Report from Dagstuhl Seminar 16141

Analysis, Interpretation and Benefit of User-Generated Data:Computer Science Meets Communication Studies

Edited by

Thorsten Quandt[1], German Shegalov[2], Helle Sjøvaag[3], and Gottfried Vossen[4]

1 **Universität Münster, DE,** thorsten.quandt@uni-muenster.de
2 **Twitter – San Francisco, US,** gera@twitter.com
3 **University of Bergen, NO,** helle.sjovaag@uib.no
4 **Universität Münster, DE,** vossen@wi.uni-muenster.de

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 16141 "Analysis, Interpretation and Benefit of User-Generated Data: Computer Science Meets Communication Studies".

Seminar April 3–8, 2016 – http://www.dagstuhl.de/16141
1998 ACM Subject Classification H. Information Systems
Keywords and phrases Communication Studies, Text Analysis, NLP, Text Mining, Topic Detection, Senitment Analysis, Machine Learning
Digital Object Identifier 10.4230/DagRep.6.4.1

1 Executive Summary

Thorsten Quandt
German Shegalov
Helle Sjøvaag
Gottfried Vossen

The success of the Internet as a communication technology and tool for human interaction in countless contexts, including production and trade, has had a dramatic impact on modern societies. With diffusion rates nearing one hundred percent in most societal groups, there is virtually no one whose life is not influenced by online communication – either directly or indirectly. Every day, private end users and business users act and interact online, producing immense amounts of data. Many disciplines, including computer science, computer linguistics, psychology, and communication studies, have identified 'big data' generated by online users as a research field. As a result, big data has become a somewhat over-hyped catch-all term for many different types of data, which are analyzed through varying methods for multiple purposes. This ranges from an analysis of (unstructured) Twitter or Facebook content to rule-structured texts as found in the professional media (i.e., news websites). The implication of value generated through sheer size of data sets is misleading, though – much of the value is based on the nature of these data sets as being user-generated, either on purpose or inevitably (and often unknowingly) as behavioral traces of actions with divergent aims.

Analysis, Interpretation and Benefit of User-Generated Data: Computer Science Meets Communication Studies, *Dagstuhl Reports*, Vol. 6, Issue 4, pp. 1–15
Editors: Thorsten Quandt, German Shegalov, Helle Sjøvaag, and Gottfried Vossen
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Big data sets generated by human users pose some challenges to the scientific disciplines that are interested in them: Typically, computer scientists have the knowledge and tools to access, extract and process big data sets. However, the analysis and interpretation of such data mirrors the interactions of users who produced the data and is not following a purely technological logic. In other words, such data has a human/social component, and in order to interpret and understand it, social-scientific theories and methods are helpful. Social scientists, however, typically do not specialize in the practicalities of online technologies and of programming. While they have theoretical approaches and empirical methods available that can be helpful in the analysis of user generated content – and this is especially true for communication scholars who specialize in the analysis of (online) media content –, their possibilities to access and process data are limited (as this is not core to their field yet).

Consequently, both disciplinary approaches will not be able to fully address the challenges of analyzing big data based on user (inter)action from the perspective of their own 'silo'. A combination of the two approaches seems fruitful, as each discipline may help in solving the problems of the other, and the sum will be more than its parts – leading to a better understanding of social interaction and human communication in a digitized world. This seminar will bring together both computer scientists interested in the analysis of (large-scale) user-generated data, and communication scholars interested in computer-assisted acquisition and processing of such data. It is intended to start a fruitful dialogue on potential approaches, methods, uses and benefits of a cooperation between the two disciplines, and it will also include the input of practitioners in the field of media and business who will offer valuable insights into practical use cases.

2 Table of Contents

Executive Summary

Working groups

Panel discussions

3 Working groups

3.1 Analyzing Text Microstructures

Christian Baden (The Hebrew University of Jerusalem, IL) and Tatjana Scheffler (Universität Potsdam, DE)

In this working group session, we discussed three major challenges in analyzing text microstructures.

