Information Visualisation, Counterterror Intelligence

Syndicate 4

INTRODUCTION

The participants in Syndicate 4 were Denis Gouin, Zachary Jacobson, "Kesh" Kesavadas, Hans Joachim Kolb, Vincent Taylor, Johan Carsten Thiis and David Zeltzer. Portions of this report were authored by Michael Towsey.

By consensus, the members of the Syndicate selected the broad area of *information visualisation* as the topic of interest. In general, it is agreed that information visualisation refers to the presentation of "non-physical" data with no obvious 3D referents, as typified, for example, by multidimensional sonar or financial data; concepts embodied in documents and the relationships among them, or the morale and readiness of military units [Card, Mackinlay et al. 1999].

It was felt that the mission of the Syndicate was to:

- identify information visualisation issues in application domains of importance to NATO,
- identify and characterize the required capabilities and available technologies that address those domains, and
- recommend research and development priorities with respect to the technologies involved.

A number of application domains were considered, but due to the short time available to the Syndicate, this list was reduced to four, and ultimately only one application domain – counterterror intelligence – was addressed. This area is clearly of high priority to NATO, and at the same time, it is largely characterized by non-physical types of data that are problematic to present, and so counterterror intelligence is well-suited to consideration by Syndicate 4.

VISUALISATION REFERENCE MODEL

Once the topic of interest *information visualisation* and the application domain *counterterror intelligence* were selected, the next steps were to identify and characterize component technologies necessary for visualisation of counterterror intelligence data, and to estimate the level of maturity of these technologies.

It was thought that a *visualisation reference model* would be helpful in order to ensure that the Syndicate agree on the visualisation process under consideration, and to consider the technologies that comprise counterterror intelligence visualisation.

The model chosen is close to the VisTG model developed by Martin Taylor, but focuses primarily on the computational engines involved in data analysis and presentation (cf. *The VisTG Model for Visualisation*, these proceedings).

Paper presented at the RTO IST Workshop on "Massive Military Data Fusion and Visualisation: Users Talk with Developers", held in Halden, Norway, 10-13 September 2002, and published in RTO-MP-105.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **00 APR 2004**	2. REPORT TYPE **N/A**	3. DATES COVERED **-**
4. TITLE AND SUBTITLE **Information Visualisation, Counterterror Intelligence Syndicate 4**		5a. CONTRACT NUMBER
		5b. GRANT NUMBER
		5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)		5d. PROJECT NUMBER
		5e. TASK NUMBER
		5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Halden Norway**		8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)		10. SPONSOR/MONITOR'S ACRONYM(S)
		11. SPONSOR/MONITOR'S REPORT NUMBER(S)
12. DISTRIBUTION/AVAILABILITY STATEMENT **Approved for public release, distribution unlimited**		
13. SUPPLEMENTARY NOTES **See also ADM001665, RTO-MP-105 Massive Military Data Fusion and Visualization: Users Talk with Developers., The original document contains color images.**		
14. ABSTRACT		
15. SUBJECT TERMS		

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**UU**	**31**	

Figure 1 illustrates this visualisation model. The Syndicate specifically assumed that consideration of sensor technologies for gathering data was outside the scope of its analysis. It was also assumed that, in general, visualisation is a multimedia and multimodal activity. That is, data must be presented to analysts and decision makers visually, aurally and perhaps, haptically, as appropriate for the application in question. Likewise, they should be able to intuitively interact with presentations naturally with voice and gestures, again, according to the requirements of the application. Furthermore, a "task level" human-machine interface (HMI) enables decision makers to interact with a computer-mediated activity in terms of interest to the human, not the machine. Humans should not be burdened with extraneous cognitive tasks required to operate a computer system – a well-designed HMI should make the computer "invisible" to its users, in the words of Donald Norman [Norman 1998].

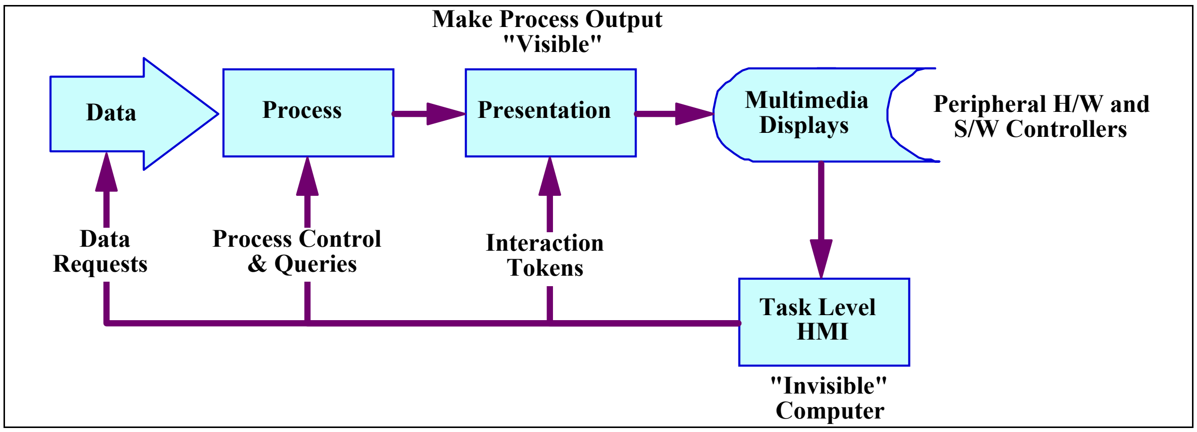


Figure 1: A Schematic View of the Computational Processes Involved in the Presentation of Data to be Visualised. The syndicate focused on characterizing the functionalities contained in the "Process" and "Presentation" modules, and identifying and rating the maturity of the technologies that address those functionalities.

COUNTERTERROR INTEL REQUIREMENTS

The Syndicate considered three main requirements areas in this application domain. Data must be

- gathered from a variety of sources,
- analyzed with a range of tools, some automated and some human-in-the-loop, and the
- analyzed data would be presented to decision makers.

