-Assady, Annette Hautli-Janisz, Verena Lyding

ISBN: 979-10-95546-13-9
EAN: 9791095546139

12 May 2018

Proceedings of the LREC 2018 Workshop
"The 3rd Workshop on Visualization as Added Value in the Development, Use and Evaluation of Language Resources (VisLR III)"

12 May 2018 – Miyazaki, Japan

Edited by Mennatallah El-Assady, Annette Hautli-Janisz, Verena Lyding

https://typo.uni-konstanz.de/vislr/

Organising Committee

- Mennatallah El-Assady, University of Konstanz, Germany*
- Annette Hautli-Janisz, University of Konstanz, Germany*
- Verena Lyding, eurac research, Bolzano, Italy*

*: Main editors and chairs of the Organising Committee

Programme Committee

- Miriam Butt, University of Konstanz, Germany
- Chris Culy, Independent consultant
- Koenraad deSmedt, Univeristy of Bergen, Norway
- Florian Heimerl, University of Wisconsin-Madison, USA
- Stefan Jänicke, University of Leipzig, Germany
- Daniel Keim, University of Konstanz, Germany
- Steffen Koch, University of Stuttgart, Germany
- Victoria Rosén, University of Bergen, Norway
- Chris Weaver, The University of Oklahoma, USA

Preface

While "traditional" language resources, ranging from multi-layered treebanks to linked ontologies, encode complex linguistic information, data analysis in the online world faces yet another set of challenges: Data is generated around the clock, comes in huge quantities and reflects events in the digital world in real-time. With this type of data at hand, we need to think about novel ways of visually encoding properties of language, not only to respond to certain trends such as hate speeches and cyberbullying, but also to investigate the development of language in a more general manner. Visualization offers new methods for processing and analysing online data, and as such, the workshop fits squarely into this year's hot topic 'LRs in the Online World'. VisLR III also promotes LREC's general objectives of bringing together related disciplines. We particularly reach out to disciplines starting to use language resources and visualization to develop their own research agendas, such as the social sciences.

M. El-Assady, A. Hautli-Janisz, V. Lyding May 2018

Programme

14.00 – 14.05	Introduction
14.05 – 15.30	Mennatallah El-Assady *Visual Text Analytics: Techniques for Linguistic Information Visualization*
15.30 – 16.00	Karolina Suchowolec, Piotr Bański, Andreas Witt *Bridging Standards Development and Infrastructure Usage by Means of Concept Graphs: the Liaison of CLARIN and ISO TC37SC4 in Practice*
16.00 – 16.30	Coffee break
16.30 – 17.00	Maria Skeppstedt, Kostiantyn Kucher, Manfred Stede, Andreas Kerren *Topics2Themes: Computer-Assisted Argument Extraction by Visual Analysis of Important Topics*
17.00 – 17.30	Michael Abounabhan, Fatima Abu Salem, Kassim Antar, Shady Elbassuoni, Woody Hojeily and Sara Najem *Bridging the Gap between Data versus Technology Producers: An Interactive Visual Interface for Data Exploration*
17.30 – 18.00	Wolfgang Jentner, Florian Stoffel, Dominik Jäckle, Alexander Gärtner, Daniel Keim *DeepClouds: Stereoscopic 3D Wordle based on Conical Spirals*
18.00 – 18.05	Closing

Table of Contents

Bridging Standards Development and Infrastructure Usage by Means of Concept Graphs: the Liaison of CLARIN and ISO TC37SC4 in Practice

Karolina Suchowolec[1], Piotr Bański[2], Andreas Witt[2,3]
[1]Technische Hochschule Köln, [2]Institut für Deutsche Sprache, [3]Universität zu Köln
[1]Campus Südstadt, Ubierring 48, 50678 Köln, [2]R5, 6-13, 68161 Mannheim, [3]Albertus Magnus Platz, 50923 Köln
karolina.suchowolec@th-koeln.de,{banski, witt}@ids-mannheim.de, andreas.witt@uni-koeln.de

Abstract

The present submission reports on a pilot project conducted at the Institute for the German Language (IDS), aiming at strengthening the connection between ISO TC37SC4 “Language Resource Management” and the CLARIN infrastructure. In terminology management, attempts have recently been made to use graph-theoretical analyses to get a better understanding of the structure of terminology resources. The project described here aims at applying some of these methods to potentially incomplete concept fields produced over years by numerous researchers serving as experts and editors of ISO standards. The main results of the project are twofold. On the one hand, they comprise concept networks dynamically generated from a relational database and browsable by the user. On the other, the project has yielded significant qualitative feedback that will be offered to ISO. We provide the institutional context of this endeavour, its theoretical background, and an overview of data preparation and tools used. Finally, we discuss the results and illustrate some of them.

Keywords: CLARIN, ISO, standardisation, terminology, concept systems, network analysis, terminology visualisation, interactive graph visualisation

1. Introduction

The present submission reports on a pilot project conducted at the Institute for the German Language (IDS), aiming at strengthening the connection between ISO TC37SC4 “Language Resource Management” and CLARIN[1]. The former is a subcommittee of an international standards body focusing on the codification of norms for various aspects of language resource creation, management, and use, whereas the latter is a federated research infrastructure, currently bringing together dozens of centres of various types in almost 20 countries, mainly but not exclusively in Europe.

A simplified approach would paint ISO as the producer of standards and CLARIN as a consumer, but in fact, the relationship holds in both directions: numerous CLARIN researchers are also ISO experts, and CLARIN centres are responsible for the emergence of best practices that often end up becoming codified into standards. This relationship has recently been strengthened also on the political plane: since June 2017, ISO TC37SC4 and CLARIN-ERIC have a formal liaison agreement. The project described here may be seen as a practical expression of the new level of cooperation that both partners have entered.

As a CLARIN centre, IDS offers access to an interactive information system about standards endorsed or simply used by CLARIN centres. In order to make this system even more comprehensive, we plan to add to it a visualisation of the concept network extracted from standards published by ISO TC37 SC4. The present submission describes steps taken towards that goal and discusses the results.

Section 2 couches the project in a theoretical background, section 3 provides a view of the current ISO offer, section 4 describes the aims of the project and the technical preparations, while section 5 discusses the results. Section 6 focuses on the evaluation of the results, and section 7 outlines further directions for research.

2. Theoretical Background: Definitions and Concept Systems in Terminology Work

ISO Standard 10241-1:2011 (“Terminological entries in standards – Part 1: General requirements and examples of presentation”) provides recommendations on how to prepare and write terminological entries in ISO standards. Those recommendations follow the general terminology approach as described in e.g. ISO 704:2009-11 (“Terminology work – Principles and methods”). This approach is concept-oriented and in the following we focus on its two elements: definitions and concept systems.

2.1 Definitions

Definitions are a central means in terminology management for concept description. Definitions used in standards or other normative terminology sources differ from definitions used in other applications.

As stated in ISO 704:2009-11 (p. 22–24), in normative terminological resources, definitions are needed to clearly identify concepts in a domain and distinguish them from one another. Preferable definitions are short and concise, giving just enough information for concept distinction. This means that a definition in a standard does not aim at providing comprehensive or complete understanding, and that distinguishes it from an encyclopaedic definition. Moreover, definitions should be as general as possible and not limited to one standard only. In particular, a definition should, whenever possible, “be appropriate for other standards within closely related subjects.” (ISO10241-1:2011 p. 26)

[1] ISO TC37SC4 stands for Subcommittee 4 (“Language resource management”) of Technical Committee 37 “Terminology and other language and content resources”, cf. https://www.iso.org/committee/297592.html. CLARIN stands for “Common Language Resources and Technology Infrastructure”, cf. https://www.clarin.eu/. CLARIN is an ERIC (European Research Infrastructure Consortium).

As for structure and wording, the so-called intensional definition is preferred. It begins with a hyperonym of the concepts and continues with delimiting characteristics. A concept is not defined in isolation, but in relation to other concepts in the domain, which might be referenced in the definition body: "*Definitions* shall be systemic in order to enable a terminologist to reconstruct the *concept system*." (ISO 704:2009-11 p. 23) Given the above-mentioned requirements, we can expect that redundancy, ambiguity and the use of underspecified terms in standard definitions should be avoided.

2.2 Concept Systems

The general terminology approach stresses the importance of arranging concepts into concept systems, which are defined as "set[s] of concepts [...] structured according to the relations among them" (ISO 10241-1:2011 p. 3). More loosely structured sets are called concept fields. Concept systems are crucial for terminology management in many ways. A concept definition should, preferably, reflect the position of a concept within the concept system by referencing other related concepts (e.g. ISO 10241-1:2011, 2011 p. 26). Also, concept systems help standardization bodies in assigning normative status (*preferred*, *admitted*, *deprecated*) to synonymous terms. For terminology users, concept systems offer additional access to the data and help to explore the structure of the given domain.

Although concept systems are generally recognized as a useful tool in terminology management for both authors and users, there has been a lack of appropriate tools to dynamically and interactively generate concept systems from terminology databases (e.g. Drewer, Massion and Pulitano, 2017 p. 21). Hence, one had to resort to static means of visualisation such as regular drawings, which made the creation and maintenance of concept systems rather cumbersome.

Because concept systems play such an essential role in creating a coherent and well-formed terminology resource, a new generation of terminology tools has recently emerged that address the gap of dynamic visualisation (cf. Früh and Deubzer, 2016). In addition, more and more projects use generic dynamic visualisation technology and complement their terminology database with an additional visual access (e.g. *Verweis Viewer*: Chiocchetti, Culy and Ralli, 2015; *EcoLexicon*: Faber, León-Araúz and Reimerink, 2016; Lang, Schwinn and Suchowolec, in press).

Various display modes for concept systems have been proposed (DIN 2331:1980-04, 1980), but from the point of view of the end user, visualisation as a graph is often considered to be the desired one. A graph can be defined as "a set of *nodes*, which are an abstraction of any entities […], and the connecting links between pairs of nodes called *edges* or relationships." (Igual and Seguí, 2017 p. 142) A graph can be undirected or directed, depending on whether the relationship between edges is symmetric or not. There are several characteristics for analysing a graph (or network). Most importantly, the so-called centrality measures compute the importance of a node within a graph. Igual and Seguí (2017 p. 147ff) mention the following measures for graph analysis:

- degree centrality: „number of edges of the node"
- betweenness centrality: "number of times a node is crossed along the shortest path/s between any other pair of nodes"
- closeness centrality: "[quantified] position a node occupies in the network based on a distance calculation"
- eigenvector centrality: "relative score for a node based on its connections"
- PageRank: an algorithm developed for Google to "rate webpages objectively and effectively measure the attention devoted to them. [...] The rank of page P_i is the probability that a surfer on the Internet who starts visiting a random page and follows links, visits the page P_i." (p. 154–156). PageRank can be applied to the nodes of any directed graph.

In terminology management, attempts have recently been made to use graph-theoretical analyses to get a better understanding of the structure of terminology resources (Falke, Lang and Suchowolec, 2017). The pilot project described here aims at applying some of these methods to potentially incomplete concepts (concept fields) produced over years by numerous researchers serving as experts and editors of ISO standards. Apart from supplying proof of concept, the project intends to provide ISO with feedback of various sorts: on what can be done to make ISO definition networks more robust, and on the existing problems that can be eliminated quickly and improve the coherence of the existing networks.

3. Terminological Data in ISO Online Browsing Platform

The present section describes the current state of the ISO Online Browsing Platform[2] – a front-end to the ISO terminological database, where a user can get a glimpse of the content of standards by browsing selected sections (see figure 1).

3.1 Terminological Entry

Terminological entries of the OBP are concept-oriented and in general follow the recommendations of ISO. They can be displayed only as a list (sorted by relevance or by term), in two different views: the basic view and the full view.

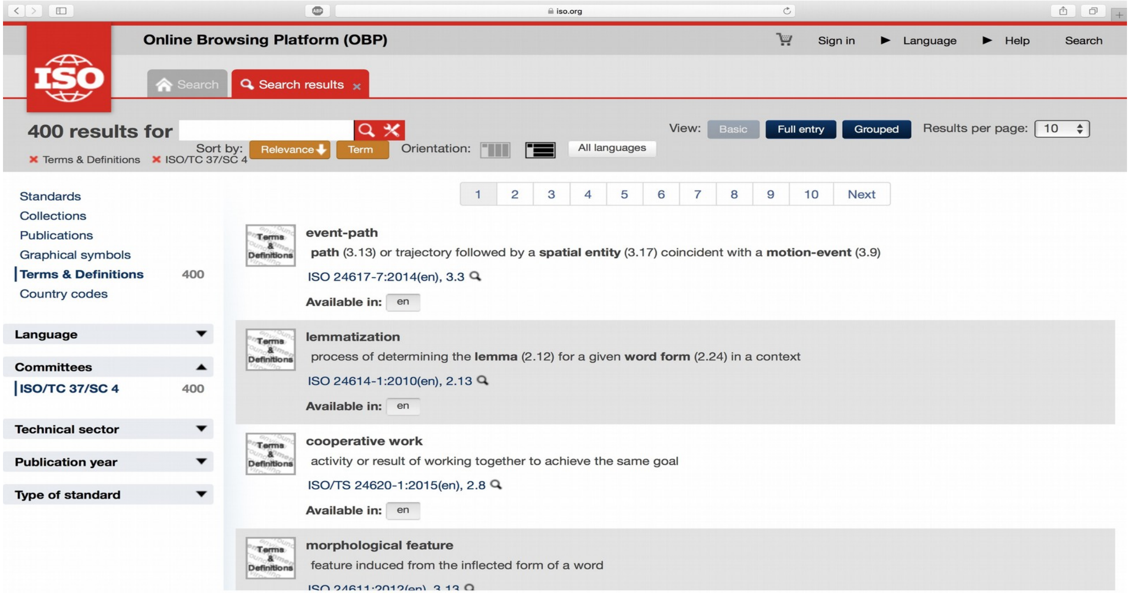
The basic view includes one term per concept (no synonyms), its definition, the name of the standard where the concept is defined and the relevant definition paragraph number.

The full view ("full entry") might additionally include all possible terms, (scope) notes, as well as examples. In case the definition is re-used from a different standard, the entry might also contain the reference to the original source of the definition.

3.2 Conceptual Relations

Within a terminological entry of a standard, other concepts are explicitly referenced in the definition body – concepts mentioned in a definition of a given concept are highlighted and cross-referenced with a hyperlink and the target paragraph number, as illustrated below:[3]

[2] https://www.iso.org/obp/ui



event-path
path (3.13) or trajectory followed by a **spatial entity** (3.17) coincident with a **motion-event** (3.9)
ISO 24617-7:2014(en), 3.3
Available in: en

lemmatization
process of determining the **lemma** (2.12) for a given **word form** (2.24) in a context
ISO 24614-1:2010(en), 2.13
Available in: en

cooperative work
activity or result of working together to achieve the same goal
ISO/TS 24620-1:2015(en), 2.8
Available in: en

morphological feature
feature induced from the inflected form of a word
ISO 24611:2012(en), 3.13

Figure 1: Screenshot of the Terminological Component of the ISO Online Browsing Platform (retrieved 2018-01-11)

> **lemmatization**
> process of determining the **lemma** (2.12) for a given **word form** (2.24) in a context

The current system has some restrictions and limitations that we enumerate in what follows.

- Cross-references are restricted in scope to the current standard only; there is no cross-referencing across standards, even within the same scientific committee.
- The relation type is not specified explicitly and is, therefore, subject to human interpretation. We observe different types of relations: hyperonymy as in [1], associative relation as in [2], and meronymy as in [3]:

  segmentation annotation
  annotation (2.3) [1] that delimits linguistic elements that appear in the **primary data** (2.1) [2]

  sentence
  related group of **word forms** (3.24) [3] containing a predication, usually expressing a complete thought and forming the basic unit of discourse structure

- Some obvious cross references are neither highlighted nor hyperlinked, as in [4] and [5]:

  graph
  set of nodes [4] (vertices) V(G) and a set of edges [5] E(G)

- By analogy to other terminological resources, a list of relations (Chiocchetti, Culy and Ralli, 2015), or even an explicit relation typing (Lang, Schwinn and Suchowolec, in press) could be provided in a dedicated section of an entry. However, as mentioned above, this information is missing.

To sum up, conceptual relations in the terminological entries of the OBP are rather ad-hoc, locally limited to a given standard, and sometimes incomplete. Within the current display options, it is impossible to grasp the conceptual *structure* of a set of standards or even of a single standard.

4. Project Description

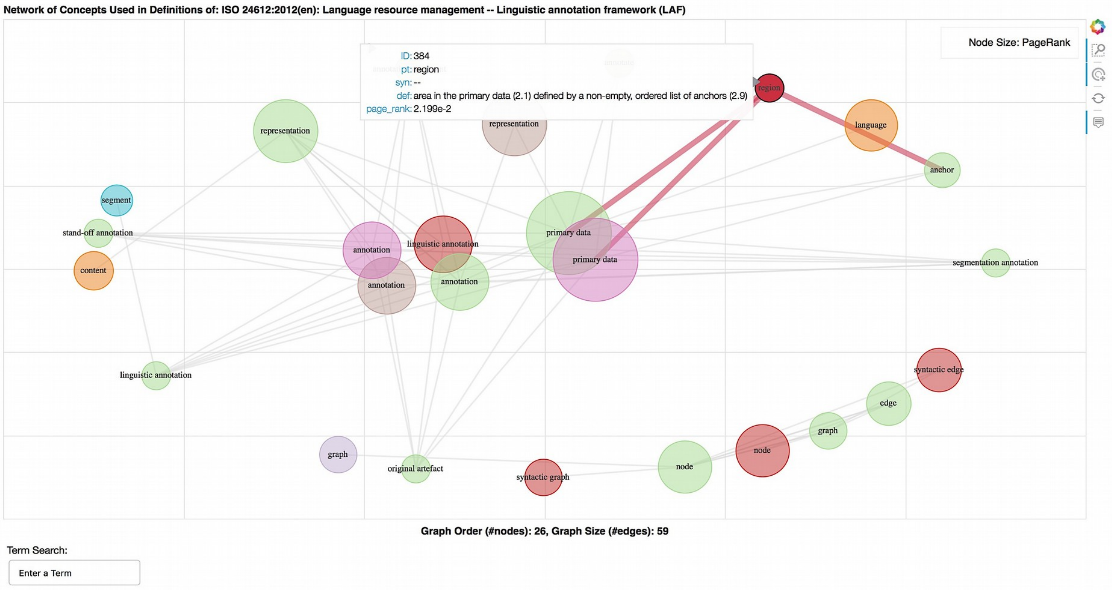
The rationale for our project rests on the following two assumptions of standard-related terminology work (see section 2):

- that the conceptual structure of a domain is implied in definitions
- that visualization techniques make conceptual structures more transparent and thus more beneficial.

We aim at making the conceptual structure of the terminological resources in the OBP explicit, by retrieving all concepts mentioned in the definitions regardless of their highlighting and then visualizing the relations as an interactive concept graph. The work is done with two user groups in mind, and thus encompasses two kinds of aims and questions:

- For end users – researchers in HLT and related areas:
  - to add a visual browsing option for the resource,
  - to give an assessment of the importance of concepts within a certain field, also for educational and training reasons,
  - to answer questions such as “Which concepts are mentioned by many definitions and are, therefore, central and need to be studied first?” or “Is there a conceptual overlap between standards?”

