FOI, Swedish Defence Research Agency, is a mainly assignment-funded agency under the Ministry of Defence. The core activities are research, method and technology development, as well as studies conducted in the interests of Swedish defence and the safety and security of society. The organisation employs approximately 1000 personnel of whom about 800 are scientists. This makes FOI Sweden's largest research institute. FOI gives its customers access to leading-edge expertise in a large number of fields such as security policy studies, defence and security related analyses, the assessment of various types of threat, systems for control and management of crises, protection against and management of hazardous substances, IT security and the potential offered by new sensors.

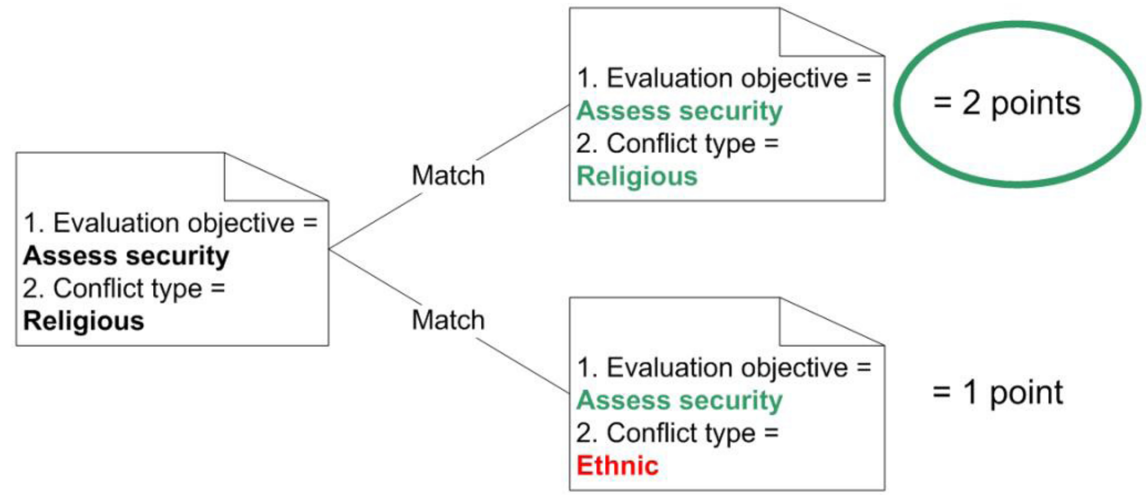
FOI
Swedish Defence Research Agency
Defence Analysis
SE-164 90 Stockholm

Phone: 08-55 50 30 00
Fax: 08-55 50 31 00

www.foi.se

Technology to Support Assessment of Operations

CHRISTIAN MÅRTENSON, PONTUS SVENSON



FOI-R--2871--SE
ISSN 1650-1942

Användarrapport
December 2009

Defence Analysis

Christian Mårtenson, Pontus Svenson

Technology to Support Assessment of Operations

Titel	Teknik för insatsutvärdering
Title	Technology for Operational Assessment
Rapportnr/Report no	FOI-R--2871--SE
Rapporttyp/Report Type	Användarrapport/ User report
Månad/Month	December
Utgivningsår/Year	2009
Antal sidor/Pages	43 p
ISSN	ISSN 1650-1942
Kund/Customer	Försvarsmakten
Projektnr/Project no	E11109
Godkänd av/Approved by	Göran Kindvall
FOI, Totalförsvarets Forskningsinstitut	FOI, Swedish Defence Research Agency
Avdelningen för Försvarsanalys	Defence Analysis
164 90 Stockholm	SE-164 90 Stockholm

Sammanfattning

Denna rapport ger en kort introduktion till några teknikområden som bedömts vara relevanta för framtida utveckling av tekniska stödverktyg för operationell utvärdering. Dessutom beskrivs kortfattat några existerande verktyg som skulle kunna anpassas för utvärdering.

Nyckelord: utvärdering, multifunktionella insatser

Summary

This report gives brief introductions to some areas of technology judged to be relevant for a possible future development of computer support tools for operational assessment. In addition, some existing tools that could be adapted for use in assessment are described.

Keywords: assessment, multifunctional operations

Table of Contents

1 Introduction

1.1 Purpose

The purpose of this report is to prepare for future work in constructing computer tools to support operation assessment and evaluation. Currently, there is no standard method for such assessment. Hence, the report surveys some technologies that are potentially interesting for planning and conducting evaluation and assessment of multifunctional operations. The report is meant to be read by researchers as well as military personnel involved in work to define standard methods for operation assessment and evaluation. It is meant more as a starting point than as a complete guide to technological areas that could be relevant to a future method for assessment and evaluation. Hence, is does not go into detail on any of the tools described.

Multifunctional operations are defined [1] as operations with a large number of actors, each having different mandates, means and methods for achieving their specific goals in the conflict area. The actors, who in addition to military units can be civilian authorities and non-governmental organizations (NGO's) most often come from several different countries and cultures and can have partially overlapping and partially disjunctive goals. Multifunctional operations are characterized by a need to understand what factors are important for collaboration between the different actors. Compared to standard military operations, it is more difficult to determine appropriate measures of effectiveness of our actions – i.e., to determine if a specific sequence of events was caused by our actions or not.

Hence, there is a need for research and development in the area of evaluations and assessments of multifunctional operations. Currently, there are no standard methods available, although there is some work in progress on the issue [2].

The goal of this report is to support the future development of standard methods for evaluation and assessment by conducting a scan of technical areas that could be of interest for this application, with the aim of providing a smorgasbord of possibly applicable technical areas. The purpose is to bring technological possibilities and limitations to attention to the methodology people working on evaluation development.

1.2 Scope and delimitations

In this report, we focus only on some of the technology areas that could be of interest for evaluation and assessment.

In section 2, a brief introduction to some of the most common methodologies for management of uncertainty is given. This is followed by a section on causality and causal analysis. These sections give important background knowledge. All information available in multifunctional operations will be uncertain, and determining the causal links between actions and effects could be seen as the major problem of evaluation and assessment – by determining the causal relations between our actions and goals, we also implicitly evaluate them. Section 4 builds on the base laid in the earlier sections and discusses several different ways of structuring the assessment process by looking for so-called indicators. This theme is continued in subsection 4.4, where some related tools are described briefly.

Visualization is always an important ingredient of any analysis process. Section 5 gives a very short description of the field of visual analytics and lists some relevant tools. Section 6 presents the field of case-based reasoning, which has the potential to allow re-use of knowledge in assessment situations. The report ends with a brief summary and some recommendations for future work.

There are large numbers of COTS[1] products available for basic data analysis and simple information management (e.g., spreadsheets and databases). Hence, the report does not include discussion of such tools.

In the cases where existing tools are listed, they are meant as examples and not as an exhaustive list of available tools.

1.3 Challenges where technology could help

There are several instances where technology tools can be very useful for assessment and evaluation. Here follows some comments from users indicating the need for technology tools in operational assessment and evaluation.

- "A member of the group should have specific skills in visualization techniques; and may require the use of additional visualization tools", a key finding in MNE 5 UK/US CIP/CIME Limited Objective Experiment (LOE), October 2007.
- "Conflict prevention and peace-building activities often lack some or all of these preconditions for a variety of reasons, especially when they are performed during and after open armed conflict, (often due to the often limited time for planning).", OECD-DAC
- "Post conflict situations teem with all types of information: rumours, conjectures, half-truths, first-hand, second-hand and third-hand information, misinformation and sometimes, too, the right information, at the right time to the right people." OECD-DAC

[1] COTS = commercial off the shelf

2 Managing uncertain data

2.1 Introduction

Handling uncertainties is a general problem for all military operations. Uncertainties arise from what is sometimes called the frictions of war as well as from inexact statements and reports in addition to the inherent uncertainty that is always associated with human intentions and actions. In multifunctional operations, the problem is exacerbated by the fact that there are several different actors, whose intents and goals are often not well-known.

There are very many different mathematical models and theories for uncertainty management. Each methodology has its strengths and weaknesses, and choosing which one to use for a specific application is not always straight-forward. For any military information management or processing system, it is important to realize that all results will have some kind of uncertainty associated to it. In addition to being able to display this uncertainty in an adequate way to the user, it is also important to be able to determine how the uncertainty may be changed by the information processing. In general, a distinction between two cases can be made: adding more information to reduce uncertainty, and extrapolating from available information, which will produce prognoses that are more uncertain than the input data.

Information can be uncertain, or defective, in several different ways. A statement can be:

- ambiguous
- uncertain
- imprecise
- incomplete
- vague
- inconsistent.

Depending on which of these characteristics is most important, different methodologies for handling the uncertainty must be chosen. For example, fuzzy logic was invented to handle vague information, while Dempster-Shafer theory is appropriate for imprecise evidence.

In situations with antagonistic opponents, it is also important to consider the possibility of deception, i.e., that the opponent will try to fool us. This raises additional questions of how to separate uncertainty due to”bad sensors/ processing” from uncertainty due to deception.

Some of the different theories for management of uncertainty that have been suggested so far are:

- Logic
- Probability theory
- Bayesian analysis
- Maximum likelihood methods
- Evidence theory
- Fuzzy sets
- Random sets

2.2 A simple maximum likelihood example

To get a taste for uncertainty, consider the following problem. You capture 1000 fishes in a lake, mark them and release them. Shortly afterwards, you capture another1000 fishes. By inspection, you determine that 100 of these, or 10%, are marked. Your task is now to estimate how many fishes there are in the lake.