Data Sources

While this is not an exhaustive listing, the Syndicate identified four primary sources of data that would be of interest to the intelligence community:

- communications, such as
  - email, phone, FAX, radio, video, . . .;
- open sources, such as
  - newspapers, WWW, newsgroups, TV, . . .;

- commercial transactions; and
- behaviour of people and organizations.

For each of these data sources, functionalities and technologies required to analyze the data were characterized, and rated as to the respective technology maturity level – *high, medium, or low*.

In addition, several main processing steps were identified. First, since it is practically impossible to attend to unfiltered streams of data of the magnitude represented by counterterror intelligence, the first step would be to rapidly analyze the incoming data streams for features of interest, which would be used to distinguish data to be analyzed further. Therefore, for each data source identified by the Syndicate, a first step would be to estimate the maturity of technologies available to recognize features in the various data streams. Once features can be identified, categorized and prioritized, the filtering of the data becomes straightforward.

Processing Engines

Filtered data becomes the input stream for further analysis. The Syndicate roughly characterized three further stages of data analysis:

- link analysis,
- data mining, and
- behaviour analysis.

In the view of Syndicate 4, each of these processes may be implemented by processing engines consisting of arbitrarily complex algorithms and software systems, some of which might be completely automated, while others may be “human-in-the-loop”. Especially for human-in-the-loop processes, each of these analysis activities will require its own visualisation and HMI components.

In the final stages, the data output of the processing algorithms must be presented to decision makers for action. Visualisation and HMI issues were identified in each of the three analysis areas.

DATA SOURCES

Feature Recognition and Communications

Email, phone, FAX, radio, video

In point-to-point communications content is arbitrary and unconstrained. This means that a robust natural language understanding (NL) capability is required to fully comprehend the content and intent of such messages, which is largely beyond current capabilities. Nonetheless, textual analysis technologies do exist to identify content features of such communication, so even though technology for understanding arbitrary NL content is not yet available, communications can still be categorized and related based on identified concepts contained therein.

In addition, many easily recognized parameters of communications can be derived, including

- Source,
- destination(s),

- length,
- encrypted(?),
- language,
- subject field,
- attachments,
- routing,
- etc.

Content analysis

Textual concept recognition	
in some languages	High
for multilingual	Low
OCR	High
Speech recognition	High
Image and video feature recognition	Low
Intent recognition	Low

Feature Recognition and Open Sources

Newspapers, WWW, newsgroups, radio, TV, . . .

These are largely broadcast media, in which the domain of discourse is largely constrained by context. In such cases, for example, newspaper articles, NL technologies have been available for some time that can interpret such media and provide reliable paraphrased interpretations.

Content analysis

Textual concept recognition	
in some languages	High
for multilingual	Low
OCR	High
Speech recognition	High
Image and video feature recognition	Low
Intent recognition technologies	Medium (NL paraphrasing technologies exist)

Feature Recognition and Commercial Transactions

Transaction signatures

Customer ID

Credit card #

Product(s) purchased

Amount of product purchased

Purchasing frequency and history

. . .

All such signature parameters are typically maintained by merchants and are subject to data mining.

Feature Recognition and Behaviours

Especially in democratic societies, for a host of reasons, it is simply not possible or desirable to monitor the behavior of all citizens. On the other hand, law enforcement and intelligence agencies have ample surveillance tools to monitor individuals and groups that have come to their attention.

Scope

Suspect entities

Behaviour signatures

Phone calls

Recipient and locations

Travel

Residence

Biographical data

Gait, action and mannerisms

. . .

Data sources

Current law enforcement surveillance methodologies collect behavior signatures.

PROCESSING ENGINES

Link Analysis

Link Analysis is a technique very useful to show relationships among people, organizations, events, incidents, behaviours and locations as shown on the left side of Figure 2 taken, from the U.S. company IntelCenter. Shown on the right side of Figure 2 is a subset of Mapping al-Qaeda v1.0, a product utilizing link analysis technology to provide visual maps of terrorist networks around the world and to help foster a better understanding of al-Qaeda's operational characteristics and organizational structure.

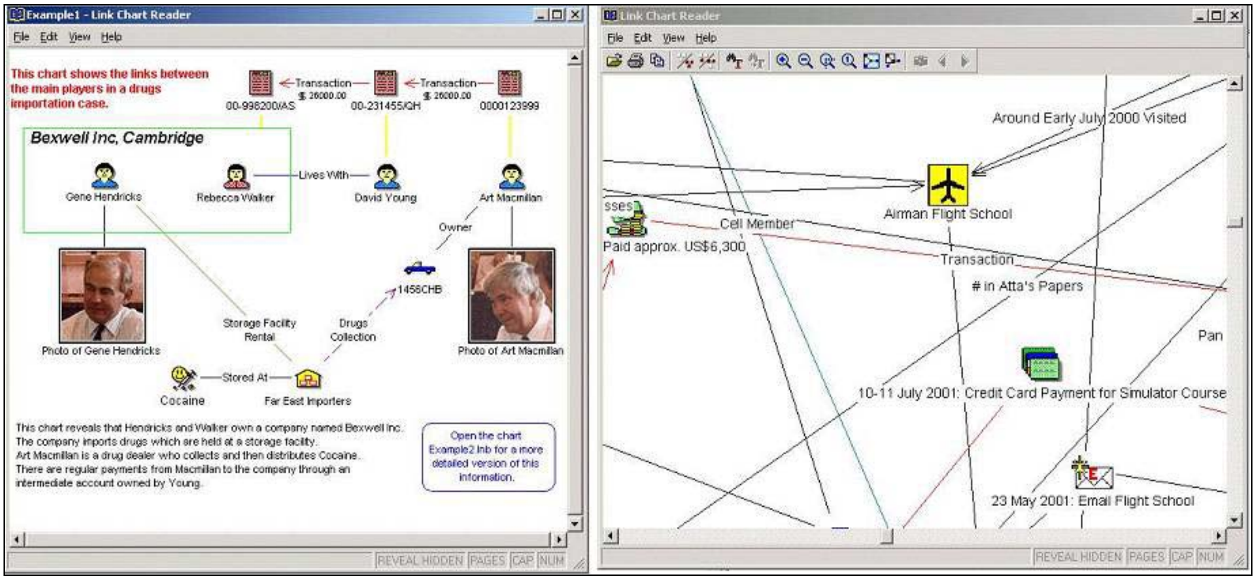


Figure 2: Examples of Link Analysis.