[3] All examples were retrieved from the OBP on January 08, 2018. The numbers in square brackets are introduced by the present authors.


Network of Concepts Used in Definitions of: ISO 24612:2012(en): Language resource management -- Linguistic annotation framework (LAF)

ID: 384
pt: region
syn: --
def: area in the primary data (2.1) defined by a non-empty, ordered list of anchors (2.9)
page_rank: 2.199e-2

Graph Order (#nodes): 26, Graph Size (#edges): 59

Term Search: Enter a Term

Figure 2: Bokeh Visualisation of ISO 24612:2012 Concepts and their Links to other Standards; hover tool on *region* for concept information and edges of related concepts. In order to give a visual hint of the importance of a node in a graph, we plot the size of the nodes according to an importance measure, here: PageRank.

- For terminology authors, as a diagnostic and evaluation tool providing answers to questions such as:
  - Is the conceptual structure of the given standard well-formed?
  - Is highlighting or hyperlinking missing?
  - Are cross-references correct and complete?
  - Are there entire concepts missing?
  - Are any definitions circular?
  - Are there doublets or overlapping definitions across standards?

The data used in the pilot project come from the ISO Online Browsing Platform and is copyrighted by ISO. Following advice from the CLARIN Legal Helpdesk, the data have been used only internally within the project, with the results planned to be displayed using HTML frames (in this way yielding all control over the data to ISO). The snippets of text mentioned in this paper and in the illustrations therein fulfill the EU copyright law conditions on research exception.

4.1 Data Preparation

Data was downloaded in the form of HTML (which most probably had been exported from TBX, TermBase eXchange, which is the standard representation and exchange format in the terminology community (ISO 30042, 2008)). We have searched for “full entry” in the facet “TC 37 SC 4”, with the option “Terms & Definitions” set. The result was exactly 400 concepts (definition entries) from 20 published standards. The resulting HTML was exploited to harvest Terms, Definitions, Standard, Paragraph and Notes; we assumed that there is always a preferred term (altogether, 405 terms were identified); we used the Python library *lxml* with the ElementTree API (*etree*) to process the HTML tree; the output was a CSV file. We created a relational database for storing the harvested terminology; the database is normalized (3rd normal form) and implemented in MySQL.

4.2 Technical Implementation

This section provides an overview of tools used in the project and an outline of the technical implementation.

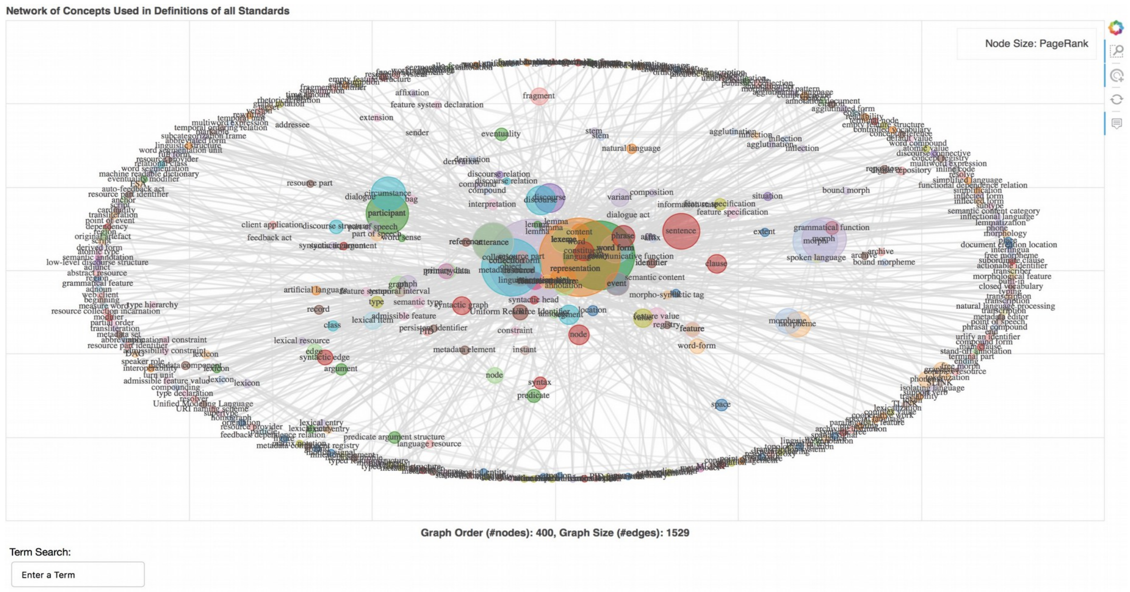
4.2.1 Tools

Preprocessing, DB management and visualisation was done in Python 3.6; next to the above mentioned lxml.etree for HTML processing, we used re (for regular expressions), pymysql (for connecting to the MySQL database and reading our terminology database), networkx (for creating basic initial graph and computing measures such as centrality, etc.), and bokeh (v. 0.12.11) and its sub-libraries (for visualising the networkx graph as an interactive HTML file); we used bokeh's hover tool for creating interactive tool tips when moving over the elements of the graph.

4.2.2 Extraction of Links Between the Concepts

The general idea was to check which ISO concepts are mentioned in a given ISO definition: if concept A is mentioned in the definition of concept B, then concept B points to concept A and a directed link (concept B → concept A) is created. All concepts form the nodes and all links form the edges of a graph. Such a graph is directed and records the links between concepts based on concept definitions.

We used Python NLTK library for linguistic pre-processing (WordNetLemmatizer with the default POS “n”). This step is necessary, because definitions are the


Network of Concepts Used in Definitions of all Standards

Node Size: PageRank

Graph Order (#nodes): 400, Graph Size (#edges): 1529

Term Search:

Enter a Term

Figure 3: Bokeh Visualisation of the Entire Concept Network of ISO TC37SC4. Node Size: PageRank

only source of information on the conceptual structure in our terminological resource and we test them against a list of all possible terms.

For every concept, we recorded the identified terms. Then, we used the information from the MySQL database to retrieve the corresponding IDs of concepts that use the identified terms. As a result, we obtained a data structure that records links between concept IDs.

In order to maximize automation, in cases where terms are ambiguous, we created links to all possible concepts. In particular, we also linked to concepts where the term in question is assigned as a non-preferred term.

4.2.3 Creation of a Graph Object

We used Python library networkx (v. 1.11) to create a graph object. First, for each concept (ID) from the database we created a node; then we added the edges using the data structure as described in the previous section.

4.2.4 Visualisation of the Graph as an Interactive HTML File

In order to visualise the graph object in an interactive HTML file with the bokeh Python library, we combined the approach of Meier (2016) with bokeh's method *from_networkx*, also customizing the ColumnDataSource by adding columns for concept definitions, preferred terms, synonyms and the source standard, and for the value of the centrality measure (or PageRank) for the size of the nodes. This is done on the fly by consulting the MySQL database for information on concepts and by computing the values of centrality measures with networkx. We have also colour-coded the nodes in such a way that each ISO standard uses a different colour.

We have used the customized columns of the ColumnDataSource in bokeh's hover tool. The hover tool provides interactive tool tips when moving the mouse over an element of the graph. In our case, when pointing to a graph node, the concept ID, definition, preferred term, synonymous terms as well as the source standard are displayed. Moreover, we make use of hover's inspection and selection policy for highlighting nodes and edges of related concepts. Finally, we added a JavaScript callback for term search.

5. Intermediate Results and Open Issues

The outcome of the project thus far is twofold. On the one hand, it comprises concept networks dynamically generated from a relational database. On the other, the project has yielded significant feedback that can be offered to ISO.

5.1.1 Concept Networks

We have been able to dynamically generate different types of interactive graphs, varying with respect to two major parameters: scope of the domain and importance of the node.

Three different scopes of graphs have been tested:

- definitions in *one* particular standard have been searched for concepts defined in the *same* standard
- definitions in *one* particular standard have been searched for concepts defined in *all* standards (see figure 2)
- definitions from *all* standards have been searched for concepts defined in *all* standards (see figure 3).

Recall that referencing concepts across related standards is justified by the ISO recommendations on standard drafting described in section 2.

Once the scope is set, we plot the graph according to the chosen importance measure (centrality measure or PageRank, as described above). At this stage of the project, the user creates the visualization in the command line. She can either call a function that, first, lists IDs for standards as well as importance measures and, then, prompts for input; or the user can call a different function and pass these IDs directly as arguments.

The resulting graph enables visual and interactive exploration of the ISO concept landscape: by clicking on graph nodes, information on a given concept (definition, used terms, standard, importance value) can be accessed. In addition, the visualisation can be zoomed in and out. Although the underlying networkx graph is directed, this directionality has not been visualized yet.

5.1.2 Qualitative Feedback to ISO

The pilot project has already produced several kinds of feedback that can be offered to ISO.

First, the evaluation of our approach provides immediate information on missing cross references that should be added to the definitions. Moreover, in order to handle issues reported in section 6.2 below, we have implemented a visualisation of explicit cross references that highlights edges which might require a closer examination: red indicates a mismatch between the term and the paragraph number, green points at the use of an unknown synonym, and blue represents a non-exact match. (see figure 4). Further, the interactive concept networks resulting from our automatic approach can be used by editors of standards to see the textually coded concept relations at a glance, thus helping to identify ill-formed structures.

Finally, processing the content of the OBP has generated several kinds of feedback that can be used for qualitative improvement of ISO terminology. We enumerate the main types of issues below.

Typos/malformedness. In addition to infrequent regular typos in terms and definitions, we have found instances of term merger, e.g. “orientation(al) relation” or trailing punctuation, e.g. “annotate,”.

Database Export. We have found instances of HTML tagging that we attribute to faulty database export, e.g. the additional POS tag “noun” in “annotation, noun” is treated as a term, in effect becoming a synonym of “annotation” (similarly with “annotate, verb”).

Inconsistency. A quick examination of terms used in definitions revealed some inconsistencies in the treatment of the normative status (i.e. preferred term=PT, admitted term, “bare” term) in the corpus of definitions. Four cases can be identified:

- PT is chosen, but a non-PT is used in a definition
- no normative status is assigned, all terms are treated as equal, “bare” terms (possibly, this is a result of faulty database export); this causes inconsistencies in definitions, because there is no guidance on which term to use as reference
- for the same set of terms, different standards choose different PTs
- it is not always clear whether a term is used as an instance of a particular concept or with a more general meaning (e.g. “content” or “text”)
- moreover, multiples of minimally modified definitions can be found across standards (sometimes, the duplication is signalled in the “Source” section, but sometimes this information is missing).

6. Evaluation

We evaluated our approach to automatically retrieving conceptual relations in terms of precision and retrieval ratio by comparing our concept networks of the smallest scope (one standard only) to the explicit cross references in the definitions of a standard as described in section 3.2. We do not use *recall* due to the lack of a gold standard; because the manually annotated cross references are incomplete (see section 3.2), we cannot regard them as a gold standard. The retrieval ratio shows then how well our fully automatic approach performs in comparison to the manual annotation.

First, we generated graph objects for explicit cross references. In order to create graph edges, we exploited HTML markup in combination with the parenthesised information on the paragraph number. Again, we used Python libraries *lxml, re, NLTK* and *networkx.*

Then, for each standard, we compared the edges retrieved by our automatic approach with the edges retrieved from explicit cross references. See Table 1 for the number of edges for each of the 20 standards by TC 37 SC 4.

As for precision, we looked at the number of edges retrieved by the automatic approach that are missing in the explicit cross references. However, because of the above-mentioned incompleteness of the cross references, a human evaluator within the project had to decide whether those edges were valid or invalid candidates. The precision is, then, the number of valid candidates to the total number of edges retrieved by our automatic approach.

$$\text{precision} = \text{e\_valid}_{\text{auto}} / \text{e\_total}_{\text{auto}}$$

As for retrieval ratio, we look at the number of edges retrieved from explicit cross references that are missing in our automatic approach. The retrieval ratio is, then, the ratio of the number of edges not missing in the automatic approach to the total number of edges from explicit cross references.

$$\text{retrieval ratio} = \text{e\_not\_missing}_{\text{auto}} / \text{e\_total}_{\text{expl}}$$

In the following, we discuss the overall precision and retrieval ratio for all standards in total.

	#nodes	*#edges*			*#nodes*	*#edges*	
		auto	*expl*			*auto*	*expl*
1	20	22	21	**2**	28	38	34
3	22	22	20	**4**	25	17	0
5	15	24	19	**6**	14	22	16
7	24	42	40	**8**	38	63	47
9	5	6	3	**10**	47	89	91
11	32	43	36	**12**	20	21	21
13	4	4	4	**14**	31	69	55
15	4	0	0	**16**	22	49	45
17	23	34	31	**18**	9	7	0
19	9	14	0	**20**	8	2	2
				∑ auto = 588		∑ expl = 485	

Table 1: Number of edges retrieved by the automatic approach (“auto”) and from explicit cross references (“expl”) in 20 standards of TC 37 SC 4

6.1 Retrieval Ratio

There are 19 edges in the explicit cross references that are not retrieved by our automatic method (i.e. 466 edges are

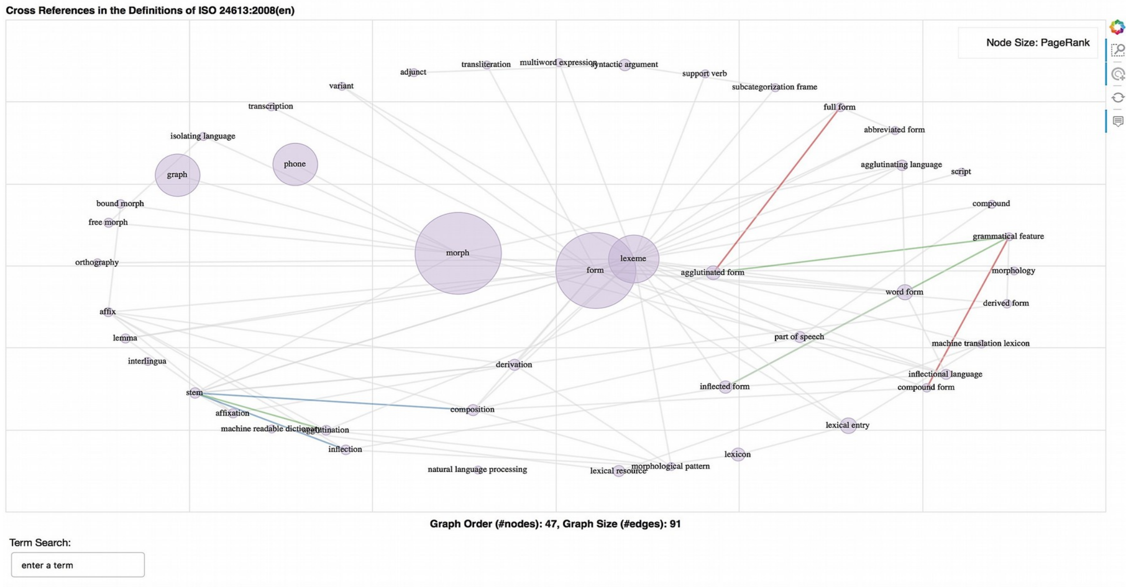


Figure 4: Bokeh Visualisation of ISO 24613:2008 as a Diagnostic Tool. Edges for revision are colour-marked: red: mismatch between the term and the paragraph number, green: unknown synonym used, blue: non-exact match.

"not missing"). 16 edges got missed because of the wording in the definitions: using unknown synonyms (e.g. "time unit" instead of "temporal unit" or "temporal interval"), referencing non-exact matches (e.g. "agglutinative" instead of "agglutination") or elliptical constructions ("compound or derived form" instead of "compound form", "derived form"). In two cases, there is a mismatch between the term used in the definition and the paragraph number ("URI (3.1.1)" instead of "URI (3.2.2)". Finally, one case involves the use of an inflected verbal concept ("annotated"), which was missed by our lemmatizer that was set for detecting nouns only. The resulting retrieval ratio is 0.96.

6.2 Precision

There are 122 edges in our automatic method that are not retrieved from explicit cross references. 86 of these edges were considered valid candidates by a human evaluator.[4] Out of the remaining 36 edges, 24 need a closer evaluation by a domain expert and 12 are invalid candidates, mostly because of the above-mentioned lemmatizer setting. Considering these 36 edges only, our precision reaches 0.94.

7. Future Work

We judge the results of the pilot project as very promising for our intended user groups, while naturally noting some issues that deserve further attention.

Already at this point, the project supplies new ways of both browsing the concept network (for the user) and of increasing its coherence (for the maintainer).

The dynamically generated interactive graphs offer additional visual access to the OBP, complementing its traditional list view. They can be used to explore the textually coded concept structure, also allowing to spot central concepts at a glance. We believe that this facilitates a better understanding of the structure of the domain, although empirical evidence is still needed.

Terminology authors benefit from this kind of visual access as well, because it helps to spot issues in definition texts that are rather difficult to find using the list display.

Already at this stage of the project, it is clear that even the simplest graph-theoretical characteristics such as graph size and graph order provide a valuable contribution to terminology management and help to improve the quality of the resource. In our future work, we therefore intend to exploit other graph-theoretical characteristics for the target user groups. In the longer perspective and with the scientific community in mind, we intend to perform exploratory network analysis in the sense of Igual and Seguí (2017) in order to gain more insight into the nature of the data at hand.

In further future work, we are going to implement more evaluation metrics for our fully automatised approach with larger scopes. At this point, we anticipate the need to address the general issue of how to deal with term ambiguity and underspecification, which, on one hand, is a notorious problem in terminology management, but might be, on the other hand, limited by the specific features of definitions in standards. The latter might be remedied with the help of user feedback.

[4] Note that this high number is partially due to the fact that some standards do not indicate any cross references at all, see standard number 4, 18 and 19 in Table 1.

Furthermore, we are going to deal with the negative impact of duplicate definitions on the readability of graph visualisations. At present, duplicate concepts are plotted as separate nodes, see figure 2, containing among others a cluster of 3 *annotation* nodes. We attribute this both to the source data (*ad hoc* modifications of definition bodies) and to the current visualisation solution. Further improvements will, therefore, deal with the link clutter of the larger-scope graphs as well as with a suitable directionality visualization for the edges (cf. Holten et al., 2011).

Next steps will also address the need to extract/add some hierarchical structure to the rather flat concept networks. We plan to look at pattern-based approaches such as Arnold and Rahm (2014), Sierra et al. (2008) or Storrer and Wellinghoff (2006), and to explore the options offered by graph-theoretical methods.

Respecting the ISO copyright restrictions, we intend to offer framed access to concept graphs to researchers from the CLARIN service offered by IDS Mannheim.

8. Acknowledgements

We would like to thank the anonymous reviewers for their detailed comments and valuable suggestions.

9. Bibliographical References

Arnold, P. and Rahm, E. (2014). Extracting Semantic Concept Relations from Wikipedia. In proceedings of WIMS '14, Thessaloniki, Greece, no pages, preprint: https://dbs.uni-leipzig.de/file/wims2014-final.pdf.