Challenge 1: Finding frames

- We use the same techniques for allegedly theoretically different purposes: LDA "finds frames" or "finds topics" depending on what we want, but really it is one technique, so this does not make sense. Frames and topics should generally be orthogonal, such that one topic can take different frames, and one frame can take different topics.
- There are existing strategies to run a topic extraction tool, then control for that, and call all further patterns frames; this works but is theoretically unsatisfying. Also, other types of structures have been analyzed in this way (speech acts, event structure, etc.). Frames have a specific theoretical structure (four frame elements, following Entman):

```
          Evaluation
              |
Cause - Focal Concern - Projection/Treatment
```

- We discussed an idea based on this structure – use this structure to construct frames from texts:
  1. use topical text contents (headlines, lead paragraphs) to identify a (narrow) topic
  2. structure remaining textual contents
  3. use syntax, sequence, connectors, grammatical information (whatever is useful/available) to figure out which other contents are related how to the focal topic

Challenge 1b: Focusing topic models (or other similar techniques)

- Models tend to perform better if more textual content is excluded (even up to excluding verbs).
- Possible cause seem to be stylistic differences.
- This is also highly dependent on the size of documents one analyses – aggregating individual texts into larger "documents" leads to different results in topic models.
- Such insights might help focus pattern finding algorithms on different text properties: topical (if highly reduced), stylistic, etc.
- There are also formal solutions, where one can run topic models in a multilevel logic: annotating authors, they can be trained to disregard author-specific variation and focus on content differences; and there is no reason why one can do this only for partialing out author-related variance, might be a great strategy for comparative research.

Challenge 2: Similarity of texts

- Determining the similarity of text has many uses: deduplication, detecting taken-over materials, assessing diversity, etc.; however, many existing approaches are unclear about what exactly they mean, as they depend on features of the text whose theoretical relevance is heterogeneous and/or unclear.

- Four basic types of similarity that are of interest: literally identical sequences (quotes, plagiarism, unedited text); content similarity (the same topic and arguments), stylistic (the same way of expressing content, sentence style, etc), and sentiment.
- Furthermore, it may be relevant to assess similarity despite different languages; there is rich work in machine translation, evaluating translation quality based on these above dimensions, and also providing avenues for interlingual comparison; however, machine translation increases comparison error in ways not yet sufficiently understood.
- Evaluation methods developed for machine translation (e.g., similarity to a (set of) reference translation(s)) may also be used in a monolingual setting to determine the similarity of sentences/texts.
- Problems related to challenge 2:

  (1) Matching (finding out which of very many possible pairs are worth comparing/have relevant similarity); for this, using meta data may be useful to consider only likely combinations (date, for instance). Parallelization is probably desirable, to avoid memory problems. In addition, internal text structure (e.g., sections, zoning, substructure of journalistic texts) can be used to presegment documents for comparison of smaller chunks. Possible detection problems:

  - Some kinds of texts (e.g., sports reporting) are structurally very similar because there are just few ways of saying something; can also be considered valid result though.
  - Similarity often does not concern entire texts, but a text may be similar to a part of another text only, or both only overlap in particular paragraphs; so text-level similarity scores might be too crude
  - However, there is a specific interest in determining what are possible elaborations/truncations of texts, so both determining if there are similar passages, and determining what is different are important.
  - The shorter the text the more likely does one find similar others just by chance.

  (2) Measuring/scoring the similarity. For each kind, there are good algorithms existing that can be developed and applied: for literal identity, plagiarism checkers; for contentual similarity, comparisons of extracted entities (rank sum, bags of words strategies, Jaccard, etc); for stylistic, bag of words & linguistic resources; for sentiment, new generation sentiment measures that take into account differentiated scores and intensifiers/negations, also some machine learning approaches. There may be a point in keeping different kinds of similarity apart and finding typical patterns of similarity (e.g., high contentual low literal/stylistic similarity → paraphrase, some literal, low sentiment → quote/challenge, etc.).

Challenge 3: Metaphors

- Metaphors matter both for sentiment (they have evaluative implications) and for framing (they structure content), but are difficult to find. Existing approaches are deductive, domain specific, and laborious, and they still detect a lot of cases that are actually literal uses, not metaphors.
- One idea to solve this: Use related content (e.g., Wikipedia, dictionaries) to determine if a word that might be a metaphor is used in a context related to its definition, or out of such context (so it is probably metaphorical).

