Linköping University Post Print

**Text-based Analysis for Command and Control
Researchers: The Workflow Visualizer
Approach**

Ola Leifler and Henrik Eriksson

N.B.: When citing this work, cite the original article.

The original publication is available at www.springerlink.com:

Ola Leifler and Henrik Eriksson, Text-based Analysis for Command and Control Researchers:
The Workflow Visualizer Approach, 2011, Cognition, Technology & Work.

<http://dx.doi.org/10.1007/s10111-010-0170-3>

Copyright: Springer Science Business Media

<http://www.springerlink.com/>

Postprint available at: Linköping University Electronic Press

<http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-67229>

Automated Text-based Analysis for Decision-Making Research

The Workflow Visualizer Approach

Ola Leifler · Henrik Eriksson

Abstract We present results from a study on constructing and evaluating a support tool for the extraction of patterns in distributed decision making processes, based on design criteria elicited from a study on the work process involved in studying such decision making. Specifically, we devised and evaluated an analysis tool for C² researchers who study simulated decision-making scenarios for command teams. The analysis tool used text clustering as an underlying pattern extraction technique, and was evaluated together with C² researchers in a workshop to establish whether the design criteria were valid and the approach taken with the analysis tool was sound.

Design criteria elicited from an earlier study with researchers (open-endedness and transparency) were highly consistent with the results from the workshop. Specifically, evaluation results indicate that successful deployment of advanced analysis tools require that tools can treat multiple data sources and offer rich opportunities for manipulation and interaction (open-endedness) and careful design of visual presentations and explanations of the techniques used (transparency). Finally, the results point to the high relevance and promise of using text clustering as a support for analysis of C² data.

Keywords command and control, text analysis, text clustering, exploratory sequential data analysis, text clustering

1 Introduction

When studying distributed decision-making, there are great technical challenges with setting up the proper instrumentation and subsequently collecting sufficient data to enable the study of the selected aspect of the decision-making process (Andriole, 1989). Furthermore, there are ontological challenges involved as distributed decision making as performed in military command and control (C²) can be understood as a continuous process for decision making (Orasanu and Conolly, 1993; Rasmussen, 1993), a process for sensing the environment (Alberts and Hayes, 2003), a joint cognitive system integrating people and machines (Hollnagel and Woods, 2005), a system for distributing functions among actors (Hollnagel and Bye, 2000), a system for communicating intent (Shattuck and Woods, 2000), a structured workflow among a set of actors (van der Aalst and van Hee, 2002), or in terms of the specific psycho-social aspects of a command team (Brown, 1993; Argyle, 1972). Depending on the perspective, different methods and tools are required to understand staff work and evaluate tool support.

Although all of the perspectives offered by all these approaches to studying and explaining command team behavior may have their distinct advantages for understanding distributed decision making in general and command and control in particular, the abundance of theoretical concepts for understanding C² may in itself present a problem for the evaluation of the effectiveness of intelligent support systems for analysis.

1.1 Research method

The study in this paper has been conducted in two parts in which we designed a support tool based on automatic pattern extraction and evaluated the resulting tool in a workshop.

1. The first part consisted of creating a specific support tool for the task of selecting information from large data sets by using automatic pattern extraction techniques. Most data sources used in the study of distributed decision making such as command and control are represented as text directly or transformed into text through the transcription of speech. Thus, using text-based classification of messages was considered an viable option for filtering the datasets used in analysis. Based on earlier work on the feasibility of using automatic pattern extraction in texts for supporting data exploration (Leifler and Eriksson, 2010b), and design criteria elicited from interviews with researchers in command and control (Leifler and Eriksson, 2010a), we designed a prototype tool to create patterns from texts.
2. The second part consisted of a workshop during which we presented a working prototype of our text clustering-based support tool to participants in the study and had discussions concerning the affordances of such a tool in their work. This discussion resulted in conclusions regarding the conditions for successfully deploying pattern extraction techniques in C² exploration tools.

1.2 Outline

The remainder of this paper is organized as follows: Section 2 provides a background on the types of data analysis tasks and tools that are in use for generating and verifying hypotheses in the research settings we have studied. In Section 3, we describe the design of the *Workflow Visualizer* approach to explore relationships in text-based data sets. We describe the affordances of the tool in two use cases that were constructed using authentic scenario data studied by the participants in our study. Section 5 presents the results from a workshop evaluation where the *Workflow Visualizer* was evaluated with respect to the possible applications of the tool and the technique it embodies, Section 6 puts the design and the results of the evaluation in context and Section 7 concludes this paper with conditions for successful deployment of automated clustering as support for the analysis of distributed decision making.

2 Background

The field of decision making has seen a paradigm shift in the last fifteen years (Cannon-Bowers et al., 1996) where new appreciations of what characterizes decision-making has resulted in an increased interest in methods for studying joint, distributed decision making scenarios in naturalistic settings such as those involved in military command and control and crisis management.

Fig. 1 Four stages of qualitative data analysis according to Miles and Huberman (1994).

To measure the key functions in decision making, and consequently measure the effects of those functions in naturalistic settings, researchers have developed and adopted methods to balance the rigor of laboratory experiments with the relevance of field studies. In the research projects we have studied, two such methods have been in use: *role-playing simulations* (Rubel, 2001) and *micro-world simulations* (e.g., (Wærn and Cañas, 2003; Johansson et al., 2003)), both of which imply certain characteristics of the data sets generates and require specific methods of analysis.

