EC Project 610829

A Decarbonisation Platform for Citizen Empowerment and Translating Collective Awareness into Behavioural Change

D4.3: Algorithms for Tracking Information Diffusion Patterns

08 February 2016
Version: 1.2

Version History

Version	Date	Author	Comments
0.1	15/09/2015	M. Goebel	Initial version
0.2	24/09/2015	M. Goebel	Restructuring and consolidation
0.21	26/09/2015	D. Maynard	Feedback and minor edits
0.3	28/09/2016	S. Zhu	Network analysis section
0.31	30/09/2016	A. Scharl	Document revision
0.32	10/11/2015	M. Goebel	Document revision
1.0	12/11/2015	A. Scharl	Minor edits
1.1	05/02/2016	M. Goebel	Update based on reviewer feedback
1.2	08/02/2016	A. Scharl	Edits and final version

Peer reviewed by Diana Maynard

Dissemination Level: PU – Public

This document is part of the DecarboNet research project, which receives funding from the European Union's 7th Framework Programme for research, technology development and demonstration (Grant Agreement No 610829; ICT-2013.5.5 CAPS Collective Awareness Platforms for Sustainability and Social Innovation).

Table of Contents

Executive Summary

This document summarizes outcomes of work performed in Work Package 4 (WP4) and constitutes Deliverable D4.3 of the DecarboNet Project – with the goal of identifying opinion leaders and prevalent dissemination paths across news and social media content streams, based on the datasets collected in WP2. We report on both the required pre-processing and knowledge extraction steps as well as on the analytic methods developed to investigate information diffusion processes:

- For the former, work has focused on story clustering, event extraction, and user modelling. This includes extensions of the system architecture of the *Media Watch on Climate Change* (knowledge integration and analytics platform developed in WP3),[1] in order to support the required new data structures.

- For the latter, we present results on opinion leader detection and centrality networks, including graph-based methods to visualize these networks and the ability to distinguish opinion holders from opinion targets (WP2) across online media channels.

As part of the WP6 use case, these analytic methods will be applied to investigate the impact of upcoming environmental events including COP21, the *United Nations Climate Change Conference* (December 2015) and *Earth Hour*, the annual global event organized by the World Wide Fund for Nature (March 2016).

[1] www.ecoresearch.net/climate

Introduction

Information Diffusion

Information Diffusion measures and predicts the propagation of information messages through an actor network along a temporal axis. The actor network may be modelled as a graph describing all persons, objects, and semiotics that take part in the message passing, either implicitly or explicitly.

The DecarboNet project (www.decarbonet.eu) studies opportunities to support and analyse community dialogues centred on the topics of *energy* and *climate*. D4.3 seeks to provide insights into the information diffusion patterns that explain how information spreads and evolves over time in these domains. In line with this goal, we formulate our research questions as follows:

1. Which topics are most widely discussed? (What?)
2. Which actors or information sources are the most influential in the community, and how can we quantify their impact? (Who?)
3. Which information channels are used to spread environmental information, with a special focus on major topics covered by traditional media channels, and the related discussions on social media platforms? (Where?)
4. How does information spread in terms of velocity and flow patterns? (How?)

To study diffusion paths and patterns of information in real-world data, we first have to transform the raw input data to make the *information diffusion* explicit for analysis by extracting information from data. The data we collect through the DecarboNet data acquisition pipeline (see Deliverable D2.1) is provided as large collections of individual documents of heterogeneous internal representation (depending on the acquisition channel such as Twitter, Facebook, RSS feeds, etc.).

In a first processing step, this low-level input data is transformed into high-level information streams to be analysed for diffusion. In particular, in this work we focus on user modelling and story modelling as the most promising information streams for insightful diffusion analysis. This choice is motivated by the natural network interpretation of both the information actor space (user network) and the information topic space (story network). For user modelling, the focus of diffusion patterns lies on identifying *opinion leaders* and tracking the distribution channels that contribute to the spread of information – e.g., it addresses questions second and third from the set of postulated research question above.

Story modelling allows investigating diffusion patterns directly on the information space, through providing better boundaries of the information itself. Information in the news data and social media context is defined to be a *message* of observational or communicational character that can be passed between agents.

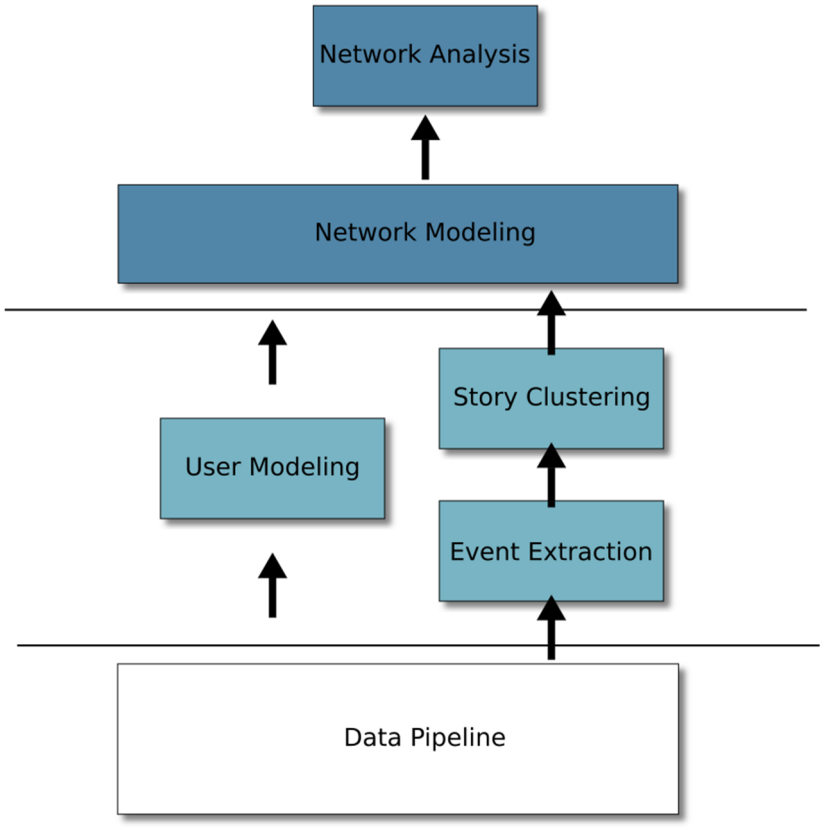


Figure 1. From data to information diffusion analysis

A *story* further defines information along temporal and factual aspects and thereby increases the acuity of the messages studied in information diffusion. In the work presented herein, we concentrated on event extraction as a means for story detection, since events are particularly prevalent in news articles as the distinguishing characteristic. Story clustering addresses questions 1 and 4 from the set of postulated research question above. Figure 1 depicts this data transformation process both for making user information and story information explicit for the network modelling and subsequent analysis.