3.2 Visions for the Computer-Assisted Identification, Analysis and Evaluation of Texts

Christian Baden (The Hebrew University of Jerusalem, IL)

The group identified three main areas of basic challenges for the computerized treatment of texts:

1. accessing texts, from various archives, via scraping, apis, etc.
2. curating texts/evaluating the quality of repositories (formatting, standardization, annotation/metadata, api transparency, etc.)
3. analyzing text (henceforth the focus of discussion)

The third challenge was discussed as being the one of upmost important, and the group approached it from two angles:

Angle 1: Finding entities & patterns in text

- There is a great potential for automation, however, fully supervised approaches are very labour intensive, while unsupervised approaches are hard to trust/use/get published. In essence, there is an urgent need for transparency (what does it do) and intellegibility (ability to theoretically evaluate the rules).
- One idea that was favored by the group was an alternative process to simplistic automatic content analysis: We propose to use machine learning not to solve the task, but to propose additional indicators and possible additional rules. The process is as follows: (1) start from set of indicators/rules and train ML to find other contents of similar or related kind; (2) generate rules that, if applied, improve performance; (3) return to human in a format that can be understood/evaluated, and if confirmed, integrate into model (iterative semi-supervised approach). This process can be applied to various problems: language fuzziness (find possible typo/variants/synonyms), entity extraction (find additional names), pattern recognition (find additional related components), etc. Another advantage is that the procedure can explicate detection rules, so we can learn not only what is in the text, but also what are underlying structures of discourse useful for analysis.
- The process of finding additional contents of similar kind can also be used to augment/-contextualize/evaluate information. Finding other texts about the same event, or other pictures of the same thing, etc. might be useful to augment journalists' information base, criticize one-sided information, detect contested information, check veracity, etc.
- One further extension is potentially interesting: Like in knowledge graphs, additional information available in discourse can be integrated by linking entities to online resources (e.g., Wikipedia, dictionaries, prior discourse) for elaboration and classification ('active intelligence'/intelligent classification/generation of background information)-

Angle 2: Doing this collaboratively

- There are lots of tools and approaches out there, but little collaboration. This leads to problems in findability of tools, documentation, standardization (of approaches, exchange formats, etc), and also referencing/crediting and related incentive systems. In short, many researchers are solving the same problems in parallel, and make the same mistakes in parallel, instead of working together.
- One idea to solve this is to build an infrastructure that facilitates collaboration (github ++): This should not be only about archiving/sorting/finding tools (possibly with some

mechanism for identifying if existing projects look related to the one you are currently working on, suggesting code/tools), but also about rendering the collaboration and use visible (so one can show that one's tool is useful, gain credit, and get references for developed tools, which makes this worthwhile career-wise).

- Furthermore, and related to this, there is a need for more interdisciplinary education between communication and computer science (iSchools, data science, communication programs that train computational skills/computer science programs that relate to social science research methodology/applications).

Overall, the group came to one central conclusion:

What is needed is not an integrated catchall solution using fancy maths and big red buttons, but an assortment of tools specialized to capture specific ingredients of social scientific concepts, which are well described, allow human intervention, and generate output formats that can be integrated analytically (so, no automated-frame-finder but tools extracting entity classes, relation classes, stylistic contents, etc.).

3.3 Interactions between Computer Science (CS) and Journalism (Studies) in the Future

David Domingo (Free University of Brussels, BE), Johann-Christoph Freytag (HU Berlin, DE), Ari Heinonen (University of Tampere, FI), and Rodrigo Zamith (University of Massachusetts – Amherst, US)

In this working group session, we discussed various questions on the potential interactions between computer science and journalism (studies). The discussion revealed that some parts of journalistic work can be substituted by computers and robots, some others cannot. However, we found that the discussion about "substitution" is misleading, as the new configurations of information distribtution will require both humans and computers, and that it's not about competition, but about new forms of journalistic work.

1. Can computer systems substitute human journalists?

   (a) If journalism is a filter between events/news and citizens/consumers then an algorithm could do the filtering task. But journalists may still be better filters, based on their intuition, judgments, and a more global view on events and their relationships. Challenge for CS: develop more complex and comprehensive algorithms/methods for performing filtering (almost) like journalists. If computer systems achieve that level of reliable filtering, journalists could focus on other, "more interesting" tasks. Already happening in some areas of journalism, such as weather/sports reports (robot journalism). However, journalists may be more than just filters; they are also "sense makers" that bring information together. In theory, computers may also be able to perform that task.