The answer is easily computed using rather pedestrian mathematics. Figure 1 below shows the resulting probability distribution for the number of fish in the lake.

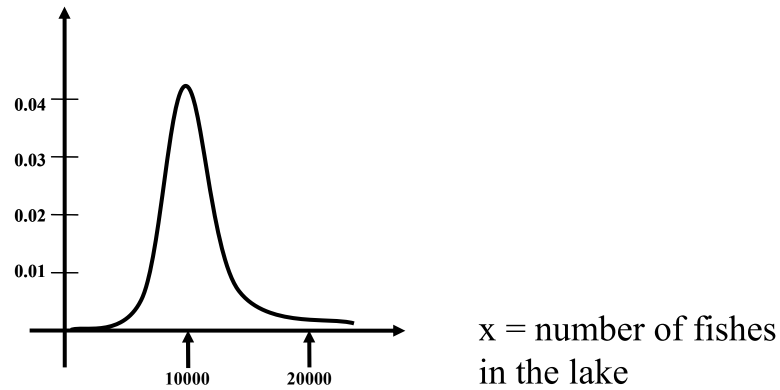


Figure 1 shows the probability distribution of number of fish in the lake.

This very simple example gives rise to several important issues that need to be considered for any system. What is the correct answer for the number of fishes? Is it the value for which the probability is maximised? Or should an interval, containing, e.g., the correct answer with 90% be given? The answer depends on the application, user and situation – uncertainty visualisation systems need to be adaptive and change displays according to the needs of the user.

2.3 Bayes rule and Bayesian analysis

We will return later to the question of causal analysis. If we know that two events A and B are causally connected, Bayes rule allows us to do some computations on the probability of one event when the other is observed. Consider a case where you know that A causes B with probability p. A could for instance be the event that somebody passes your house, while B corresponds to your dog barking. If we now hear the dog bark (i.e., event B occurs), can we then compute the probability that this was caused by somebody passing the house? There might of course be several different reasons for why the dog would bark. How can we be sure that it really was caused by somebody passing the house?

Bayes rule, which gives a relation between the *conditional probabilities* P(A|B) and P(B|A), makes this possible.

We can deduce Bayes rule quite easily by starting from the equation

$$P(A \text{ and } B) = P(A) * P(B \mid A)$$

for the intersection of two events. Now the same rule also applies if we switch A and B, and hence

$$P(A \text{ and } B) = P(B) * P(A \mid B)$$

By equating the two right hand sides, we get

$$P(A|B) = P(B|A) * P(A) / P(B)$$

which gives us the desired relation and allows us to compute the reverse probability P(A|B) given P(B|A).

Bayes rule is often used in conjunction with so-called *influence diagrams*. These are graphical representations of events and causal links between them. The figure below shows one example.

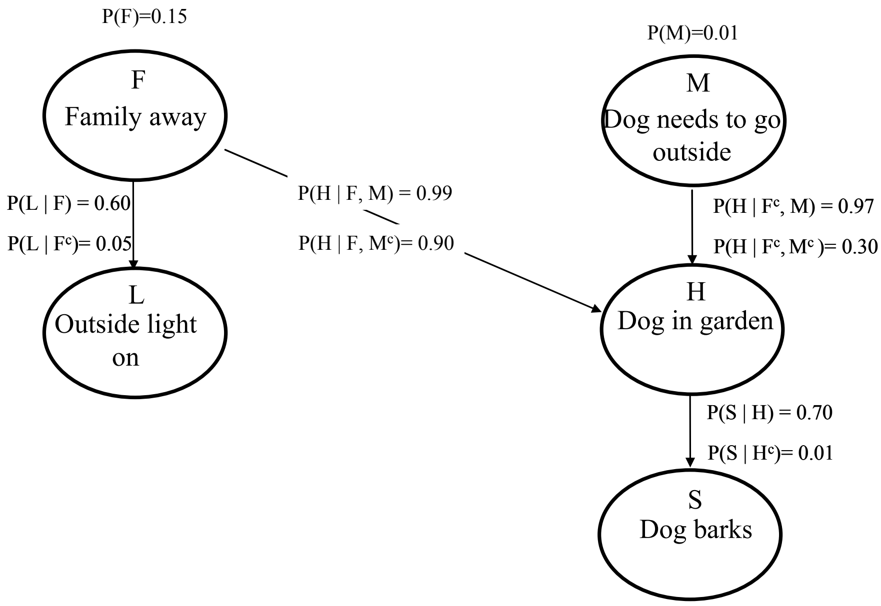


Figure 2 shows an example of a Bayesian belief network.

The information presented in the figure tells us what events are causally related and also the conditional probabilities.

Given the information in the network, often called a graphical model, we can calculate, for instance, the probability that the family is away given that the outside light is on and the dog is not barking, to be

$$P(F \mid L \;\&\; \text{not } S) = 0.500551 \sim= 0.50$$

Bayes rule and the Bayesian network thus give us the possibility to calculate probabilities of events even though we are not observing them, but rather consequences of the events of interest. Uses of Bayesian networks are ubiquitous, as will be seen in later sections. Figure 3 shows an example Bayesian network used for modelling movements of a tank platoon.

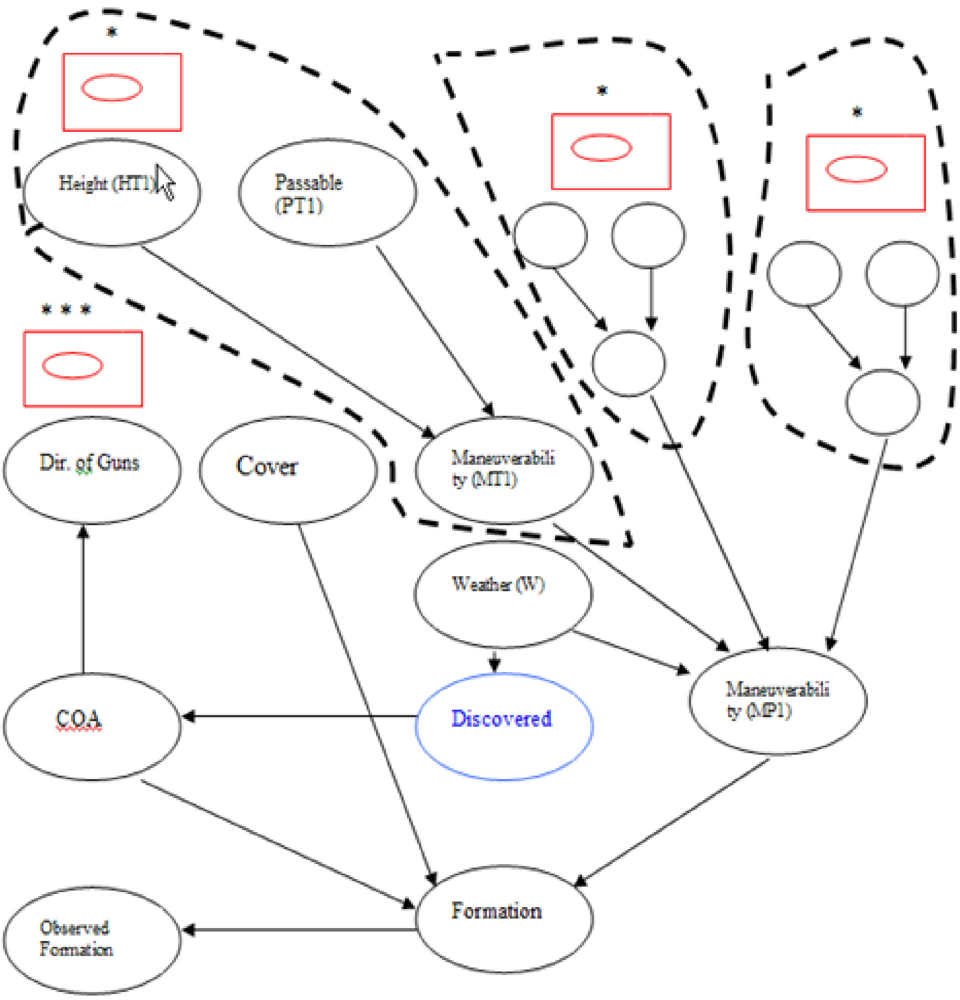


Figure 3 shows an example of a Bayesian network.

Bayesian decision analysis can be summarised in the following bullets

- Model *beliefs* about a *parameter of interest* through a prior probability which, in presence of further information, is updated to the posterior
- Model *preferences* and *risk attitudes* about (possibly multi-criteria) consequences with a (multi-attribute) function
- Associate with each alternative its (multi-attribute) *posterior expected utility*
- Propose the alternative which *maximises* the posterior expected utility

An important extension of Bayesian analysis is *robust* Bayesian analysis. Here, we allow several different prior probabilities and/or conditional probabilities. This is sometimes also called second order probability, and allows us to have uncertainty not only in the variables but also in the probability distributions.

2.4 Dempster-Shafer theory or evidence theory

Dempster-Shafer theory [3] is an extension of probability theory that allows us to easily reason about imprecise or unspecific events. Standard probability theory only deals with specific events, such as “the object is a tree”. Dempster-Shafer theory provides formalism for discussing unspecific events, such as “the object is a tree *or* a bird”. This is useful for example when we are unable to determine exactly what a specific piece of evidence points at. Dempster-Shafer theory can be understood in terms of the so-called basic probability assignment or mass function m.