Prior to and during the analysis of simulation-based scenarios, researchers typically follow a set of procedures for determining appropriate instrumentation and analysis tools (Morin, 2002). As the data sets are usually rather large and diverse, with both observer logs, simulation logs, communication data, screen recordings and possibly even video recordings of the participants, researchers must use exploratory, data-driven methods of analysis in their studies. They may be either qualitative or quantitative as regards the end results, but primarily data-driven.

2.1 Data analysis methods

In the characterization of qualitative data analysis by Miles and Huberman (1994), there are basically four stages: the initial collection of data, which leads to three stages that can be performed iteratively: data reduction, data display, and conclusions drawing (Fig. 1). For researchers to know what data to focus on, they need some means of visualizing possible connections in data, and for them to produce such visualizations from large sets of data, they typically need to perform some a priori reductions of the data set. Following data reduction and data display comes a stage of drawing conclusions and verifying results, which typically comes at the end of an analysis phase.

Another description of data-driven research which captures many central aspects of distributed decision-making research is exploratory, sequential data analysis (ESDA), (Sanderson and Fisher, 1994) which describes data collec-

tion and analysis as a process of iteratively selecting parts of the available data sets (typically logs and recordings), conducting analyses of the transformed products that are derived from the data sets (typically transcribed speech and annotated events) and finally using the results to guide the selection of another set of data to study closer until conclusions can be drawn.

The ESDA process of data-driven hypothesis generation closely resembles one from the data mining community, where Cutting has characterized the process as iterating between scattering and gathering (Cutting et al., 1992). In text mining community, Rosell and Velupillai (2008) has proposed to describe data analysis when studying texts as consisting of four stages of analysis, with our interpretations of how to reconcile this description with Cutting's in parentheses:

1. Cluster the text set (scatter)
2. Identify interesting clusters (gather)
3. Explore cluster contents (scatter)
4. Formulate potential hypotheses (gather)

Also, both the scatter/gather paradigm and the four stages of Rosell and Velupillai (2008) encompass the iterated work process in ESDA and qualitative data analysis by Miles well: when analyzing logs from a team scenario, researchers may use transcribed communications as an entry point to further analysis and annotate the transcribed text according to a certain annotation schema. The annotation schema creates a set of distinct objects of study in the form of a set of episodes, which in turn direct further analysis by making video logs or observer reports at specific points in time relevant to study. The four stages of Rosell and Velupillai could similarly be described as re-formulations of the stages in qualitative data analysis introduced by Miles.

Due to the similarities between the paradigm of data mining research and data analysis in decision making research, we had previously conducted a study on using data mining methods in text analysis (Leifler and Eriksson, 2010b). In our evaluations, we found that of all the meta-data found in communications between members of a group of decision makers, the message texts were the most significant factors when attempting to emulate human classifications by machine classification. Also, automatic text classification emulated human classifications well enough that we believed it to be justified to incorporate it in a support tool for scenario analysis.

2.2 Data analysis tools

In qualitative content analysis, researchers seek to categorize data from interviews and other sources in a common framework. The framework can either be taken from previous literature from the research field the study concerns

Fig. 2 An overview of F-REX, an ESDA tool used for analysis of C² scenarios by the participants in the study.

Fig. 3 An overlay of the tool support provided by ESDA tools similar to F-REX on the four stages of qualitative data analysis by Miles.

(a priori coding), or be developed as part of the analysis (emergent coding) (Lazar et al., 2010). Especially when an emergent coding is needed for analyzing data, it can be very labor-intensive. To support the coding and analysis of communication data from command teams, ESDA tools can be used for understanding patterns in the sequence of messages exchanged between the members of a group.

Some researchers use ESDA tools for merging and viewing many different data sources. Figure 2 displays some of the capabilities of one such tool, F-REX, a re-implementation of the earlier MIND system (Thorstensson et al., 2001). In MIND and similar ESDA support tools, a number of data sources are made available as a series of events along a common scenario timeline. For every scenario, particular configurations can be made to emphasize a particular data source of importance to analysis by providing a specific layout of the graphical components. In the middle of Figure 2, a screenshot displays how screen-captured video, radio communications, text messages and other data sources are available through a graphical interface with a timeline at the bottom.

If we compare the work process of qualitative data analysis presented by Miles with the affordances of an ESDA tool such as F-REX/MIND (Figure 3), there are two main activities supported by the tool: data collection and data dis-

Fig. 4 Infomat, an Information Visualization GUI by Rosell and Velupillai (2008)

play. To facilitate the process of reducing data sources to manageable and comprehensible chunks, researchers have devised tools for visual exploration of patterns (Albinsson and Morin, 2002) to find critical incidents by using explicitly available attributes of communications to elicit patterns. From an earlier study on the work process involved in C^2 research (Leifler and Eriksson, 2010a), we had concluded that there were issues related to focusing a study and drawing conclusions from data that could probably benefit from support tools built on foundations from the data mining community.