Mapping al-Qaeda v1.0 was produced by the private company, IntelCenter. The focus of IntelCenter is on studying terrorist groups and other threat actors and disseminating that information in a timely manner to those who can take action on it. Its primary client base is comprised of military, law enforcement and intelligence agencies in the US and other allied countries around the world (http://www.intelcenter.com/linkanalysis.html).

Medium technology maturity

Data Mining

Data mining has been a technology applied to marketing data bases for some time. The Syndicate felt that a number of off-the-shelf products are available to perform data mining of information contained in commercially-available data base systems. Such information was considered to be *structured data* stored in categories and formats defined *a priori*. It was felt that, on the whole, such tools have suboptimal HMIs and visualization components, and therefore, require further development for use in the counterterror domain. Finally, data mining of new types of data, which would be largely *unstructured*, remains a technological challenge.

Mining *structured* data

E.g., commercial transaction data

Off-the-shelf technologies available but difficult to use

High maturity but visualisation and HMI development required

Mining *unstructured* data

Low maturity

Data representation and association, automation tools, HMI and visualisation require major R&D

Behaviour Analysis

Behaviour analysis is an emerging technique that allows investigators to identify suspect behaviours by comparing events with 'normal' information stored in a knowledge base. An example that could be drawn from a drug interdiction scenario is shown in Figure 3. In this case, the behaviour of a ship in terms of the itinerary, ports visited, time spent in each port is compared with information describing normal activities of the same category of vessel / ship Kluchert 1998].

Scope

Suspect entities

Technology maturity <u>Low</u>

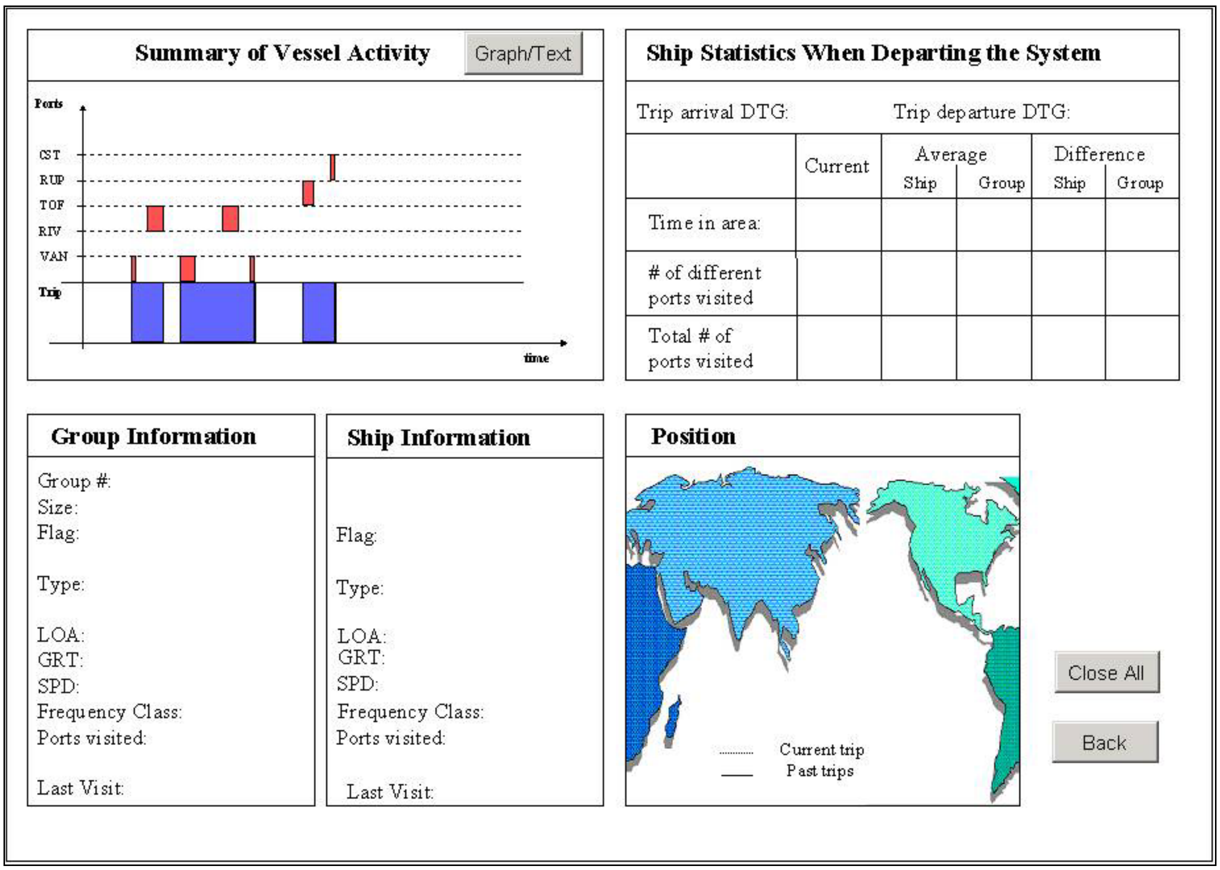


Figure 3: Examples of Behavior Analysis.

In order to exploit behavior signature collection, it is necessary to find patterns in recognized features. Some tools are available, some are automated systems, and some depend on human-in-the-loop processing. The latter have, of course, human/machine interaction and visualisation requirements. Since the scope and scale of behavior signature collection is so large, and because information visualization is so strongly application dependent, further R&D will be required to define and develop interaction techniques and metaphors, as well as visualization solutions.