Identifying Opinion Leaders

Information flow analysis helps to gain a better understanding of the role of actors that participate in the spreading of information. In particular, we are interested if there is a relevant set of users (individuals or corporate Twitter accounts) that have a significant impact on climate- and energy- related information flows.

We distinguish between opinion leaders in terms of being the *focus of media attention* (e.g. CEOs, politicians, actors, etc.; see Section on “Opinion Holders vs. Opinion Targets” below) and opinion leaders in terms having an *impact* on evolving online discussions. For the coverage of *news media sites*, we calculate impact by multiplying the frequency of the coverage on a specific topic with the reach of a Web site estimated via Alexa traffic statistics.[2] For the postings on *social networks*, we use common node centralities and other network metrics.

[2] www.alexa.com

Analysing Threaded Dialogs

In addition to the impact of a source, T4.3 also investigates information diffusion and the factors influencing a story as it unfolds through the news landscape, from a technical rather than social perspective – i.e., a document-driven approach that identifies document metrics to characterize the diffusion of a story by means of the documents that it entails.

Information diffusion is traditionally studied through the lens of network analysis, and many different approaches are proposed in the literature to detect information diffusion patterns in social media (Guille et al., 2013). Due to the explicit graph structure of social networks such as Twitter, Facebook and LinkedIn, we can trace links between users and messages through follow, share and reply activities. Yet these approaches are not directly applicable if one wants to track the latest topics covered by online news media – one the one hand, news media do not contain an explicit graph structure; on the other hand, the sheer volume of social media coverage does not allow clustering of the complete content streams ("firehose").

Deliverable Structure

To address this challenges, T4.3 pursues a two-step approach: (i) analyse news media coverage in real time to extract the major topics of interest, and (ii) query the collected archive of domain-relevant social media postings for this topic, creating and graph structure based on explicit *@$user* references. This document is therefore organized into the two distinct sections of *Network Modelling* and *Network Analysis*:

In *Network Modelling*, we present the work performed to transform low-level input data into high-level network data. We discuss methods of document clustering, feature extraction, and user modelling in order to build explicit data flow networks for the subsequent analysis.

In the second part, *Network Analysis*, we present state-of-the-art methods to identify and exploit patterns in the network data generated by the modelling process. We conclude this document with a short discussion of the problems we encountered during our research and outline research avenues for the near and medium future of this ongoing work.

Network Modelling

Data Collection

WP2 and WP3 have established a large document corpus that is extended with real-time content feeds from a wide range of sources and distribution channels, both from classic news channels as well as social media. These documents are the low-level input to all data processing steps that allow us to make explicit a network data structure to analyse diffusion patterns (network modelling).

Event Extraction

D2.2 introduced the *Recognyze* component for Named Entity Recognition (NER), which supports disambiguation based on Linked Open Data (LOD) corpora such as DBpedia[3]. We have successfully deployed Named Entity profiles for persons, organizations, and geolocations to annotate documents from the *Media Watch on Climate Change* (MWCC).

[3] www.dbpedia.org

During this reporting window, we have extended *Recognyze* by a GEMET and an event profile. GEMET[4] offers an open thesaurus of curated by the European Environment Agency (EEA) that defines a general terminology for the environment with multi-language support. The event corpus was extracted semi-automatically from Wikipedia for the climate domain. This step was necessary due to no such data being available in structured format in the public domain. While DBpedia does contain an event category, the quality of said data is heavily biased towards esoteric events and therefore effectively unusable for our purposes. Our scraper extracted more than 1000 distinct events from various climate relevant Wikipedia categories such as climate conferences and natural climate events. Since these events are extracted from an existing knowledge base, we refer to this process as *explicit event extraction* (in contrast to implicit event extraction, where events are extracted from free text, and which will be discussed in the next section).

Story Clustering

In order to build information diffusion networks, we cluster documents around stories. A *story* is a chronologically ordered sequence of co-related events. For example, an *Earth Hour*[5] story may contain summaries of all the local campaigns and sub-events organized around the annual event, and the various follow-up interactions in social networks.