   (b) Currently not all events/information on events is digital and available online. Therefore, sensors (such as cameras), which are currently used for surveillance, could also be used for capturing events that can then be evaluated and filtered by humans and/or machines. Challenge: how to get more sensors integrated into event-generating networks? Risk: event-generating networks might also be used for other purposes than news generation, with a less democratic goal.

(c) Computers could produce more efficient multi-dimensional news reports that show the information more like a process rather than and "end product". In this way we can better represent/conserve the complexity of reality and make the process of news report generation more transparent (tracing the provenance of information, see below).

2. Some deeper reflection on Aspect 1: We may not be able to substitute human journalist with computer systems completely with current technology, neither may it be desirable due to possible manipulations of automated systems. Journalist may in any case still be needed as safeguards of the process of news generation.

3. We can also improve journalistic tasks with computer-based systems, without substituting humans with robots. We developed two ideas in this direction:

   (a) Enriching news reports with information about the newsgathering process. This could be done by semi-automatically logging actions, documents and sources that are used during that process, thus making it more transparent for consumers and other journalists. One of the practical ways to give access to the newsgathering log data is to link it to individual elements of news reports. (For example, automatically storing a list of the documents a journalist accessed or keywords used in searches, and allowing the journalist to select the trace data to make public.)

   (b) Improving journalistic memory by better structuring news archives with time series and algorithmic calculations, thus allowing to answer searches and queries with time dimension, showing the evolution of actors, topics, contexts. Example: how has the relationship between Syria and the US changed over the last 20 years? Algorithm could highlight the sentiment in reports of relations between main actors and the topics usually discussed in those reports, presenting it longitudinally as a timeline.

4. In a more near term it is advantageous/desirable to simplify the interaction between the journalist (or journalism scholar) and the set of computer tools that he/she uses. (Put differently: make the current state of the art more accessible to end-users.) Using the paradigm of SQL as a declarative language, it could be possible for the user to simply express what is the desired outcome of an analytical process, together with possible sources and filters/constraints, thus freeing the uses of technical cumbersome details about the algorithms and methods used during that process. (For example: SELECT NamedEntities FROM doclist AND SENTIMENTANALYZE (NamedEntities. Obama AND NamedEntities.Putin) AND ExtractTopics. That query would automatically apply an NER tool and Topic Modeling tool to extract information from a set of unstructured documents and save it in as elements in a database.) At the same time, users may have the option of exploring the tasks/steps that the computer system may apply, in order to give more experienced/advanced users the ability to fine-tune the analytical process, or to let new users understand the operations that are being performed.

3.4 What is not there yet? Dreams and Visions

Martin Emmer (FU Berlin, DE), Elisabeth Günther (Universität Münster, DE), Wiebke Loosen (Hans-Bredow-Institut – Hamburg, DE), Alexander Löser (Beuth Hochschule für Technik – Berlin, DE), and Gottfried Vossen (Universität Münster, DE)

The group discussed several "levels" of visions.

Vision level 1: It would be desirable to make the differences between "forums" of public debate visible: like comment sections of tabloids vs. quality papers, or the papers itself. Furthermore, one might want to look for argument structures, types of authors, audiences etc. The goal here would be a sensor for "public opinion", delivering data that can be compared to results of public opinion surveys. This would offer further insight in public communication processes.

Vision level 2: The second 'level' is lifting the first approach to a more global/macro level, as the group identified one major challenge today: organizing political debates under conditions of extreme speed, heterogeneity, and ambiguity (i.e. fragmentation, filter bubbles, increasing masses of information). So it is not a small set of tv-news and quality papers that organizes this discussion anymore.

Based on this, the overall goals would be a system that analyzes the mediated public sphere online, in order to provide the society – citizens AND elites – with information about the issues, arguments and opinions that are currently debated in society.