Given atomic statements a,b,c,…, in Θ, the mass function is defined on all combinations, i.e. all subsets of Θ, and must fulfil some requirements:

- $m(\emptyset) = 0$
- $0 \leq m(x) \leq 1$
- $\Sigma m(x) = 1$

The set of all subsets of Θ for which the mass function is non-zero is called the *focal elements* of m. The mass function is used to represent uncertainty. For instance, we can define a mass function for the question “who will win the world cup?” as

- m({Germany}) = 0.4
- m({Sweden }) = 0
- m({Germany, Italy}) = 0.3
- m({Germany, Italy, France}) = 0.1
- m({Θ}) = 0.2

The interpretation of this function is that we have evidence that indicates that Germany will win, but also have some evidence that points to either Germany or Italy winning. It is important to realise that this does not mean that the probability that Germany will win is 0.4. We can only say that the probability that Germany will win is at least 0.4.

Mass functions can be combined in much the same way as probability distributions, and it is in fact possible to base influence diagrams on them. Combination of two mass functions is done by Dempster’s rule, which is quite easy and intuitive. If we have evidence pointing towards (A, B, C) and evidence pointing towards (B, C, D), the combination should point to (B, C), i.e., the intersection between the two sets.

We can also argue that the mass for the intersection should be the product of the masses pointing towards (A, B, C) and (B, C, D). Dempster's rule is the generalisation of this where we sum over all possible combinations. It can be written as

$$m(A) = K \ \Sigma_{x \cap y = A} \, m_1(x) \, m_2(y)$$

where K is a normalisation constant, which is necessary in order to ensure proper normalisation.

To illustrate Dempster's rule and evidence theory, we will work through a classical example. Consider a scenario where there is a murder investigation and four suspects: A, a business associate to the victim O; B, O's butler; S, O's son; and V, a friend of O.

There are several clues found by the police investigators:

1. Cigarette butt found on scene of crime;
2. Somebody has overheard an argument between the victim and either the butler or the son;
3. A shoe-print is found;
4. And a blond hair has been found on the scene of the crime.

After some initial questions, the detectives also establish some background knowledge that will help them match the clues to the right persons:

- V does not smoke;
- Many witnesses heard the argument;
- Shoeprint-size matches B and V;
- Only V is blonde.

Now comes the most crucial step in any application of Dempster-Shafer theory. Guided by her experience, the detective must determine numerical weights for each of these evidences, in the range 0 (completely certain not true) to 1 (completely certain true). These weights will be used as basic probability assignments for the sets consistent with the different clues. One separate mass function will be specified for each clue, and the clues will then be combined using Dempster's rule. The detective establishes the following mass functions

- $m_1(\{A, B, S\})$ = 60% (cigarette; V non-smoker)
- $m_1(\Theta)$ = 40%;

- $m_2(\{B, S\}) = 80\%$ (argument)
- $m_2(\Theta) = 20\%$;
- $m_3(\{B, V\}) = 60\%$ (shoeprint)
- $m_3(\Theta) = 40\%$;
- $m_4(\{V\}) = 30\%$ (hair)
- $m_4(\Theta) = 70\%$

These functions represent four mass functions or *basic belief assignments* defined on the space $\Theta = \{A, B, S, V\}$. Note particularly the inclusion of mass for Θ in each of the evidences, and recall that this means the basic probability mass that cannot be assigned to any smaller set of suspects. Evidence 4, for instance, points directly to the friend with 30% likelihood. This does however not mean that the probability that the friend is the perpetrator is 30% – the friend is also included in the set Θ, so the only thing we can say is that there is at least a probability 30% that the friend did it and at most a 100% probability that the friend did it.

This is the crucial realisation needed to understand Dempster-Shafer theory. Comparing to normal probabilities, it would not have been possible to also include the friend in the complementary probability, i.e., the clue would have had to be translated into a statement that the friend did it with 30% likelihood and there is a 70% likelihood that the friend did not do it.

Now we will use Dempster's rule to combine the evidences. First combine m_1 and m_2 (cigarette and argument), which gives us the mass function m_5. The combination is done using a simple graphical short-hand: construct a table whose rows and columns are indexed by the focal elements of the two mass functions we wish to combine. For each cell in the table, write both the intersection of the focal elements of the row and column and the product of their masses.

m_5	m_2	0.8	0.2
m_1	X	{B,S}	Θ
0.6	{A,B,S}	**{B,S}, 0.48**	**{A,B,S}, 0.12**
0.4	Θ	**{B,S}, 0.32**	**Θ, 0.08**

The results of the combination can be read off from the bottom right elements of the table. In our case, we get

- $m_5(\{A,B,S\})= 0.12$
- $m_5(\{B,S\}) = 0.48+0.32 = 0.8$

- $m_5(\Theta) = 0.08$

The evidence processed so far seems to implicate the son or the butler. The next step is to combine m_5 and the shoeprint m_3:

m_6	m_3	0.6	0.4
m_5	X	{B,V}	Θ
0.8	{B,S}	**{B}, 0.48**	**{B,S}, 0.32**
0.12	{A,B,S}	**{B}, 0.072**	**{A,B,S}, 0.048**
0.08	Θ	**{B,V}, 0.048**	**Θ, 0.032**

Again, we read off the result from the bottom right, getting:

$m_6(\{B\})$ = 0.552

$m_6(\{B,S\})$ = 0.320

$m_6(\{B,V\})$ = 0.048

$m_6(\{A,B,S\})$ = 0.048

$m_6(\Theta)$ = 0.032

We have one clue left; m_4 which pointed to the friend V. Combination with m_6 gives us

m_7	m_4	0.3	0.7
m_6	X	{V}	Θ
0.552	{B}	**Ø, 0.1656**	**{B}, 0.3864**
0.32	{B,S}	**Ø, 0.096**	**{B,S}, 0.224**
0.048	{B,V}	**{V}, 0.0144**	**{B,V}, 0.0336**
0.048	{A,B,S}	**Ø, 0.0144**	**{A,B,S}, 0.0336**
0.032	Θ	**{V}, 0.0096**	**Θ, 0.0224**

Resulting in m_7:

$m_7(\emptyset)$ = 0.1656+0.0960+0.0144 = 0.2760

$m_7(\{B\})$ = 0.3864

$m_7(\{V\})$ = 0.0144+0.0096 = 0.0240

$m_7(\{B, S\})$ = 0.2240

$m_7(\{B, V\})$ = 0.0336

$m_7(\{A, B, S\}) = 0.0336$

$m_7(\Theta) = 0.0224$

We seem to have a problem. m_7 has mass on the empty set Ø, which is not permitted. To remove this, we need to use the normalisation constant K from Dempster's rule. We'll simply take the mass assigned to Ø and redistribute it to all other sets by dividing m_7 with 1-0.276 = 0.724, resulting in

$m_7(\{B\}) = 0.3864 / 0.724 = 0.5337 =$ c:a 53 %

$m_7(\{V\}) = 0.0240 / 0.724 = 0.0331 =$ c:a 3 %

$m_7(\{B, S\}) = 0.2240 / 0.724 = 0.3094 =$ c:a 31 %

$m_7(\{B, V\}) = 0.0336 / 0.724 = 0.0464 =$ c:a 5 %

$m_7(\{A, B, S\}) = 0.0336 / 0.724 = 0.0464 =$ c:a 5 %

$m_7(\Theta) = 0.0224 / 0.724 = 0.0309 =$ c:a 3 %

Here we have seen another of the crucial steps in applying Dempster-Shafer theory: the occurrence of probability mass for the empty set is a sign of conflicting evidences, and needs to be handled. There is a variant of the Dempster-Shafer theory where instead of distributing this conflict mass to the other sets, one simply discards it or keeps it on Ø. This is called the transferable belief model, and corresponds to an open world assumption.

The presence of a large mass of conflict could also be a sign that we haven't included all possible hypotheses in our set Ø. There could be another suspect who matched all the clues.

In order to reach conclusions regarding the evidence, we will introduce two other basic concepts of Dempster-Shafer theory: the belief interval consisting of a *possibility* and a *plausibility* for each element. The possibility (also called belief) of a set is all the mass that explicitly indicates some elements of that set, while the plausibility is all the mass that does not explicitly exclude the elements of the set. For instance, for the set {B,S} we get the belief $Bel(\{B, S\}) = m_7(B) + m_7(B,S) = 0.5337 + 0.3094 = 0.8431$, and the plausibility $Pl(B, S) = 1 - Bel(\Theta \setminus \{B,S\}) = 1 - Bel(\{A,V\}) = 1 - m7(V) = 1 - 0.0331 = 0.9669$.

For this set we thus get the belief interval [0.8431, 0.9669] or approximately [0.84, 0.97]. We can do the same calculation for all the different subsets of Θ, resulting in

{A}: [0.00, 0.08]

{B}: [0.53, 0.97]

{S}: [0.00, 0.39]

{V}: [0.03, 0.11]

{B, S}:	[0.84, 0.97]
{B, V}:	[0.61, 1.00]
{A, B, S}:	[0.89, 0.97]

The interpretation of the results is another tricky part of Dempster-Shafer theory. The detective chooses to arrest the butler, since by combining the evidences it can be seen that there is 53% that explicitly points to the butler, and only 3% that excludes the butler.