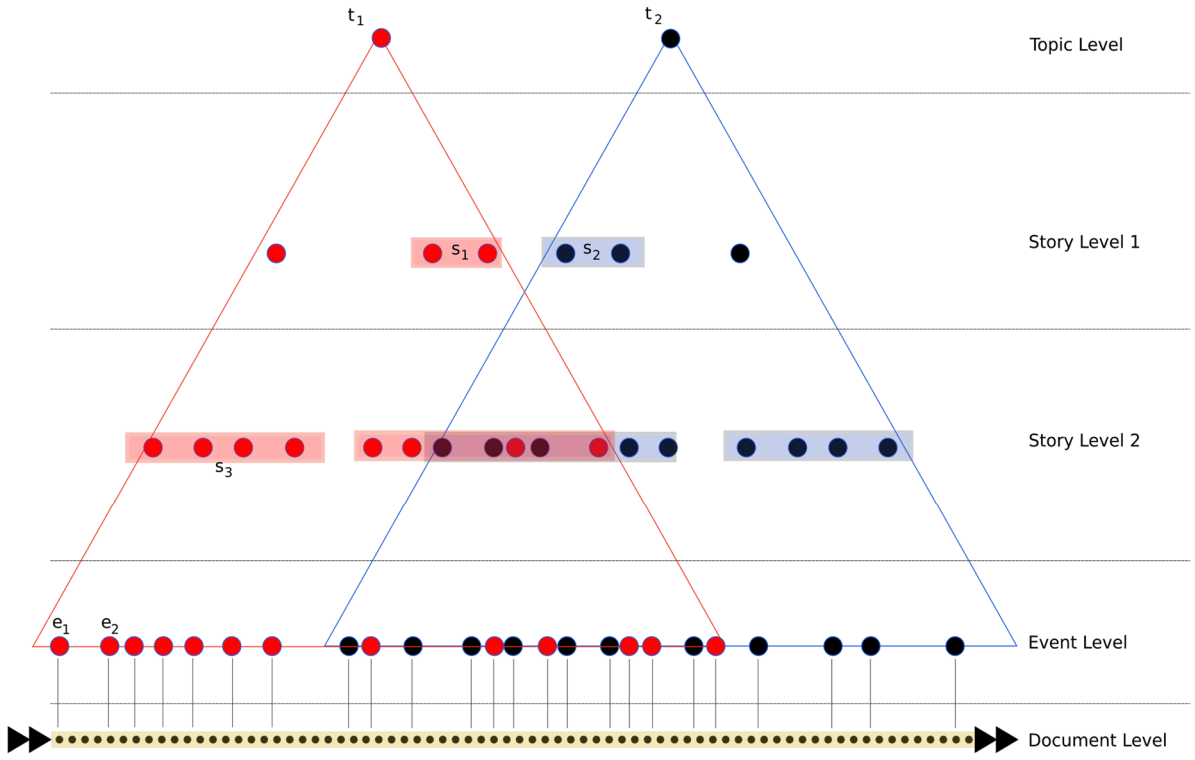


Figure 2. Story clustering as a hierarchical approach

[4] www.eionet.europa.eu/gemet
[5] Earth Hour is a global annual event organized by the World Wide Fund for Nature (WWF), and the main use case analyzed in WP6.

Stories can be told at a continuously changing degree of granularity, with the most granular story being a single event, and the most general story being a topic (c.f. Figure 2). For technical reasons, this continuity constraint will be relaxed to an n-level degree of granularity in practice.

News clustering is a special case of text clustering that aggregates articles into stories based on their textual content and publication date (Wu et al., 2015). This approach allows to group similar articles together regardless of whether they contain copycat fragments (citations) or not. The state-of-the-art approaches to news clustering represent the news content using keywords (Lu et al., 2014), paraphrases (Petrovi et al., 2012), named entities (Montalvo et al., 2015), topics (Xia et al., 2015) or event models (Zhu and Oates, 2013; Hu et al., 2014; Wu et al., 2015) and further use them as features for clustering. In our work, we follow the approach presented in (Wu et al. 2015) and extract events to produce relevant features for story clustering.

Document Features

A pre-processing step for story clustering is the definition of document feature transforms that map input documents to a well-formed set of individual features. The goal of feature extraction is to find the optimal transform such that the resulting features best represent the gist of the document, i.e. they are the *fingerprint* of the document. In this work, we have considered the following set of document feature transforms as input to document clustering:

- **Stopword transform**, removes the stopwords from the document's content
- **Keyword transform**, the most significant keywords of a document
- **Quotation transform**, set of all quotes in a document
- **Entity transform**, set of most significant named entities found in a document
- **Event transform**, set of most significant events found in a document
- **Mixture transform**, a combination of all transforms described above

While most of the above document transforms are straightforward to implement given the prior work reported in deliverables D2.1 and D2.2, some transforms required the development of additional natural language processing components.

The Keyword transform uses a model of the statistical significance of all n-grams (after stopword removal) for a given document corpus (e.g. climate social media, climate NGO, etc.) to extract the most significant keywords from a document. The transform returns a vector of the top ten most frequent keywords in a document.

The idea behind the event transform is that most documents can easily be reduced into the core *events* they mention without losing too much of the document's meaning (e.g. a bag of events). In this work, we understand as events *actions* performed by *agents* (persons, organizations, countries, etc.). Note that in contrast to the explicit event extraction discussed in prior section, we are dealing with *implicit event extraction* here, since we do not make use of any external knowledge base to query or disambiguate extracted events. To reduce the overall amount of individual events per document, we only consider a document's title together with the first four sentences for event extraction. The event transform therefore yields a highly compact representation that still provides a meaningful fingerprint of the document.

In the context of news articles, an *event* is the major topic of an article, e.g. a conference opening, an earthquake or a soccer match. At the same time, such an event may have several levels of granularity and contain multiple sub-events. We refer to such higher-level events as *complex events*.

We extract events using relation extraction and further represent the documents (i.e. news articles and social messages) using the bag-of-events model as features for clustering. We apply dependency parsing to obtain the parse trees of the sentences and then extract relations using these trees. The relations are modelled as triples of the form

s (subject) - **p** (predicate) - **o** (object),

where **p** is a verb phrase (VP) representing an *action*; **s** is the *agent* that performs the action described in the predicate, and **o** is an optional object which specifies further any additional details concerning the action, such as its place, time, manner, etc.

Locality Sensitive Hashing

Locality Sensitive Hashing (LSH) uses a hash function that maps similar documents to similar hashes. The hashing function takes as input a vectorised document representation and computes a fix-bit hash key. We use the popular SimHash hash as an LSH algorithm and the Hamming distance to measure similarity between the hashes and produce story clusters (Rajaraman and Ullman, 2011). The similarity threshold was empirically inferred based on the manual evaluation on a subset of documents. We cluster news articles into stories using as features the extracted relations representing events.