Many components of behavior analysis technology are currently available but major integration engineering is required to develop a complete behavior analysis toolset. Emerging technologies typically suffer from prohibitively high false alarm rates. Human-in-the-loop signal detection requires visualisation and HMI R&D. Figure 4 illustrates a schematic view of the purported components of a distributed, behavioral analysis toolset, with regional, local, and on-site components.

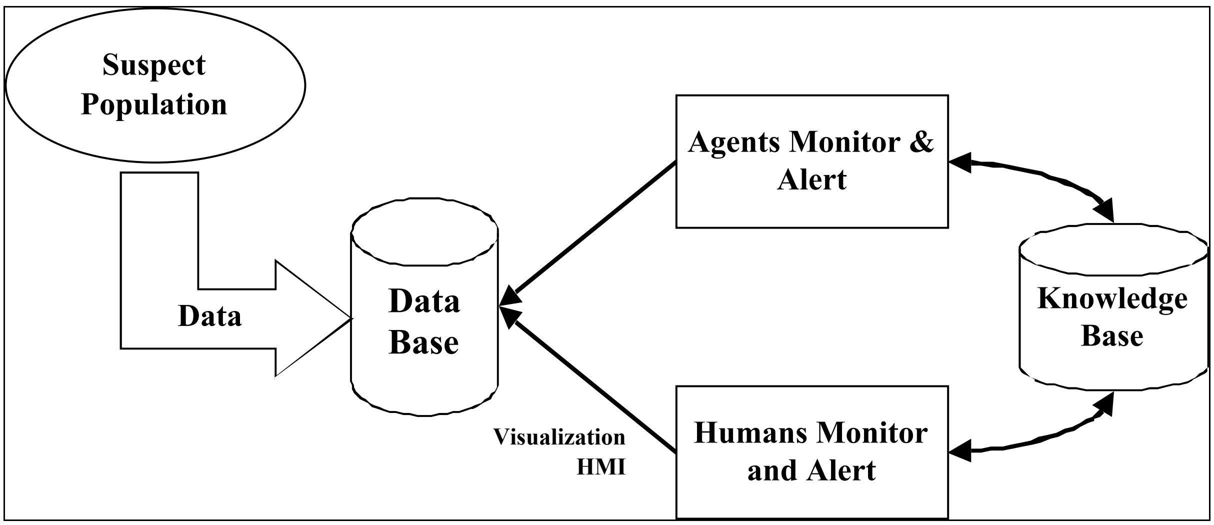


Figure 4: A Schematic View of a Technology for Monitoring a Suspect Population.

SOME ISSUES IN INFORMATION VISUALISATION

It is clear from the above examples that different kinds of data require different kinds of visualisation algorithm and different kinds of human interaction. Human intervention can take place at three levels: first to decide on the appropriate visualisation algorithm, second to set the algorithm parameters and finally real-time interaction with the displayed information.

In the case of mining structured information, typically each datum can be represented as a point in an n-dimensional feature space. It is possible to visualise large amounts of information by projecting a high-dimensional space onto a 2 or 3 dimensional space. As an example, Figure 5 displays a large number of communications between multiple senders and receivers over time. Despite the high volume of data, the representational framework is a simple cube, which can be easily manipulated to offer different views.

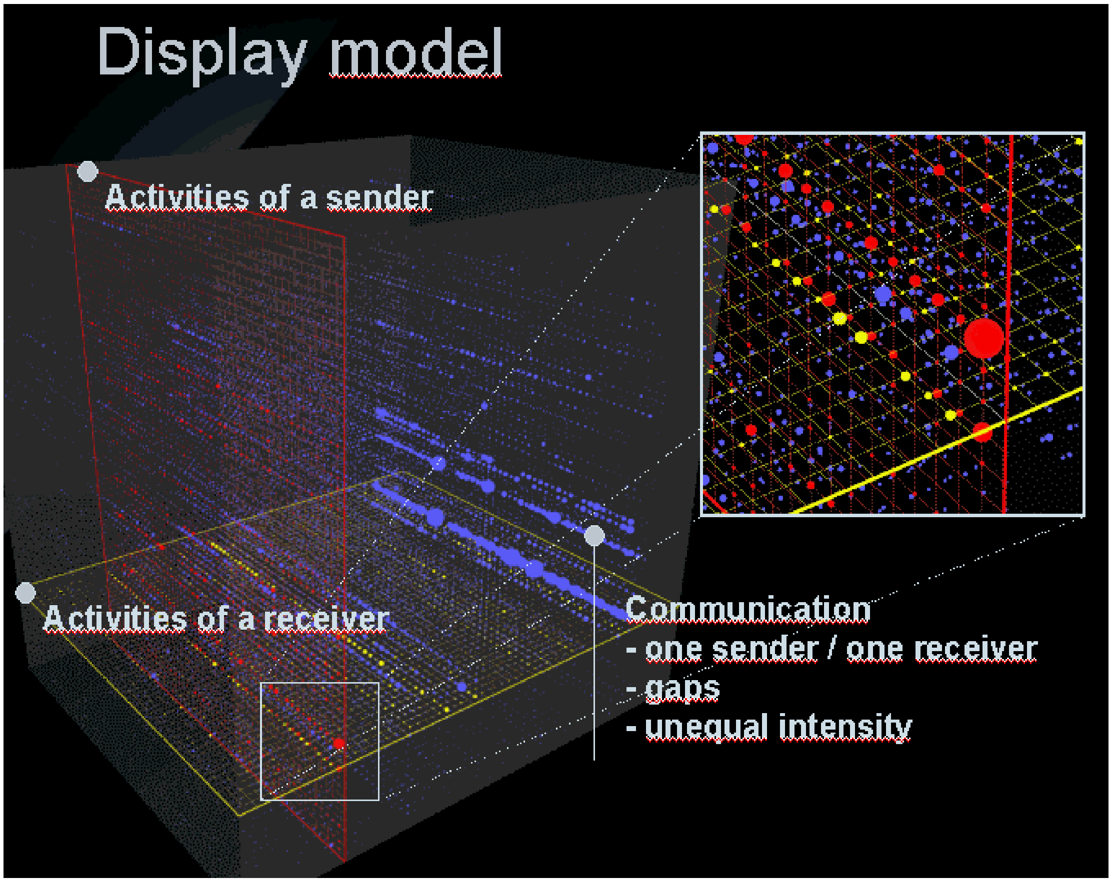


Figure 5: Visualisation of Communication Channels Over Time.