In this simple example we have seen both some of the strengths and weaknesses of Dempster-Shafer theory. Like in all uncertainty theories, the modelling of the problem is a crucial step. In order to make use of Dempster's rule for combination, it is necessary to first find an appropriate space Θ to work in, and then to translate all the different evidence into mass functions defined on Θ.

2.5 Data collection

One of the most challenging issues when designing a system to handle uncertainty is how to determine the specific values of uncertainty in the first place. For sensors or other automatic systems, we can use Bayesian reasoning to determine the probability that, e.g., a vehicle of a specific type has passed an acoustic sensor. This is done by collecting large amounts of statistics on what signals are emitted by different vehicles passing a sensor of the same type and then using Bayes rule to determine the inverse probability that a specific type has passed given that a certain emission was detected.

For mathematical reasoning with uncertain data, it is often necessary to determine specific numbers for each data. It is often argued (see, e.g. [4]) that these can be determined by so-called betting arguments.

For more soft data, such as results from interviews of input from humans, the situation becomes more complicated. There exist standard measures of uncertainty in intelligence reports [5], where each intelligence item is classified according to its *reliability* and the *credibility* of the reporting source. However, there are no strict guidelines for when to assign the different grades to a piece of evidence.

Social scientists, who rely on interviews and other qualitative means of acquiring data rather than output of sensors and signal processing, are often faced with this problem.

For the purpose of evaluation and assessment of multifunctional operations, there are several different kinds of data that are relevant. Since state-building is often part of the goal in the operations, economic and cultural indicators such as the rate of unemployment or the number of girls in school are often very important.

Such figures can be analysed using various statistical softwares. Perhaps the most common tool for such analyses is Excel.

There are also other kinds of input data that can be important in multifunctional operations. These types come from interviews and field studies of the situation in the area of operations. The interviews can be either written or, more common, taped conversations.

For such material, it is important to have software tools that enable the analyst to annotate and structure this often unstructured material.

One such software tool is atlas.ti[2], which is used by many social scientists to code interviews and other data. The tool has options for simple text mining and searches on the data that has been input to it, and also intuitive interfaces for tagging data, both manually and semi-automatically, using queries. Advanced features include support for transcription, multimedia data sets, and connection to Google Earth for geo-mapping.

[2] http://www.atlasti.com/

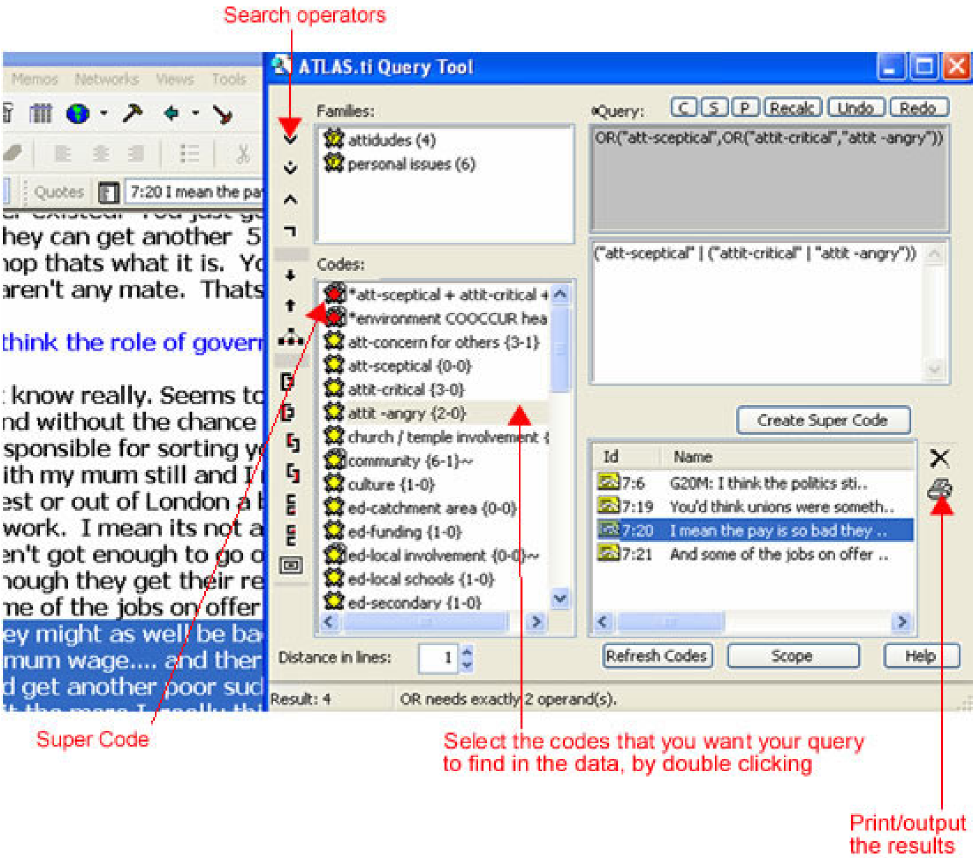


Figure 4 shows the interface of atlas.ti.

3 Causal analysis

The question of attribution and causal analysis is intimately connected to evaluation and assessment of operations. There is a rich body of work in statistical theory on the meaning of causality, and several different competing methodologies for how best to test causal connections. Perhaps somewhat surprisingly, there is still considerable debate on the correct way to think about causal relations, as witnessed for example by the web log debate[3]. There are several different perspectives. Here, we will briefly describe two of them, due to Rubin and Pearl, respectively. A recent survey focusing on Pearl's model is [6]; see also the book [7]. Yet a third perspective is given by Shafer in [8]

The Rubin causal model starts with the assumption that different interventions will lead to different outcomes. This is called the "potential outcomes" framework, and simply means that in order to test the causal relation between an action A and an effect E, we must look at two different cases: one in which we apply the action A and one in which we do not. If the outcome is the same, then A has no causal effect, whereas if the outcome is different, we can quantify the causal effect of A. This example, while seemingly trivial, also captures the so-called fundamental problem of causality: we cannot test both outcomes, since by applying the action A, we could irreversibly change the state of the world. To take a concrete example, we can't test if a certain medicine works for a disease by first subjecting the same patient to the treatment and then later expose the same patient to non-treatment.

In order to overcome this difficulty, it is necessary to use multiple subjects when testing medicine. For similar reasons, simulations used for decision support for military commanders must use several different simulations in order to produce reliable results.

Using different subjects to test a medicine, however, also results in other difficulties. When comparing treatment A on patient X with no treatment of patient Y, we cannot be sure if the difference in the result stems from the application of the medicine or from other differences in the patients.

Statistics overcomes these problems by using many subjects and by trying to control for different causes.

[3] http://www.stat.columbia.edu/~cook/movabletype/archives/2009/07/disputes_about.html,
http://www.stat.columbia.edu/~cook/movabletype/archives/2009/07/philip_dawids_t.html.
http://www.stat.columbia.edu/~cook/movabletype/archives/2009/07/more_on_pearls.html,
http://www.stat.columbia.edu/~cook/movabletype/archives/2009/07/more_on_pearlru.html,
http://www.stat.columbia.edu/~cook/movabletype/archives/2009/07/rubinism_separa.html,
http://www.stat.columbia.edu/~cook/movabletype/archives/2009/07/where_to_start.html,
http://www.stat.columbia.edu/~cook/movabletype/archives/2009/07/pearls_and_gelm.html

The problem an analyst is facing when doing an assessment of an operation is considerably harder. If the task is to determine if action A caused a specific effect in, for example, Afghanistan, it is not possible to do several independent experiments to determine the statistics. If a sufficiently detailed simulation model of the area of operations could be built, then the problem could in principle be solved by running many simulations and noting if the causal relationship one is looking for is present in them. However, modelling any area of conflict in sufficient detail to make this possible seems impossible. It is of course possible to simplify the model and look for qualitative causal relations, but one must then be wary of the conclusions one can draw.

One other important model of causality is based on structural equation modelling [7]. In this framework, which was extended and formalised considerably by Pearl, one uses both experimental data and hypotheses about the qualitative causal structure as inputs. This comes in the form of hypotheses that A causes B that are embedded in a network structure. Structural equation modelling can be compared to Bayesian belief networks, but are more general and allow for instance causal cycles (i.e., loops where A causes B and B causes A).

The advantage of this model is that the causal assumptions used very often have implications that can be tested directly on the experimental data.

4 Modelling and reasoning

4.1 Conflict analysis and theories of change

An important part of operational assessment is to assess to what degree the assumed *theories of change* are correct. The theories of change are causal relationships connecting the inputs, outputs and effects of activities that are believed to lead towards mission goals. The following is an example of a theory of change from [9], where the implementation of a Counter Insurgency doctrine, COIN, is suggested to lead to a number of improvements for the stability in Afghanistan.

"If we employ a suitable COIN approach, then the insurgency will be quelled/ reduced/ weakened and the GoA (Government of Afghanistan) legitimacy and popular support will increase, thereby enabling conditions for economic development, good governance and the sustained provision of essential services in Afghanistan"

As a basis for identifying theories of change a conflict analysis should be performed. This can include identifying the profile of the conflict, its causes and potentials for peace, the involved actors and the conflict dynamics and future trends [10].