Such a system should have various features and functions:

- It should make the public debate "visible" und understandable.
- It may be designed as central platform (liquid democracy) – or maybe as distributed, self-organizing systems?
- The role of state remained unclear. The group agreed that there should be a separation of government and media. So the group argued for a self-organizing system.
- Data protection: It remained disputed what info to include in analysis and presentation of data.
- Public broadcasting: Maybe there could be a new role for this type of actor.
- Fact checking should be included.
- All types of media should be included as well: text, video, pictures.

What form could a project like that have? The group collected various features:

- assistant system for journalists, giving overview over state of debate, facts
- app, usable by everybody
- bot that participates in debates, enriching it
- target groups: reaching highly-involved and less interested citizens at the same time
- low threshold-strategy to get many people using the system
- dealing with problems: user selectivity, instrumental use of results etc.

Vision level 3: Finally, we added a global and ethical dimension, asking: Can such a system be used in multiple countries (Europe)? The discussion revolved around two aspects:

- Would it be useful to build such a system in hardware? This would allow democratic values to be encapsulated in a system. Having independent systems for Europe in order to secure data security would be crucial, then.

- Dystopian pictures of future (science-fiction) remained: Often, we refer to capitalism as the main agent of acceleration and content multiplication. So we asked: Are there possible features to de-accelerate debates?

3.5 Methods on Obtaining Curated Data (such as for trend detection, or for understanding social problems, for power-law distributed data)

Alexander Löser (Beuth Hochschule für Technik – Berlin, DE)

The group identified two differing goals for the two disciplinary fields involved in the workshop:

1. Computer studies: Train a smart machine (super intelligence) that does a task (spotting terrorists, products, winning strategies for playing "GO").
2. Social sciences: Learn from human behavior and abstract it into a report.

Several methods in the two fields were identified by the group:

A. Observing + Transforming (done in most cases by CS people) This includes text mining from samples (https://aclweb.org/anthology/), but also transforming image representation into text, transforming tables into text (robot journalism).

B. Asking people/Survey methods This includes micro-task crowd sourcing and active learning (for sampling strategies see http://burrsettles.com/pub/settles.activelearning.pdf)

C. Controlled Experiments There is a huge body of research in SS, often ignored by CS, because they set up experiments which are focused around machines. Furthermore, there is a lot of potential in "Games with a purpose". There are some issues to be solved with that approach, though. One major question is often not answered: "What is the stimulus?" The preferential method here is to eliminate all other factors that might obfuscate the outcome of your experiments. Additional problems may arise from fatigue (one solution may be taking more people, but avoiding long game time).

D. Simulation This may include creating a machine that is "creatively" creating curated, labeled data (Dynamic programming, Monte Carlo simulations).

E. Ensemble methods At the end, the major solution may be learning an ensemble from these methods (Boosting, Bagging, DNN:CNN, RNNs or LSTMs?....) and iterate (go back to data sampling and curation) until "good enough".

3.6 Funding Workshop

Helle Sjøvaag (University of Bergen, NO)

The last workshop focused on funding schemes for possible joint future applications. Horizon 2020 and ERC were discussed as EU funding schemes. Other schemes mentioned include EURA (collaboration between certain EU countries); COST Actions (networking scheme); RISE (new funding for research exchange); UNESCO (but this is policy oriented). In the US, foundations are the most likely source of funding, including the Knight Foundation, The

Democracy Fund, Google, The Spence Foundation, TOW Centre, and Reinhold's Journalism Institute. Other funding schemes include the Dutch Press Fund, Tekkis, a Finnish agency. The question of industry funding was also raised.

3.7 Workshop on Data Journalism

Helle Sjøvaag (University of Bergen, NO)

The breakout workshop on data journalism met in three sessions during the week. Martin Emmer (FU Berlin, DE),Gottfried Vossen (Universität Münster, DE), Seth C. Lewis (University of Oregon, US), Rodrigo Zamith (University of Massachusetts – Amherst, US), amian Trilling (University of Amsterdam, NL), Ralf Schenkel (Universität Trier, DE), Ari Heinonen (University of Tampere, FI) , Jukka Huhtamäki (Tampere University of Technology, FI), Raul Ferrer Conill (Karlstad University, SE) and Helle Sjøvaag (University of Bergen, NO) were the core participants.