By contrast, the visual representation of textual or behavioural data is difficult because each of the data items sits in a different ‘space’. The appropriate data model is a directed graph but the effective visualisation of graphs remains an area of active research. Even a small graph as in Figure 2 can be difficult to display in a way that clarifies rather than obfuscates the data. One computational difficulty is to position the nodes so as to preserve spatial proximity between closely related objects and yet to prevent a confusion of crossing arcs. An example of a fully automated visualisation of a complex text is illustrated in the left side of Figure 6. The text is the 130,000 word Nixon-Watergate transcripts. The nodes represent important terms or concepts in the transcripts. The arcs show the strength of the associations between the concepts. An important feature of the software is that it offers multiple (visual and textual) views into the one text and it allows the user to manipulate the representation if desired.

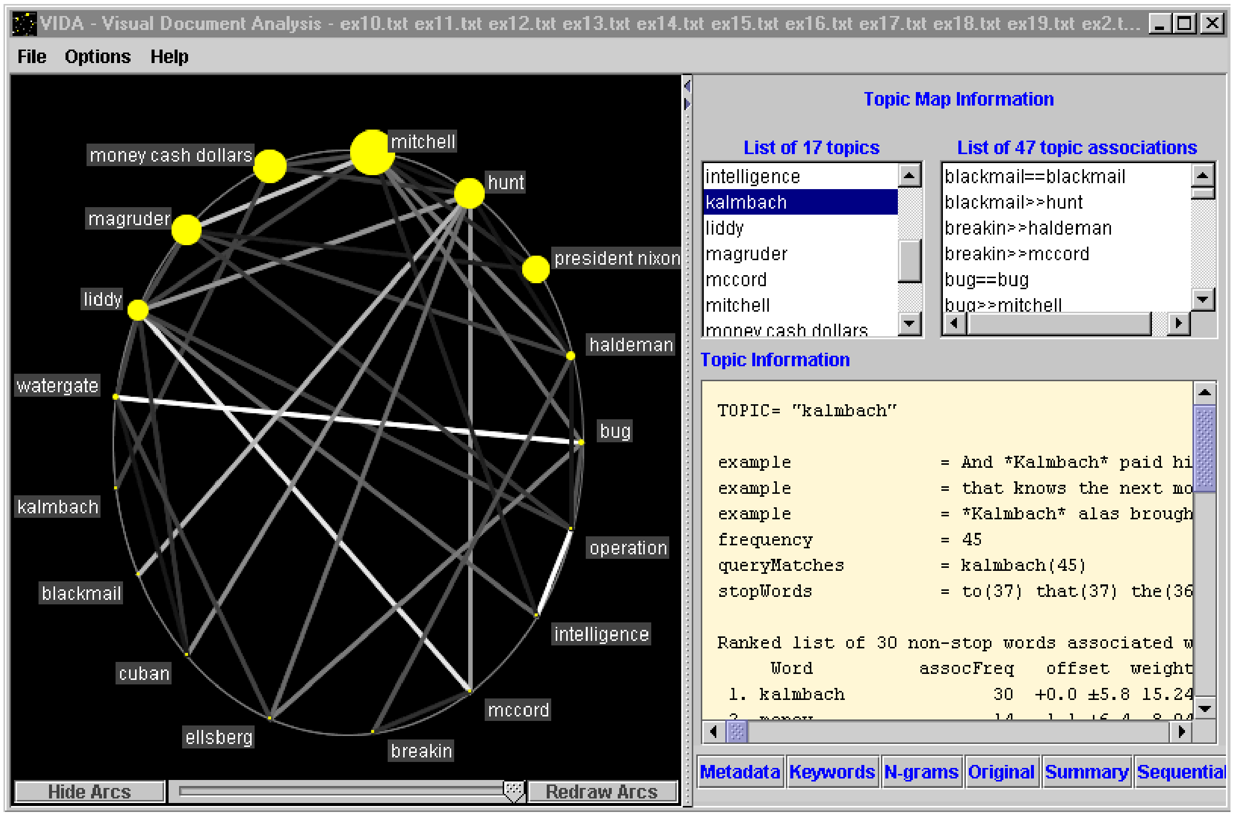


Figure 6: A Visualisation of Topics in the Nixon-Watergate Transcripts.

Another strategy used to visualise and navigate large amounts of information is *zoom and focus*. The total data space is first represented on the screen in low resolution. The user selects one part of the space which is then enlarged and displayed in higher resolution. This approach is particularly meaningful if the data structure is a tree. However a difficulty with *zoom and focus* is that information tends to loose its meaning when taken out of context. This is true regardless of whether the information items are textual or visual. Current research is attempting to get around this problem by using a variety of *fish-eye views* on data. That is, user-selected data is enlarged in the centre of view, while connected information is shown less prominently on the edges.

The issues in information representation are not only algorithmic. Physiological and psychological considerations are also important. The *grok box* project (http://vader.mindtel.com/concepts.html) is exploring ways of conveying information through sensory modalities other than vision. Just as the geometry, color, texture and dynamics of a visual icon are all meaningful features which can be encoded with information, so too, an auditory or 'sonified' icon can convey information through its tone, pitch, timbre, duration and location. And tactile stimuli which impinge on hands, fingers, arms, or skin and muscle sensations of the body can convey information through touch, felt position, motion, and force. The principle is to push all the sensory modalities to their 'margin' in order to interpret large volumes of information. But an important question is how much information can be and should be loaded into icons representing abstract entities? Iconography is itself a language which must be learned and if the language is not intuitive, the effort to learn it may become an obstacle. Figures 6, 7 and 8 illustrate examples from the spectrum of possibilities.