It is important that the theories of change are made explicit and formulated in a concrete and non-ambiguous manner. One way of doing this is to use some sort of formal syntax, e.g. by drawing influence diagrams as in Figure 5.

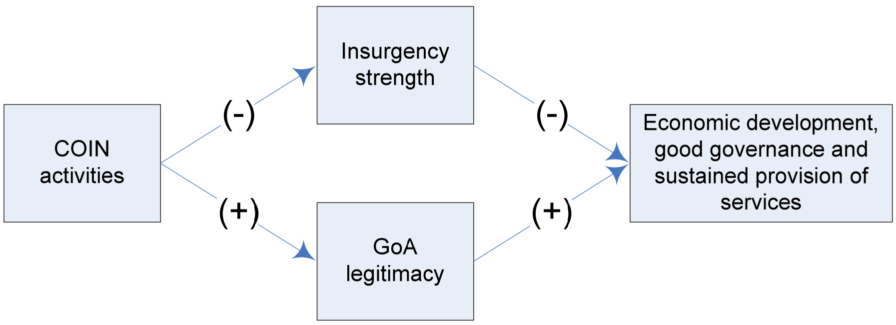


Figure 5 shows the previous example as an influence diagram. The arrows represent causal relationships and the plus and minus signs give a coarse description of how the nodes influence each other. If the COIN efforts are increased the GoA legitimacy will increase (positive correlation (+)) and the Insurgency strength will decrease (negative correlation (-)).

This also enables complicated relationships to be captured and lucidly presented. As a consequence, a greater part of the conflict analysis can be included as theories of change relationships in a model that could be seen as the product of a system analysis, see Figure 6. An introduction to the methods of system analysis and its applications to conflict analysis can be found in [11].

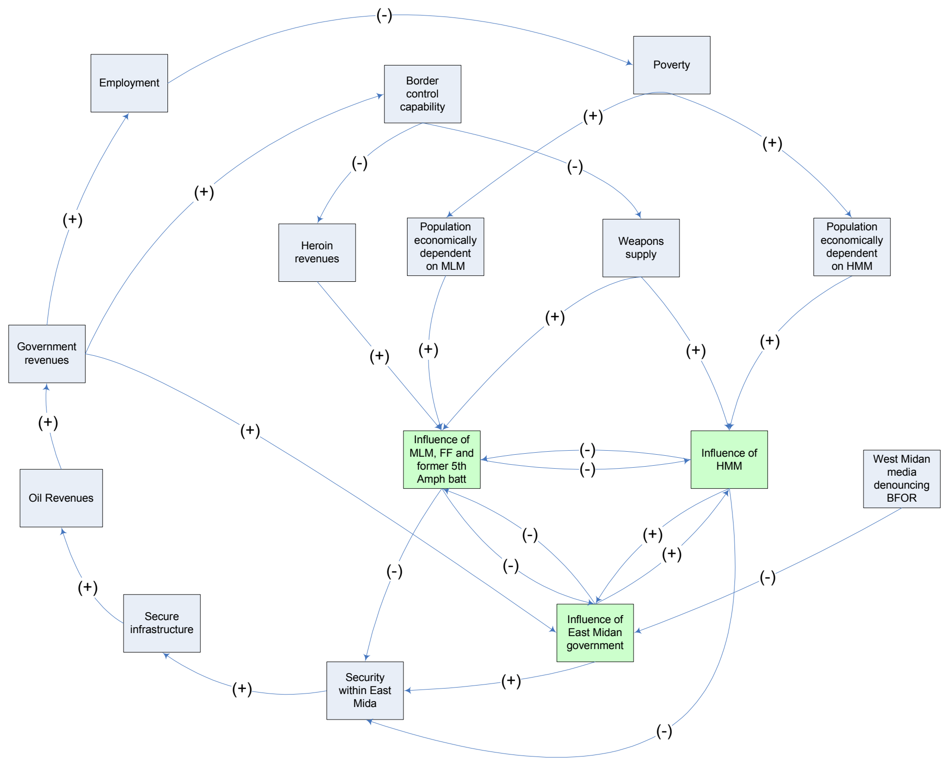
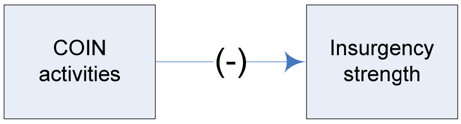


Figure 6 gives an example of a complex, but due to the benefits of influence diagrams, still comprehensible conflict analysis. The example is the result of a systems analysis of a demo scenario at FOI [12].

This type of conflict/system analysis can gain a lot by using appropriate technological tools. Simple tools can be used to smoothly create, edit and visualize influence diagrams. More advanced tools can be used to combine the influence diagrams with more detailed models and data sources for quantitative analysis and simulations to study conflict/system behaviour when some parameters are changed.

How to evaluate a theory of change when expressed as an influence diagram depends on the kind of data that is available. If the values of the nodes in the diagram can be measured in numbers, a purely *quantitative* evaluation of the relationship is possible. The current values of the nodes A and B are compared to their *baseline* values, the values they had at the beginning of the operation. Then it is determined whether the differences are compatible with the predictions of the influence diagram or not.



	COIN activities	Insurgency strength
Baseline	347	239
Current	385	298
Change	+	+

Figure 7 shows a casual relationship and how it can be quantitatively evaluated. The amounts of both nodes have increased (in an unspecified measure), which contradicts the prediction of negative correlation given by the influence diagram. This indicates that the conflict analysis is incorrect.

In many cases, quantitative data can be hard to obtain. The reasons for this can be plentiful: baseline data is missing, the security situation does not allow necessary data collection, or the resources to perform data collection are not sufficient. In such cases the assessment of the theories of change can be made in a more qualitative manner. This is done by estimating the differences directly through e.g. interviews where statements like "A is much worse now than last year" or "B is continuously being improved" are collected. These are then compared to the predictions of the influence diagram, see Figure 8.

	COIN activities	Insurgency strength
Current	"Little more than baseline"	"More than baseline"
Change	+	+

Figure 8 shows how the example from the previous figure would look like in the case of qualitative data without explicit baseline measurements.

4.2 Indicator breakdown

When preparing an operational assessment the main tasks often boil down to a list of specific questions that must be answered. Some might have to do with the success-rate of operations, others with the assessment of the theories of change. These questions are in many aspects similar to knowledge requests (KR) that occur in the intelligence domain. Thus they can often be treated in an analogous manner.

A KR is issued when information is lacking in a decision situation. The request is often formulated as a simple yes/no question (or true/false statement), but can be on a high level of abstraction. This makes it necessary to decompose the request into multiple sub-queries, each on a lower level of abstraction than the original query. Sub-queries are further decomposed until they reach a level where it is possible to find a direct answer to them, either by consulting a knowledge base or by directing sensing capabilities (sensors or human observers) to observe some part of the area of interest. Within the Swedish Armed Forces (SwAF), this kind of observable sub-queries are called *indicators*. When the answers to the indicators have been collected, they are combined to give the answers to their super-queries, which in turn are combined to answer the queries on the next level above, and so on until the initial knowledge request is answered. This procedure is called structured argumentation, and is further described in the next section.

The MPICE[4] framework [13] is an example of an indicator breakdown structure in the domain of conflict assessment. It is a hierarchical system of goals, indicators and measures that tries to capture the essential aspects of conflict drivers and stabilizing factors. The goals are grouped in five different themes:

1. Political Moderation and Stable Governance
2. Safe and Secure Environment
3. Rule of Law
4. Sustainable Economy
5. Social well-being

Each theme has two types of goals, Drivers of Conflicts, which should be diminished, and Institutional Performance, which should be strengthened. The goals are divided into a number of indicators, which are further decomposed into measures, see Figure 9. The indicators consists of yes/no queries and the measures of statements that can be fulfilled to different (measurable) degrees. Note that the terminology differs slightly between MPICE and SwAF; in SwAF the indicators are the lowest level observables while in MPICE it is the measures.

[4] MPICE = Measuring progress in Conflict Environments

Goal A: Competition for Exclusive Power Diminished (Driver of conflict)

To what extent do political elites/leaders and identity groups perceive the political process in exclusive (i.e., "zerosum") terms?

- Perception among identity group members that loss of power (e.g. to other identity groups) will eliminate the prospect of regaining power in the future. (S/PD)`-
- Perception among identity group members that loss of power (e.g. to other identity groups) will eliminate the prospect of progressing economically in the future. (S/PD)`-
- Public rhetoric from political elites/leaders asserting that their rivals have negotiated the peace settlement in bad faith (i.e. that the settlement is a trick or that their rivals will manipulate the peace settlement to assert control over security forces). (CA)`-
- Number of assaults and assassinations perpetrated by members of one of the former warring factions against leaders of other identity groups. (QD, EK) `-
- Number of assaults and assassinations perpetrated by members of one of the former warring factions against other members of their own identity group. (EK)`-
- Revisions to the Constitution or governance principles document to permit continuation in power of the incumbent. (EK)`-
- Revision of the electoral code to favor the incumbent. (EK)`-

To what extent are political elites/leaders polarized on the basis of their identity?

- Importance of identity group membership as a requirement for political leadership. (S/PD)`-
- Prominence of inflammatory and exclusionary rhetoric in the discourse of political elites/leaders. (CA)`-

Figure 9 shows an example of a “Driver of conflict” goal from the MPICE framework [5]. The goal consists of two indicators with seven and two measures respectively. Each measure is marked with the associated data collection type (S/PD=Survey/Polling Data, CA=Content Analysis, QD=Quantitative Data, EK=Expert Knowledge).