2.3 Pattern extraction

Many statistical analyses of data from simulated command and control scenarios concern text-based sources, whether these are primary sources from written communications or secondary sources such as transcribed speech or annotations from video logs. Statistical analysis techniques require numerical representations of data sources, which means that video recordings, screen recordings and audio data need to be transformed into products that can be used for such analyses. Not all analyses of command teams involve statistics and qualitative studies are very important in understanding team dynamics. However, irrespective of the type of desired end results in a study (for example categorization schemas of communications or sociometric status diagrams), numerical methods may be used as support for navigation and exploration of data sets (see e.g., (Albinsson and Morin, 2002)).

We had previously found text clustering in particular to be useful as a pattern extraction technique compared to other techniques for extraction information from communications and observer reports (Leifler and Eriksson, 2010b). Text clustering can be performed to relate texts to one another based on *distance metrics* and when suitably used in a framework for clustering texts based on them, they can be used to guide a manual search for patterns between texts and terms in texts, as has been demonstrated in the Infomat information exploration tool by Rosell and Velupillai (2008).

Figure 4 presents the Infomat tool in which users have a matrix-based view of a dataset where they have direct control when exploring possible patterns in the dataset: clusters, important terms and word co-occurrences. The Infomat can be considered a support tool for experts in data mining who have in-depth knowledge of both how to perform clustering, represent terms and texts in a grid-like representation and understand the significance of graphical patterns that emerge as part of manipulating the different clustering settings in the tool. One of the conclusions from our earlier interview study (Leifler and Eriksson, 2010a) was that transparency is crucial to success for advanced support tools in analysis. The Infomat tool, though capable in clustering and one of the more interactive tools available for text clustering, still required a radically different set of interactions and manual translations to the types of data normally encountered in analysis when tested on data from the scenarios used. However, by combining several of the steps normally performed in sequence with the Infomat to reach conclusions on possible patterns in data and creating a radically different interaction mechanism for exploring clusters, we created one of the parts of the Workflow Visualizer exploration support system to test whether text clustering could be used successfully as a support option for decision making research.

3 The Workflow Visualizer

In our design of the support tool, we built on results from an interview study which indicated that researchers spend much of their effort on narrowing research questions and looking for patterns in data (Leifler and Eriksson, 2010a) with little use from automatic tools for pattern extraction. Leifler and Eriksson (2010a) provided two main observations related to the design of intelligent support tools in analysis: the requirements for open-endedness and transparency in tool design. The design of the Workflow Visualizer was based on concrete interpretations of these two observations in the following way with respect to the use of text clustering:

Open-endedness implies the ability to choose how to use computer-based models of a scenario depending on trust in the model and user needs. In our design, text clustering can be used for two distinct purposes, or not at all. Users can choose to inspect parts of the communication flow based on key terms occurring in the messages, or inspect clusters of messages based on their proximity to each other according to the clustering model. The former method requires users to only rely on the computer model for selecting the most relevant terms to select messages by, whereas the latter requires users to trust the vector-based model to produce contextually significant clusters of messages. The option not to use clustering at all (but still use the tool) means that users can select messages through directly available attributes and

Fig. 5 An architectural overview of the Workflow Visualizer, capable of managing a set of different text-based data sources and manipulating them in views related to the exploration of large data sets.

metadata instead, in a selection component where all explicitly available attributes of messages, such as the participants in the communication, the timeframe of the communication, are represented graphically.

Transparency implies that computer-generated models of data sets, such as vector-based models of the texts in a communication flow, should be made directly accessible to users by exposing the defining features through a graphical interface. If possible, the process used to create the computer-based representation should also be directly comprehensible. Transparency depends much on the conceptions users have of the underlying techniques that the computer uses. In our design, we hypothesized that *making key terms* extracted by the vector-space model into *part of the interface* for selecting groups of messages would help make the clustering process more transparent.

3.1 System Description

The Workflow Visualizer support tool is described in Figure 5, which can be considered a three-layered architecture for managing scenario data which is in some respects similar to ESDA tools. The Workflow Visualizer was built to test specifically the utility of text clustering and auxiliary techniques for filtering large data sets of texts. To make the tool useful as a stand-alone support tool navigating large data sets, it incorporates functionality for importing, viewing, and manipulating scenario data, similar to what is offered by tools such as MIND and F-REX. There are three main views available for users, a *timeline view* (shown in Figures 10, 11, 13, and 14) in which messages are shown in the row corresponding to the original sender of a message at the time the message was sent, *manual filtering view* in which users can select subsets of the communications and observer logs based on manual filtering options of keyword occurrences, time periods, and participants. The manual filtering view was implemented based on the requirement for open-endedness as it provided another way of interacting with the same underlying clustering technique: the keywords panel and keyword-based communications chart (see Figure 8) are based on the most significant keywords as extracted by the

Fig. 6 A comparison between the four stages of qualitative data analysis by Miles and the Workflow Visualizer support tool.

clustering engine, but no clusters are produced automatically. Finally, the *clustering view* which is shown in Figure 12 provides a visual representation of text clusters, along with the option of representing individual clusters as sets of color-coded messages in the timeline view (see Figure 13 and 14), that we believed could be more transparent to users than a matrix-like interface similar to Infomat.

The Workflow Visualizer can import and manage a variety of data formats such as simulation logs, observer reports, and written communication in the form of both e-mail and chat logs. It is also easily extendable to new formats by having a modularized architecture where new import components can be constructed either for new data formats or for new scenario types, where there could be use for displaying static events in the timeline such as the duration of individual sessions in a series of command team training sessions as shown in Figures 13 and 14. It was not intended as a replacement for any existing system, but rather intended to serve as a demonstration platform for the techniques studied in this paper.