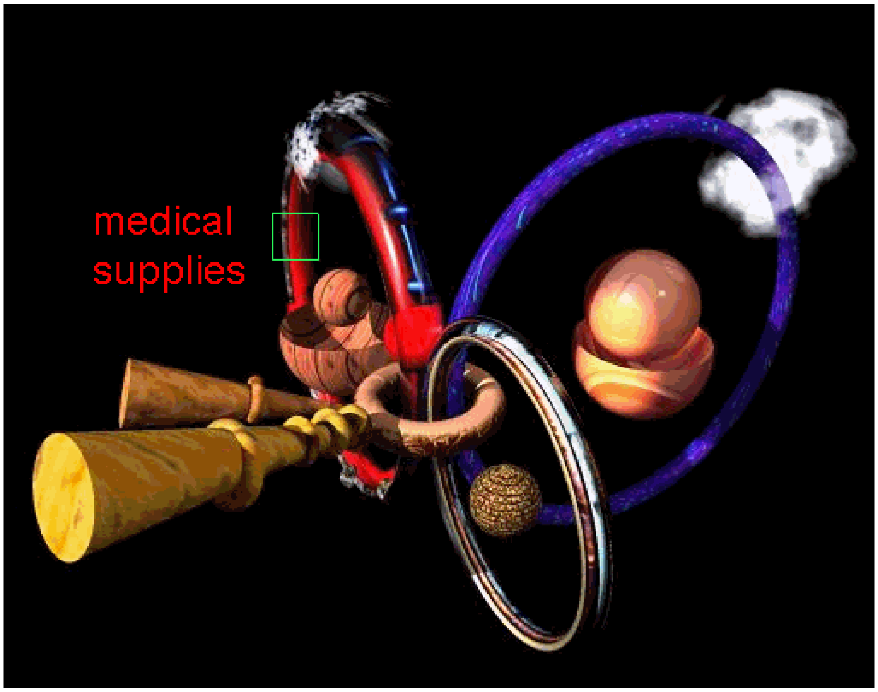


Figure 7: Visual Representation of a Refugee Camp in Africa.
Accessed through http://vader.mindtel.com/concepts.html (click on applications)

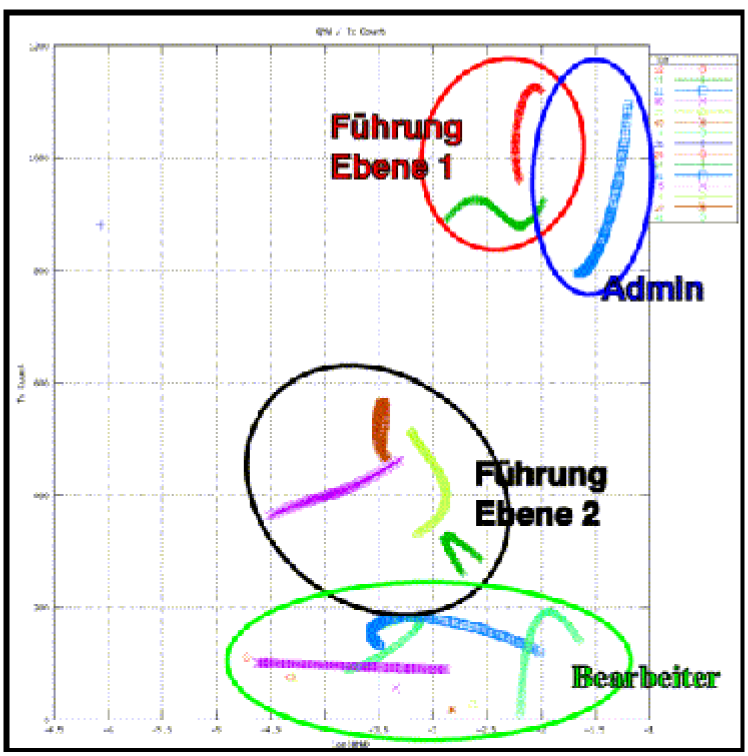


Figure 8: Visualisation of Topics in the Nixon-Watergate Transcripts.

Figure 7 is an abstract representation of a refugee camp in Africa. The iconography is complex because all the features of each icon are encoded with information. The wooden ring in the centre represents the camp itself and the other icons are influences acting on or interacting with the camp. The red ring represents medical supplies but it has several attached features which must also be interpreted. The language of the icons in Figure 7 is complex and would take some time to learn.

In Figure 6, by contrast, the iconography remains limited but intuitive and therefore immediately comprehensible. The size of a node represents the frequency of a term in the text and the intensity of the arc represents the strength of the association. Figure 8 illustrates an intermediate example of iconic complexity. It is a representation of a command structure obtained by cluster analysis of communications. The command hierarchy is illustrated in the vertical dimension while the colours and shapes are easily interpreted by reference to the index at top-right.

SUMMARY

It is said that "a picture tells a thousand words" but this wisdom does not automatically extend to the representation of *non-physical* data having no obvious 3D referents. Future progress in information visualisation will depend both on algorithmic developments and on a population of users gradually learning an iconic language, much in the same way that PC users have learnt Windows iconography over the past 20 years. Much depends on the application. In mission-critical applications, one might expect iconography to remain conservative. Non-critical applications will allow increasing exploration of iconography in its broadest sense.

Currently, link analysis and data mining are the "low hanging fruit." These technologies are "almost there" and potentially may be most productive in the short term for generating useful intelligence. However, in current systems, visualisation capabilities, and the human/machine interface are poor. The most difficult challenge, moreover, is scaling the algorithms to handle the vast quantity of data that must be processed.

Behaviour analysis is a promising application and it is based on existing information technologies. Matching behavior to known parameters based on biometrics can also be a potential area of great promise. Proof-of-concept prototypes, however, need to be developed and evaluated.