As the MPICE framework aims at being general enough to encompass a very large number of conflict types, it is only a fraction of all goals, indicators and measures that are relevant for a specific operational assessment. At the moment, the lowest level of the hierarchy consists of over 800 measures. This means that it might be useful to have a software tool that uses past experiences to suggest relevant subsets of the hierarchy, cf. Chapter 6.

4.3 Structured argumentation

As mentioned above, when a knowledge request (or goal) has been decomposed into observable indicators (or measures), information can be collected by consulting a knowledge base or by exploiting available sensing capabilities. The collected information is analyzed to answer the low-level queries, which then can be propagated upwards in the decomposition tree and be fused to answer the queries of the higher level. The fusion can either be made manually, based on the experience of the analyst, or if suitable it can utilize a supporting mathematical framework. The propagation and fusion is repeated until the initial query is answered.

The described procedure is known as structured argumentation, and is a standard method in intelligence analysis. There are many potential benefits to follow such a procedure, although seemingly little scientific effort has been put into proving it.

- **Cognitive biases**. By explicitly expressing the decomposition structure, analysts and evaluators are reminded of the full spectrum of indicators to be considered, hence encouraging a careful analysis and avoiding a narrowly focused analytic mindset [14, 15].
- **Explanation.** The important task of rigorously explaining the analysis or evaluation results to the customer is made easier when the logic of the analysis can be delineated, and it can be shown what assumptions were made and how evidence was used [15].
- **Collaboration**. As the previous bullet also applies to explaining the analysis to colleagues, structured argumentation also facilitates collaboration. A further argument in favour of this is that the explicit decomposition also allows for independent sub-tasking and a shared workload [14].
- **Corporate memory**. If a suitable software tool is chosen to support the structured argumentation, the decomposition can be stored for re-use in similar tasks in the future [14, 15]. A framework for such knowledge re-use will be described in Section 6.1.

There are a number of research and commercial software tools to help analysts working with indicators and structured argumentation. Some of them will be described in Section 4.4.

4.4 Tools

This section will briefly present a number of tools that support the modelling and reasoning aspects of operational assessment discussed earlier in the chapter.

4.4.1 Consideo Modeler

The Consideo Modeler is a commercial tool for constructing system models using influence diagrams [16]. The tool also supports quantitative model analysis through the use of simulations.

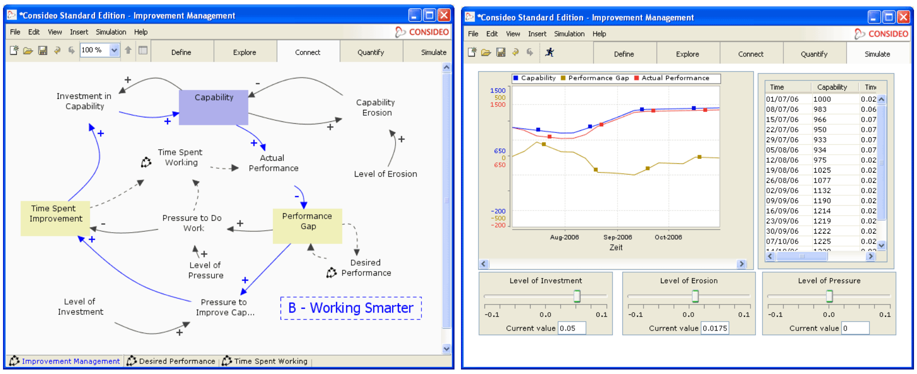


Figure 10 shows a pair of screenshots of Consideo Modeler.

4.4.2 Disciple-LTA

Disciple-LTA is a computer-based cognitive assistant, with the aim to help intelligence analysts in their daily work [15]. The system is a research prototype being developed at George Mason University. LTA means “Learner, Tutor and Assistant” which refers to the system’s ability to register procedures of experienced users and then use this knowledge to assist novice analysts. The system employs the method of indicator breakdown and structured argumentation described in the previous sections. It can also assist the user in *believability* analysis, by asking clarifying questions and keeping track of the chain of custody of evidence.

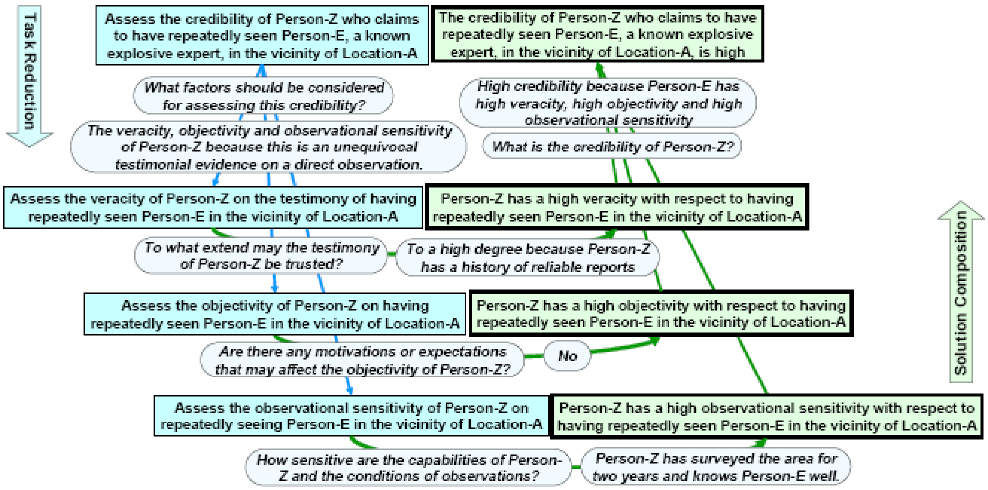


Figure 11 shows a screenshot of structured argumentation in Disciple-LTA.

4.4.3 High SEAS

High SEAS is a commercial web based tool for structured argumentation and visualization, developed by SRI International [17]. It supports the use of templates for best practice problem decompositions, and serves as a corporate memory as previous work is indexed and stored in a knowledge base.

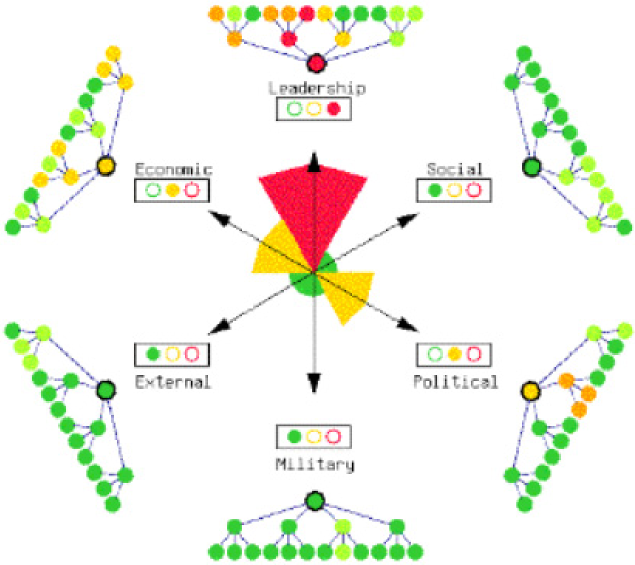


Figure 12 shows a visualization of problem decomposition in High SEAS.

4.4.4 Impactorium

Impactorium is a research prototype under development at FOI for modelling, structuring, fusing and analyzing intelligence information [18]. The tool supports problem decomposition through indicator breakdown and structured argumentation. It can also connect to different information sources, such as a command and control system or the web, to search for evidence relevant for determining the status of the indicators.

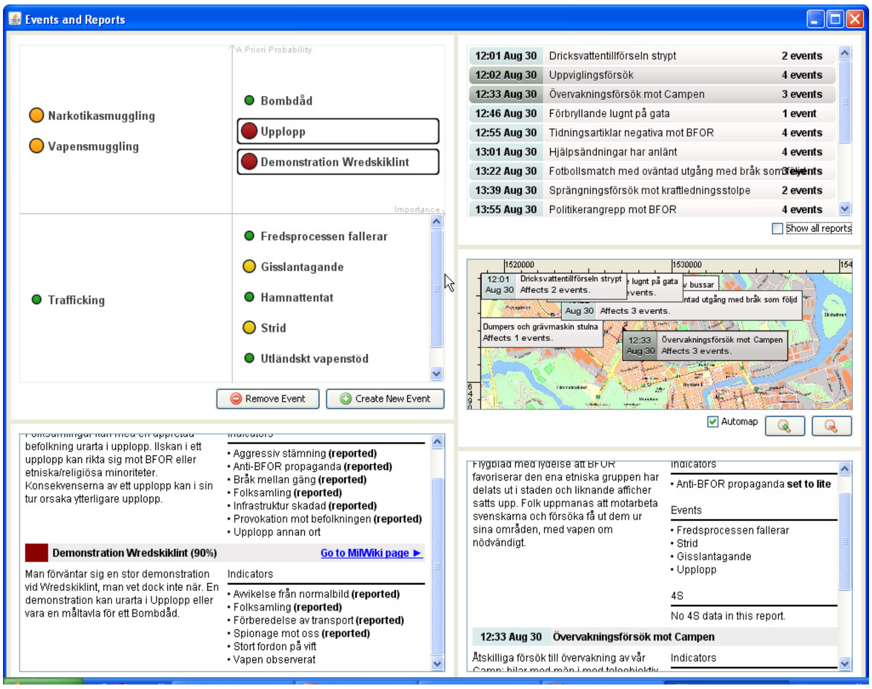


Figure 13 shows a screenshot from the main visualization view of Impactorium.