Chiocchetti, C., Culy, C. and Ralli, N. (2015): Visualising conceptual relations in the domain of law: Verweis Viewer. In Roche, C. (Ed.): TOTh 2013. Actes de la septième conférence TOTh – Chambéry – 6 & 7 juin 2013. Annecy: Institut Porphyre, Savoir et Connaissance, pages 135–154.

DIN 2331:1980-04 (1980). Begriffssysteme und ihre Darstellung. April 1980. Berlin: Beuth.

Drewer P., Massion F. and Pulitano, D. (2017). Was haben Wissensmodellierung, Wissensstruktur, künstliche Intelligenz und Terminologie miteinander zu tun? Deutsches Institut für Terminologie e.V., pages 1–28 http://dttev.org/images/img/abbildungen/DITeV_org_Terminologie_und_KI_2017_03_22_v2.pdf.

Faber, P., León-Araúz P. and Reimerink, A. (2016). EcoLexicon: New Features and Challenges. In Proceedings of GLOBALEX 2016 – Lexicographic Resources for Human Language Technology Workshop, pages 73–80, Portorož, Slovenia.

Falke, S., Lang, Ch. and Suchowolec, K. (2017). Entwicklung eines graphentheoretischen Terminologie-Analysetools. Talk given at ars grammatica: Grammatische Terminologie – Inhalte und Methoden, June 2017, IDS Mannheim, Germany.

Früh B. and Deubzer F. (2016). Von der Terminologieverwaltung zur Wissensorganisation. *Edition* 16(1):27–32.

Holten, D., Isenberg P., van Wijk, J. J. and Fekete, J.-D. (2011). An extended evaluation of the readability of tapered, animated, and textured directed-edge representations in node-link graphs. In Di Battista, G., Fekete, J.-D. and Huamin, Q. (Eds.): IEEE Pacific Visualization Symposium 2011. Proceedings, pages 195–202.

Igual, L. and Seguí, S. (2017). Network Analysis. In L. Igual & S. Seguí (Eds.), *Introduction to Data Science. A Python Approach to Concepts, Techniques and Applications*. Cham: Springer, pages 141–164.

ISO 704:2009-11 (2009). Terminology work – Principles and methods. November 2009.

ISO 10241-1:2011 (2011). Terminological entries in standards – Part 1: General requirements and examples of presentation. April 2011.

ISO 24612:2012 (2012). Language resource management – Linguistic annotation framework (LAF). June 2012.

ISO 30042:2008 (2008). Systems to manage terminology, knowledge and content – TermBase eXchange TBX. First edition

Lang, Ch., Schwinn, H. and Suchowolec, K. (in press). Grammatische Terminologie am IDS – Ein terminologisches Online-Wörterbuch als ein vernetztes Begriffssystem. In *Sprachreport* 34(1):16–26.

Meier, B. (2016). NetworkX Visualization Powered by Bokeh. Talk given at EuroPython 2016. video: https://av.tib.eu/media/21112; jupyter notebook: https://ep2016.europython.eu/media/conference/slides/networkx-visualization-powered-by-bokeh.ipynb, Bilbao, Spain.

Sierra, G., Alarcón, R., Aguilar, C. and Bach, C. (2008). Definitional verbal patterns for semantic relation extraction. *Terminology* 14(1):74–98.

Storrer, A. and Wellinghoff, S. (2006). Automated detection and annotation of term definitions in German text corpora. In Proceedings of LREC 2006, pages 2373–2376 http://www.lrec-conf.org/proceedings/lrec2006/pdf/128_pdf.pdf.

Topics2Themes: Computer-Assisted Argument Extraction by Visual Analysis of Important Topics

Maria Skeppstedt[1,2], **Kostiantyn Kucher**[1], **Manfred Stede**[2], **Andreas Kerren**[1]

[1]Department of Computer Science and Media Technology, Linnaeus University, Växjö, Sweden
{maria.skeppstedt,kostiantyn.kucher,andreas.kerren}@lnu.se
[2]Applied Computational Linguistics, University of Potsdam, Potsdam, Germany
stede@uni-potsdam.de

Abstract

While the task of manually extracting arguments from large collections of opinionated text is an intractable one, a tool for computer-assisted extraction can (i) select a subset of the text collection that contains re-occurring arguments to minimise the amount of text that the human coder has to read, and (ii) present the selected texts in a way that facilitates manual coding of arguments. We propose a tool called Topics2Themes that uses topic modelling to extract important topics, as well as the terms and texts most closely associated with each topic. We also provide a graphical user interface for manual argument coding, in which the user can search for arguments in the texts selected, create a theme for each type of argument detected and connect it to the texts in which it is found. Topics, terms, texts and themes are displayed as elements in four separate lists, and associations between the elements are visualised through connecting links. It is also possible to focus on one particular element through the sorting functionality provided, which can be used to facilitate the argument coding and gain an overview and understanding of the arguments found in the texts.

Keywords: Argument extraction, topic modelling, text analysis, argument visualisation, stance visualisation, text visualisation, information visualisation, interaction

1. Introduction

The large amount of opinionated text that is constantly being produced online might help us to better understand why certain opinions are held. For instance, opinions related to consumer behaviour, health decisions or to political and religious radicalisation. These online texts often include arguments to why a particular stance is taken. By extracting these arguments, new insights into reasons and motives for taking this stance might be gained.

It is an intractable task to manually extract arguments from the vast amount of opinionated text that is being produced, e.g., online text in the form of tweets or discussion forum posts, or text in the form of free text answers to survey questions. Fully automatic argument extraction, on the other hand, has been shown difficult, e.g., by Boltužić and Šnajder (2015). Computer-*assisted* methods for argument coding might, however, be a feasible option. For instance, methods that (i) automatically extract the parts of the text collection that contain re-occurring arguments, to minimise the amount of text that the human coder has to read, and (ii) visualise the automatically extracted information in a way that facilitates manual coding.

We here present Topics2Themes, a visual analysis tool for computer-assisted coding of arguments in collections of short, opinionated texts, e.g., online texts.

Topics2Themes is based on previous related research on qualitative text analysis and argument extraction, which will be described more closely in Section 2. This research has shown that coding a subset of a text collection, selected by topic modelling, produces results similar to those obtained by coding the entire text collection (Baumer et al., 2017). There is also previous research on argument extraction that has shown topic modelling to be suitable for semi-automatic extraction of arguments from opinionated texts (Sobhani et al., 2015).

Building on results from these previous studies, we chose topic modelling as the method for selecting which subset of texts in a document collection to manually code. This was implemented through (i) a back-end that uses topic modelling to automatically extract important topics, as well as the terms and texts with which each topic is most closely associated, and (ii) a front-end that presents these texts for manual coding of arguments and visualises associations between topics, terms, texts, and manually coded arguments. This functionality is described in Sections 3.1 and 3.2, respectively.

The design of the graphical user interface is based on previous visualisation research, as well as on an evaluation of the back-end functionality, as described in Section 3.3. Sections 4 and 5 provide a comparison to previous visualisations and suggestions for future extensions of Topics2Themes. Finally, Section 6 concludes this paper.

2. Background

The main functionality of Topics2Themes is based on previous research on qualitative text analysis, argument extraction, and stance detection.

2.1. Qualitative Text Analysis

Among the large body of research on qualitative text analysis (Myers, 2009, pp. 163–180), we here focus on one particular study, conducted by Baumer et al. (2017), on which the main ideas for Topics2Themes are based.

Baumer et al. (2017) used free-text survey responses for performing two totally independent analyses: (i) a grounded theory-based study and (ii) data analysis based on the output of topic modelling. When comparing the output from the topic modelling and the grounded theory analysis, the authors found that “The topic modeling results captured to a surprising degree many of the themes identified in grounded theory, and vice versa." Each topic produced by topic modelling was, however, often aligned with several grounded theory themes, and each grounded theory theme was typically aligned with several topics produced by topic modelling.

The algorithm used for topic modelling was Latent Dirichlet Allocation (LDA). In addition to the input in the form of a collection of text documents, this algorithm also requires an input parameter in the form of the number of topics that are to be identified in the collection. Ten topics were here requested from the algorithm. The output given from the LDA algorithm is (i) a set of terms from the collection that represents each identified topic, and (ii) a ranking of the text documents according to the probability that an identified topic is present in the document.

As LDA produces different results depending on what random number is used for the initialisation, the LDA algorithm was run 10 times with different initialisations, to verify the consistency of the produced topics.[1]

The results of the topic modelling were presented by showing the 25 most representative terms and the 50 most representative survey texts for each topic. A manual analysis was then performed of this output, through assigning high-level descriptors for each topic. One researcher had to allocate a few hours over two days to perform the topic modelling-based analysis. The analysis based on grounded theory, in contrast, took two researchers several hours of work per week over about two and a half months.

Topics2Themes includes functionality for supporting the procedure of topic modelling-based text analysis that is described by Baumer et al. (2017). The aim is, however, to construct a more generally applicable tool, which can be applied for text analysis of any collection of short texts. The tool should, in addition, provide a graphical user interface with which the results of the topic modelling can be analysed, and which does not require any knowledge of, e.g., programming or topic modelling.

2.2. Argument Extraction

Another addition to the functionality described by Baumer et al. (2017) is that Topics2Themes is mainly meant to be used on opinionated texts, with the aim of extracting arguments.

Topics2Themes can be applied on a text collection for which there is no previous knowledge of what arguments are present. In contrast, most previous studies on argument extraction assume that a set of pre-defined arguments are known, and take on the task of detecting in which debate posts these arguments are used. That is, the argument extraction task is modelled as a standard text classification task. F-scores for the task that range from 0.5 to 0.8 have been reported (Hasan and Ng, 2014; Boltužić and Šnajder, 2014; Sobhani et al., 2015). The results are, however, difficult to compare, since the granularity of the argument categories varies between the different studies.

In the study by Sobhani et al. (2015), topic modelling was applied for carrying out the classification task. Topic modelling was applied to unlabelled data and the extracted topics were then manually mapped to eight pre-defined arguments. The unlabelled data was, thereafter, clustered based on the extracted topics, i.e., a post was assigned to a topic cluster if its probability of containing that topic was above a certain threshold. When the topic modelling approach Non-Negative Matrix Factorization (NMF) was used, an F-score of 0.5 was achieved, which was six percentage points better than the supervised baseline classifier. In contrast, the use of LDA-based topic modelling gave very low results. The authors attribute the difference in results between the two topic modelling approaches to that LDA is better adapted to longer texts than the short discussion posts that were used in the study.

The work by Boltužić and Šnajder (2014) has been extended by performing a hierarchical clustering of argumentative sentences, based on text similarity measured by bag-of-word features and by word embedding features (Boltužić and Šnajder, 2015). The aim of this clustering was to group similar arguments in an unsupervised fashion and thereby automatically come up with the set of arguments that were assumed as pre-defined in the other argument extraction studies. Results achieved for this approach were low, and the authors note that computer-assisted argument extraction, i.e., what we aim for in this study, might be more feasible than a fully automatic extraction.

2.3. Stance Detection

Stance detection is related to the task of argument extraction, but instead of extracting arguments, it is detected which stance is taken. The stance detection task is thus also modelled as a text classification task, typically with the aim of determining whether a text expresses a stance *for*, *against* or whether it is *neutral/undecided* towards a pre-defined proposition or target (Mohammad et al., 2017). Stance classifiers have been trained for detecting stance towards a number of different targets, and on different text genres, including Internet discussion forums (Walker et al., 2012; Hasan and Ng, 2013) and tweets (Mohammad et al., 2017).

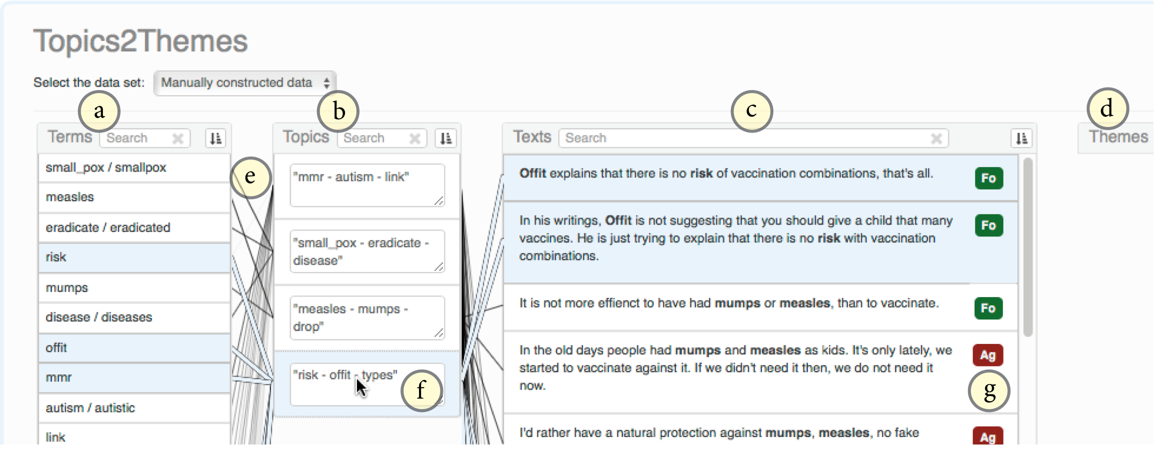
As the knowledge of which stance is taken in a text might be useful when performing text analysis for finding arguments, Topic2Themes includes functionality for importing stance information, i.e., each text in the text collection can have a tag attached to it that states if the text is *for*, *against* or *undecided* towards the stance target of interest.

This tagging could either be done automatically by a stance classifier, or manually by annotators. In the case of the texts fed into the tool being free text answers to surveys, the stance tagging could be provided by supplementary closed-ended questions given to the respondents. In the case of two-sided debate texts from online debate forums, stance information is typically provided as meta-data for the debate posts.

3. Functionality of Topics2Themes

Topics2Themes consists of two main parts, (i) the back-end which produces an automatic analysis of a text collection with the help of topic modelling, and (ii) the front-end

[1] Nine topics appeared in all runs. Among them were seven retained, as one of the topics consisted of terms in a non-English language and another was a meta-topic about the survey.



Offit explains that there is no **risk** of vaccination combinations, that's all.

In his writings, **Offit** is not suggesting that you should give a child that many vaccines. He is just trying to explain that there is no **risk** with vaccination combinations.

It is not more effienct to have had **mumps** or **measles**, than to vaccinate.

In the old days people had **mumps** and **measles** as kids. It's only lately, we started to vaccinate against it. If we didn't need it then, we do not need it now.

I'd rather have a natural protection against **mumps**, **measles**, no fake

Figure 1: The initial state of the interface, when the topic modelling has been carried out for a text collection, but before any manual coding of themes has been performed by the analyst. The interface includes the following components: (a–d) the panels containing lists of terms, topics, document texts, and user-created themes, respectively; (e) links between the related elements of the respective lists (e.g., terms belonging to a topic); (f) a topic highlighted by the user by hovering; and (g) a stance symbol assigned to the corresponding text.

which consists of the graphical user interface, where the results are presented and where the analysis is carried out.

3.1. Back-end

Functionality to be able to perform the procedure described by Baumer et al. (2017) was implemented in the back-end, but the exact parameters were made user-configurable. That is, the following parameters were made configurable: (i) the maximum number of topics that are to be identified in the text collection, (ii) the maximum number of salient terms to include for each identified topic, (iii) the maximum number of text documents to associate with each topic, (iv) the number of times to re-run the topic modelling algorithm to make sure that the extracted topics are stable, and (v) the amount of overlap between the returned term sets of the different re-runs for a topic to be considered stable.

Some configurable additions to the procedure described by Baumer et al. (2017) were also provided. Since Sobhani et al. (2015) showed that the NMF algorithm is better suited for topic modelling-based argument extraction of short discussion posts than LDA, it was also made configurable whether LDA or NMF is to be used.

In addition, the option to include an English-specific text pre-processing was also provided. This pre-processing consists of (i) a concatenation of collocations that occur frequently in the text collection into one term, and (ii) a replacement of term instantiations of the same concept (morphological variations, synonyms and related terms) with a string that represents the concept. The latter was achieved by applying clustering of word embedding vectors associated with the terms and assigning terms that belong to the same cluster to a joint concept. DBSCAN clustering (Ester et al., 1996) was used, and the maximum distance between two vectors for them to be considered as belonging to the same cluster was also made user-configurable. Examples of the results of the pre-processing is shown in the terms panel in Figure 1(a), where a underscore indicates collocation, and a slash indicates different term instantiations of the same concept.

A standard stop word list is used to remove stop words when constructing the topic models. This list can, however, be extended by the user with domain-specific stop words. In addition, it is possible to configure the automatic removal of frequently or infrequently occurring terms. There is also a user-constructed exception list with terms that are not to be included among the automatically constructed concept clusters. The user can thereby ensure a high quality of the clusters that are used by inspecting a register of automatically constructed clusters and adding terms that have been incorrectly associated with a cluster to the exception list.

Similar to what is described by Baumer et al. (2017), the input to the tool is a collection of text documents. To adapt to the genre of opinionated text, it is also possible to provide a pre-tagging of each text with any of the three stance categories *for*, *against*, or *undecided*.

The back-end was implemented as a RESTful API using the Flask web development framework for Python. The implementations of DBSCAN, text to vector transformation and the two topic modelling algorithms[2] that are available in Scikit-learn were used (Pedregosa et al., 2011). The word embedding vectors were accessed through the Gensim library (Řehůřek and Sojka, 2010), and an out-of-the-box word2vec model[3] trained on Google news was used.

3.2. Front-end

The presentation and user interaction is to support the procedure of qualitative analysis based on topic modelling that is described by Baumer et al. (2017). The tool is mainly aimed to be used on larger collections of short, opinionated texts for extracting arguments. For each of the arguments extracted, it should also be indicated to which stance category the argument is associated. The front-end was im-

[2] Partly following suggestions by Bakharia (2016).

[3] `code.google.com/archive/p/word2vec/`

plemented as a web application with D3[4] to make it easily accessible for the end users.

To provide an example of the functionality of the front-end, we manually authored 50 short texts that were similar in content to discussion posts from British online debates on vaccination, and applied NMF to extract four topics. The example user scenario could thus be that a researcher who studies vaccine hesitancy would like to know what reoccurring arguments *for* and *against* vaccination that are used in online discussions.

We have identified three main user tasks. The user is first to gain an *overview* of what has been extracted by the automatic topic model, thereafter the user is to *analyse the texts* by manually extracting arguments, and finally, the user is to *explore the arguments* that have been extracted.

3.2.1. An Overview of the Topic Modelling Results

After the topic modelling has been carried out, an overview of the model output is provided. Figure 1 shows this initial view which is presented to the user before any manual coding has been carried out. The first panel shows the salient terms, the second one shows the extracted topics, and the third shows the texts (see Figure 1(a–c)). The terms and texts are sorted according to their summed salience for the extracted topics, and could therefore be described as giving an indication of what is generally important in the text collection. The associations between terms/topics/texts and the strengths of the associations are shown through connecting links with different widths, as displayed in Figure 1(e). The figure only shows the top-ranked terms and texts, but lower-ranked terms/texts can be reached in the tool by scrolling down.