The intended use of the Workflow Visualizer was to support researchers in exploring relations in a visual manner familiar to them but based in part on text clustering, and thereby draw better conclusions from C² research data (Figure 6). Two specific use cases had been elicited from interviews with researchers (Leifler and Eriksson, 2010a), and they were based on datasets from authentic C² research scenarios. These use cases were developed as realistic scenarios in which the features specifically based on text mining techniques could be tested. In the subsequent workshop evaluation, we used the use cases to reason about both the design criteria in general and if the specific Workflow Visualizer approach could offer enough by way of navigation and data reduction to support the claim that text clustering can provide benefits in research and analysis of distributed decision making scenarios such as command and control.

4 Use Cases

The two use cases came from C² scenarios that the interview participants were familiar with but where they had not used their own tools for eliciting patterns in data. The first scenario concerned performance analysis of staff engaged in information warfare and specifically their reactions to radio interferences in a scenario. The use case that builds on this scenario consisted of how to search for patterns in data with respect to specific events and extract text messages based on prominent terms in the communication flow that were related to those events. The second scenarios concerned a series of ten C² exercises for the rescue services where analysts wished to explore differences in the communication and performance of the teams between the exercises. In the second scenario, we presented a use case with the Workflow Visualizer for finding patterns between exercises by automatically clustering data according to distance metrics imposed by the text clustering engine. Both scenarios and the accompanying data sets were provided by researchers who participated in our study.

4.1 Performance analysis

The first scenario concerned information warfare, in which a group of commanders were responsible for securing transportation of VIP's via helicopter in a hostile, fictive area. They coordinated their efforts with their higher command through e-mail and were monitored by human observers who took notes of their actions during the scenario. They also logged their own perceptions of threats in their environment during the course of the experiment.

The data sources from the first scenario (text messages, observer logs and simulation logs) were imported in the Workflow Visualizer using an appropriate import component. Each import component creates a data model of each scenario that is based on the concepts of *message* and *event*. A *message* contains participants, a timestamp, text, possibly a manual classification and other scenario-specific metadata. An *event* has a timestamp, a description and possibly other metadata. Based on these messages and events, the Workflow Visualizer enables after-action reviewers with means of sifting through the information through direct manipulations for selecting, visually presenting and clustering data according to the general stages of qualitative data analysis (Figure 1). Also, the importer can be configured to consider certain parts of the scenario data to be static (in the information warfare case, radio interferences they should react to) by which to select messages.

Users can choose to select a subset of messages based on keywords or other attributes. In Figure 7, the user has made a selection for messages that contain the keyword “störning” (interference). The number of matched messages is

Fig. 7 Selection of messages based on communication participants, the time frame for the communication and the keywords present in the messages.

indicated before the user chooses to populate the timeline view (*Create Process*). In the first scenario, the simulation logs have information about all actual interferences during the scenario. When analyzing team performance, the user of Workflow Visualizer wants to inspect the team's reactions to interferences and therefore selects all messages sent among the staff mentioning interference.

Much written text from exercises has no explicit context structure such as threads that are available in e-mail conversations. Therefore, it can be difficult to identify which messages relate to one another as responses to earlier questions for instance. When creating a visual overview of the communication between different participants in the exercise, we chose to implement a color scheme based on the hypothesized context of a message. Based on a selection of keywords in the keyword list, a set of messages is drawn as communication arrows (Figure 8). The color of each arrow depends on whether there has been any earlier conversation between the sender and recipient. If the last message received was from the intended recipient of the current message and contained the selected keyword(s), then the same color as that last received message is chosen. Otherwise, a color is chosen at random. This scheme was intended as guidance to testing whether a set of messages was to be considered a conversation on the same topic or not. All actions are not available as messages, however, which is why the user needs to triangulate messages with logs from human observers who monitor and log the team's behavior.

The observer reports available in the dataset from the first scenario consist of reports categorized according to a hierarchy of possible conditions for exerting command. When importing the observer reports, the import component provides these observer categories as a parameter on which to make selections. This parameter, along with all other param-

Fig. 8 Hypothesized threads in the communication indicated by arrows of different colors.

Fig. 9 When users select one of the observation categories in the tree (“communication interference/minor” above), they can choose a color when plotting observations with that category along the timeline.

eters, is used to create a graphical selection component in the selection view. Figure 9 shows the selection components for observer categories that can filter reports depending on the category of the report. Each category in the scenario represents an enabling condition (“förutsättning”) for C^2 that the report is concerned with. In the scenario, there are two different categories of reports related to interferences, labelled “communication interference/minor” and “communication interference/major”. When selecting one of these in the tree, the user gets to select a color that will be associated with the observer reports indicating interference. With a selection of messages and events that the researchers believe may capture situations during the scenario when the staff has reacted to interferences, the researcher can turn to the timeline view for better understanding how they have reacted to the interference.

In Figure 10, we see the timeline of events that is displayed when observers noted that the staff believed they were subjected to a major interference. At the top, two true interference episodes are listed and it is clear from this representation that the staff did not react in time on interferences they had been subjected to.