Data journalism started as data assisted reporting in the 1960s and has been described as precision journalism or data assisted journalism. It involves journalism practice using social science methods, databases, and using data to do journalism. The data used to do data journalism is typically government data, leaks, and open data. Sometimes this data comes in unmanageable form, such as PDFs or even printouts. Data journalism today is primarily practiced in big newsrooms with the resources to allocate staff to data journalism processes, such as design, data science, statistics and journalism. Because of the nature of the data, data journalism typically involves scraping and visual analytics, and the work frequently requires teamwork.

Challenges to data journalism include acquiring the skills needed to handle tools for data journalism research and presentation. Furthermore, most data journalism projects are ad-hoc projects, with few reproducible workflows. Hence, contingency in data storage and scalable workflow models is a problem. For journalists, the challenge is how to better turn unstructured documents into structured documents. For research the challenge is to look beyond the text as object of study. Data journalism is more than text, which challenges the way we look at societal communication. For journalists and researchers alike, a common challenge is how to treat journalism as data over a large repository beyond archiving, as semantic networks. Part of a solution to this problem is to create transparent workflows.

The discussion in the workshop developed into an effort to establish a collaborative research design for a project on data journalism, based on the interdisciplinarity necessitated by the research object: as visuals, background data, data bases, hidden or licensed data, and text. To study data journalism, the tools as well as the analysis object requires a mixture of social science and computer science approaches. Conceptually, thinking about data journalism in computer science concepts will better facilitate a research design. By using the computer science workflow approach, we then broke the data journalism process into a reference process model, from which the methodology can be derived. And as data journalism is largely about application development, the empirical focus includes both practice (workflow) and 'text' (data).

The data journalism sessions resulted in a rudimentary research design, a collaborative document from which the project can be further developed, and allocation of project leadership.

3.8 Workshop on Methods

Helle Sjøvaag (University of Bergen, NO) and Thorsten Quandt (Universität Münster, DE)

This joint session revolved around mapping computer science methods appropriate for communication science research. The methods were divided into three strands:

1. Methods for data access/access to sources/data clearing
2. Language based methods, sentiment analysis, NLP
3. Methods for relation analysis: pattern detection, temporal flows

A collaborative table overviewing the available tools was created. What emerged from the discussion is not only a list of available approaches, but also that most of us using hybrid analysis approaches use the same tools, or the same types of tools. As most researchers need to write their own scripts for scraping, sharing these workflows for static content would be a good idea. The issue of hiring companies to extract the information needed in communication science research was raised, to which the black box problem was discussed.

3.9 Workshop on Relation Analysis

Helle Sjøvaag (University of Bergen, NO)

The session included Thorsten Quandt (Universität Münster, DE), David Domingo (Free University of Brussels, BE), Martin Emmer (FU Berlin, DE) , ohann-Christoph Freytag (HU Berlin, DE), Jukka Huhtamäki (Tampere University of Technology, FI), Alexander Löser (Beuth Hochschule für Technik – Berlin, DE), Ralf Schenkel (Universität Trier, DE), Gera Shegalov (Twitter – San Francisco, US), Helle Sjøvaag (University of Bergen, NO), Hendrik Stange (Fraunhofer IAIS – St. Augustin, DE), Martin Theobald (Universität Ulm, DE) and Gottfried Vossen (Universität Münster, DE).

The session revolved around analysis of relations between actors in a textual context. For most research in social science, an actor is an individual person or persons that act in a coordinated way. An institution can be an actor, but an actor can also be a technology – an algorithm, for instance. Actors are actors because they do things – or the actors interact. What is of interest to social scientists is what happens when actors interact. In texts, actors in the text are also actors (e.g. politicians and countries can act). An actor is someone who has an intention, has agency. In computer science, actor means something else. One can have actor-based computations, programs that exhibit certain behavior, for instance programs that are self-regulating, self-organizing, can take in and send out messages, modeling dynamic behavior (in the macro aspect). Computer science calls actors AGENTS.

Actor network theory can be useful in looking at networks. Methods involving pattern detection, network analysis, similarity measures, time and dynamics can be used for analyzing relations/relationships in networks. Networks are useful to measure distance, so distance measures must be defined in the research design. The closeness of nodes can visually represent this. Or size can carry information. There are two levels of networks metrics: network density, how many of the possible connections exist in a network; and the structural position of a node in a network. Network analyses are node centric, focused on betweeness.