As described by Baumer et al. (2017), the set of salient topic terms is not enough to determine the content of a topic. The analyst also needs to be provided with typical text examples. The interface, therefore, gives equal importance to presenting the texts that are associated with a topic as to presenting the terms with which it is associated. When the user lets the mouse hover over a term/topic/text element (for instance, the "risk—offit—types" topic element in Figure 1(f)), its associated elements within the other two categories are highlighted, which makes it possible for the user to explore connections between the three categories. The user can also select any element by mouse click. This has the effect that the elements that belong to the other categories and that are associated with the selected one are sorted as the top-ranked elements within their respective panels. In Figure 2(a), the user had previously clicked on the topic element that is named "smallpox—eradicate—disease", which has had the effect that the associated elements in the other panels are shown on top.

The number of elements in the topics panel shows the user how many topics have been extracted by the topic modelling. Each topic is given a default name that is made up of the three terms with which the topic is most closely associated. The name can, however, be changed by the user to one that better describes the topic.

The "stance symbol" in the right upper corner of a text element indicates the stance category of the text: for instance, in Figure 1(g) the label "Ag" with red background represents the *against* stance. By scrolling through the texts that are connected to a topic, an overview of its associated stances is achieved.

3.2.2. Extract Arguments from Texts

Baumer et al. (2017) found that grounded theory-based themes and topic modelling-based topics did not correspond one-to-one, but with the relation many-to-many. To be able to follow this procedure, the user must be able to define additional categories to the ones automatically extracted by the topic model. A functionality to add an additional category of elements, which are user-defined, is therefore included in Topics2Themes. When referring to these categories that the user creates as elements in the right-most panel, we adhere to the grounded theory-inspired vocabulary used by Baumer et al. (2017) and call these categories "themes". However, since the main purpose of the tool is to extract arguments, these manually extracted themes typically correspond to arguments detected in the texts. Figure 2 shows the tool after the manual user coding has started and themes have also been added. With the "+" button on the themes panel displayed in Figure 2(b), the user can create a new theme. The theme can then be given a description, and texts can be associated with it by drag-and-drop of a text element onto a theme element, as shown in Figure 2(c–d).

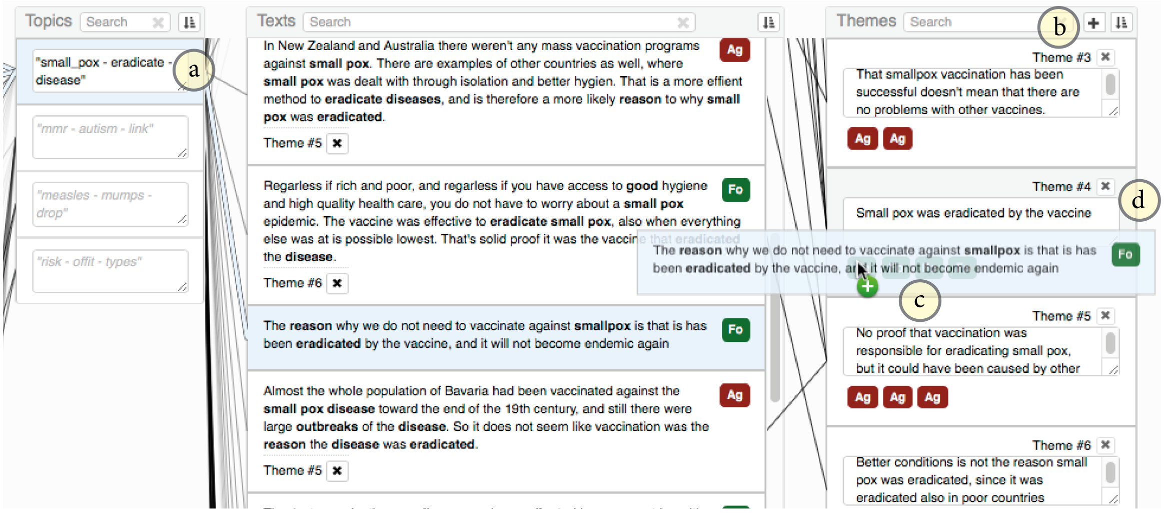
In the typical use case, the user extracts arguments from each one of the topics in turn. In the example of Figure 2(a), the user analyses texts that are associated with the topic "smallpox—eradicate—disease", and this topic element is selected. The procedure of detecting arguments is then to analyse each of the texts that are associated with the topic. Terms that are associated with a topic are written in a bold-faced font, which makes them stand out from the rest of the text and which facilitates the analysis (see Figure 2(c)). If an argument is found in a text, the user has two choices: (i) if it is an argument that has previously occurred in the analysis, the text should be assigned to the matching theme that contains this argument, or (ii) if the argument is new, a new theme should be created for this argument, and the user should assign the text to this new theme.

3.2.3. Explore the Arguments

The final task is to explore the arguments that have been created. This task needs to be carried out during the analysis in order to find out whether a text contains an argument that has previously been created. It can also be carried out when the analysis is finished to gain an overview and understanding of the arguments found in the text collection.

Figure 3 shows when the user has selected a theme in order to investigate the argument for which this theme was created. The terms, topics and texts that are associated with this theme are then sorted as the most high-ranked elements in their respective panels. The descriptive text of the theme, as well as the information of with which topic(s) and term(s) it is associated, gives a high-level understanding of the theme. Reading the texts with which the theme is connected, on the other hand, gives a deeper understanding of the theme.

[4] d3js.org



In New Zealand and Australia there weren't any mass vaccination programs against **small pox**. There are examples of other countries as well, where **small pox** was dealt with through isolation and better hygien. That is a more effient method to **eradicate diseases**, and is therefore a more likely **reason** to why **small pox** was **eradicated**.

Theme #5 ✖

Regarless if rich and poor, and regarless if you have access to **good** hygiene and high quality health care, you do not have to worry about a **small pox** epidemic. The vaccine was effective to **eradicate small pox**, also when everything else was at is possible lowest. That's solid proof it was the vaccine that **eradicated** the **disease**.

Theme #6 ✖

The **reason** why we do not need to vaccinate against **smallpox** is that is has been **eradicated** by the vaccine, and it will not become endemic again

Almost the whole population of Bavaria had been vaccinated against the **small pox disease** toward the end of the 19th century, and still there were large **outbreaks** of the **disease**. So it does not seem like vaccination was the **reason** the **disease** was **eradicated**.

Theme #5 ✖

Theme #3 ✖

That smallpox vaccination has been successful doesn't mean that there are no problems with other vaccines.

Theme #4 ✖

Small pox was eradicated by the vaccine

The **reason** why we do not need to vaccinate against **smallpox** is that is has been **eradicated** by the vaccine, and it will not become endemic again

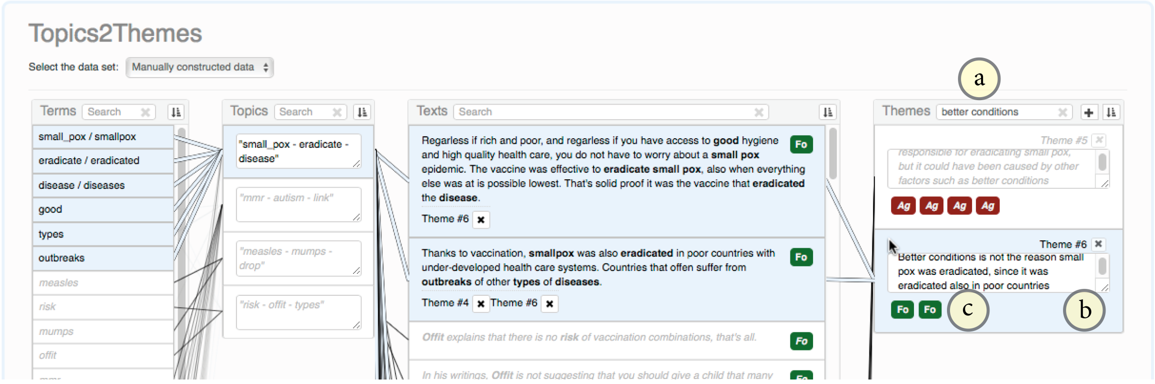
Theme #5 ✖

No proof that vaccination was responsible for eradicating small pox, but it could have been caused by other

Theme #6 ✖

Better conditions is not the reason small pox was eradicated, since it was eradicated also in poor countries

Figure 2: The user interface during the analytical session: (a) the user has focused on a specific topic by clicking; (b) the user has then created several themes by using the button in the themes panel; (c) the user is dragging a document text element to a theme element to create an association; (d) the target of the drag-and-drop operation is theme #4.



Regarless if rich and poor, and regarless if you have access to **good** hygiene and high quality health care, you do not have to worry about a **small pox** epidemic. The vaccine was effective to **eradicate small pox**, also when everything else was at is possible lowest. That's solid proof it was the vaccine that **eradicated** the **disease**.

Theme #6 ✖

Thanks to vaccination, **smallpox** was also **eradicated** in poor countries with under-developed health care systems. Countries that offen suffer from **outbreaks** of other **types** of **diseases**.

Theme #4 ✖ Theme #6 ✖

Offit explains that there is no risk of vaccination combinations, that's all.

In his writings, Offit is not suggesting that you should give a child that many

Theme #5 ✖

responsible for eradicating small pox, but it could have been caused by other factors such as better conditions

Theme #6 ✖

Better conditions is not the reason small pox was eradicated, since it was eradicated also in poor countries

Figure 3: Exploring arguments: (a) Certain themes are searched for; (b) theme #6 is selected by clicking to show with which terms, topics and texts it is associated; (c) stance symbols indicate stances taken in the texts connected to a theme.

For each of the texts that are associated with a theme, the stance symbol of the text is displayed at the bottom of the theme element, as shown in Figure 2(e). These symbols have two functions: (i) as indicators of to which stance category the argument is typically associated, and (ii) to show how frequently the argument occurs in the texts that have been analysed.

3.3. Design Decisions for the Front-end

The main inspiration for the front-end presentation of Topics2Themes was the List View visualisation of the Jigsaw system (Stasko et al., 2008). We also carried out a manual evaluation of the functionality of the back-end part of Topics2Themes in order to gain further inspiration for how the front-end was to be designed.

3.3.1. The Jigsaw System

Jigsaw aims at helping analysts to search, review, and understand the content of a text collection, e.g., a collection of case reports in a police investigation. The main functionality of Jigsaw consists of an interactive visualisation that focuses on identifying and highlighting connections between entities present in the documents. Typical entities are people, places and dates that are automatically detected by a named entity recognition system. A connection between two entities means that they co-occur in the same document. The List View displays these automatically extracted entities in the form of vertical lists. Connections between different entities are displayed by lines that connect the list elements and by highlighting of the elements. Jigsaw also provides several ways of sorting the lists.

One of the main cues for exploring the output of a topic modelling algorithm, as well as for exploring the result of the text coding, is to explore associations, i.e., connections between terms, topics, texts and the themes created. Since the Jigsaw List View is focused on displaying connections, we assess that a similar functionality is suitable for visualising the results of the topic modelling and the manual coding. Instead of using the List View design for displaying entities, we thus display terms, topics and themes, as well as the actual texts, and how they are connected.

3.3.2. Evaluation of the Back-end Functionality

In order to come up with further ideas of how to design the user interface, we first implemented the back-end functionality of Topics2Themes. We thereafter used the topic modelling output of the back-end to perform a manual coding, i.e., a coding without the help of a graphical user interface, according to the procedure described by Baumer et al. (2017). As data, we used a previously compiled resource with debate posts from the British online debate forum Mumsnet[5] (Skeppstedt et al., 2017). The resource consists of 1,190 debate posts from six discussion threads on the topic of vaccination, where the posts are manually classified as expressing a stance *for* or *against* vaccination or as being *undecided*. The back-end was configured to use NMF to extract ten topics, with a required overlap of 70% between ten re-runs of the algorithm in order for a topic to be considered stable. This resulted in six stable topics. The 50 most typical texts, for each one of the six topics identified by the topic model, were coded, one topic at a time. Thus, a set of re-occurring arguments was identified for each one of the topics (38, 23, 51, 61, 40, and 33 arguments, respectively). There was a large semantic coherence between the texts that were selected for analysis for a topic, and there were only occasional occurrences of arguments that were identified as associated to more than one topic (Skeppstedt et al., 2018).

The main difficulty of the manual analysis was the cognitive load of remembering which arguments had been previously identified. To be able to go through a set of semantically coherent texts, i.e., those that belonged to the same topic, limited the set of possible arguments to look for, and, thereby, also led to a decreased cognitive load.

It can thus be observed that the use of topic modelling for text selection and sorting has two main benefits. First, only a subset of the document collection has to be read in order to find re-occurring arguments, which makes it possible to analyse large text collections also when the time available is limited. Second, the possibility to focus on one topic at a time facilitates the analysis, as the user only has to remember the limited set of arguments that have previously been extracted for the topic that is currently being analysed.

The cognitive load of remembering previously defined arguments was, however, still too large, despite the support from the topic modelling. With the themes panel in Topics2Themes, we therefore aim to further decrease this cognitive load. The extracted arguments/themes are not only listed with a description, but it is also easy to use the interface to explore the arguments that have been extracted previously. These arguments can, for instance, be sorted according to whether they are associated with other texts that contain certain terms, e.g., the terms of the text that is currently being analysed. In addition, in order not to add to the cognitive load required for the task of extracting and remembering arguments, we aimed to construct a clean interface with a minimum of distractions.

From the explanation given by Baumer et al. (2017), we had interpreted the coding task as a task of associating topics to a corresponding theme. Although the authors emphasise the importance of reading the texts for finding themes, our draft interface provided the functionality of associating a topic to a theme. The manual coding showed, however, that although the topics are important for sorting and selecting texts and for giving a topical focus for the analysis, the user does not associate topics directly with themes. Instead, when analysing the texts, the coder associates the *text* to a theme. In Topics2Themes, the task of the user is, therefore, to associate a text to a theme.

The design choice of associating texts to themes also has the advantage of providing a better traceability of the analysis performed. It is still easy to obtain high-level information in the form of which themes are associated with a particular topic, i.e., a topic can be selected and all associated elements in the themes list can be examined. If the analyst instead would like to read the original text in order to trace the reasons to why a certain theme has been identified, the current design enables the texts, from which the theme originated, to be easily identified.

4. Comparison to Previous Visualisations

There are a number of examples of related visualisation tools, but to the best of our knowledge, there is no previous tool that offers the support for computer-assisted argument extraction and the same type of overview of associations between elements that is provided by Topics2Themes.

The output of topic models has previously been visualised and made available for user interaction. The Termite tool (Chuang et al., 2012), for instance, aims at enabling quality assessments and improvements of produced topic models. Salient terms and topics are shown in a grid that indicates which terms belong to which topics. The Serendip tool, in contrast, has exploration of texts as the main aim, and the grid view design is used there instead for displaying topics and their connected documents, which can be sorted according to a number of different criteria (Alexander et al., 2014). Visualisations for showing alignment between two sets of topics (Chuang et al., 2013), as well as non-interactive visualisations (Tangherlini et al., 2016), are other examples of topic model visualisations.

The focus of these previous approaches is, however, not to use the output of the topic models as support for text coding. Therefore, the functionality of Topics2Themes that supports computer-assisted argument extraction is not available. For instance, the functionality of adding new categories in the form of themes, displaying stance, or the functionality of exploring data through selecting an element to show with which other elements it is associated.

There are also different types of visualisations of stance taking in text, as well as of the related concept of sentiment in text. For instance, visualisations to show changes in sentiment over time, the text elements that have resulted in the stance/sentiment classifications carried out by the underlying natural language processing tools, or the topics/targets towards which the sentiment is directed (Kucher et al., 2018). There is also work on the use of topic modelling for extraction of topics from text, as well as visualisations of sentiment towards these extracted topics (Hoque and Carenini, 2014). The aim of extracting topics in these previous studies is, however, very different from the aim

[5] `www.mumsnet.com/Talk`

of the topic modelling in Topics2Themes. These previous studies extract topics to find out towards what the stance or sentiment is directed, whereas Topics2Themes is to be applied on texts where stance related to a pre-defined issue is expressed. The aim of topic modelling for Topics2Themes is instead to facilitate the process of argument extraction from these texts.

There are also previous tools that are specifically aimed at visualising argumentative text. A tool constructed by Wyner et al. (2015) helps an analyst to extract important arguments by highlighting information in the text that might be relevant for this task. This information includes named entities, terms important for the domain of the text and for expressing sentiment, as well as terms belonging to topics derived from LDA-based topic modelling. In contrast to Topics2Themes, the tool by Wyner et al. (2015) is not focused on analysing the text on the basis of extracted topics. Therefore, the tool does not include the functionality of ranking the text sections according to their relevance for the extracted topics. The potential of saving time in the analysis through only reading the most important parts of the text, as in the process described by Baumer et al. (2017), is, thereby, not achieved.

The VisArgue framework is another example of visualisation for argumentative text (El-Assady et al., 2016). VisArgue uses topic modelling (among other techniques) for visualising arguing patterns. In contrast to Topics2Themes that aims at facilitating extraction of re-occurring arguments, VisArgue aims at providing an overview of an entire debate or of an entire single argumentative document with respect to topic changes throughout the document and in the course of the debate. There are also other aspects of the argumentative genre that can be visualised, and that are not included in the aims for Topic2Themes, for instance, the quality of the argumentation (Gold et al., 2017).

5. Future Directions

The next step to generate new ideas for improvement is by extensively using the front-end part of Topics2Themes for text coding. After the initial use of the tool, we already have a number of ideas for possible extensions.

A subset of a text collection might consist of longer texts, even in a text genre that mainly contains shorter texts, e.g., the genre of discussion forum posts. These texts are not suitable to show “as-is” in the text panel of Topics2Themes, since this panel is meant to give an easily scrollable overview of several texts from the collection. We therefore aim to instead show summaries of the longer texts when the user scrolls through the text panel. The user should also be provided with the option to associate a theme to a substring of a longer text. This would give the user the possibility to trace the origin of an extracted argument to an exact text snippet in a longer text.

Another important functionality to add is the possibility to set topic modelling parameters through the user interface. These parameters are currently changed through text-based property files, which is not optimal from a usability perspective. In addition, the functionality for visualising the connections between terms and their corresponding topics and texts might be useful for determining some of the topic modelling parameters, e.g., whether a term should be included in the stop word list. The visualisation of term associations could also be useful for determining which terms to exclude from the automatically constructed concept clusters. The user interface should, therefore, also be extended by functionality for fine-tuning of the concept clusters.

The assignment of themes might be further facilitated by automatically sorting previously created themes according to the likelihood of them being associated with the text that is to be analysed. For a typical use case, only a small number of texts are manually associated with each theme, which makes it difficult to train a high-precision machine learning classifier to automatically associate texts to themes. To train a classifier to rank themes might, in contrast, be feasible even with a small training data set.

Finally, since Topics2Themes is meant to be applied on online texts, it could also be relevant to add newly produced texts to an existing text collection. Support for including newly detected topics into an existing analysis must then be added, as well as support for visualising changes in the prevalence of different topics over time.

6. Conclusion

We here presented Topics2Themes[6], a tool for carrying out computer-assisted coding of arguments in opinionated text. The tool is meant to be applied on larger text collections consisting of short texts, for instance, online texts in the form of tweets or posts from Internet discussion forums.