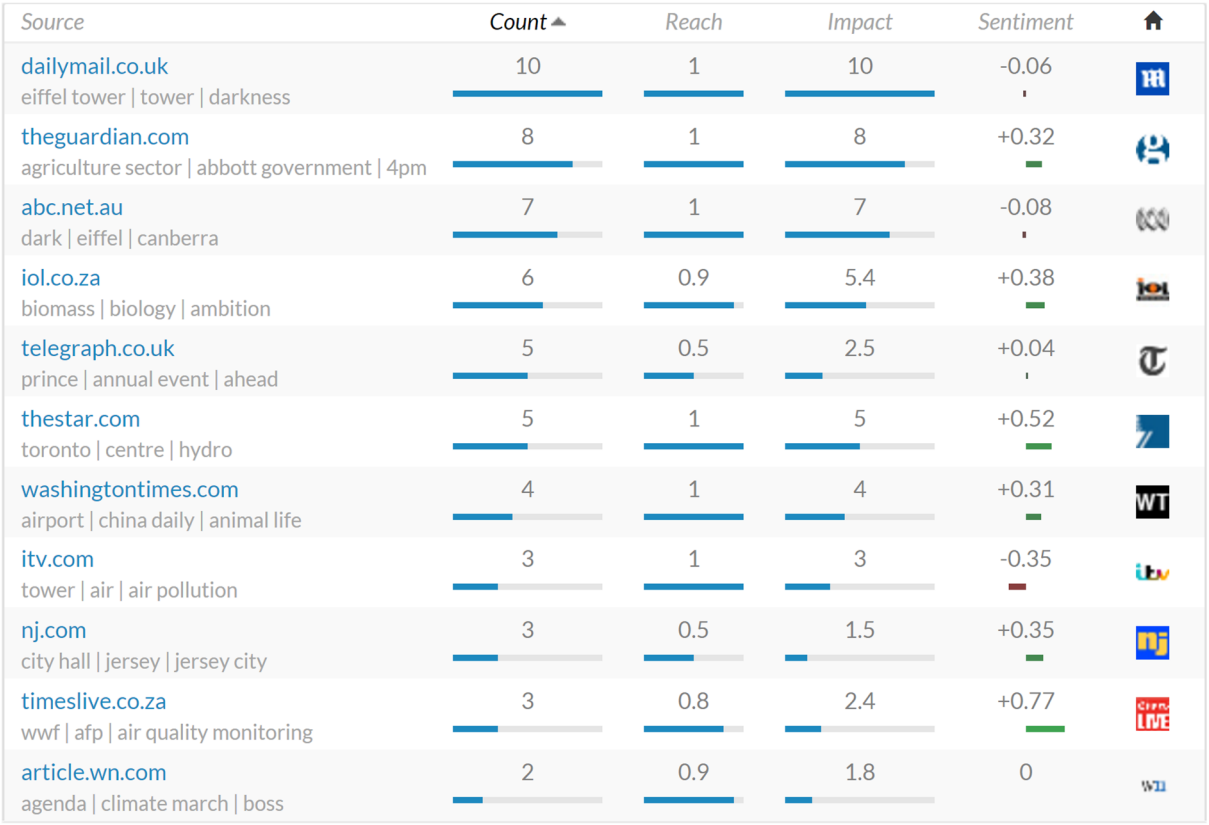
Our approach gives good results for identifying news stories from the news data set. Overall, we have identified ten distinct news stories from the *earth hour* data from within the last year (about 10000 documents), and about 20 distinct news stories from *earth day* data set within the last year (about 8000 documents).

Hierarchical Dirichlet Modelling

Various clustering approaches were assessed by our colleagues at USFD, who were clustering documents for story and burst detection on Twitter data. Their extensive evaluation of several state-of-the-art story clustering methods on large Twitter corpora for recent event coverage (story detection and sub-story detection) has favoured the generative Hierarchical Latent Dirichlet Allocation (HLDA) approach over spectral clustering and the LSH approach described above. HLDA is a probabilistic method that assumes the existence of multiple latent topics (modelled as individual Dirichlet distributions) in a data set, which it then fits given a training data set to produce a mixture model of overlapping topic distributions. In the hierarchical case, the general LDA is extended to add a second level of topics to describe nested topics. We have chosen LSH over HLDA for the following reasons:

1. Available data is not limited to Twitter streams; a considerable part of the data is news articles, which feature much larger textual segments. This issue can be addressed by applying a text summarisation algorithm that reduces large documents to short paragraphs or single-sentence summaries.

2. The current pipeline structure does not easily lend itself to a model-based approach with considerable training overhead from machine learning; future work will experiment with HLDA for story modelling and ways to overcome computational bottlenecks.



Source	Count	Reach	Impact	Sentiment
dailymail.co.uk eiffel tower | tower | darkness	10	1	10	-0.06
theguardian.com agriculture sector | abbott government | 4pm	8	1	8	+0.32
abc.net.au dark | eiffel | canberra	7	1	7	-0.08
iol.co.za biomass | biology | ambition	6	0.9	5.4	+0.38
telegraph.co.uk prince | annual event | ahead	5	0.5	2.5	+0.04
thestar.com toronto | centre | hydro	5	1	5	+0.52
washingtontimes.com airport | china daily | animal life	4	1	4	+0.31
itv.com tower | air | air pollution	3	1	3	-0.35
nj.com city hall | jersey | jersey city	3	0.5	1.5	+0.35
timeslive.co.za wwf | afp | air quality monitoring	3	0.8	2.4	+0.77
article.wn.com agenda | climate march | boss	2	0.9	1.8	0

Figure 3. Overview of international news media coverage about the Earth Hour 2015 campaign (02-04/2015), including f*requency, reach, impact* and *sentiment* values

Figure 3 shows major international news outlets leading the Earth Hour 2015 coverage. *Reach* is a proxy of Web site popularity, normalized from *Alexa* traffic statistics. *Impact* is the frequency of coverage multiplied by reach. Sentiment indicates whether the coverage tends to be positive or negative. The search-specific keywords underneath the URL reflect the focus of the respective outlet's coverage. When computed for shorter intervals – e.g., the last 24 hours or the last week – these keywords reflect recent trends and can be used as seed terms for a query to analyse related discussion on social media platforms.

Threaded Dialogues

As mentioned above, combining the heterogeneous document types from social media and news media channels comes with a number of technical issued to be solved. Social media streams are of a very different nature in terms of content characteristics (e.g. length) and the availability of meta-data, as compared to traditional news articles. Combining the two into an integrated story model has introduced the challenge of identifying the connection between the two data sets, as well as finding the relevant social content for a given news article. Sometimes we missed the most relevant social content for existing news articles in our data

sets due to not following the right social user accounts. To remedy this issue, we have extended our social channels with search-by-URL for the most influential news articles encountered (e.g. the core stories). To address the problem of modelling threaded dialogues of heterogeneous content channels, we had to extend our data structure to reflect typical social network actions such as comment, re-tweet, like, etc. Making such actions and references explicit allows to model stories as complex nested hierarchies, reflecting the actual dialogue thread as it unfolded in the online world.