REFERENCES

Card, S.K., J.D. Mackinlay and B. Schneiderman (1999), *Information Visualization*, in **Information Visualization: Using Vision to Think**, S.K. Card, J.D. Mackinlay and B. Schneiderman, eds. San Francisco CA, Morgan Kaufmann, pp. 1-34.

Kluchert, R. (1998). *Innovations Used to Improve the Recognized Maritime Picture (RMP) in the Maritimes Force Pacific (MARPAC)*, Operational Research Division, Department of National Defence, Ottawa Canada, November 25, 1998.

Norman, D.A. (1998), **The Invisible Computer**, Cambridge MA, MIT Press.

Syndicate 4: Information Visualization

Counterterror Intelligence

Syndicate 4

Denis Gouin

Zack Jacobson

“Kesh” Kesavadas

Hans-Joachim Kolb

Vincent Taylor

Johan Carsten Thiis

David Zeltzer

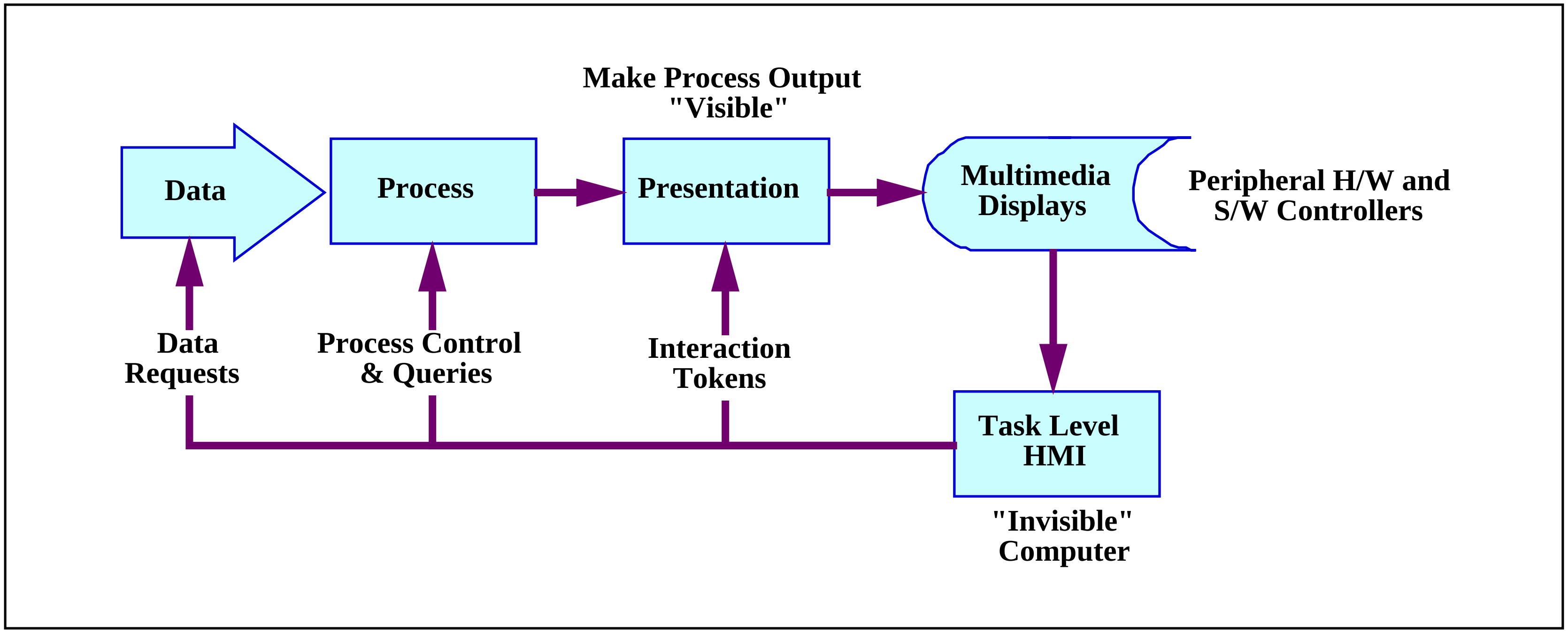
Overview

- Approach
- Visualization Reference Model
- Application Requirements
- Capabilities and Technologies
- The Matrix

Approach

- Information Visualization
  - How to present "non-physical" information with no straightforward mapping to 3D metaphor?
- Visualization Reference Model
- Apply to Specific Domains of Interest to NATO
  - Counterterror Intelligence
  - Requirements
  - Functionalities and technologies
- Indicate R&D Directions
  - Rate technology maturity
  - Encapsulate in matrix form

Visualization Reference Model



Counterterror Intel Requirements

- Collect Data
  - Data Sources
    - » Communications
    - » Open source
    - » Commercial transactions
    - » Behavior of people and organizations
- Process
  - Feature recognition $\Rightarrow$ filter $\Rightarrow$ content analysis
  - Arbitrarily complex algorithms and software
    - » Automation
    - » Human-in-the-loop
- Presentation
  - Identify visualization (+HMI) issues

Counterterror Intel Requirements (cont'd)

- **For Each Data Source**
  - **Characterize functionalities and technologies**
  - **Rate technology maturity level (high, medium,low)**
  - **Identify visualization (+HMI) issues**

Counterterror Intel Data Sources

- Communications
  - Email, Phone, FAX, Radio, Video, . . .
- Open Sources
  - Newspapers, WWW, Newsgroups, TV, . . .
- Commercial Transactions
  - Individuals
  - Organizations
- Behaviors
  - Individuals
  - Organizations

Counterterror Intel Data Processing

- Pair Capabilities & Technologies with Data Sources
- Feature Recognition
  - Communications
  - Open Sources
  - Commercial Transactions
  - Behaviors
- Link Analysis
- Data Mining
- Behavior Analysis