Topics2Themes uses topic modelling to automatically extract frequently occurring topics in the text collection. A subset of the collection, formed by texts most likely to discuss the extracted topics, is presented for manual coding. The coding of large text collections using limited manual resources is thereby made possible.

The coding is further facilitated by the graphical user interface provided by Topics2Themes. The user is able to create theme elements for each detected argument type, and to associate these elements to the texts in which the arguments occur. Associations between the automatically extracted elements and the manually coded elements are visualised, i.e., associations between the topics, terms representing the topics, texts and themes. Thereby, an overview of the text collection content is provided, as well as of the arguments that are used in the text collection.

7. Acknowledgements

We would like to thank the reviewers for their valuable input. We would also like to thank the Swedish Research Council (Vetenskapsrådet) that funded this study, mainly through the project “Navigating in streams of opinions: Extracting and visualising arguments in opinionated texts” (No. 2016-06681) and partly through the StaViCTA project, framework grant “the Digitized Society – Past, Present, and Future” (No. 2012-5659).

8. Bibliographical References

Alexander, E., Kohlmann, J., Valenza, R., Witmore, M., and Gleicher, M. (2014). Serendip: Topic model-driven

[6] A supplementary video demo of Topics2Themes is available at `https://vimeo.com/257474950`

visual exploration of text corpora. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*, VAST '14, pages 173–182. IEEE, October.

Bakharia, A. (2016). Topic modeling with Scikit Learn. https://medium.com/@aneesha/topic-modeling-with-scikit-learn-e80d33668730 (Accessed January 10, 2018), September.

Baumer, E. P. S., Mimno, D., Guha, S., Quan, E., and Gay, G. K. (2017). Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology*, 68(6):1397–1410, June.

Boltužić, F. and Šnajder, J. (2014). Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58, Stroudsburg, PA, USA, June. ACL.

Boltužić, F. and Šnajder, J. (2015). Identifying prominent arguments in online debates using semantic textual similarity. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 110–115, Stroudsburg, PA, USA, June. ACL.

Chuang, J., Manning, C. D., and Heer, J. (2012). Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, AVI '12, pages 74–77. ACM.

Chuang, J., Hu, Y., Jin, A., Wilkerson, J. D., McFarland, D. A., Manning, C. D., and Heer, J. (2013). Document exploration with topic modeling: Designing interactive visualizations to support effective analysis workflows. In *NIPS Workshop on Topic Models: Computation, Application and Evaluation*.

El-Assady, M., Gold, V., Hautli-Janisz, A., Jentner, W., Butt, M., Holzinger, K., and Keim, D. A. (2016). VisArgue: A visual text analytics framework for the study of deliberative communication. In *Proceedings of the International Conference on the Advances in Computational Analysis of Political Text*, PolText 2016, pages 31–36. University of Zagreb.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, KDD '96, pages 226–231, Palo Alto, California, USA. AAAI Press.

Gold, V., El-Assady, M., Hautli-Janisz, A., Bögel, T., Rohrdantz, C., Butt, M., Holzinger, K., and Keim, D. (2017). Visual linguistic analysis of political discussions: Measuring deliberative quality. *Digital Scholarship in the Humanities*, 32(1):141–158, April.

Hasan, K. S. and Ng, V. (2013). Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the International Joint Conference on Natural Language Processing*, IJCNLP '13, pages 1348–1356, Stroudsburg, PA, USA. ACL.

Hasan, K. S. and Ng, V. (2014). Why are you taking this stance? Identifying and classifying reasons in ideological debates. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Stroudsburg, PA, USA, October. ACL.

Hoque, E. and Carenini, G. (2014). ConVis: A visual text analytic system for exploring blog conversations. *Computer Graphics Forum*, 33(3):221–230, June.

Kucher, K., Paradis, C., and Kerren, A. (2018). The state of the art in sentiment visualization. *Computer Graphics Forum*, 37(1):71–96.

Mohammad, S., Sobhani, P., and Kiritchenko, S. (2017). Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):1–23, June.

Myers, M. D. (2009). *Qualitative research in business & management*. SAGE, London.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Řehůřek, R. and Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Paris, France, May. European Language Resources Association (ELRA).

Skeppstedt, M., Kerren, A., and Stede, M. (2017). Automatic detection of stance towards vaccination in online discussion forums. In *Proceedings of the International Workshop on Digital Disease Detection using Social Media*, pages 1–8, Stroudsburg, PA, USA, November. ACL.

Skeppstedt, M., Kerren, A., and Stede, M. (2018). Vaccine hesitancy in discussion forums: Computer-assisted argument mining with topic models. Accepted for publication at Medical Informatics Europe.

Sobhani, P., Inkpen, D., and Matwin, S. (2015). From argumentation mining to stance classification. In *Proceedings of the 2nd Workshop on Argumentation Mining*, Stroudsburg, PA, USA. ACL.

Stasko, J., Görg, C., and Liu, Z. (2008). Jigsaw: Supporting investigative analysis through interactive visualization. *Information Visualization*, 7(2):118–132.

Tangherlini, T. R., Roychowdhury, V., Glenn, B., Crespi, C. M., Bandari, R., Wadia, A., Falahi, M., Ebrahimzadeh, E., and Bastani, R. (2016). "Mommy blogs" and the vaccination exemption narrative: Results from a machine-learning approach for story aggregation on parenting social media sites. *JMIR Public Health Surveill*, 2(2):e166, November.

Walker, M. A., Anand, P., Abbott, R., and Grant, R. (2012). Stance classification using dialogic properties of persuasion. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 592–596, Stroudsburg, PA, USA. ACL.

Wyner, A., Peters, W., and Price, D. (2015). Argument discovery and extraction with the Argument Workbench. In *Proceedings of the 2nd Workshop on Argumentation Mining*, Stroudsburg, PA, USA. ACL.

Bridging the Gap between Data versus Technology Producers: An Interactive Visual Interface for Data Exploration

M. Abounabhan, F. K. Abu Salem, K. Antar, S. Elbassuoni, W. Hojeily, S. Najem
Computer Science Department, American University of Beirut, National Council for Scientific Research (CNRS)
P. O. Box 11-0236, Riad El Solh, Beirut 1107 2020, Lebanon, Riad al Solh, Beirut 1107 2260, Lebanon
mea32@mail.aub.edu,fa21@aub.edu.lb,kha13@mail.aub.edu,se58@aub.edu.lb,weh01@mail.aub.edu, snajem@cnrs.edu.lb

Abstract

With the accumulation of large amounts of vital information surrounding the refugee communities in Lebanon, the current operational apparatus is facing unprecedented challenges in making sense of the deluge in data. In the majority of those instances the governmental sector in Lebanon is still largely lacking in the information systems expertise required to extract knowledge from this data in order to inform policy making. In this paper, we present an open, interactive visual interface to assist non-technical users in the exploratory data analysis of various types of data that are of vital significance. The intended purpose is to reduce the gap between producers of data and users of technology, and encourage evidence-based strategies for the refugee crisis management. In contrast to existing open visual platforms, our tool requires minimal technical expertise and effort, to the extent that not a single download or installation is needed. It also provides a user-friendly frontend and connects the user to data visualisation and analytics tools in the backend developed using Python. Our case studies reveal interesting revelations as a result of temporal and spatial phenomena, following the Syrian war and the influx of refugees into Lebanon.

Keywords: Visual Interface, Exploratory Data Analysis, Primary Care Health Data, Refugee Crisis

1. Introduction and Background

The status quo surrounding data in the Middle East and North Africa (MENA) region is that there is a general lack of studies and information around the current humanitarian crises following the widespread wars. In light of current events surrounding the Syrian crisis, there has been an uptake in data-driven research in the Middle East specifically as it is an issue of global concern. The MENA region is traditionally known to be a data-desert where not enough data is gathered, or data is not gathered in proper form, or at best, data is gathered in proper form but not efficiently used. The regional instabilities have exposed a large fraction of the population to food and nutrition insecurity in addition to undernutritions comorbidities, and numerous communicable and non-communicable diseases (Taleb et al., 2015; Hwalla et al., 2016). Thus, the internally displaced individuals and refugees have become a burden of the countries respective economies and infrastructures. The need for producing evidence-based strategies to face the emerging sociodemographic changes has become pressing (Breisinger et al., 2012). With the accumulation of large amounts of vital information surrounding the refugee communities in Lebanon, the current operational apparatus is facing unprecedented challenges in making sense of the deluge in data, mostly spread out across various uncentralised sources. In the majority of those instances, the governmental sectors are still largely lacking in the information systems expertise required to extract knowledge from this data to inform policy making, and there seems to be a need to develop scalable and efficient frameworks within reach of users of various levels of technical expertise (Koliopoulos et al., 2015).

Lebanon has still been reeling from the devastation incurred by its own civil war when the war in neighbouring Syria broke out. The operational staff in its ministries are stretched beyond limits and do not seem to be equipped to hop on the data analytics revolution overtaking the developing world. Concomitantly, the lack of a political will to invest in data-intensive research for managing the refugee crisis in Lebanon leads us to think that a minimalist, technically non-invasive platform is more likely to encourage owners of data to begin thinking along data-driven lines. The global burden of disease[1] is a massive initiative that provides a tool to demonstrate health loss from hundreds of diseases, injuries, and risk factors; however, this is not a platform which native owners of data can use at their whim, and personalise the way they see fit. RapidMiner (Fischer et al., 2002), Knime (Berthold et al., 2009), Weka (Hall et al., 2009) and Gephi (Bastian et al., 2009) are some of the leading platforms with the highest potential for scalable big data analytics. These visual tools allow code free mining of big data, but they still require non-trivial installations and a learning curve before one is able to use them. This may be a hindrance for people with extremely limited quantitative skills, who also have no access to technical support within their working environment.

In this paper, we present an open, interactive visual web interface to assist non-technical users in the exploratory data analysis of various types of data that are of vital significance. The intended purpose is to reduce the gap between producers of data and users of technology, and encourage evidence-based strategies for the refugee crisis management. In contrast to existing open visual platforms, our tool requires minimal technical expertise and effort, to the extent that not a single download or installation is needed. It also provides a user-friendly frontend and connects the user to data visualisation and analytics tools in the backend developed using Python. Our case studies reveal interesting revelations as a result of temporal and spatial phenomena, following the Syrian war and the influx of refugees

[1] http://www.healthdata.org/gbd

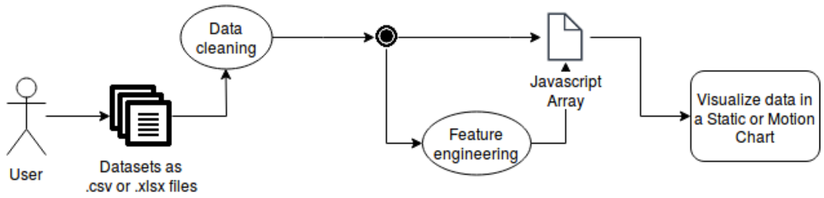

Data cleaning
Javascript Array
Feature engineering
Visualize data in a Static or Motion Chart
User
Datasets as .csv or .xlsx files

Figure 1: System Architecture

into Lebanon. These revelations become essential if one is to embark on predicting demand on the primary health care system, for example, assuming the crisis will persist. A plethora of work in the literature tackles geospatial and temporal visualizations of social data relevant to the MENA region (see for example, (Ghanem et al., 2014; Jänicke et al., 2013)). They rely on online textual and social media open data. Our work stands out by exploring publically available and private datasets that have been collected surrounding vital population statistics.

The rest of our paper is organised as follows. We first describe the design process and the software prototype information using a system's architecture. We also describe the data formats that our tool is able to process, as well as the anomalies in the data that are permissible and which our tool is able to clean up automatically. We further describe the pivotal datasets showcasing vital information relating to refugees in Lebanon, and some feature engeering we performed to make sense of the time series data representing demand on the primary health sector in Lebanon. Some of the datasets are publicly available – for example, UNHCR's persons of concern[2] and the global terrorism data [3]. One other dataset provided by the Ministry of Public Health in Lebanon cannot be made public. We finally present insights from the data that our interface is able to munge, highlighting the temporal and spatial dependencies of the observations found in the tackled datasets. Particularly, we are interested in the temporal effects of the Syrian refugee influx and its effect on the primary health care centers in Lebanon, as well as the spatial correlations existing between locations of major traumatic events and primary health care centers of high demand. Indeed, these basic manipulations that any novice user can perform using our web interface confirm the authors' viewpoint that any future forecasting of the demand on Lebanese infrastructure should be tackled as a spatiotemporal phenomenon. Readers can interact with our tool from here: `http://104.238.170.191/`.

2. The System

Figure 1 depicts the system architecture for our interface. Our system takes as input a dataset to be analysed, which typically consists of one or more comma-separated variable (csv) files or Excel (xls) files. The dataset is then passed through a data cleaning module, which detects anomalies in the data that are automatically eliminated. The cleaned data is then passed through a feature engineering module that extracts temporal and spatial features in the data, which can then be used to perform visual data analytics using various modes such as motion (dynamic) or static charts.

The frontend of our interface uses the Flask microframework to link the frontend to Python. Under the "Graph" option, "Browse" is a frontend option using JavaScript, and "Use Data" is a backend option using Python. Under the "Patient" option, "Browse" and "Data Sample" are both backend options using Python. "Static Visuals", seen in Figure 2, is also backend using Python. More on these options follow below as we describe our case studies.

We explain each module (component) in our system separately next.

2.1. Input Files

Our system assumes the input to be one or more csv (or xls) files. The columns in the files correspond to attributes such as dates, geographical locations, gender, age, etc. Each row corresponds to a data instance. Figure 3 shows a dummy snapshot representing what the Ministry of Public Health dataset looks like[4].

2.2. Data Cleaning

The data cleaning process accounts for anomalies such as null or missing values in numerical attributes. Currently, we perform two basic data cleaning operations to deal with such anomalies, depending on the type of attribute. For categorical attributes like gender or nationality, for example, we replace a null or missing value with a zero and for continuous attributes, we replace the missing value with the average value of all the categories. In future releases, we plan to utilise third-party data cleaning tools such as OpenRefine[5] or DataWrangler[6] to perform further cleaning operations such as data standardisation and duplicate elimination.

2.3. Feature Engineering

The data imported to the "Patients" page goes a step further past cleaning; it undergoes feature engineering process to generate additional features that are needed for spatiotemporal exploration. For instance, we use gazetteers to generate geo-coordinates of locations in the dataset that can be used to perform spatial analysis on the datasets as we explain in our second case study.

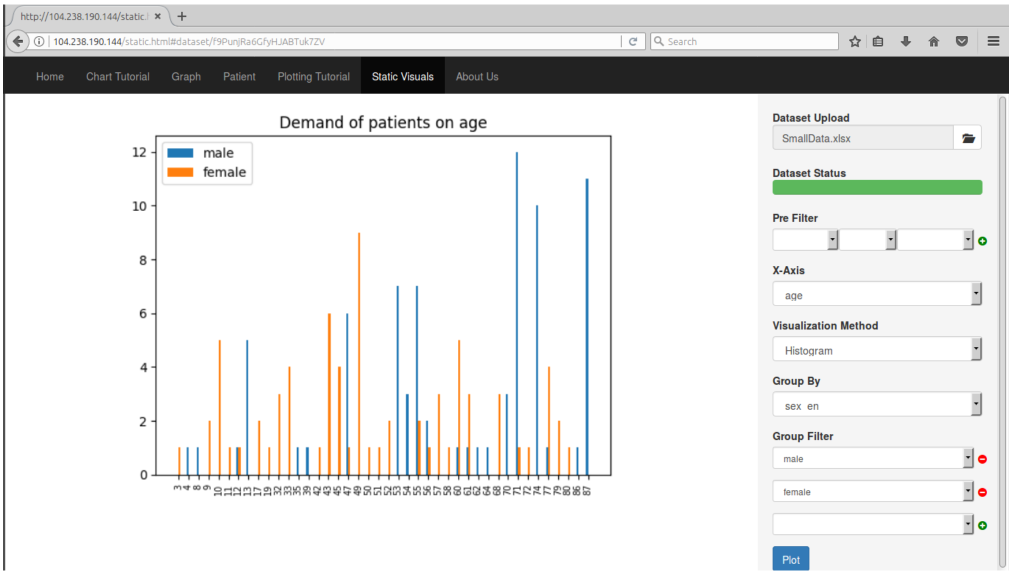
[2] Available from `popstats.unhcr.org/en/persons_of_concernpopstats.unhcr.org/en/persons_of_concern.csv`

[3] Available from `http://apps.start.umd.edu/gtd/downloads/dataset/globalterrorismdb_0616dist.xlsx`