User Modelling

In the context of information diffusion, we define a user to be any participant in a communication of information, e.g. a publisher of an article, the author of a comment, a tweet, etc. A user therefore need not necessarily correspond to a natural person, since publishing organizations also represent valid users in our model. As a first step, we make the users of our media content explicit, allowing us to:

- associate unique IDs to users
- aggregate over multi-profile users
- understand the cultural and geographic reach of users
- aggregate periodic metrics on users
- describe implicit user mentions

The chosen approach to user modelling is to use the existing information we have on publishers of documents:

1. HTML pages provide rich user information in the HTLM header
2. Social media provides user information via the API (e.g. open graph)

Together, this information allows us to link users from the static and the dynamic channels of our data acquisition pipeline into a universal user model. In particular we utilize metadata where provided in the HTML header to interlink existing profiles.

Additionally, user metrics measured periodically allow us to understand trends in performance and reach of the individual users as a function over time. Initially, user metrics of interest include:

- global reach (geopolitical spread)
- topical interest (topic modelling)
- influence (opinion leaders, story influencers)
- controversy (variation in text sentiment)

These may be extended at a later stage to also include a user's community, and also their needs (e.g. to understand why a user participates in a given communication). As part of this deliverable, we have integrated periodic user metrics into the back-end pipeline of the *Media Watch on Climate Change*, enabling statements on the dynamically evolving network of monitored users as per the outlined user model.

Network Analysis

Overview

Information diffusion in social networks is driven by four main factors (Gonzalez-Bailon, 2011): (i) the propensity of individuals to follow actions determines their likelihood to participate in the diffusion process; (ii) The number of connections to influencers also determines the likelihood of participation; (iii) acquiring a number of “low threshold actors” facilitates reaching a critical mass of followers; (iv) connections to multiple sources is more important than repeated exposure to the same source. Opinion leaders sit at the core of the network. While they do not necessarily need a high number of connections, they do, however, possess connections to other highly influential agents in the network.

The network topology itself is another important factor for the velocity of information diffusion. Random networks, i.e. networks with long ties, do not support diffusion well, which stands in contrast to the common theory of “the strength of weak ties” (Granovetter 1973). The core idea of this theory is that individuals are more likely to be influenced by agents not belonging to their close circle of friendship. The rationale behind this is that closer friends share mutual information, while remote friends tend to introduce more novel, and thus more interesting knowledge. However, a study simulating the spread of medical knowledge in a network of artificially designed topology refutes this theory (Centola 2010). The study shows, that highly clustered networks, i.e. networks where the clusters represent strong relations between their members, foster the spread of information. This goes in line with the second of the before-mentioned factors driving information diffusion. The strong connection within the cluster exposes individuals to multiple actions of their peers. This repeated exposure “convinces” the individual of the transmitted fact and in return triggers the further spread of the fact by the individual.

Diffusion Metrics

Once the document flow is transformed into a graph structure, it is possible to apply established algorithms from network analysis to detect the opinion leaders, e.g. HITS (Kleinberg, 1999) or PageRank (Page et al., 1999). Del Corso et al. (2005) rank articles proportional to the cluster size they belong to and inversely proportional to their publication time. Further, they compute the news source rank as a function of the ranks of its articles. Zhang et al. (2013) utilize social media posts to rank news articles by computing TF-IDF for each tweet-article pair, which does not scale well. (Weng et al., 2010) propose to identify the most influential Twitter users for a specific topic using Latent Dirichlet Analysis (LDA).

Other metrics used on information diffusion networks include:

- *Influence*, as applied to Twitter users in (Cha et al., 2010, Bakshy et al., 2011)
- *Power* (Hanneman et al., 2005), using network centrality
- *Popularity* (Application: web pages)

Given an explicit information diffusion network, we can measure the following structural properties on the network (Valente 1995, Valente 2012): Network density, network reciprocity, network centralization (measuring the extent to which a network is centred around a single node), a network’s assortativity coefficient, and the network’s centrality (e.g. Freeman’s

Network Centrality). Furthermore, we can measure properties on a network's nodes and edges, such as the Freeman edge betweenness (Wasserman et al., 1994) to make statements about the individuals within a given diffusion network. Such analysis allows us to determine significant nodes (e.g. actors or stories) or edges (relations between actors or stories). A node's position in a diffusion graph is typically referred to as its *centrality*. Network centrality is an indicator of importance of nodes and edges in graph data structures, and it is typically defined over the graph's topology within the proximity of the node/edge of interest.