At another stage of the scenario (Figure 11), a selection of the messages sent by members of the staff indicate that they were consistently late at recognizing both the presence

Fig. 10 A set of message clusters and observations concerning the staff’s beliefs about interferences.

Fig. 11 Clusters of messages and a set of observations regarding interference as displayed in the timeline view of the scenario at the time when there is simulated radio jamming.

and absence of interferences. In the timeline, we have chosen to group messages in clusters according to when they were sent, and color them with a color gradient between two colors according to a metric (here, the same metric as used by the Random Indexing clustering). We can notice that the clusters appear with an offset in time from actual changes in interference. At the end of a period of interference, the first cluster indicates that the staff begins to talk about interference (as corroborated by human observations) and decide to act on the interference at about the time when the interference ceased. Their reaction to the absence of interference comes much later as indicated by the second cluster of messages.

4.2 Exploration

The second scenario contained data from 10 runs of minor command and control exercises (lasting approximately 4 hours each) with the rescue services, where the researchers were interested in finding whether any one of the exercises had been different from the others based on the contents of the messages between members of staff. In this scenario, we demonstrated the exploratory use of Random Indexing for clustering messages. Our approach with finding patterns in

large text sets using Random Indexing was inspired by the Infomat information exploration tool by Rosell and Velupillai (2008) (see Section 2.3). Infomat was created to support exploration patterns of large text corpora through the use of vector-based representations of terms and texts. The Infomat representation of the texts and terms was a sparse visual matrix representation, with dots indicating occurrences of terms in a particular text. The Infomat is a powerful, multipurpose vector-space exploration tool, but unfortunately with a very steep learning curve for analyzing specific content such as communications and an interface that is very far from what is commonly used in exploration and analysis when the raw data sources themselves are displayed (see Figure 2).

Our adaptation of the Infomat interface is shown in Figure 12 where a set of message cluster is shown using the key terms occurring in each, where stop words have been filtered out but no stemming or other preprocessing has been conducted. The user can select a certain number of clusters to create, which causes a clustering algorithm to generate the best partitioning of the set of messages in the selected number of cluster. By selecting a color for a cluster, the user can differentiate clusters from each other when showing them in the scenario timeline. Also, the user can reason about their relevance to the scenario by inspecting the key terms deduced for each cluster of messages. A cluster that contains related terms the user is interested in can be selected for further inspection. For example, Cluster 5 in Figure 12 contains a reference to “incoming” (“inkommer”), “radiatorroom” (“radiorummet”) and “plotting” (“plottar”), which are related terms that describe the process of managing new information arriving at the staff and entering the information on a common situation overview map. Cluster 2 is described by terms that denote calm (“lugn”, “lugnt”) periods. To explore the differences in perceived stress, the user hypothesizes that messages in cluster two are concerned with status reports regarding low workload. He chooses that cluster together with cluster one that is concerned with messages regarding a traffic accident (“trafikolycka”).

Figure 13 displays the timeline of the second Sand scenario. Only a few messages from the selected clusters appear in the timeline. On the other hand, in the third scenario depicted in Figure 14, we see a larger number of messages. The higher frequency of messages could indicate better communication, increased workload or other differences between the scenarios, but by navigating through the timeline, and possibly by using other data sources such as video recordings, a researcher can probe the dataset and search for explanations. By selecting individual messages, he can see the message texts and additional information, and when moving along the timeline, there is a vertical time indicator moving with the cursor.

Fig. 13 The two clusters of messages during scenario two of the Sand material.

Fig. 14 The same cluster of messages at scenario three in the Sand material. There is a clear difference compared to scenario two according to the clustering.

5 Prototype Evaluation Workshop

These two scenarios, highlighting the support for performance analysis and exploration, were used as the basis for a workshop during which we presented both scenarios to five participants: four from the original group of interview participants and one external communication analysis expert. The workshop was conducted over the course of half a day as a set of sessions in which the participants got to discuss our interpretations of the challenges in communication analysis as expressed in the scenarios presented, and the two use cases of the Workflow Visualizer tool in each scenario. Each session was conducted with a presentation of the scenario and use case followed by a five-minute individual reflection where each participant wrote down their own impressions on paper and then a 45 minute joint discussion.

Both scenarios were discussed according to four topics: whether the scenario or intended tool use seemed *obscure*, whether the task described was *similar* to scenarios that the participants had taken part in themselves, whether there were parts of the presentation of the tool that were *surprising or unfamiliar* and last, if they had *general remarks* on the described scenario and use of the Workflow Visualizer.

These four topics were chosen to provide options for the participants to organize both their critique and their reflections regarding current work practices (similar tasks) and the presented tool support. All participants were actively involved in discussing all four topics regarding both scenarios. They were directed to bring notes on all four topics to the joint discussion and share them with each other in turn.

Fig. 12 Representation of a set of message clusters, as hypothesized by the Random Indexing approach to cluster message texts.

5.1 Team performance scenario

The first use case concerned analyzing team performance when the performance metric was decided in advance as the time and character of the team's reactions to communication interferences. In the discussion of the proposed method for analyzing team performance using Workflow Visualizer, the participants stressed the importance of transparency, especially in the communication chart (Figure 8). They found that the most obscure part of the presentation concerned the chart and the basis for constructing it.

These keyword-based threads are rather opaque for me as a communications expert so I get unnerved when a program does that for me.