[4] For privacy and data sharing agreements, this dataset cannot be revealed

[5] `http://openrefine.org/`

[6] `http://vis.stanford.edu/wrangler/`



Dataset Upload: SmallData.xlsx

Dataset Status

Pre Filter

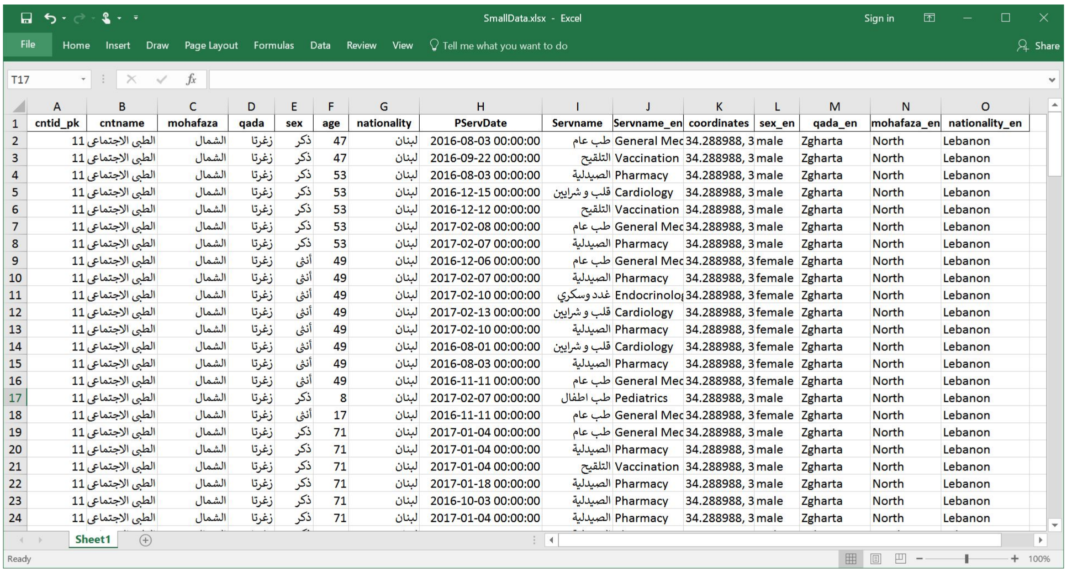
X-Axis: age

Visualization Method: Histogram

Group By: sex_en

Group Filter: male; female

Figure 2: Static Visuals page



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	cntid_pk	cntname	mohafaza	qada	sex	age	nationality	PServDate	Servname	Servname_en	coordinates	sex_en	qada_en	mohafaza_en	nationality_en
2	11	الطبى الاجتماعى	الشمال	زغرتا	ذكر	47	لبنان	2016-08-03 00:00:00	طب عام	General Mec	34.288988, 3	male	Zgharta	North	Lebanon
3	11	الطبى الاجتماعى	الشمال	زغرتا	ذكر	47	لبنان	2016-09-22 00:00:00	التلقيح	Vaccination	34.288988, 3	male	Zgharta	North	Lebanon
4	11	الطبى الاجتماعى	الشمال	زغرتا	ذكر	53	لبنان	2016-08-03 00:00:00	الصيدلية	Pharmacy	34.288988, 3	male	Zgharta	North	Lebanon
5	11	الطبى الاجتماعى	الشمال	زغرتا	ذكر	53	لبنان	2016-12-15 00:00:00	قلب و شرايين	Cardiology	34.288988, 3	male	Zgharta	North	Lebanon
6	11	الطبى الاجتماعى	الشمال	زغرتا	ذكر	53	لبنان	2016-12-12 00:00:00	التلقيح	Vaccination	34.288988, 3	male	Zgharta	North	Lebanon
7	11	الطبى الاجتماعى	الشمال	زغرتا	ذكر	53	لبنان	2017-02-08 00:00:00	طب عام	General Mec	34.288988, 3	male	Zgharta	North	Lebanon
8	11	الطبى الاجتماعى	الشمال	زغرتا	ذكر	53	لبنان	2017-02-07 00:00:00	الصيدلية	Pharmacy	34.288988, 3	male	Zgharta	North	Lebanon
9	11	الطبى الاجتماعى	الشمال	زغرتا	أنثى	49	لبنان	2016-12-06 00:00:00	طب عام	General Mec	34.288988, 3	female	Zgharta	North	Lebanon
10	11	الطبى الاجتماعى	الشمال	زغرتا	أنثى	49	لبنان	2017-02-07 00:00:00	الصيدلية	Pharmacy	34.288988, 3	female	Zgharta	North	Lebanon
11	11	الطبى الاجتماعى	الشمال	زغرتا	أنثى	49	لبنان	2017-02-10 00:00:00	غدد وسكري	Endocrinolo	34.288988, 3	female	Zgharta	North	Lebanon
12	11	الطبى الاجتماعى	الشمال	زغرتا	أنثى	49	لبنان	2017-02-13 00:00:00	قلب و شرايين	Cardiology	34.288988, 3	female	Zgharta	North	Lebanon
13	11	الطبى الاجتماعى	الشمال	زغرتا	أنثى	49	لبنان	2017-02-10 00:00:00	الصيدلية	Pharmacy	34.288988, 3	female	Zgharta	North	Lebanon
14	11	الطبى الاجتماعى	الشمال	زغرتا	أنثى	49	لبنان	2016-08-01 00:00:00	قلب و شرايين	Cardiology	34.288988, 3	female	Zgharta	North	Lebanon
15	11	الطبى الاجتماعى	الشمال	زغرتا	أنثى	49	لبنان	2016-08-03 00:00:00	الصيدلية	Pharmacy	34.288988, 3	female	Zgharta	North	Lebanon
16	11	الطبى الاجتماعى	الشمال	زغرتا	أنثى	49	لبنان	2016-11-11 00:00:00	طب عام	General Mec	34.288988, 3	female	Zgharta	North	Lebanon
17	11	الطبى الاجتماعى	الشمال	زغرتا	ذكر	8	لبنان	2017-02-07 00:00:00	طب اطفال	Pediatrics	34.288988, 3	male	Zgharta	North	Lebanon
18	11	الطبى الاجتماعى	الشمال	زغرتا	أنثى	17	لبنان	2016-11-11 00:00:00	طب عام	General Mec	34.288988, 3	female	Zgharta	North	Lebanon
19	11	الطبى الاجتماعى	الشمال	زغرتا	ذكر	71	لبنان	2017-01-04 00:00:00	طب عام	General Mec	34.288988, 3	male	Zgharta	North	Lebanon
20	11	الطبى الاجتماعى	الشمال	زغرتا	ذكر	71	لبنان	2017-01-04 00:00:00	الصيدلية	Pharmacy	34.288988, 3	male	Zgharta	North	Lebanon
21	11	الطبى الاجتماعى	الشمال	زغرتا	ذكر	71	لبنان	2017-01-04 00:00:00	التلقيح	Vaccination	34.288988, 3	male	Zgharta	North	Lebanon
22	11	الطبى الاجتماعى	الشمال	زغرتا	ذكر	71	لبنان	2017-01-18 00:00:00	الصيدلية	Pharmacy	34.288988, 3	male	Zgharta	North	Lebanon
23	11	الطبى الاجتماعى	الشمال	زغرتا	ذكر	71	لبنان	2016-10-03 00:00:00	الصيدلية	Pharmacy	34.288988, 3	male	Zgharta	North	Lebanon
24	11	الطبى الاجتماعى	الشمال	زغرتا	ذكر	71	لبنان	2017-01-04 00:00:00	الصيدلية	Pharmacy	34.288988, 3	male	Zgharta	North	Lebanon

Figure 3: A dummy snapshot representing potential MoPH Dataset

2.4. Data Analytics and Visualisations

Our data analytics and visualisations are focused on spatiotemporal exploration of the data. Our first step is to allow the user to filter the data using one or more attributes in the file. For example, the user can specify a certain date range or a particular location. Once the dataset is filtered, the user can generate a set of visualisations to get a summary of the data. Our system supports both motion (dynamic) and static charts. Particularly, we support *histograms*, *line plots* and *box plots*, which are all generated using Python's Pyplot library[7] in the "Static Visuals" page seen in Figure 2.

In addition to static visualisations, our system supports motion charts powered by Google Charts[8]. Motion charts are a natural way to explore data in a spatiotemporal fashion. The motion chart is dynamic and thus enables users to understand the change in several indicators as a function of time.

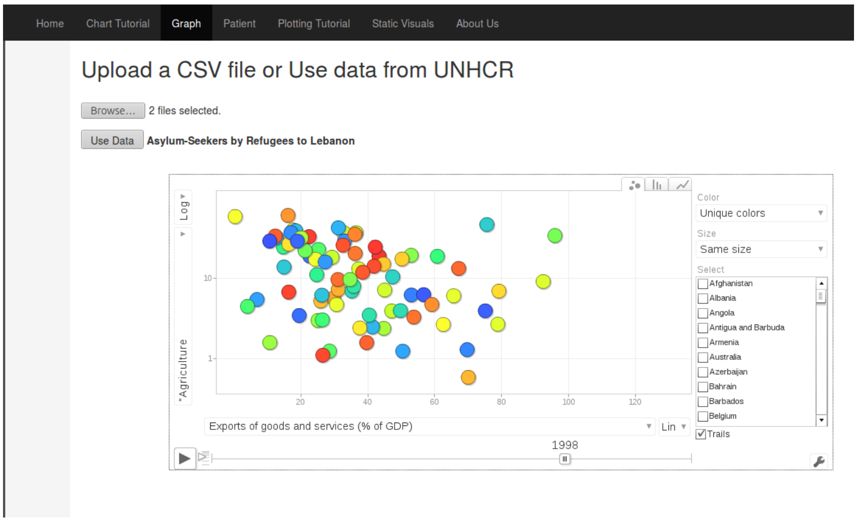
3. Case Studies

3.1. UNHCR's Persons of Concern

The "Graph" page has a "Browse" component in the top left corner of the page, seen in both Figures 4 and 5, that allows the user to import two csv files. Each of these files should be in such a format that it features an indicator with

[7] `https://matplotlib.org/`

[8] `https://developers.google.com/chart/interactive/docs/gallery/motionchart`



Upload a CSV file or Use data from UNHCR

Browse... 2 files selected.

Use Data Asylum-Seekers by Refugees to Lebanon

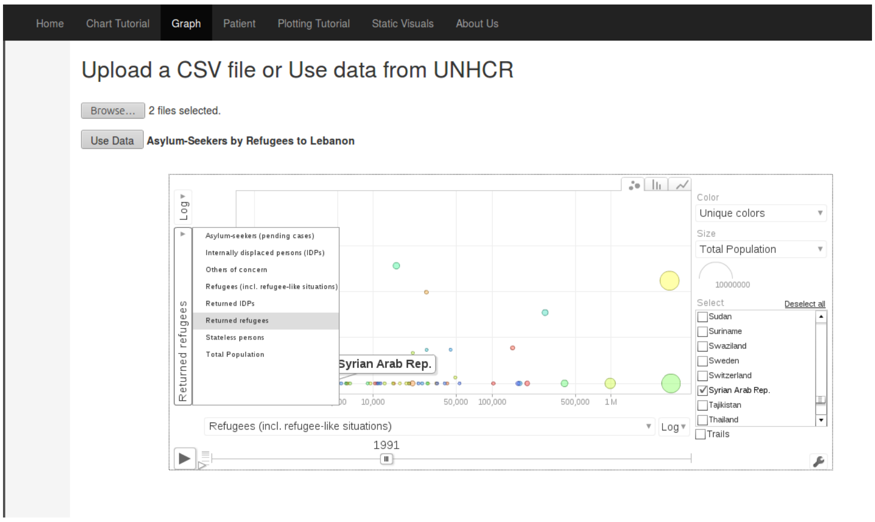
Exports of goods and services (% of GDP)

Color: Unique colors

Size: Same size

☑ Trails

Figure 4: Motion Chart from the imported indicators



Upload a CSV file or Use data from UNHCR

Browse... 2 files selected.

Use Data Asylum-Seekers by Refugees to Lebanon

Refugees (incl. refugee-like situations)

Color: Unique colors

Size: Total Population

☑ Syrian Arab Rep.

□ Trails

Figure 5: Motion Chart from UNHCR's Persons of Concern dataset: Number of Returned Refugees versus Total Number of Refugees

the columns in years and the rows as locations. In order to retrieve the best results, the years and locations should be common in both indicators. Otherwise, the plots generated and the resulting correlations may not make sense. Once these files are imported, the datasets go through a cleaning process and are displayed in a Motion Chart for the user to view.

The "Use Data" component, located under "Browse", loads a dataset from the United Nations High Commissioner for Refugees (UNHCR) named "Persons of Concern". This dataset is a csv file where the first column denotes the years, second column denotes locations, and the successive columns are the observations which the user would like to represent. Once loaded into the Motion Chart, the user can choose from a collection of columns to represent either axes and the size of the bubble as seen in Figure 5 in the drop-down list of indicators from the UNHCR dataset on the Y-axis. For example, Figure 5 plots the number of Returned refugees against the number of Refugees, across an interval time, versus the total population, indicated by the size of the bubbles. Motion charts like these help shed light on the extent to which refugees are able to integrate, as opposed to return to places of conflict. It also helps shed light on the load borne by the host community in terms of population size.

The format of the import files for the "Browse" component is quite simple to replicate. Users can make quick modifications to their own datasets and replicate the required format in order to visualize their own data on our platform. From Figure 4, the users can observe how their data evolve with time and such, deduce a correlation from their data if one exists. In this figure, one can investigate a correlation between gross GDP and agricultural net output, for example, also across a given span of time. To facilitate studying the

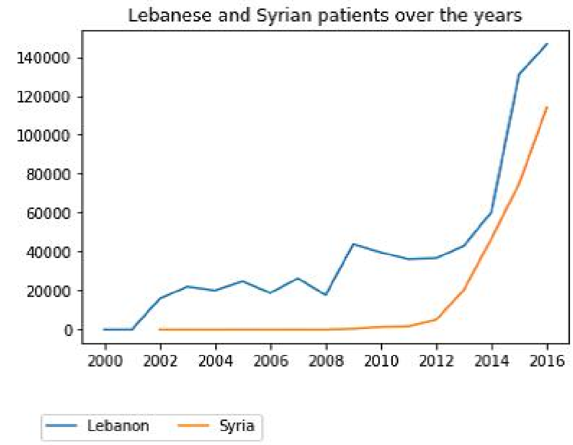


Figure 6: Overall Demand

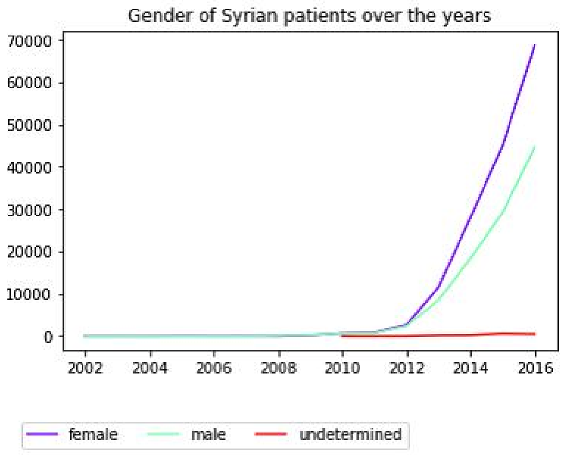


Figure 7: Gender Distribution

dynamic graph, the user is provided with a couple of options found on the left of the graph: the user can select one or more locations to follow their movement with time, and the user can check the “Trail” box found at the lower right corner of the graph to track the development of the selected locations. Such visualizations can be useful for personnel working on the refugee crisis in Lebanon.

3.2. Datasets Released by the Ministry of Public Health in Lebanon

In addition to the global challenges common to all, Lebanon is bearing an additional burden associated with its hosting of Syrian, Iraqi, and formerly, Palestinian refugees. It is timely to try and understand the impact that this continuing crisis is bearing on the primary health care system in Lebanon. The pivotal dataset that we have experimented with was received by another party on behalf of the Ministry of Public Health (MoPH) in Lebanon. The data received by the current authors is in aggregated rather than individualised form, and so does not represent human subjects per se. Instead, it constitutes a time series depicting daily demand on the primary health centers in Lebanon since 2009 by any given nationality on Lebanese soil. The locations of the centers are in terms of districts. Using our website, one can manipulate filters to perform exploratory and spatial analysis that help quantify the trends. Particularly, one can choose to observe how trends in demand have been changing by gender, by nationality, by medical departments named in the dataset, by location, by season, by year, and cues into the trends before the Syrian war and afterward.

The “Static Visuals” option of our interface, seen in Figure 2, allows us to make a selection of various exploratory plots. Once someone has produced a csv file similar to the one in Figure 3, they can put in some filters on some of the attributes. For example, one can choose to plot observations for age groups where age is greater than 50, or to plot observations where data is greater than 2011. One can also choose to subset the MoPH dataset by governorate and district. Currently, our tool shows these two terms in Arabic transliteration for ease of use by public servants, respectively shown as “Mohafaza” and “Qada”. The x-axis is an independent attribute against which the user can choose to plot the magnitude of demand on given medical centers. The “visualization method” allows the user to choose any of histograms, line plots, or box plots. The “Group by” option allows the user to aggregate demand across, say, more than one district, or more than one governorate.

The initial exploration of demand we conducted reveals that there has been a clear increase in the number of Syrian patients over the past couple of years whereas the number of patients from other nationalities seems to be insignificant. Even though our data spans the years 2000 to recent 2017, we only see a real impact of Syrian patients around 2012 onwards, to the extent that their numbers are equalling that of the Lebanese patients (Figure 6). In Figures 8 and 9, we look at the age distribution of Syrian patients in 2009 and 2016. While in both cases the skewness is positive signaling the highest demand for infant care, the number of Syrian children treated in 2009 was around 100 while it exceeds 30000 in 2016. We now take a look at the geospatial distribution of Syrian patients around the various districts. Choosing the histogram option from our tool and the filtering according to Syrian patients in both 2009 and 2016, we observe the following. While in 2009 Syrian patients were mostly aggregated around the regions of Baabda, Metn and Tripoli, which are mostly industrial and economical hubs that provided jobs for the Syrian population residing in Lebanon at that time. In 2016, the distribution shifted to span multiple districts in the Northern part of Lebanon, which is attributed to the geographical proximity of these regions to hotspots of conflict along the northern borders with Syria.

Filtering at the level of medical services, we observe that the high number of patients going into Pediatrics in 2009 makes sense since the age distribution at that time was skewed towards infants. However, in 2016, while the Pediatrics services are the most visited, there seems to be a high number of people getting into the Pharmacy, Gynecology, General Medicine, and Vaccination services. Filtering at the level of gender, the results in Figure 7 indicate that

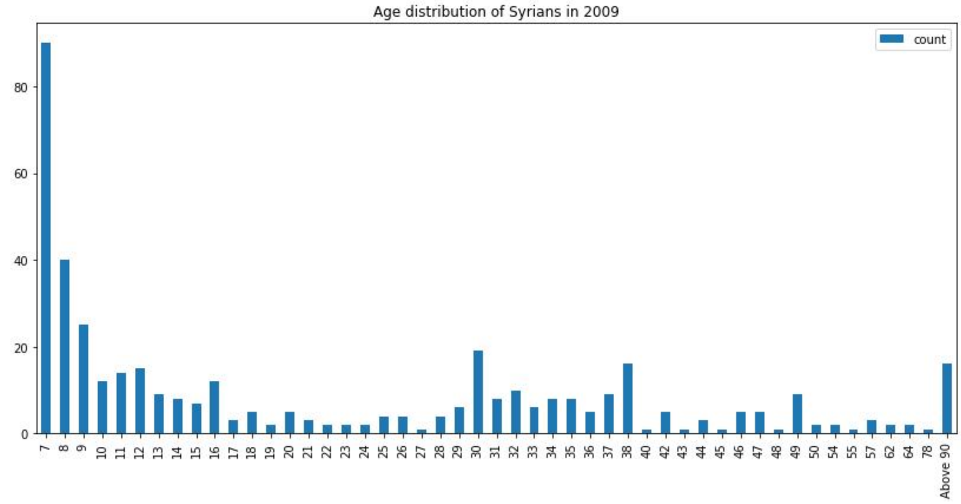


Figure 8: Age distribution 2009

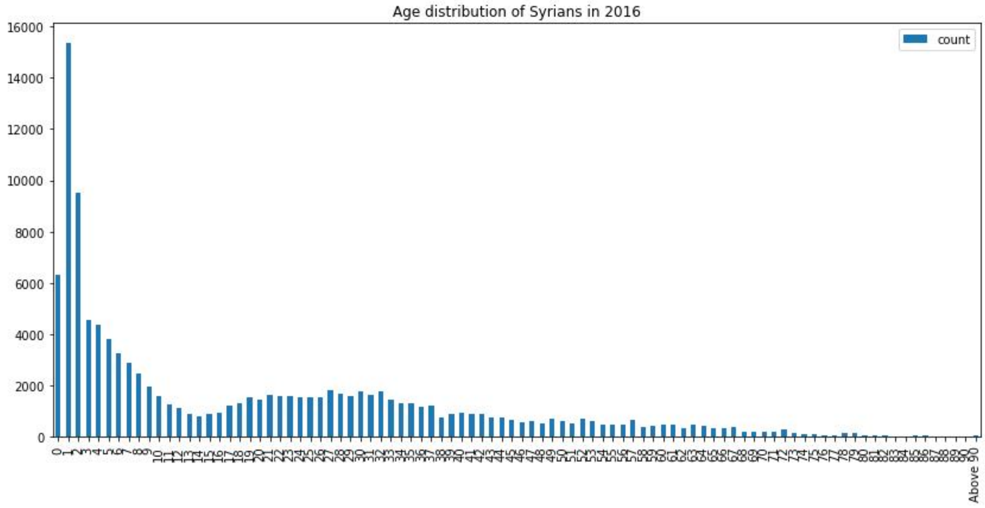


Figure 9: Age distribution 2016

the number of female patients has increased tremendously over the years, significantly faster than that of the males. This seems to confirm earlier official reports that females and children constitute the largest part of Syrian refugee population in Lebanon.