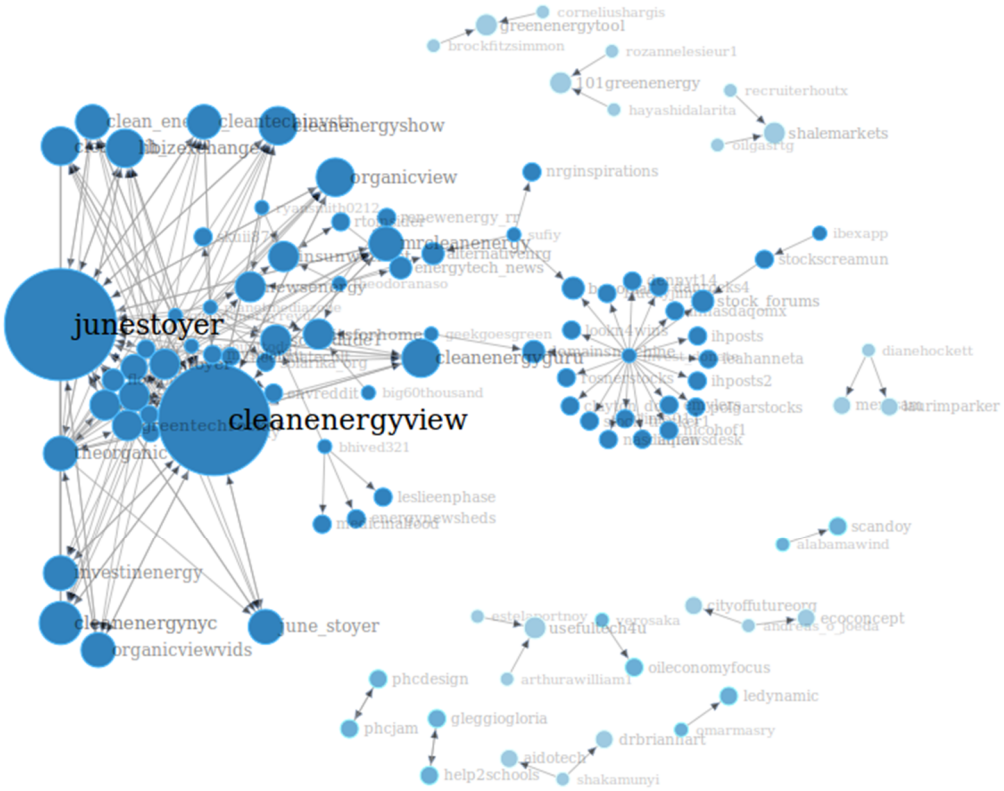


Figure 4. Degree node centrality on the Twitter Earth Hour data set for the topic 'energy', where nodes represent users and edges their cross-references[6]

Common node centralities in the literature include:

- *Degree centrality* (both in-degree and out-degree), defined as the number of links incident upon a node, is the earliest and also the most fundamental centrality measure used to describe node characteristics in a graph;
- *Closeness centrality* emphasizes "the distance of an actor to all others in the network by focusing on the distance from each actor to all others" (Hanneman et al., 2005), based on the length of the average shortest path between a vertex and all vertices in the graph;

[6] The planned integration of network centrality metrics into the *Media Watch on Climate Change* (WP3) will make use of the graph-based visualization tools of the FP7 Project Pheme (www.pheme.eu), which develops automated methods to assess the veracity of social media content.

- *Information centrality* is a variation of the closeness centrality that favours symmetric subgraphs by using the harmonic mean of weighted shortest paths rather than the arithmetic mean as in the general case;
- *Eigenvector centrality*, a node has high score if connected to many nodes are themselves well connected. It is used as indicator of popularity since it tends to identify centers of large cliques;
- *Katz centrality* (Katz et al., 1955) allows measuring the degree of influence of a node (actor) through counting the total number of walks between each pair of nodes
- *Betweenness centrality*, yields the fraction of shortest paths between node pairs that pass through the node of interest (Newman, M., 2004):
    - Freeman betweenness centrality,
    - Flow betweenness,
    - Random walk betweenness, which counts how often a node is traversed by a random walk between two other nodes.

As part of D4.3, we have measured degree node centrality (in-degree) on the Twitter user data set for collected for the Earth Hour data. Other centralities are currently implemented to get more insight into the diffusion patterns of selected topics. We use the user mention relation available in the Twitter API to construct the information diffusion network. Figure 4 shows a visualization of this centrality for the top 300 users based on the topic 'Energy'. Larger nodes correspond to a higher centrality score of the respective nodes. The visualization allows us to quickly identify those regions of the information diffusion network with the highest user connectedness.

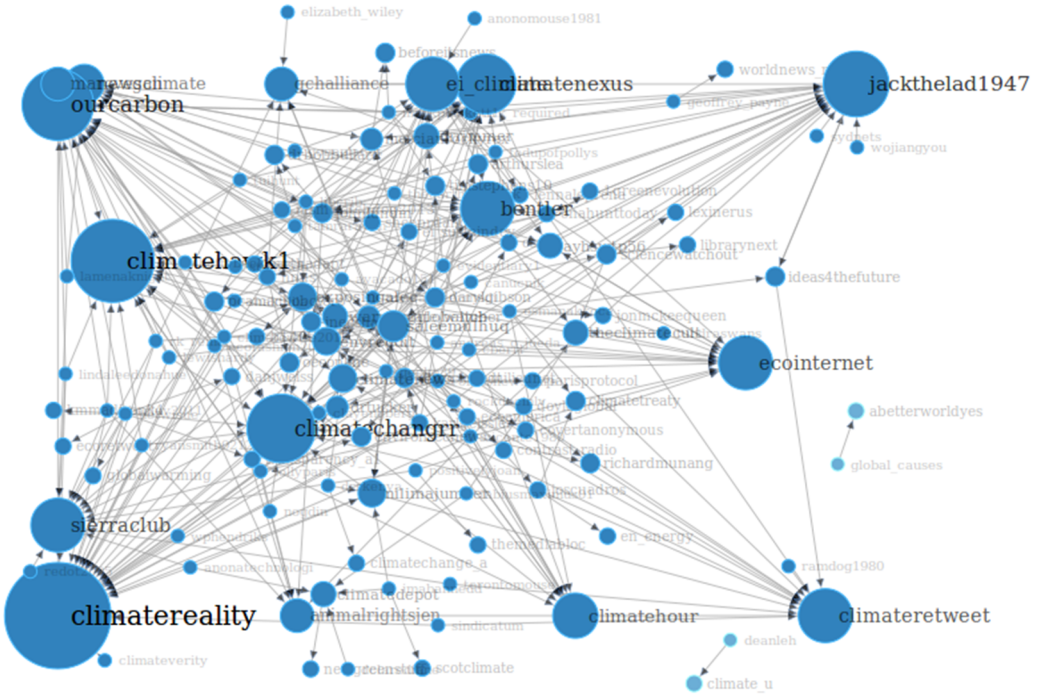


Figure 5. Node centrality on the Twitter earth hour data set for the topic 'climate change'; nodes represent users, edges represent user mentions

These larger node clusters in the centre of the graph have the highest data throughput where the data is shared among a maximum set of users in the network, which roughly correspond to high-volume discussions.

We can see from the graph that there exist few users with a very high node centrality in the diffusion network. Also worth noting is the connectedness of the overall graph, with one large connected component in the centre of the figure, and several isolated and small components around the edge of the figure. Figure 5 shows the diffusion graph visualization for the topic 'climate change' (also top 300 users). Here, the overall connectedness is even stronger than in the 'energy' topic graph in Figure 4.