Feature Recognition and Communications

- Email, Phone, FAX, Radio, Video
  - Many easily recognized parameters
    - » Source, destination(s), length, encrypted(?), language, subject field, attachments, routing, etc.
  - Content analysis
    - » Textual concept recognition
      - High in some languages
      - Low for multilingual
      - High OCR
      - High speech recognition
    - » Low image and video feature recognition
    - » Low intent recognition

Feature Recognition and Open Sources

- Newspapers, WWW, Newsgroups, TV, . . .
- Domain of Discourse Constrained by Context
- High Concept Recognition Technologies
- Low Intent Recognition Technologies

Feature Recognition and Commercial Transactions

- Transaction Signatures
  - Customer ID
  - Credit card #
  - Product(s) purchased
  - Amount of product purchased
  - Purchasing frequency and history
  - . . .
- All Signature Parameters Maintained by Merchants
- Subject to Data Mining

Feature Recognition and Behaviors

- Scope
  - Suspect entitities
- Behavior Signatures
  - Phone calls
    - » Recipient and locations
  - Travel
  - Residence
  - Biographical data
  - . . .
- Data Sources
  - Current Law Enforcement Surveillance Methodologies

Link Analysis

- Find Patterns in Recognized Features
- Some Tools Available
  - Automated
  - Human-in-the-loop ⇒ visualization
- Medium Technology Maturity
- Both Automated and Human-in-the-Loop Link Analysis Tools Require Further R&D Including Visualization and HMI

Data Mining

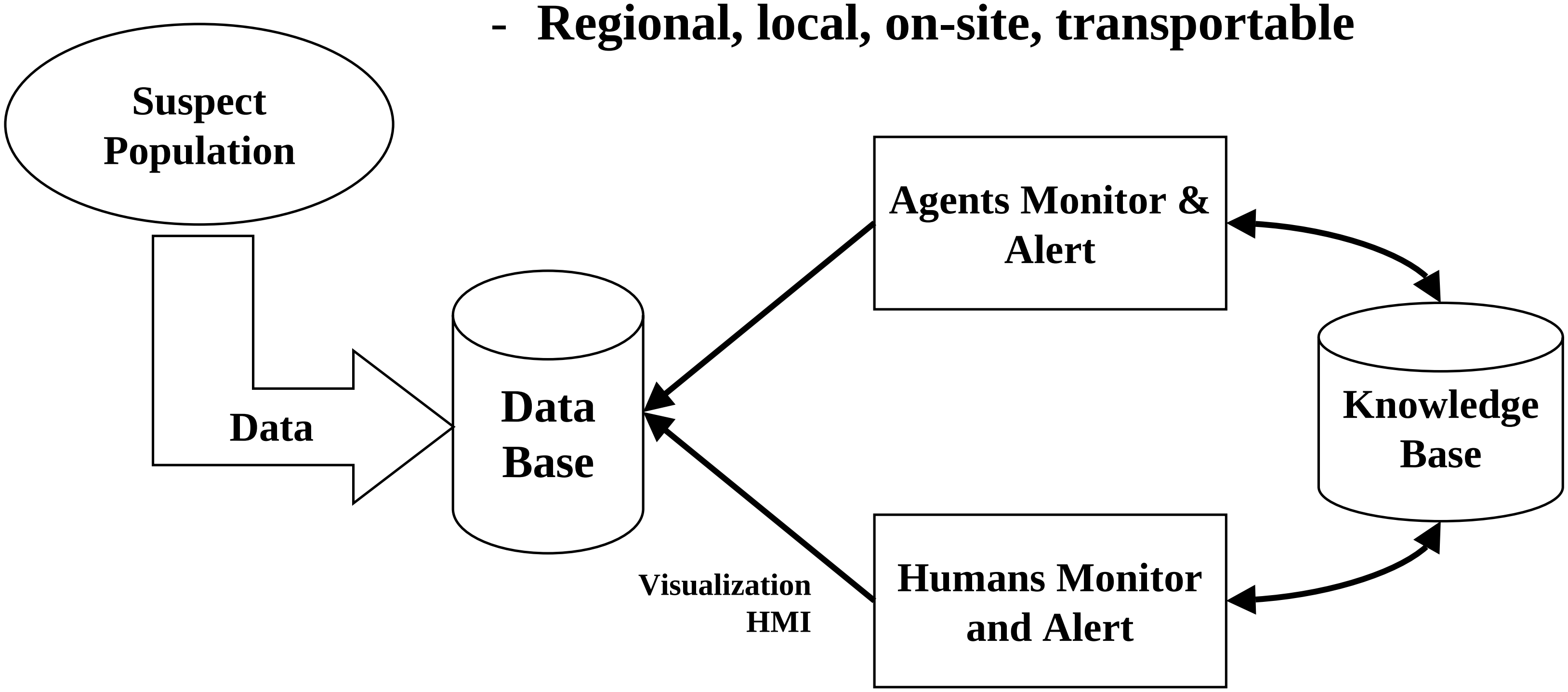
- Search DBs
  - Recognized features
  - Others . . .
- Mining *Structured* Data
  - E.g., commercial transaction data
  - Off-the-shelf technologies available but difficult to use
  - High maturity but visualization and HMI development required
- Mining *Unstructured* Data
  - Low maturity
  - Data representation and association, automation tools, HMI and visualization require major R&D

Behavior Analysis

- Scope
  - Suspect entitities
- Low technology maturity
  - Many components available but major integration engineering required
  - Robust and reliable monitoring technology not available
    - » Prohibitively high false alarm rate
    - » Human-in-the-loop signal detection
    - » Visualization and HMI R&D

Behavior Analysis (cont'd)

- **Objective Distributed Technology**
  - Regional, local, on-site, transportable



Summary

- **Link Analysis and Data Mining Are "Low Hanging Fruit"**
  - **Technologies "almost there" and potentially most productive in generating useful intelligence**
  - **Technology components exist but visualization and HMI are poor**
  - **Most difficult challenge is algorithm "scaling"**
  - **Technologies are evolving and may be influenced by working groups**
- **Matrix to be developed later**

Questions?