The communication chart quickly became the focus of attention for the workshop participants when discussing this scenario. The graphical nature of the chart, along with an opaque reasoning for creating it, generated many questions regarding exactly how the threads were generated. The threading of sentences was originally intended for different use and was not presented as a central piece of the software for supporting communication exploration, but it immediately caught the attention of the participants. One of the participants, however, was not familiar with the material used in the scenario and the issues raised during the manual analysis of the material. He was therefore hesitant to comment:

I feel that I have too little knowledge to say what it is exactly that is unclear. [...] I can see a few screens but I have a hard time to get a feel for it.

Another participant noted that he felt the purpose of creating threads of communications in the first place was not clear to him, at least not in the scenario in which it was described:

[The timeline view] seems much more important for the analysis you intended to do, whereas the previous [keyword-based communication chart], it seemed like a very good diagram [...] but I cannot really see the point.

They continued to note that when specialized tools, that make specific assumptions about data in order to perform deeper analyses, are used in settings where conditions and data collection methods may change over time, researchers need to be careful. One participant gave an example from a series of experiments:

An important aspect of the [X] dataset is that the method used may not be 100 percent logical. [...] They have developed and changed [the method used in the trials] between trials which means that you mix many different kinds of data. [A specialized tool such as F-REX] becomes very sensitive to what you enter in it. If you are aware of what you are entering and are also aware of what you are looking at it may be very helpful and very good. If you have a vague understanding of the logic behind it, it can be very confusing.

Following an explanation that the keyword-based chart was to be considered as a hypothesis and support for exploration, the participants expressed positive views regarding the possibilities for using it for similar tasks that they were engaged in. One of the participants noted that, in particular:

I think you could use this tool not just for communication analysis but for cause-and-effect analysis in general. I see reasonably large similarities with what we do in [tools such as F-REX]. You can see a thread in a course of events.

Also, when there are no prior categories to apply for categorizing textual data collected from a C² experiment, they thought it would be helpful to have access to a tool for efficient exploration:

In our [microworld] study, we have huge quantities of data and we're not clear exactly what we are looking for. [...] For the exploratory phase if we're using your terminology, I think this is most useful.

With understanding and transparency in focus, they went on to state that human analysts always have knowledge that

can not be encoded in the tool and that the tool must allow close human control over how messages are grouped and how terms are treated:

If I know that this word is really important, can I tag it [as such]? I think that would be a very important feature. I know that *Klippan* is [an important term], and when we evaluate team performance there are certain terms that we look for. That is central to evaluating the outcome of the scenario.

The participants recognized the specific features of the scenario in which the tool had been deployed and could reason concretely about how they would like to use clustering support.

5.2 Pattern exploration scenario

The second scenario, in which we demonstrated how to use clustering as a means to guide hypothesis generation and search for patterns in communication, sparked even more discussions. Here, the workshop participants engaged in discussions of what they would want the tool to do in the future and asked questions regarding how the work process would look like given the tool:

can you use this to look for interesting sequences and identify points in time that you want to zoom in on, do a new clustering on and eventually arrive at something that is manageable for an analyst in the end?

Working with clusters as hypotheses for communication patterns was considered valuable for understanding team performance. The mathematical foundations of the clustering approach were considered difficult, though, and although the messages in each cluster could be plotted along a timeline and the key terms defining each cluster were given in the interface, they were hesitant to use a model that they did not have any prior understanding of. However, when discussing general remarks regarding clustering, the participants began constructing scenarios in which they would like to use the tool for exploratory purposes:

It could say something about the development of a scenario—in the beginning, you talk a lot about “danger” or “risk” and as time goes by you talk more about [other issues]. You could extract a graph of that particular word or that [cluster in which the word is central] along a timeline.

Other participants filled in and discussed how the color of each cluster could be used to separate them according to keywords in them. They also gave examples in this use case of how the tool could be used with better defined user control to achieve a better workflow during the exploration phase:

I would like to edit these clusterings and remove certain keywords that are not relevant and merge certain keywords that were essentially the same like calm, calmly.

6 Summary and Discussion

The construction of the Workflow Visualizer support tool was guided by three observations in prior research (Leifler and Eriksson, 2010a). First, there were two broad requirements for support tools in analysis; open-endedness and transparency. These requirements were also consistent with observations from our earlier work on support tools for C² (Leifler, 2008), and were interpreted in the context of studying command teams. Second, in the workflow of analyzing distributed decision-making scenarios, reducing data sets to manageable sizes had received comparatively less support from tools in analysis compared to the other stages (see Figure 3). Third, the similarity between the processes of analyzing research data and performing data mining suggested that there could be a sufficient overlap between the affordances of data mining approaches and the requirements of the research analysis process to justify using a data mining approach in C² research.

Guided by these three observations, we crafted a prototype tool for selecting and representing team communications based on text clustering (Rosell and Velupillai, 2008) and other features which we tailored to the purpose of selecting parts of team communications for analysis. Our development focused on providing a transparent representation of message clustering and several components for managing communications, simulation data and observations using a common set of manipulations, thereby permitting an open-ended use of the tool.