4.4.5 ACH

Analysis of competing hypotheses (ACH) is a process for evaluating which hypothesis best explains observation data [19]. It consists of several steps. First, different possible hypotheses that could explain the data are generated. This is normally done by a group of analysts. For each hypothesis, evidence and arguments for and against it are listed and attempts are made to disprove the hypotheses. The process can involve collection of further evidence in order to be able to falsify a hypothesis.

5 Visualisation tools

An important feature of any tool system for analysis is its visualization capabilities. Here, the new research area of visual analytics[5] can provide several important inputs to a tool for evaluation and assessment. Visual analytics is defined as the science of analytical reasoning facilitated by interactive visual interfaces, and can be seen as the natural combination of scientific visualisation as used in, e.g., physics or biology, and analysis. The field relies on results from cognitive psychology and various visualisation theories and seeks to enable users to understand more about their data by appropriate visualisations of it.

Visual analytics is related to the field of business intelligence, where companies try to find out as much as possible about their internal economic processes in order to optimize their operations[6]. One commercial tool in this area which has many potential uses for evaluation and assessment is Qlikview[7]. This tool allows users to quickly integrate and visualise data from many different sources. The tool is based on a new query language, AQL (Associative Query Logic) and technology that allows association of data between different database tables without first defining OLAP cubes. The tool also has sophisticated compression methods that allow it to operate on large amounts of data in working memory rather than on disk. The benefit for the user of this is that they can quickly formulate new queries and combine data sources on the fly.

[5] http://en.wikipedia.org/wiki/Visual_analytics

[6] This is the traditional definition of business intelligence. Today, the term is often also used for what is more properly named competitive intelligence, where the aim is to determine what possible competitors are doing and what happens in markets.

[7] http://www.qlikview.com/

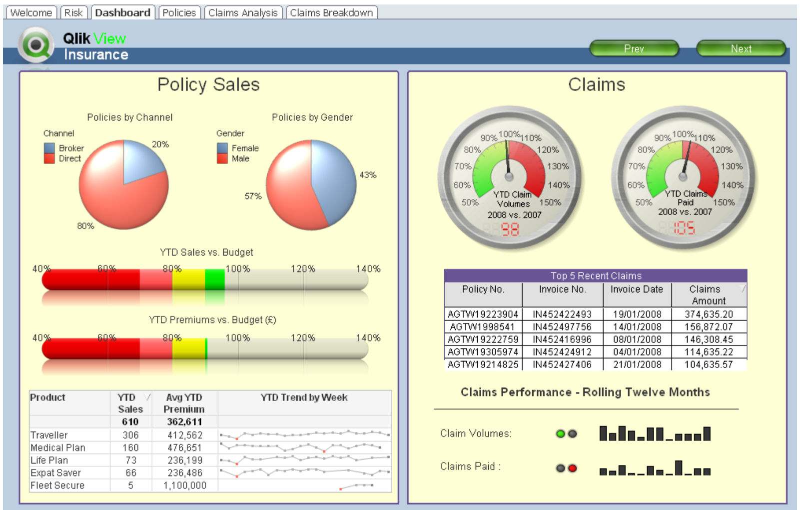


Figure 14 shows the dashboard user interface of Qlikwiew.

Another interesting tool for visualisation is the open web service ManyEyes[8], which allows users to upload data sets and apply many forms of visualisation to them. ManyEyes would of course not be appropriate for operational use, but making analysts play with it could be an important contribution to their visualisation education. Swiwel[9] is a similar web service.

[8] http://manyeyes.alphaworks.ibm.com/manyeyes/
[9] http://www.swivel.com/

6 Re-using knowledge

Modern use of information technology has in many ways altered the conditions for corporate knowledge re-use. Electronically stored information can with a single mouse-click be distributed to an entire organization, resulting in risk of knowledge losses due to information overload. Technology can help, but one major obstacle is that the vast majority of the produced information consists of unstructured (natural language) texts. This means variations, inconsistencies and inherent ambiguities which hinder the direct processing of machines. Accordingly, the challenge of today is how to efficiently add structure to information in order for information consumers to fully exploit the capabilities of computers when searching.

In an operational assessment context, the first step of knowledge re-use is simply to be able to find relevant documents. The most elementary search is text search, based on an index of all words in a document repository. The next level would be to include search on document metadata, such as document format, author or creation time, things you in search engines normally find under the name "advanced search". There are international standards for fundamental metadata types, e.g. the Dublin Core Metadata Initiative [20], but for more domain and application specific metadata this is not always the case. Deriving a standard for the assessment domain, covering for instance assessment objectives and methods, would be a step towards more effective knowledge re-use.

So far we have only discussed adding structure to a document's meta-layer; the content itself will still be unstructured. However, the aim of the *Semantic Web* effort is to change this. The Semantic Web is an evolving development of the World Wide Web which will enable both humans and machines to "understand" web content. This is accomplished by the development of technologies and procedures for capturing domain knowledge in semantic models (ontology), which then can be used to annotate document content. As the semantic models are expressed in (standardized) formal languages, such as the Resource Description Framework (RDF) and the Web Ontology Language (OWL), the meaning of the content can be processed by machines. This will allow very fine-grained information retrieval and improved conditions for knowledge re-use. For an introduction to semantic technologies, see e.g. [21].

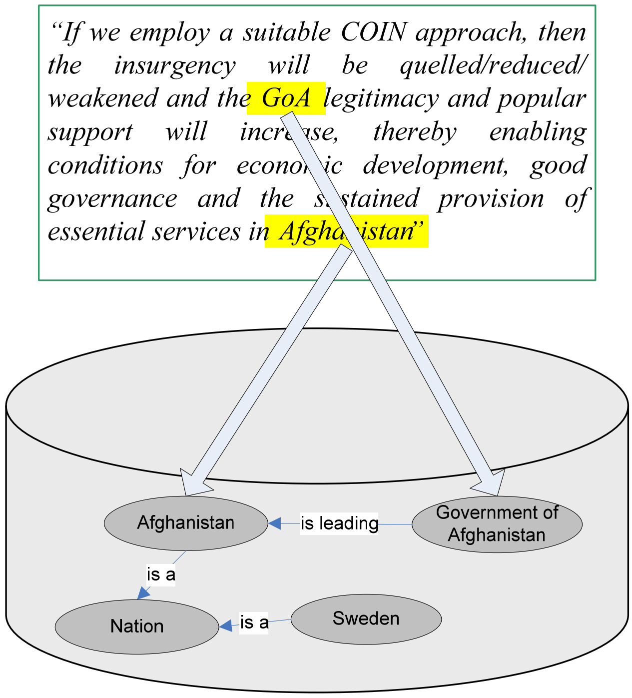


Figure 15 illustrates how semantic annotation transforms unstructured natural language text to machine processable structured statements. The annotation can be made manually or by automatic entity extraction tools.

Structuring document content in part removes the need for advanced metadata. Or rather, it replaces metadata in the traditional sense, where metadata refers to an entire document, with metadata referring to individual sentences or statements. This type of statement level metadata is not limited to the annotation of text documents, but is applicable to almost all kinds of knowledge representation. In an assessment perspective, this could be useful for annotating database schemas and content, and for making the content and structures of analysis models such as those described in section 4 retrievable.

In the following section we will have a look at a framework for knowledge re-use based on retrieval through similarity matching of structured problem descriptions. The structuring could be made using semantic technologies.

6.1 Case-Based Reasoning

Case-based reasoning (CBR) is a method for knowledge re-use that in line with human problem-solving and decision making is based on *analogy* for the identification and evaluation of alternative courses of action.

Figure 16 illustrates the process cycle of CBR. Given a new problem you search a case database, containing historical problems and their solutions, to retrieve the most similar problem. The solution of the retrieved problem will, as it is or revised, be suggested as a solution to the new problem. Finally, the new problem and the (revised) solution will be added as a new case in the case database and retained for future use.

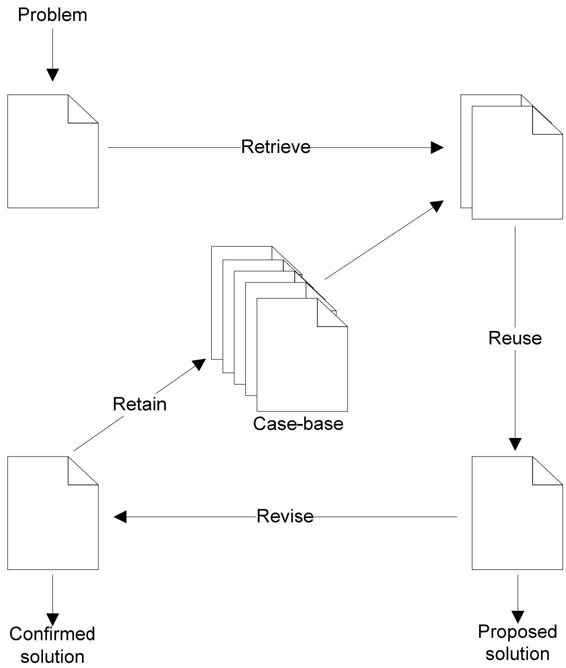


Figure 16 shows the CBR-cycle, as described in [22].