Opinion Target Detection			**Related Work**		
	Based on JDPA Corpus	**Based on MPQA Corpus**	**(Zhuang et al., 2006)**	**(Kessler & Nicolov, 2009)**	**(Ginsca, 2012)**
Precision	0.9	0.9	0.48	0.75	0.86
Recall	0.89	0.92	0.59	0.65	0.88
F-measure	0.9	0.91	0.53	0.7	0.87

Table 1. Evaluation results of opinion target detection, including a comparison with reference systems reported in the literature

Opinion Holders vs. Opinion Targets

When identifying opinion leaders, the traditional notion of defining opinion leaders as "a minority of members in a society [who] possess qualities that make them exceptionally persuasive in spreading ideas to others." (Cha et al., 2010) requires adaptation before being applied to social media. The literature uses opinion leaders (Katz and Lazarsfeld, 1955) and innovators (Rogers 1962, Valente 1995) or hubs, connectors, mavens (Gladwell 2002), synonymously. The idea of a few highly persuasive individuals influencing large minorities has the advantage of reaching a wide audience by just dealing with a manageable quantity of individuals. However, applying this theoretical principle to social media can be misleading. Decision-making on social media is strongly peer-driven, i.e. people are highly responsive to the suggestions of their close peers and friends (Domingos and Richardson, 2001). Large-scale analysis of tweets empirically backed this theory (Cha et al., 2010) and showed that simple in-degree, i.e. the number of followers, serves only as a limited measurement to identify opinion leaders in networks. This calls for complementary strategies to investigate the drivers of public opinion, such as tracking the number of re-tweets or the number of mentions across various media channels.

For analysing the number of media mentions, recent advances in sentiment analysis allow connecting opinion targets with sentiment-expressing terms (see Deliverable D2.3.1). This goes beyond calculating the sentiment of sentences or documents. It enables us to determine who is talked about in conjunction with a specific topic, to extract what is being associated with this person, and to distinguish specific statements about an influential person from mere co-occurrences.

To identify opinion targets, a set of predefined features such as POS tag sequences or term proximity serve as input for a classifier. After evaluating those features, the classifier connects sentiment terms with their targets, if valid, and transfer the given polarity onto them. We plan to activate opinion target detection within the *Media Watch on Climate Change*, complementing the existing ability to analyse the statements of opinion holders across sources (news media outlets, social media platforms, etc.). Prior to this activation, we have conducted a formal evaluation based on standardized corpora. Table 1 presents the results of this evaluation, showing that the implemented system outperforms similar approaches reported in the literature. Zhuang et al (2006) evaluate their work on the well-known MPQA corpus (Wilson 2008), a collection of movie reviews. Kessler and Nicolov (2009) compile and use a first corpus of blog posts, the JDPA corpus, for evaluation, while Ginsca (2012) evaluate on the full JDPA corpus.

Summary and Outlook

This deliverable outlines the progress made on identifying and analysing information dissemination paths for news and social media content, collected as part of the *Earth Hour* use case (WP6). We have presented the steps undertaken to cluster stories, extract events, and to model users as important prerequisites for modelling diffusion processes in electronic networks. When computing the various diffusion metrics to enable further analyses, D4.3 has successfully overcome previous limitations of the content processing pipeline with respect to the developed network model.

Most notably, the deep integration of news media articles and social media postings into a unified data store via threaded dialogs required a major overhaul of the existing system architecture. This multi-channel integration enables diffusion processes to be tracked and analysed in a holistic manner, delivering a more complete snapshot of digital stakeholder communication – going beyond previous work that has largely ignored multi-channel processes, and often did not distinguish between opinion holders and opinion targets.

Another area where new grounds were covered is document fingerprinting based on event extraction from free text (among other document feature transforms), which allows us to define more meaningful features for clustering similar stories. Results have been showcased using network centrality visualizations, and comparative studies on opinion leader extraction.

Future work will compare LSH and HLDA clustering in terms of accuracy and throughput. We also plan to extend the developed diffusion networks to better capture emerging stories, since the diffusion networks, fed by real-time data of the *Media Watch on Climate Change* (WP3), are a core resource for the WP6 use case to investigate how the characteristics of participating users (e.g., node centrality) affect the spreading of a story during global events such as COP21 and Earth Hour 2016.

References

Bakshy, E., Hofman, J. M., Mason, W. A. & Watts, D. J. (2011). Everyone's an Influencer: Quantifying Influence on Twitter. In Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011.

Centola, D. (2010). The Spread of Behavior in an Online Social Network Experiment, Science, Issue 329, pages 1194-1197.

Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. P. (2010). Measuring User Influence in Twitter: The Million Follower Fallacy. In Proceedings of the Fourth International Conference on Weblogs and Social Media (ICWSM), Washington DC, USA, May 23-26.

Domingos, P., and Richardson, M. (2001). Mining the Network Value of Customers. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2001.

Freeman, L. C. (1977). A set of measures of centrality based upon betweenness. Sociometry 40, pages 35-41.

Freeman, L. C. (1979). Centrality in social networks: Conceptual clarification. Social Networks 1, pages 215-239.

Gladwell, M. (2003). The Tipping Point: How Little Things Can Make a Big Difference. in Journal of Consumer Marketing, 2003, pages 71-73.

Ginsca, A. L. (2012). Fine-Grained Opinion Mining as a Relation Classification Problem. In Imperial College Computing Student Workshop, volume 28, pages 56-61.

Gonzalez-Bailon, S., Borge-Holthoefer, J., Rivero, A., and Moreno, Y. (2011). The Dynamics of Protest Recruitment through an Online Network. Scientific Reports 1.

Granovetter, M. S. (1973). The Strength of Weak Ties. American Journal of Sociology, volume 78(6): 1360-1380.

Hanneman, R. A., and Riddle, M. (2005). Introduction to social network methods. Riverside, CA: University of California, Riverside http://www.faculty.ucr.edu/~hanneman/nettext/C10_Centrality.html

Hu, P., Huang, M., and Zhu, X. (2014). Exploring the Interactions of Storylines from Informative News Events. J. Comput. Sci. Technol., 29(3), pages 502-518.