In the 'mental model' networks consist of several layers (questions, representations, relationships), to which is needed approximation through nodes. Distance and position are relational, while boundaries require fixity to establish a starting point to further establish centrality, density, and activity in the network. Hence, network analyses consist of metrics that quantify structural properties. Patterns in networks are communities or clusters – sub-graphs with certain properties. Patterns can be detected through clustering measures (distance for instance), identifying where clusters emerge, representing activity. Patterns can be associations that are repeating or in combination, over time, like sequences. Technical tools to collect, analyze and visualize networks include Gephi, Snappi.pi, NetworkX, and Node XL, which is an excel extension (from the work of Mark Smith). Similarity in networks indicates connection.

The group also discussed concepts like homophily, strength of weak ties, assortativity, connection principles, page rank, and bias, sampling, labeling and classification, training data, as well as ethical issues. Large part of the discussion used the problem of finding terrorists as illustrating vantage point. This angle spurred topics such as outliners and black swans, the Go Game, deep learning, the Cynefin model, gamification as incentive mechanisms to create training data, and survey design.

3.10 Workshop on What is not there yet?

Helle Sjøvaag (University of Bergen, NO)

The breakout session consisted of Ralf Schenkel (Universität Trier, DE), Damian Trilling (University of Amsterdam, NL) , Tatjana Scheffler (Universität Potsdam, DE) and Helle Sjøvaag (University of Bergen, NO). The overarching questions revolved around imaginative futures for how computer science can contribute to communication science. In terms of impossible futures, the group outlined five areas for future developments:

1. Comparative, multilingual framing analysis;
2. Diverse recommender systems;
3. Automatic validation of fact statements;
4. Speech/video to text; and
5. Bot-human interaction (automated communication).

Multilingual comparative framing analysis was the most concrete envisioned future. The 'dream scenario' would involve a system that can map different frames (aspects of a story) of the same topic across outlets, for instance to look for diversity. This would involve computational tools that could a) identify events; b) track events; c) identify similar and diverse sources; d) group sources/themes/events into different perspective and/or link the elements; to f) predict future events. This process entails automatic translation of multiple languages, understanding different argument structures that require language independent NLP. Developments in AI/deep learning in combination with big data could serve to fulfill these dream scenarios of communication scientists.

4 Panel discussions

4.1 Data Access

Thorsten Quandt (Universität Münster, DE)

The group was discussing typcial problems connected to data access. We found that most of the participants very using some form of scraper, extracting (mostly) textual information from unstructured web resources. Based on the shared experiences, we tried to systematize the various access types and formats in an overview table, primarily populated by what has been done already by the workshop participants (pdf is attached to the report).

Participants

- Christian Baden
The Hebrew University of Jerusalem, IL
- David Domingo
Free University of Brussels, BE
- Martin Emmer
FU Berlin, DE
- Raul Ferrer Conill
Karlstad University, SE
- Johann-Christoph Freytag
HU Berlin, DE
- Elisabeth Günther
Universität Münster, DE
- Krishna P. Gummadi
MPI-SWS – Saarbrücken, DE
- Ari Heinonen
University of Tampere, FI
- Jukka Huhtamäki
Tampere University of Technology, FI
- Seth C. Lewis
University of Oregon, US
- Alexander Löser
Beuth Hochschule für Technik – Berlin, DE
- Wiebke Loosen
Hans-Bredow-Institut – Hamburg, DE
- Truls Pedersen
University of Bergen, NO
- Thorsten Quandt
Universität Münster, DE
- Tatjana Scheffler
Universität Potsdam, DE
- Ralf Schenkel
Universität Trier, DE
- German Shegalov
Twitter – San Francisco, US
- Helle Sjøvaag
University of Bergen, NO
- Hendrik Stange
Fraunhofer IAIS – St. Augustin, DE
- Eirik Stavelin
University of Bergen, NO
- Martin Theobald
Universität Ulm, DE
- Heike Trautmann
Universität Münster, DE
- Damian Trilling
University of Amsterdam, NL
- Gottfried Vossen
Universität Münster, DE
- Rodrigo Zamith
University of Massachusetts – Amherst, US