The workshop evaluation indicated that the approach taken by the Workflow Visualizer tool was highly relevant to the tasks performed in C² analysis. The workshop participants correctly understood the tasks presented and how the tool was intended to support their work. They could clearly articulate several possible applications and desirable features of a tool for selecting parts of a communication flow based on text clustering or other techniques for extracting communication patterns. They also mentioned how the tool could be used in more general settings where not only communication data is analyzed but possibly also video streams and voice communications. The relationship between the timeline-based representation and the selection of events and messages was considered straight-forward, and they could readily reason about how to extract information about messages given their appearance in the timeline.

However, the workshop evaluation also indicated that the speculative nature and opaque reasoning of both the communication chart (Figure 8) and the message clustering

(Figure 12) were obstacles for the participants in evaluating the utility of the tool. Although the communication chart was simple to understand once we explained the details of how the colored threads were constructed, it was something that seemed obscure due to the lack of a direct representation of *how the threads and clusters had been created* from the messages in the communication flow. They also gave indications that the proposed approach to analyzing messages by using mathematical models of text similarity was difficult to relate to as a general method. They did not use vector-space models for other purposes and had difficulties reconciling the concepts introduced by them to their understanding of communication data. Related to these remarks, Charlotte had explained in her interview how it was difficult to grasp the mathematical underpinnings of the LISREL models she used and that several possible models could be inferred in situations where only one might make sense in the context of the scenario. Thus, even if all the requirements for conducting a mathematical analysis are presented in the graphical user interface and the researcher understands well how to use the tool, introducing an analysis tool with no firm knowledge of the theoretical foundations underlying it introduces the risk that conclusions drawn with it may be frail. In the case with Random Indexing, the conclusions possible with small datasets may not depend on the features of the vector representation but on domain-specific artefacts (Rosell, 2009).

One option for presenting the rationale behind the clusters would be to animate the clustering by showing words and messages in an Infomat representation and demonstrate how the co-occurrences of words represent the total similarity measure of messages. Although each word is represented as a vector with high dimensionality which defies representations in 3D, pairs of words could be presented visually with colors representing the scalar cosine distance between the word vectors. Such a distance could give indications of how the randomly assigned word vectors contribute to the distance measures of messages. Similarly, an animation of how the system believes messages can be organized in threads could use a timeline, along which the selected messages are plotted and compared to others in sequence to highlight the features (e.g. when messages are sent) that determine how threads are constructed.

6.1 Managing issues in automated clustering techniques

When automated clustering techniques are used as tools for exploring data sources, the algorithms used to extract clusters will be evaluated in general on the basis of at least four criteria:

Level of trust How well does the user understand the domain model created by the system, the technical precon-

ditions for using the system correctly and the limitations of the modeling technique? Given this understanding, how much does the user trust that the patterns presented reflect differences worth considering and are not artifacts of the modeling technique?

Acceptable error rate What is the minimum acceptable error rate when using a certain pattern detection technique for a particular purpose? What are the relative costs of false positives to false negatives?

Usefulness in directing search: How much easier will it become to find a given, context-specified pattern in data?

Over-reliance/under-use on the tool. Do users cease consider other options for finding meaningful patterns and what are the costs of making them less likely to use other tools? Will they become over-reliant on a technique?

Pattern extraction techniques are always based on an assumption that some natural patterns occur in our data, whether they be rule-based, hierarchical, sequential, with distance functions that are linear and monotonic or based on some other abstract notion that we assume hold. With these abstract notions, a pattern extraction technique may build an internal model of data that is used to make predictions about classifications of instances in a dataset. If the model is used to make predictions, those predictions can be evaluated according to the criteria above. Such an evaluation may in itself be problematic in three ways.

First of all, the problem is one of relevance, in that defining concepts such as *over-reliance*, *usefulness*, and *level of trust* requires a thorough understanding of the relationship between each concept and the overall performance of the human-computer configuration in which the support tool acts. For example, the *level of trust* may need to be described in more detail for us to understand the relationship between trust and performance, or maybe we find out that trust is not the most fundamental concept at all but instead understanding the theoretical foundations for the system's function is. An understanding of the theoretical underpinnings may create trust, whereas trust may not be assumed to precede understanding.

Second, the problem concerns *semantics*, in that usefulness evaluations require a clear definition of usefulness that results in a scale according to which a support system can be evaluated. The activity of interpreting research results in inherently subjective, in the sense that it requires a human subject with his or her own set of ideals and beliefs performing the activity. Subjective is not to mean biased or unfounded, however.

Third, the problem concerns the *measurability* of using the tool: for example, how do we assign a number to a scale related to trust when using the tool? Is the definition of the semantics of each concept crisp enough that it can be justified to perform such an assignment?

One way of managing these issues in evaluating machine learning-based pattern extraction techniques is to completely disregard the precision attained by classification, and instead consider the affordances of the model constructed and conduct qualitative studies of how researchers reason about their data sets with and without the availability of support tools that incorporate clustering techniques.

7 Conclusions

Our objective in this paper was to study options for using automatic pattern extraction to support researchers analyze text-based data sources from simulation-based scenarios. We based our design on two criteria (open-endedness and transparency) previously described as potentially important for analysis tools in interviews with researchers, and interpreted these for the construction of an analysis tool specifically based on text clustering.

The workshop participants evaluating the Workflow Visualizer considered it highly relevant, with several potential applications in their work. In particular, they expressed appreciation of the *open-endedness* of the tool: it could be used for different kinds of data, with little assumptions about data and with multiple modes of selecting and manipulating data.