Katz, E., and Lazarsfeld, P. (1955). Personal Influence: The Part Played by People in the Flow of Mass Communications. New York: The Free Press.

Kessler, J. S., and Nicolov, N. (2009). Targeting sentiment expressions through supervised ranking of linguistic configurations. In proceedings of the International Conference on Weblogs and Social Media, ICWSM.

Kessler, J. S., Eckert, M., Clark, L., and Nicolov, N. (2010). The ICWSM 2010 JDPA sentiment corpus for the automotive domain. In International AAAI Conference on Weblogs and Social Media Data Challenge Workshop.

Lu, M., Qin, Z., Cao. Y., Liu, Z., and Wang, M. (2014). Scalable news recommendation using multi-dimensional similarity and Jaccard-Kmeans clustering. Journal of Systems and Software, 95, pages 242-251.

Montalvo, S., Martnez, R., Fresno, V., and Delgado, A. D. (2015). Exploiting named entities for bilingual news clustering. JASIST, 66(2), pages 363-376.

Newman, M. E. J. (2003). A Measure of Betweenness Centrality Based on Random Walks, *arXiv:cond-mat/0309045*, www.arxiv.org/abs/cond-mat/0309045.

Petrovi, S., Osborne, M., and Lavrenko, V. (2012). Using paraphrases for improving first story detection in news and Twitter. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 338-346. Association for Computational Linguistics.

Rajaraman, A., and Ullman, J. D. (2011). Mining of Massive Datasets. Cambridge University Press, New York, NY, USA.

Rogers, E. M. (1962). Diffusion of Innovations. Free Press.

Valente, T. W. (1995). Network models of the diffusion of innovations (Vol. 2, No. 2). Cresskill, NJ: Hampton Press.

Valente, T. W., & Davis, R. L. (1999). Accelerating the diffusion of innovations using opinion leaders. Annals of the American Academy of Political and Social Science, *566*(1), pages 55-67.

Valente, T. W., Coronges, K, Lakon, C., and Costenbader, E. (2008). How Correlated Are Network Centrality Measures?, *Connections (Toronto, Ont.)* 28, no. 1, pages 16-26.

Valente, T. W. (2012). Network interventions. *Science*, *337*(6090), pages 49-53.

Wasserman, S., and Faust, K. (1994). *Social network analysis: Methods and applications* (Vol. 8). Cambridge university press.

Watts, D. (1999). Small Worlds: The Dynamics of Networks Between Order and Randomness. Princeton University Press.

Wilson, T. (2008). Fine-grained subjectivity and sentiment analysis: recognizing the intensity, polarity, and attitudes of private states, Chapter 7, ”Representing Attitudes and Targets”. Ph.D. Dissertation, University of Pittsburgh.

Wu, Z., Chen, L., and Giles, C. L. (2015). Storybase: Towards Building a Knowledge Base for News Events. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, ACL 2015, July 26-31, 2015, Beijing, China, System Demonstrations, pages 133-138.

Xia, Y., Tang, N., Hussain, A., and Cambria, E. (2015). Discriminative Bi-Term Topic Model for Headline-Based Social News Clustering. In The Twenty-Eighth International Flairs Conference.

Young, S. D., Belin, T. R., Klausner, J., & Valente, T. W. (2015). Methods for Measuring Diffusion of a Social Media-Based Health Intervention. *Social Networking*, *4*(02), 41.

Zhu, X., and Oates, T. (2013). Finding news story chains based on multi- dimensional event profiles. In Open research Areas in Information Retrieval, Lisbon, Portugal, pages 157-164.

Zhuang, L., Feng, J., and Zhu, X. (2006). Movie review mining and summarization. In Proceedings of the 15th ACM international conference on Information and knowledge management, pages 43-50.

Project Management Acronyms

Acronym	Description
CA	Consortium Agreement
DoW	Description of Work, i.e. GA - Annex I
EC	European Commission
GA	Grant Agreement
IP	Intellectual Property
IPR	Intellectual Property Rights
PC	Project Coordinator
PMB	Project Management Board
SC	Scientific Coordinator
PO	Project Officer
PSB	Project Steering Board
DM	Data Manager
AB	Advisory Board
WP	Work Package

Technical Acronyms

Acronym	Description
API	Application Programming Interface
CSV	Comma-Separated Values
FOAF	Friends of a Friend
EWRT	Extensible Web Retrieval Toolkit
URL	Uniform Resource Locator
XML	Extensible Markup Language

DecarboNet Consortium

The Open University
Walton Hall
Milton Keynes MK7 6AA
United Kingdom
Tel: +44 1908652907
Fax: +44 1908653169
Contact person: Jane Whild
E-mail: jane.whild@open.ac.uk

MODUL University Vienna
Am Kahlenberg 1
1190 Wien
Austria
Tel: +43 1320 3555 500
Fax: +43 1320 3555 903
Contact person: Arno Scharl
E-mail: scharl@modul.ac.at

University of Sheffield
Department of Computer Science
Regent Court, 211 Portobello St.
Sheffield S1 4DP
United Kingdom
Tel: +44 114 222 1930
Fax: +44 114 222 1810
Contact person: Kalina Bontcheva
E-mail: k.nontcheva@dcs.shef.ac.uk

Wirtschaftsuniversität Wien
Welthandelsplatz 1
1020 Wien
Austria
Tel: +43 31336 4756
Fax: +43 31336 774
Contact person: Kurt Hornik
E-mail: kurt.hornik@wu.ac.at

Waag Society
Piet Heinkade 181A
1019HC Amsterdam
The Netherlands
Tel: +31 20 557 98 14
Fax: +31 20 557 98 80
Contact person: Tom Demeyer
E-mail: tom@waag.org

WWF Schweiz
Hohlstrasse 110
8004 Zürich
Switzerland
+41 442972344
Contact person: Christoph Meili
E-mail: christoph.meili@wwf.ch

Green Energy Options
Main Street, 3 St Mary's Crt
Hardwick CB23 7QS
United Kingdom
+44 1223850210
+44 1223 850 211
Contact person: Simon Anderson
E-mail: simon@greenenergyoptions.co.uk