An example of where CBR could be a relevant technique to use is when choosing which of all indicators in the MPICE framework to use in a specific assessment. The problem description could in this case consist of a list of attributes that specifies the assessment premises, such as the assessment objectives, the objectives of the operation to be assessed, the types of conflicts in the operation area, etc. When retrieving the most similar historical case, the list of attributes is matched against all attribute lists in the case-base. The matching can be made in many different ways very much depending on how the attributes are structured (they can be categorical, numeric, probabilistic, etc.). A simplistic approach in an example with only categorical attributes would be to step through all attributes and compare them one by one. Each attribute value that coincides generates one point and the case that gets most points is the one retrieved. See Figure 17 for an example.

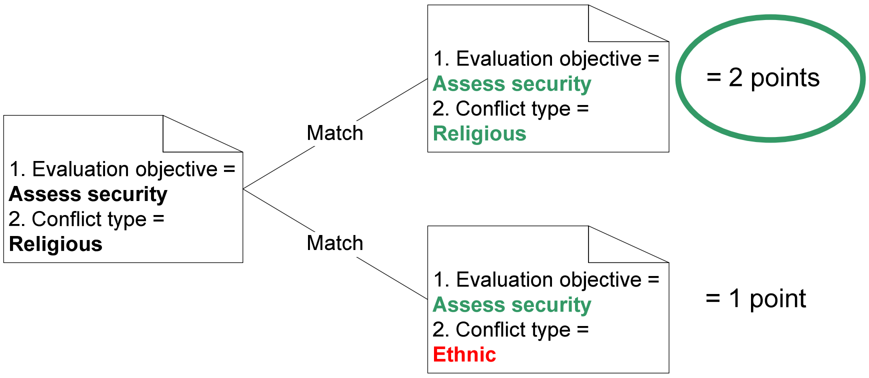


Figure 17 shows a simple example of case matching.

The solution of the retrieved case consists of a list of the indicators that were used in the corresponding historical assessment. The user will most likely have to revise this in order for it to match the specifics of the current situation. The revised list is then stored together with the initial problem description as a new case in the case-base.

From a technical view point, the main advantage of CBR is that it can be used to retrieve relevant knowledge even in domains that is data scarce, where other machine learning algorithms such as rule induction will fail. For a thorough introduction to CBR and its application to military early warning systems, see [23].

7 Summary and Recommendations

This report has presented some technical areas that can be relevant for evaluation and assessment in multifunctional operations. The purpose was not to present finished solutions for assessment and evaluation, but rather to present an overview of some results that could be useful when designing computer support tools for a standard assessment and evaluation method, when this has been defined.

We briefly mentioned some of the challenges seen by evaluators that we believe could be helped by the introduction of technical tools. Modelling as a tool for conflict analysis and assessment query decomposition was described and an introduction to several different techniques for handling uncertainty in data was given. Further, a brief overview of information structuring and its importance for knowledge re-use was given, exemplified through the description of case-based reasoning and its possible application.

We also presented some mature tools that could be of use in the near future. Of these, we particularly recommend atlas.ti and Qlikview which could be immediately useful for analysts today.

Other research areas presented are not as mature. However, we believe that they constitute valuable background knowledge when designing and implementing support tools for evaluation and assessment. We do not believe that it will be possible (or desirable) to construct a single tool that will help users perform all phases of evaluation and assessment. Rather, we think that a toolbox approach is needed, which builds upon commercially available tools and is, when needed, complemented by prototypes based on further research in specific areas. This approach is similar in spirit to what we propose for intelligence analysis in [24]. We also believe that some of the components for an intelligence analysis toolbox could be re-used for evaluation and assessment. This holds in particular for tools related to information structuring and visualization, and of course for the modelling tool.

We think such a toolbox for assessment and evaluation would consist of at least the following components:

- atlas.ti and similar tools for collecting unstructured information, if possible extended with support for uncertain data
- text mining and structuring tools for quickly getting an overview of large amounts of text and relating it to other data
- hypothesis modelling and reasoning tools similar to ACH [19] and Impactorium [18], extended with better support for uncertain data and with support for better causal reasoning
- data visualization tools such as Qlikview

- case-based reasoning tools for finding relevant earlier hypotheses that the evaluation and assessment could be based on
- case-based reasoning tools for helping the analyst determine what indicator framework and which specific indicators to use

We think these tools should be implemented gradually, starting with commercially available tools. An important aspect of analysis support for assessment and evaluation is also education. Evaluators and assessors should be given adequate training in statistics, management of uncertainty and above all methods for determining causal relations.

Before work can be started on implementing computer support tools for assessment and evaluation, a draft description of a method for assessment and evaluation must be available. It is also necessary to critically evaluate the potential benefits of implementing tools compared to the costs of implementing them. We believe that commercial tools such as atlas.ti and Qlikview would be immediately useful to analysts performing assessments.

Modelling tools that allow the analysts to break down an assessment task would also be useful (e.g., ACH). Reasoning tools that semi-automate the process of doing such breakdowns automatically could, if they can be made to work, have significant impact on the performance of assessment. However, it is not clear whether such tools can be made to work good enough, and it is likely that developing them would require a significant amount of resources and take a number of years. For such tools, it is thus better to wait until more research has been performed on similar tools for other applications, and exploit the work done in other fields (e.g., intelligence analysis) for assessment and evaluation.

8 References

1. Nilsson, C. & Derblom, M., ”*Uppföljning och utvärdering i multifunktionella insatser*”, FOI Memo 2855, 2009
2. Frelin, J., *“Purposeful Assessment”*, FOI-R—2718—SE
3. Shafer, G., *“A Mathematical Theory of Evidence”*, Princeton University Press, Princeton, NJ, 1976.
4. Halpern, J.Y., *“Reasoning about uncertainty”*, MIT Press 2003
5. Swedish Armed Forces, *“Grundsyn Underrättelsetjänst [Intelligence Service Basic View]”*, M7739-350003, Stockholm, Sweden, 2008
6. Pearl, J., *"Causal inference in statistics: An overview"*, Statistics Surveys, Vol. 3, 96--146, 2009
7. Pearl, J., *“Causality: Models, Reasoning, and Inference”*, Cambridge University Press 2000
8. Shafer, G., *“The Art of Causal Conjecture”*, MIT Press, 1996
9. Derblom, M., Frelin, J., Lindén, K., Nilsson, C., Tejpar, J.., *”Design av utvärdering för PRT Mazar-e-Sharif”*, FOI-R--2802--SE, 2009
10. OECD/DAC, *“Guidance on Evaluating Conflict Prevention and Peacebuilding Activities”*, 2008
11. Wikberg, P., Westin, J., Lindoff, J., Lundin, M., Hammervik, M., Hansson, L-Å., *”System analysis for Knowledge Support”*, FOI Memo 2600, 2008
12. Andersson, M., Dalberg, E., Grahn, P., Gundmark, T., Hansson, A., Lantz, F., Kylesten, B., Linder, S., Lindgren, D., Pihl, J., Sjöberg, E., Svenson, P.,*”The FOI C4ISR Demonstration 2008*”, Proceedings of 12th Int. Conf. on Information Fusion, 2009
13. Dziedzig, M. et al. (eds), *”Measuring progress in Conflict Environments (MPICE) – A Metrics Framework for Assessing Conflict Transformation* and *Stabilization”*, Version 1.0, Draft, 2008
14. Lowrance, J.D., Harrison, I.W., Rodriguez, A.C..*,”Structured argumentation for analysis”*, Proceedings of the 12th International Conference on Systems Research, Informatics, and Cybernetics: Focus Symposia on Advances in Computer-Based and Web-Based Collaborative Systems, 2000
15. Tecuci, G., Boicu, M., Marcu, D., Boicu, C., Barbulescu, M., Ayers, C., Cammons, D., *“Cognitive Assistants for Analysts”*, Journal of Intelligence Community Research and Development, 2007
16. http://www.consideo-modeler.de/english/
17. http://www.ai.sri.com/~seas
18. Forsgren, R., Kaati, L., Mårtenson, C., Svenson, P., Tjörnhammar, E..*,”An overview of the Impactorium tools 200*8”. Proc. Skövde Workshop on Information Fusion Topics (SWIFT 2008), Skövde, 2008
19. Heuer, R.J., *“The psychology of intelligence analysis”*, CIA 1999
20. http://dublincore.org/

21. Asadi, H., Garcia Lozano, M., Horndahl, A., Tjörnhammar, E..*,"Introduction to Technologies and Methods for Semantic Information Management"*, FOI-R—2858—SE, 2009
22. Watson, I.,*"Applying case-based reasoning: techniques for enterprise systems"*, Morgan Kaufmann Publishers Inc., 1998"
23. Malm, M. & Neider, G.,*" Erfarenhetsbaserad tidig förvarning om upploppshot. Slutrapport beslutsstöd vid internationella operationer.",* FOI-R--2041—SE, 2005
24. Brynielsson, J., Horndahl, A., Kaati, L., Mårtenson, C., Svenson, P., *"Development of Computerized Support Tools for Intelligence Work"*, Proceedings 14th International Command and Control Research and Technology Symposium, 2009