Concerns and issues raised during the workshop illustrated the relevance of *transparency* as a design criterion for support tools, but also the challenge of interpreting it in the context of tool design. The participants could clearly see how messages were grouped by threads or clusters, yet they felt that the models used for creating them were too opaque. The visual representations did not connect well enough to their own models of how messages relate to others in episodes or categories.

Taken together, the evaluation results could be directly related to the design criteria elicited from the interviews, which corroborate their importance as criteria in analysis tool design. The use cases suggested by the participants in which they would like to use tools for automatic pattern extraction point to the relevance and promise of text clustering for understanding C² data. We believe that, with the appropriate visual cues and representations of how an automatic approach groups messages, text clustering could become a valuable asset for research on distributed decision making, both in command and control and elsewhere.

8 Acknowledgments

This work was supported by the Swedish National Defense College. We would like to thank the participants at the Swedish Defense Research Agency and VSL Systems AB for participating in this study and generously providing material used for the scenarios in this article.

References

- Alberts, D. and Hayes, R. (2003). *Power to the Edge: Command, Control in the Information Age*. CCRP Publication Series.
- Albinsson, P.-A. and Morin, M. (2002). Visual exploration of communication in command and control. In *Proceedings of the Sixth International Conference on Information Visualisation*, London, UK.
- Andriole, S. J. (1989). *Handbook of Decision Support Systems*. TAB Books Inc.
- Argyle, M. (1972). *The Social Psychology of Work*. The Penguin Press, London, UK.
- Brown, R. (1993). *Group Processes - Dynamics Within and Between Groups*. Blackwell, Cambridge, Massachusetts.
- Cannon-Bowers, J. A., Salas, E., and Pruitt, J. S. (1996). Establishing the boundaries of a paradigm for decision-making research. *Human Factors*, 38(2):193–205.
- Cutting, D. R., Karger, D. R., Pedersen, J. O., and Tukey, J. W. (1992). Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.
- Hollnagel, E. and Bye, A. (2000). Principles for modelling function allocation. *International Journal of Human-Computer Studies*, 52:253–265.
- Hollnagel, E. and Woods, D. A. (2005). *Joint cognitive systems: foundations of cognitive systems engineering*. CRC Press, Boca Raton, Florida, USA.
- Johansson, B., Persson, M., Granlund, R., and Mattsson, P. (2003). C3fire in command and control research. *Cognition, Technology & Work*, 5(3):191–196.
- Lazar, J., Feng, J. H., and Hochheiser, H. (2010). *Research Methods in Human-Computer Interaction*. Wiley.
- Leifler, O. (2008). Combining Technical and Human-Centered Strategies for Decision Support in Command and Control — The ComPlan Approach. In *Proceedings of the 5th International Conference on Information Systems for Crisis Response and Management*.
- Leifler, O. and Eriksson, H. (2010a). Analysis tools for studying in decision-making—a meta-study of command and control research. Manuscript.
- Leifler, O. and Eriksson, H. (2010b). Message classification as a basis for studying command and control communications - an evaluation of machine learning approaches. Submitted for publication.
- Miles, M. B. and Huberman, A. M. (1994). *Qualitative data analysis: an expanded sourcebook*. SAGE.
- Morin, M. (2002). *Multimedia Representations of Distributed Tactical Operations*. PhD thesis, Institute of Technology, Linköpings universitet.
- Orasanu, J. and Conolly, T. (1993). The reinvention of decision making. In Klein, G. A., Orasanu, J., Calderwood,

-
- R., and Zsombok, C. E., editors, *Decision Making in Action*. Ablex Publishing corporation, Norwood, New Jersey.
- Rasmussen, J. (1993). Deciding and doing: Decision-making in natural contexts. In Klein, G. A., Orasanu, J., Calderwood, R., and Zsombok, C. E., editors, *Decision Making in Action: Models and Methods*. Ablex Publishing corporation.
- Rosell, M. (2009). *Text Clustering Exploration - Swedish Text Representation and Clustering Results Unraveled*. PhD thesis, KTH School of Science and Communication.
- Rosell, M. and Velupillai, S. (2008). Revealing relations between open and closed answers in questionnaires through text clustering evaluation. In *Proceedings of LREC 2008*, Marrakesh, Morocco.
- Rubel, R. C. (2001). War-gaming network-centric warfare. *Naval War College Review*, 54(2):61–74.
- Sanderson, P. and Fisher, C. (1994). Exploratory sequential data analysis: Foundations. *Human-Computer Interaction*, 9:251–317.
- Shattuck, L. G. and Woods, D. D. (2000). Communication of intent in military command and control systems. In McCann, C. and Pigeau, R., editors, *The Human in Command: Exploring the Modern Military Experience*, pages 279–292. Kluwer Academic/Plenum Publishers, 241 Borough High Street, London.
- Thorstensson, M., Axelsson, M., Morin, M., and Jenvald, J. (2001). Monitoring and analysis of command post communication in rescue operations. *Safety Science*, 39:51–60.
- van der Aalst, W. and van Hee, K. M. (2002). *Workflow management: models, methods, and systems*. MIT Press, Cambridge, MA, USA.
- Wærn, Y. and Cañas, J. J. (2003). Microworld task environments for conducting research on command and control. *Cognition, Technology & Work*, 5(3):181–182. 10.1007/s10111-003-0126-y.