Certain figures can help us pave the way for time series forecasting by studying the trends in the data. For example, Figure 10 shows that the demand on a center in the Akkar district follows a strong increasing trend with “peak-and-trough” patterns and cyclic components since the last three years. The peak-and-trough trends are attributed to the natural fluctuation of demand according to seasons. The cyclic effects can be attributed to the fact that the volatility of the situation causes data to exhibit rises and falls that are not of fixed period. This information is important to gather before one attempts to forecast the expected increase should current influx continues. In contrast, this same center features a different time series for serving the Iraqi refugees. Particularly, Figure 11 shows that the demand has no strong patterns that would help with developing a forecasting model. The mere increases and decreases in secular trend are a result of a less volatile situation with Iraqi refugees whose mobility into Lebanon was not as much of a direct result of traumatic events across shared borders as in the case with the Syrian refugees.

3.3. The global Terrorism Dataset: A Spatiotemporal Perspective onto the Primary Health Care Dataset

As mentioned in Section 2., the dataset imported to the “Patient” page underwent feature engineering. Specifically, we used another dataset from the Global Terrorism Database

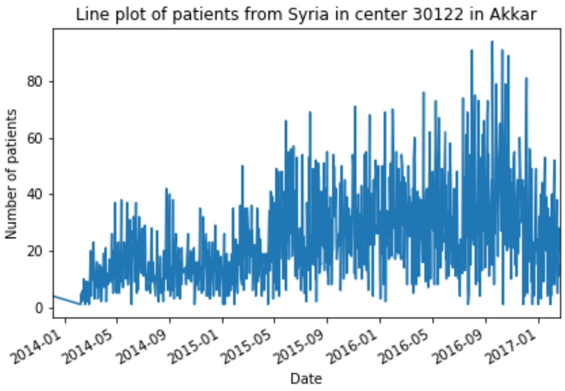


Figure 10: Time series trend – Syrian Patients in Akkar

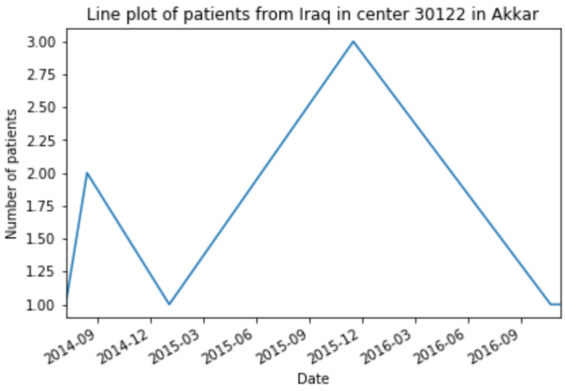


Figure 11: Time series trend – Iraqi Patients in Akkar

(GTD) in order to extract the proximity of the health centers found in the MoPH dataset to the trauma events found in the GTD. The rationale behind this suggested approach is that geographical proximity of violence hotspots to certain geolocations in Lebanon might explain the increase in demand upon centers in the implicated Lebanese districts.

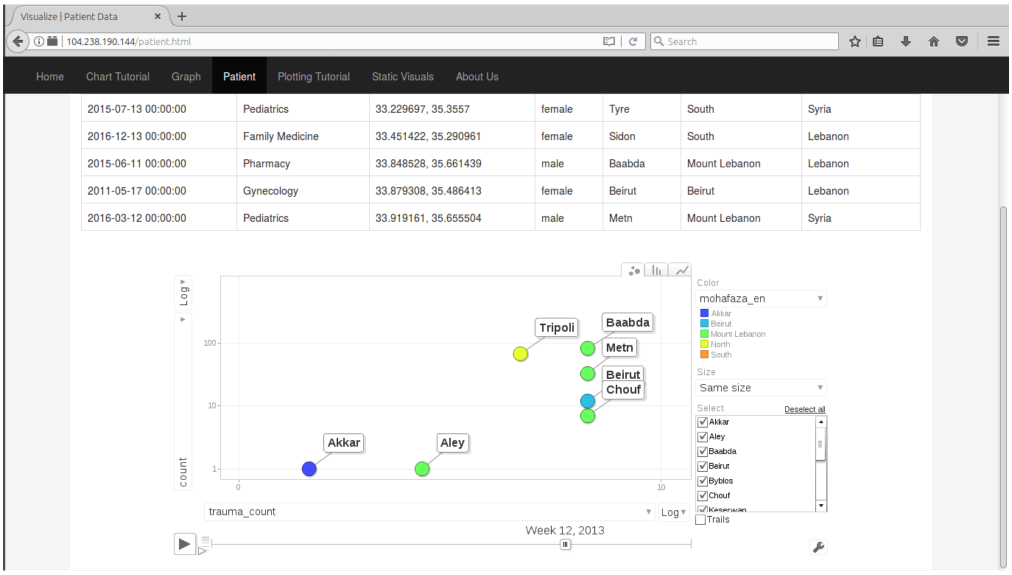
We proceeded by grouping events in the GTD into weeks, then we clustered the events near each other in each week using Python libraries such as Pandas and Scikit-learn (sklearn.cluster). This was done to avoid artificially distinguishing among traumatic events that are near each other or overlapping in space. We created a new DataFrame with the number of clustered events per week, then calculated the number of clustered events nearest to the hospital centers. This information was then appended to the original dataset featuring demand on medical centers. The final product is a dataset accumulating for time observations both the number of patients entering a medical center per week as well as the number of clustered trauma events nearest to a hospital center. Figure 12 shows the effect of high-frequency trauma events to centers around the Beirut and Mount Lebanon area. The centers were colour-coded by the district they are in. The motion chart in Figure 12 shows the demand by time. The x-axis represent the number of trauma clusters within a specified radial distance of about 277 km from the medical center. This choice of distance was arbitrary. The y-axis represents the demand on the particular center. The size of each circle can at this stage either indicate nothing ("same size" option), but it can also be modified by the user to have it correspond to the number of trauma clusters or the actual demand on the center, whichever is more visually appealing to the user. The screen-shot of the motion chart in play showing in Figure 12 reveals an interesting observation: almost all of the centers experiencing highest demand are those which have been detected to be closest to high-frequency trauma centers within about 277 km radial distance from them.

Potential users of this option can have the choice of updating a merged csv file that represents both the demand on medical centers plus their own accumulation of the frequency of traumatic events happening near a specific center. They also have the option of entering their own list of traumatic events provided it is in a format compliant with that of the GTD. For future work and amendment, we plan on giving the user the option of also manipulating the predefined distance now being at 277 km.

4. Conclusion and Future Work

The proliferation of data science techniques has generated innovative, timely, and cost efficient ways of capturing actionable intelligence in low-resource, high-risk settings. A data science approach based on machine learning and spatio-temporal analytics can contribute towards more precision for policy making and population interventions at the level of primary health care services offered by the government (Stevens and Pfeiffer, 2011). Particularly, one needs to derive computational insight into the impact and challenges surrounding the primary healthcare system in Lebanon through data provided by the primary health care unit at the Ministry of Public Health. With the help of automated techniques, one can develop real-time forecasting systems of spatiotemporal demand on the primary health sector, as a result of an abnormal rate of population growth in Lebanon, and especially if current trends of refugee influx continue. Our developed platform will help owners of data with no assumed knowledge in code development or software deployment to perform the exploration needed before embarking on any spatiotemporal modeling of the data. Using our tool, we were able to dig deeper into the temporal and spatial features affecting demand on the primary health care centers in Lebanon as a result of the Syrian refugees crisis. In the future, we also plan to allow users to visualize semi-structured and unstructured data by employing Natural Language Processing (NLP) techniques as well.

Our tool is still in its infancy and there are already ample ways it can be improved and made available to the public. Particularly, we are interested in adding a "GIS" component to it that helps users visualise spatial analysis of trends and understand spatial correlations – for example, understanding demand both as a result of seasonality as well as a result of geographical proximity to traumatic events in Syria. We anticipate our tool will be extremely useful for public servants working at the MoPH or staff at the various NGO's in the refugee aid sector with no adequate data analysts or programmers among their ranks.



2015-07-13 00:00:00	Pediatrics	33.229697, 35.3557	female	Tyre	South	Syria
2016-12-13 00:00:00	Family Medicine	33.451422, 35.290961	female	Sidon	South	Lebanon
2015-06-11 00:00:00	Pharmacy	33.848528, 35.661439	male	Baabda	Mount Lebanon	Lebanon
2011-05-17 00:00:00	Gynecology	33.879308, 35.486413	female	Beirut	Beirut	Lebanon
2016-03-12 00:00:00	Pediatrics	33.919161, 35.655504	male	Metn	Mount Lebanon	Syria

Figure 12: Motion Chart from MoPH dataset

5. Acknowledgments

The authors are grateful for the Primary Care Unit at the Ministry of Public Health in Lebanon for proving datasets that record the demand on the primary health care centers before and after the Syrian war.

6. Bibliographical References

Bastian, M., Heymann, S., Jacomy, M., et al. (2009). Gephi: an open source software for exploring and manipulating networks. *Icwsm*, 8:361–362.

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Thiel, K., and Wiswedel, B. (2009). Knime-the konstanz information miner: version 2.0 and beyond. *ACM SIGKDD explorations Newsletter*, 11(1):26–31.

Breisinger, C., Ecker, O., Al-Riffai, P., and Yu, B. (2012). *Beyond the Arab awakening: policies and investments for poverty reduction and food security*. Intl Food Policy Res Inst.

Fischer, S., Klinkenberg, R., Mierswa, I., and Ritthoff, O. (2002). Yale: Yet another learning environment–tutorial. *Colloborative Research Center*, 531.

Ghanem, T. M., Magdy, A., Musleh, M., Ghani, S., and Mokbel, M. F. M. (2014). Viscat: Spatio-temporal visualization and aggregation of categorical attributes in twitter data. In *Proc. of ACM SIGSPATIAL*, pages 537–540. ACM Press.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Hwalla, N., Weaver, C. M., Mekary, R. A., and El Labban, S. (2016). Public health nutrition in the middle east. *Frontiers in public health*, 4.

Jänicke, S., Christian, H., and Gerik, S. (2013). Geotemco: Comparative visualization of geospatial-temporal data with clutter removal based on dynamic delaunay triangulations. In *Communications in Computer and Information Science*, volume 359. Springer.

Koliopoulos, A.-K., Yiapanis, P., Tekiner, F., Nenadic, G., and Keane, J. (2015). A parallel distributed weka framework for big data mining using spark. In *Big Data (BigData Congress), 2015 IEEE International Congress on*, pages 9–16. IEEE.

Stevens, K. B. and Pfeiffer, D. U. (2011). Spatial modelling of disease using data-and knowledge-driven approaches. *Spatial and spatio-temporal epidemiology*, 2(3):125–133.

Taleb, Z. B., Bahelah, R., Fouad, F. M., Coutts, A., Wilcox, M., and Maziak, W. (2015). Syria: health in a country undergoing tragic transition. *International journal of public health*, 60(1):63–72.

DeepClouds: Stereoscopic 3D Wordle based on Conical Spirals

Wolfgang Jentner, Florian Stoffel, Dominik Jäckle, Alexander Gärtner, Daniel Keim

University of Konstanz
Universitätsstr. 10, 78467 Konstanz

{wolfgang.jentner, florian.stoffel, dominik.jaeckle, alexander.gaertner, daniel.keim}@uni-konstanz.de

Abstract

Word clouds are a widely-used technique to visualize documents or collections of documents that are arranged in a space-efficient 2D layout. Their popularity is based on the intuitive understanding and interpretation. 3D computer graphics are available in hand-held devices, on desktop computers, and in the form of specialized hardware, such as Microsofts' HoloLens (Augmented Reality) or the HTC VIVE (Virtual Reality). The wide availability of today's affordable 3D capable devices poses the question, how a 2D word cloud layout can be transferred into 3D space. In this paper, we discuss a prototypical 3D Wordle-based word cloud layout named DeepClouds that generates 3D word cloud layouts by introducing the depth of the position of words as an additional variable in the layout generation algorithm. Hereby, the algorithm exploits the z-buffer to efficiently generate an overlap-free layout from the camera's perspective. Besides introducing the DeepClouds technique, we discuss emerging problems as well as possible future areas of research and applications with respect to 3D word clouds.

Keywords: 3D, Word Cloud, Wordle, Augmented Reality, Virtual Reality, Immersive Visualization

1. Introduction

Word clouds, such as proposed by Koh et al. (Koh et al., 2010) or Cui et al. (Cui et al., 2010), are a popular choice whenever the main concepts of text-based data collections have to be visualized. They are easy to perceive, to interpret, and have further advantages, such as efficient computability, space efficiency, and typically a visually pleasing appearance, among others. Concept-wise, word clouds are an overlap-free, two-dimensional (2D) arrangement of a set of words, typically ordered by word frequency or an application dependent importance score. The spatial position is determined by a layout algorithm that positions words along a path, such as Archimedean or Euclidean spirals. At the same time, the position is optimized with respect to visual overlap and amount of non-utilized space caused by the 2D word alignment.

Recent advantages in technology enable an affordable access to 3D environments for a wide range of people. Movies are produced to be displayed on modern TVs and cinemas with depth information, leading to a more realistic, immersive experience for the viewer. Lately, mobile devices support techniques called augmented or virtual reality (AR/VR), resulting in many new applications that leverage real-world or artificial environments to present information. Additionally, dedicated hardware for VR and AR is being developed. There exist applications for 3D environments, for example, in the design industry, medical industry, and engineering (Van Krevelen and Poelman, 2010). One commonality is the fact that 3D is mostly used to display 3D information or to create an immersive feeling. An openly discussed research question is whether 3D is useful for information visualization or in other words: are there advantages of presenting abstract information in 3D as opposed to 2D (Butscher et al., 2018). While we cannot give a final answer to this question, we believe that 3D certainly increases the immersive feeling. This is important when dealing with VR and AR technologies and especially of importance for fields such as marketing and advertising. On the other hand, the written language remains our main medium of encoding information and word clouds, besides their drawbacks, remain a popular technique to represent and summarize text content.

A number of different techniques to generate word clouds have been around for some time, but technically their core is similar to what has been already described: an algorithm to place 2D elements on a 2D plane. With recent advances in available processing power and computer graphics technology, we think it is natural to expand the 2D design space of state of the art word clouds to the third dimension, e.g., to include depth in the visualization. In consequence, the word cloud layout has to be generated not only based on the width and height of the visual representation of the words to include, but also by incorporating the third dimension (depth, z-axis). This imposes new problems to solve, for example how the layout-algorithm should compute the overall position, as typically used geometric shapes such as the aforementioned spirals don't take the depth of the 3D space into account. Besides new challenges, there are also new possibilities opening up because of the additional third dimension. The depth, completely unused before, could be used to map another data value in the visual representation of the cloud. Practical use of an additional dimension in the layout can be the addition of another data attribute to be visualized without being interfered by the layout constraints, or to realize streaming word clouds that place recent nearer to the viewer, and older ones fade out after they have been pushed back.

In this paper, we present a first approach of a stereoscopic, 3D word cloud layout called DeepClouds, and elaborate on its technical details. Additionally, we discuss open questions and challenges for advanced 3D layouts, as well as interaction possibilities, which is of interest for applica-

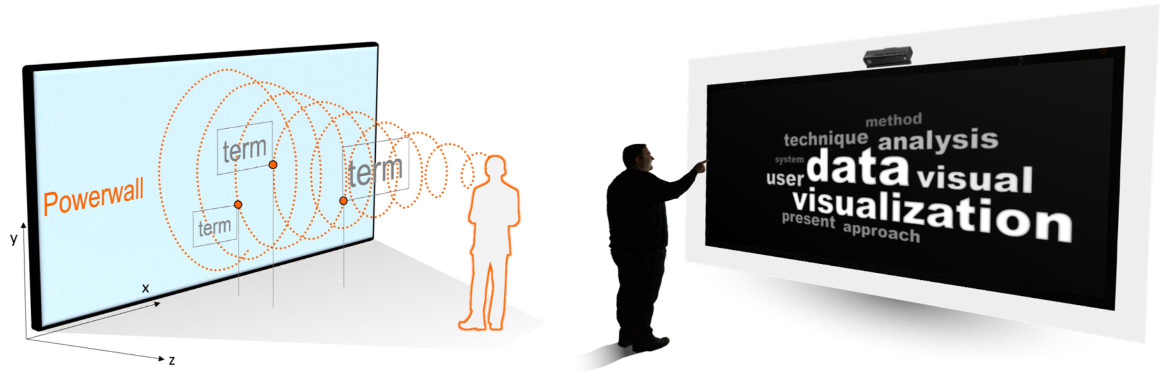


Figure 1: On the left: an illustration of the DeepClouds technique based on conical spirals, exemplifying the computation of the position of words in 3D space. Right: depiction of an envisioned DeepClouds use-case with wall-sized displays and gesture interaction.

tions in immersive 3D environments or large, wall-sized displays, which is the primary environment targeted by DeepClouds.

2. Related Work

The origin of tag clouds (also called word clouds) can be dated back far before the computer age (Viégas and Wattenberg, 2008). Flickr, amazon, and others have boosted the popularity of tag clouds tremendously in order to provide an overview of tags and their popularity (Brusilovsky, 1996). Although the tag clouds were discontinued in both cases, the simplicity of tag clouds provides a flat learning curve for users and, typically, no further explanation is necessary to describe the interpretation. Hearst et al. question the usability of tag clouds and suspect that the popularity origins from their visual aesthetics, their popularity among certain design circles, and that word clouds are perceived as being fun, popular, and hip (Hearst and Rosner, 2008). Online services like wordle (Viégas et al., 2009) allow a vast amount of non-expert users to create word cloud visualizations. Word cloud visualizations are also used for information visualization and visual analytics tools to display and summarize documents or corpora (e.g. (Viégas et al., 2007)).

A lot of research has been conducted in developing algorithms that provide space efficient and overlap free layouts, e.g. (Strobelt et al., 2012; Viégas et al., 2009; Seifert et al., 2008)). Other approaches place words according to their semantics, for example given by co-occurrences (Barth et al., 2014), word embeddings (Xu et al., 2016) or common prefixes (Burch et al., 2013). A similar problem setting emerges by animating word clouds. Here, the challenge is to keep the position of the words persistent during animation time (Cui et al., 2010), or during user interactions with the word cloud (Wu et al., 2011). Interactions include the splitting and merging of word clouds whereas the words in the resulting word cloud stay on a similar position. VCloud (Lira et al., 2016) provides the possibility to exclude and join words. Furthermore, it is possible to compare two word clouds of two data sets which is also represented as a word cloud using different colors. ManiWordle (Koh et al., 2010) and WordlePlus (Jo et al., 2015) add interactions to select, add, remove, merge, move, rotate, and resize words. In ManiWordle the user interacts with the word cloud using the mouse. WordlePlus uses pen and touch interactions.

Eventually, other work focuses on the evaluation of word cloud layouts (Lohmann et al., 2009) or the impact of the different visual property mappings (Bateman et al., 2008; Alexander et al., 2017). Rivadeneira et al. identified four main tasks that can be performed with the help of word clouds (Rivadeneira et al., 2007).

Although a lot of research can be found for compact, overlap-free layouts in a 2D space, little amount of work is available for 3D word clouds. WP-Cumulus (Tanck, 2013) is a Wordpress plugin that displays an animated word cloud based on the content of a site. To mimic a 3D effect, words that should be perceived as further in the back are displayed smaller and with a lower opacity. A similar approach is used by JS Tag Sphere (Gork, 2013) where the words are mapped onto a rotating sphere. Itoh et al. use a multi-layer spatio-temporal word-clouds where 2D word clouds are mapped into a space-time cube (Itoh et al., 2016). However, to the best of our knowledge, there is no existing work that presents a word cloud in a (stereoscopic) 3D environment.

3. DeepClouds Technique

DeepClouds is inspired by Wordle (Viégas et al., 2009), a well-known visualization technique that arranges words based on their frequency on the display. Thereby, words are positioned overlap-free and space efficient on a 2D canvas. The size and color of the words typically encode the frequency that is the importance of a specific word. We extend Wordle with a third dimension, adding another, unused dimension to the visual variable position. Furthermore, DeepClouds is designed for stereoscopic 3D graphics to emphasize the 3D effect. In the classical Wordle algorithm, the words are successively positioned along an Archimedean spiral. For each word, the algorithm starts at the same origin, and lets the word wander along the spiral until a free spot is found. Free spots on the canvas are determined by testing the bounding box of the word to position with all other, already positioned words, until the bounding boxes do not overlap The main advantage of using a spiral is that unoccupied space is searched for in a radial expanding manner originating at the same reference point, that proves to be an effective search strategy for free space. We apply the very same principle of Wordle in the DeepClouds technique, but additionally encode the importance of a word on the z-axis. This means, the farther away a word is placed, the more dispensable a word is. This brings in two core challenges regarding the positioning of the words and the infinity of space on the z-axis. Following, we discuss both challenges and describe our solutions.

3.1. Word Placement

Our placement algorithm for words is based on a hit test between bounding boxes using the z-buffer and stacked conical spirals. The z-buffer is typically a 2D array, each element representing the depth information of one pixel. After an object is rendered, the depth information (z-value) is stored in this special buffer. Analogous to the placement algorithm of Wordle, we perform for each succeeding word a hit test based on its bounding box with all other already placed words. Using the third dimension, a 2D hit test is not enough, which is why we test a hit in the z-buffer: If the bounding box of a word hits another word in the z-buffer, we need to test for a new position in x- and y-direction but also further back to encode the importance. Performing the hit test in the z-buffer is essential for preserving an overlap-free placement using perspective in a stereoscopic environment.

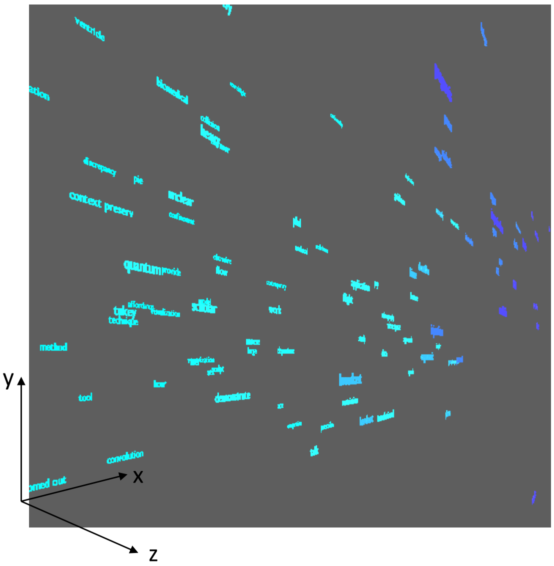


Figure 2: Detail of a DeepClouds word cloud in 3D space shown from the side.

Next, we elaborate on the layout strategy. The overall aim of the layout strategy is to preserve the classical Wordle layout as good as possible while extending it to the third dimension. However, the aforementioned described standard Wordle placement algorithm cannot be applied. You can imagine the problem as follows: The words are ordered according to their frequency, or importance, in advance. Word by word, the algorithm successively determines a free spot without overlap in the canvas. For each word, the algorithm searches for a free spot starting at the origin of the spiral. Following this strategy, a word with lower importance can be placed nearer to the user than a word with higher importance. Summing up, if we would start at the origin for each succeeding word, the concept of depth would not correspond to the importance of a word. Thus, we cannot start from the origin of the spiral for each succeeding word.

In order to preserve the impression of a classical radial Wordle layout and further encode the depth as additional visual variable, we restart the placement algorithm at the z-position of the most recently placed word. This way, we can ensure that less important words are placed further away, however, for the positioning there exist different strategies. Following, we describe three layout strategies depicted in Figure 3 that build on top of each other: a) *Continuous conical spirals.* When the algorithm finds a free spot, it stores the position along the conical spiral (x-, y-, and z-position). For the next word, the search starts at the most recently placed word. The major disadvantage is that the words are spread spaciously on the canvas (see Figure 2). The resulting layout is spacious and does not result in a compact representation that is desired. b) *Continuous stacked conical spirals.* This strategy builds upon the idea of continuing the placement at the position of the most recently placed word. To increase the compactness, word positioning does not continue along the spiral, instead a new conical spiral originating at the most recent position is created. Then, new conical spirals are created recursively until all words are placed. While words are placed closer to each other, there can occur visual artifacts such as the layout spreading into a certain direction. It cannot be guaranteed that the final word cloud hold a shape similar to the spiral. c) *Centered stacked conical spirals.* This strategy follows the concept of recursively originating new spirals at the position of the most recent word. To suit the layout best possible to the classical Wordle layout, we only story the depth value and originate each new conical spiral at the same reference point. All used spirals share the same reference point from which they originate resulting in a Wordle-alike layout.

We argue that the centered stacked conical spirals resemble the classical Wordle layout best in 3D stereoscopic space, as from the presented layout strategies it most closely resembles the spiral shape, as well as is likely to produce layouts with less free space. In the next step, we discuss the concept and usage of a virtually infinite z-axis.

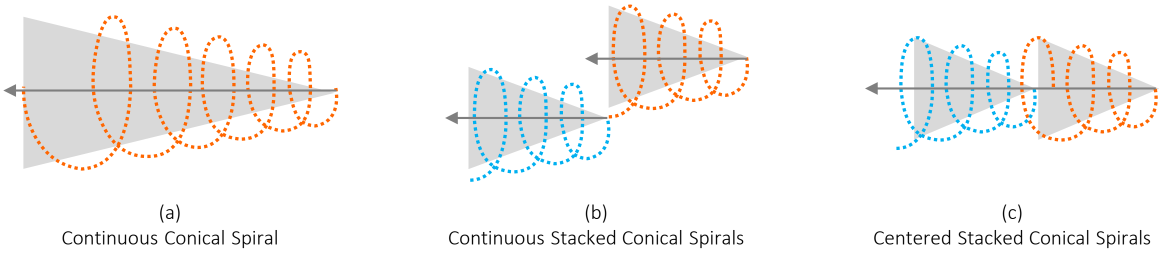

(a)
Continuous Conical Spiral
(b)
Continuous Stacked Conical Spirals
(c)
Centered Stacked Conical Spirals

Figure 3: Word placement alternatives in a 3D environment. Using a (a) continuous conical spiral, words spread overall the canvas without minimizing distances between words. (b) Minimizes the distance by starting a new spiral at the position of the most recently placed word at the cost of a skewed overall word cloud shape. The most promising approach to mimic the Wordle layout is (c) that stacks the spirals to assure they share the same origin.

3.2. Inverse Zooming

State-of-the-art zooming systems such as maps, mobile applications, and interactive visualizations (e.g. scatter-plots) apply the zooming concept based on the space-scale framework introduced by Furnas and Bederson (Furnas and Bederson, 1995). The authors describe objects as rays leaving the field of view when focusing on specific objects. It goes hand in hand with the concept of geometrically scaling objects; at some point the object is so huge that it can't fit the display anymore. In a stereoscopic 3D environment, we can adapt this concept to the task of navigating through an importance-driven word cloud. In 2D representations, the field of view is restricted by the display dimensions. In contrast, a unique characteristic of 3D is, that the field of view is restricted by the so-called view frustum. The view frustum spans a cropped pyramid into 3D space and removes everything that is not contained in the frustum, from the field of view, which is also known as *view frustum culling*. While moving through 3D space, objects may move within the frustum so that the view is updated automatically. This interaction in 3D space relates to a classical zoom interaction when moving in z-direction.

Because we use the z-axis to encode the importance of words, many words are widely spread further back on the z-axis outside of the frustum. Therefore, we introduce the concept of inverse zooming. The third dimension, that can be zoomed, causes words to literally fly through the viewer position. Compared to classical zooming, we want words to move inside the view frustum towards the center of the field of view, instead of outside. Inverted zooming enables us to infinitely zoom in z-direction and important words to pass us, and are therefore nearer to the viewer, while we focus on less important words, that are farer away. Figure 4 illustrates zoom concepts: (a) Classic zoom, where words move from the center to the display to off-screen on the left, right, bottom, and center. (b) The inverted zoom reverses this behavior and makes words located further back to move towards the center of the field of view. This way, one can successively iterate through all words that are partially visible in the plotted word cloud. Technically, we re-apply the centered stacked conical spirals to all words that surpass the far plane of the view frustum. In other words, once the user zooms the space, words in the center move beyond the virtual viewer position making room for the words that enter the view frustum. In order to fill the opening space, we calculate a new conical spiral at the last known position within the view frustum. The effect we obtain during continuous zooming is that the center of the field of view is iteratively re-filled with less important words, until they become very important, and then surpass us to make room for new words.

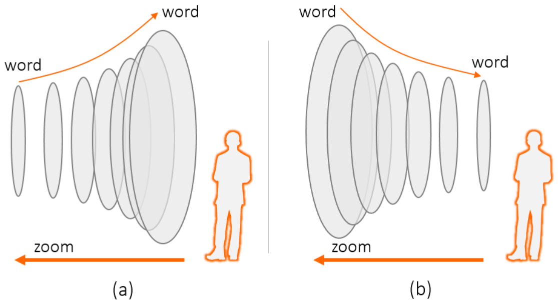


Figure 4: Zoom strategies. (a) state-of-the-art systems: objects in the middle of the focus are magnified and distances between objects increase, objects leave the field of view space while zooming. (b) Inverted zoom: words placed further in the back move into the field of view.

4. Discussion and Conclusion

In this paper, we presented a first approach of a word cloud layout algorithm for 3D space. In the following, we discuss some thoughts that arised during design and implementation of a system supporting the DeepClouds technique intended for wall-sized displays with stereoscopic 3D support.

Overlap and Perspective. In our view, the most pressing issue of 3D visualization with stacked objects is their perception. Because of that, DeepClouds is designed to produce an overlap-free layout for all possible front-view angles within stereoscopic 3D visualization using the technique described in Section 3.1.. This solves the problem of variable and changing perspective of the viewer caused by the stereoscopic display technology, but at the same time introduces a drawback compared to state of the art 2D wordles: the layout can be described as only *loosely packed*, and lacks space efficiency. While 2D word cloud layout algorithms utilize the available space very efficiently, DeepClouds has to take potentially different view angles as a layout constraint into account, which results in a larger amount of space that is perceived as free. Possibilities to overcome this drawback could be opened up by integrating user interaction and/or user tracking. For example, a refined version of DeepClouds could automatically adjust the positioning of words according to different perspectives caused by the turn of the head or body position of the viewer. This will allow to reduce the free space creation during the layout phase.

Interaction. Our current prototype implements selection of words by utilizing a Microsoft Kinect to track the index finger and a pull gesture to select words in the displayed word cloud - or with a modifier on the keyboard apply the inverted zooming concept described in Section 3.2.. After a word has been selected, corresponding documents containing the selected word are shown in an overlay in the upper right part of the screen. We integrated this feature because during the development of the layout algorithm, we recognized that the 3D view seems to make it natural to interact with objects on the screen. Connecting to the previous paragraph about overlap and perspective, gestures to rotate or adjust the representation of the displayed words seem to be a useful addition. Additionally, further operations in the data space besides the described selection of a single word are possible, e.g. to map view space manipulations such as zoom and pan directly to data filters. This is an area where the so called *inverse zooming*, as described in Section 3.2. could be applied and map different operations on the data to the different zooming techniques.

Virtual Reality. If the extended interaction possibilities and stereoscopic visualization techniques are combined, the transfer of the resulting technique to a virtual reality environment seems the next logical step. Besides the ability to visualize the word cloud in an immersive 3D environment, the direct interaction with the cloud contents will make the word cloud virtually tangible. As a result, the visualized document collection, as well as user tasks, such as document space exploration or overview, are transferred to a virtual space that can be controlled completely by the developer or designer providing the system. Besides the word cloud visualization, this could mean that the user could be presented with information augmenting the current view in 3D space. This could be beneficial for environments such as libraries or archives to provide a memorable and easily navigable experience for their users, while exploring a potentially large pool of data.

Augmented Reality. A sample application for a 3D environment with augmented reality in combination with 3D word clouds could be a content summarization of books in libraries. A user may roam through the shelves and gets information about the books she is looking at. Here, a 3D word cloud combined with a dynamic level of detail could be an effective tool to present the summarized content of the books. In preliminary experiments with Microsofts' HoloLens, we experienced that 2D visualizations appear less immersive whereas even abstract 3D visualizations feel more natural. This is a strong indication that the DeepClouds technique could be useful for AR and VR environments.

We discussed some aspects of the DeepClouds technology and possible realizations, mostly from a technical perspective, as we see this as a first step to the realization of 3D word clouds. Nevertheless, it is important to test our assumptions the resulting design with real users, in order to be able to compare our technique to 2D word clouds, and find areas where our technique needs to be improved, or to have a clear picture of areas where DeepClouds excels the current state of the art of word cloud visualizations. We are optimistic to find such areas, whether they are caused by 3D visualization, new interaction possibilities, or application scenarios that benefit from DeepClouds.

5. Bibliographical References

Alexander, E. C., Chang, C.-C., Shimabukuro, M., Franconeri, S., Collins, C., and Gleicher, M. (2017). Perceptual biases in font size as a data encoding. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–1.

Barth, L., Kobourov, S. G., and Pupyrev, S. (2014). Experimental comparison of semantic word clouds. In *SEA*, volume 8504 of *Lecture Notes in Computer Science*, pages 247–258. Springer.

Bateman, S., Gutwin, C., and Nacenta, M. A. (2008). Seeing things in the clouds: the effect of visual features on tag cloud selections. In *Hypertext*, pages 193–202. ACM.

Brusilovsky, P. (1996). Methods and techniques of adaptive hypermedia. *User Model. User-Adapt. Interact.*, 6(2-3):87–129.

Burch, M., Lohmann, S., Pompe, D., and Weiskopf, D. (2013). Prefix tag clouds. In *IV*, pages 45–50. IEEE Computer Society.

Butscher, S., Hubenschmid, S., Müller, J., Fuchs, J., and Reiterer, H. (2018). Clusters, trends, and outliers: How immersive technologies can facilitate the collaborative analysis of multidimensional data. In *2018 ACM SIGCHI Conference on Human Factors in Computing Systems, CHI '18, Montreal, Canada, April 21 - 26, 2018*.

Cui, W., Wu, Y., Liu, S., Wei, F., Zhou, M. X., and Qu, H. (2010). Context-preserving, dynamic word cloud visualization. *IEEE Computer Graphics and Applications*, 30(6):42–53.

Furnas, G. W. and Bederson, B. B. (1995). Space-scale diagrams: Understanding multiscale interfaces. In Irvin R. Katz, et al., editors, *Human Factors in Computing Systems, CHI '95 Conference Proceedings, Den-*

ver, Colorado, USA, May 7-11, 1995., pages 234–241. ACM/Addison-Wesley.

Gork, R. (2013). Js tag sphere.

Hearst, M. A. and Rosner, D. K. (2008). Tag clouds: Data analysis tool or social signaller? In *HICSS*, page 160. IEEE Computer Society.

Itoh, M., Yoshinaga, N., and Toyoda, M. (2016). Wordclouds in the sky: Multi-layer spatio-temporal event visualization from a geo-parsed microblog stream. In *IV*, pages 282–289. IEEE Computer Society.

Jo, J., Lee, B., and Seo, J. (2015). Wordleplus: Expanding wordle's use through natural interaction and animation. *IEEE Computer Graphics and Applications*, 35(6):20–28.

Koh, K., Lee, B., Kim, B. H., and Seo, J. (2010). Maniwordle: Providing flexible control over wordle. *IEEE Trans. Vis. Comput. Graph.*, 16(6):1190–1197.

Lira, W. P., Gama, F., Dall'Agnol, H. M. B., Alves, R., and de Souza, C. R. B. (2016). Vcloud: adding interactiveness to word clouds for knowledge exploration in large unstructured texts. In *SAC*, pages 193–198. ACM.

Lohmann, S., Ziegler, J., and Tetzlaff, L. (2009). Comparison of tag cloud layouts: Task-related performance and visual exploration. In *INTERACT (1)*, volume 5726 of *Lecture Notes in Computer Science*, pages 392–404. Springer.

Rivadeneira, A. W., Gruen, D. M., Muller, M. J., and Millen, D. R. (2007). Getting our head in the clouds: toward evaluation studies of tagclouds. In *CHI*, pages 995–998. ACM.

Seifert, C., Kump, B., Kienreich, W., Granitzer, G., and Granitzer, M. (2008). On the beauty and usability of tag clouds. In *IV*, pages 17–25. IEEE Computer Society.

Strobelt, H., Spicker, M., Stoffel, A., Keim, D. A., and Deussen, O. (2012). Rolled-out wordles: A heuristic method for overlap removal of 2d data representatives. *Comput. Graph. Forum*, 31(3):1135–1144.

Tanck, R. (2013). Wp cumulus.

Van Krevelen, D. and Poelman, R. (2010). A survey of augmented reality technologies, applications and limitations. *International Journal of Virtual Reality*, 9(2):1.

Viégas, F. B. and Wattenberg, M. (2008). Timelines - tag clouds and the case for vernacular visualization. *Interactions*, 15(4):49–52.

Viégas, F. B., Wattenberg, M., van Ham, F., Kriss, J., and McKeon, M. M. (2007). Manyeyes: a site for visualization at internet scale. *IEEE Trans. Vis. Comput. Graph.*, 13(6):1121–1128.

Viégas, F. B., Wattenberg, M., and Feinberg, J. (2009). Participatory visualization with wordle. *IEEE Trans. Vis. Comput. Graph.*, 15(6):1137–1144.

Wu, Y., Provan, T., Wei, F., Liu, S., and Ma, K. (2011). Semantic-preserving word clouds by seam carving. *Comput. Graph. Forum*, 30(3):741–750.

Xu, J., Tao, Y., and Lin, H. (2016). Semantic word cloud generation based on word embeddings. In *PacificVis*, pages 239–243. IEEE Computer Society.