

Establishing conclusive proof in Forensic Data Analysis

SBV Forensics and Morpheus Software
Gabriel Hopmans and Peter-Paul Kruijsen

- Research (IGP) at Dutch Police concluded that more than 80% of the time is spent on Information management
 - Less then 20% on analysis
- One of main reasons (according to Morpheus):
 - the application-centric view of working

Handout

Index of this presentation

1. Introductions
2. Forensic Data Analysis
 - And need for Meaning Based Computing
 - Why do we need it?
 - How is it used already ?
3. Project use cases
 - A. Pilot project for Dutch Police
 - TopicView : Semantic search on BlueView
 - Pilot project in Counter-terrorism
 - B. Project in Forensic Data Analysis domain
 - Finding facts of Fraud in Terabyte world
4. Lessons learned



Introductions

Morpheus software started few years ago a project for Dutch Police, TopicView

- Developer of Analysis tool (P. Kruijsen)

Forensic Researcher (G. Hopmans) at:
SBV Forensic Data Analysis (SBV-Forensics)

- Directors
 - M. de Gunst
 - C. Schaap – former chief of the fraud bureau of the Dutch police, former District Attorney, expert on money laundering,
- SBV Forensics is a professional expert in the field of investigations primarily within a financial-administrative context (fraud),
- Uses Knowledge Technology to support knowledge processes in projects with large amounts of data



Handout

Forensic Data Analysis: Why topic maps?

- Analysts, forensic accountants and related business users spend much of their time looking for evidence and uncovering relationships.
- In many traditional forensic applications it is possible to visualize networks of relations and perform complex searches
 - but the real domain knowledge cannot be represented
 - And new related terms/variants are not in the ‘information push’



Topic Maps as knowledge structure

- The “knowledge workers” know a lot about relevant indicators of behavior
 - and they make a lot of assertions
 - and assertions about assertions,
 - all of which can be captured in a topic map.
- The necessary contextualization (for example, who made which assertion) is easily represented in the topic map.
 - And these assertions, facts, insights, first observations needs to be double checked!



Handout

Introduction Forensic Domain

- Terabytes Projects of data
- Bankruptcies of organizations, Fraud in Financial domain
 - Often years of business history
 - > Thousands of employees
 - Millions of data, documents
 - Structured and unstructured
 - Somewhere there is also
 - ‘digital’ evidence of fraud

Problem: not enough time
for (traditional) analysis

Solution : Ontologies



Meaning based computing

There is a big gap
between user and all
the systems!



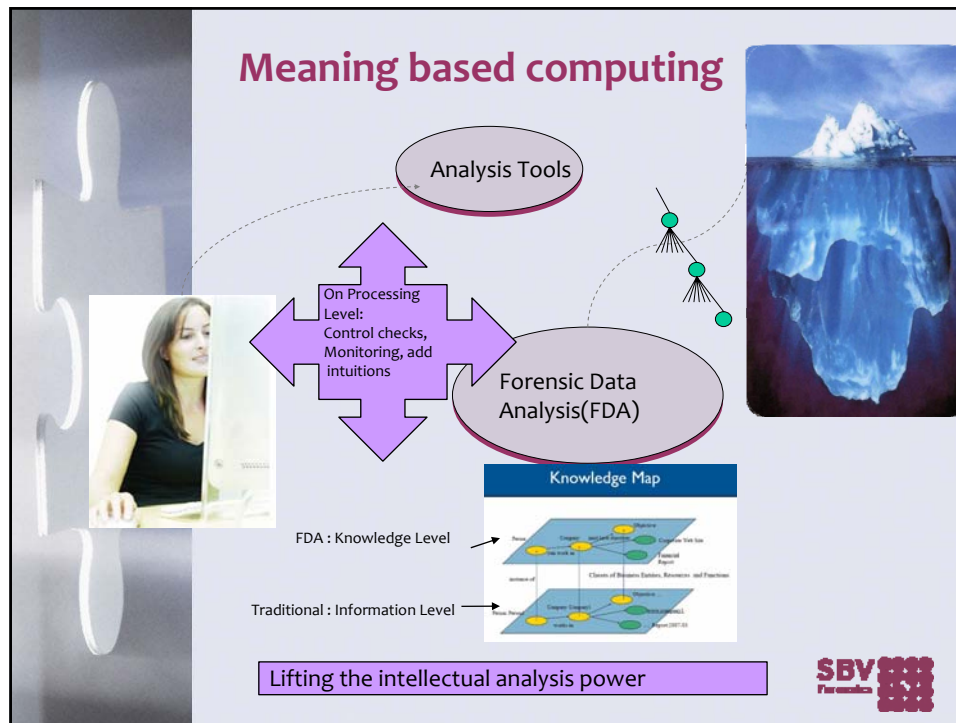
Make a ‘close loop process’ between:

- Users - Knowledge Structures - Systems

Subject centric view and combination of
bottom up/top down approaches



Handout




**TOPICVIEW
AMSTERDAM POLICE**

Peter-Paul Kruijsen
Morpheus Software

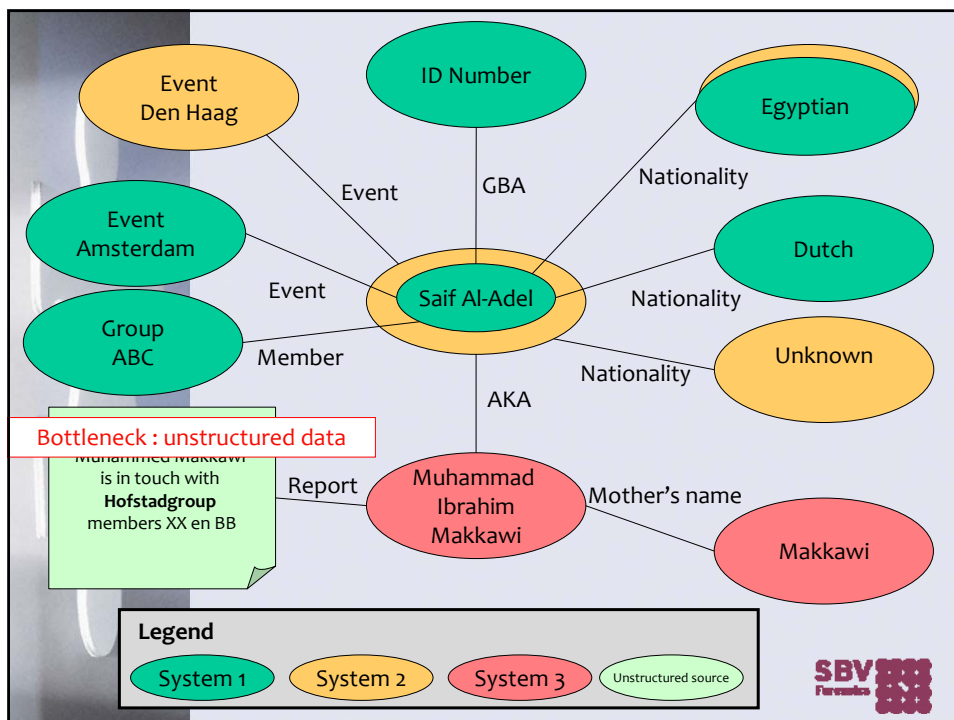

SBV Forensics

Handout

TopicView: Connecting info



- A typical real life example (not based on real facts):
- Police System 1: **Saif Al-Adel**
 - With an identification number
 - Registration event Amsterdam
 - Nationality: Egyptian and Dutch
- Police System 2: **Seif Al Adel**
 - No identification number
 - Registration Den Haag
 - Nationality: Egyptian and Unknown
- Police System 3 : Mohammed Ibrahim Makkawi
 - Might be an alias or nickname
 - Not registered as possible member of suspected terrorist group
 - However, a text is available where person is mentioned as member of the same group
- Our goals:
 - Connect three identities from different systems as one person
 - Include unstructured sources to infer new facts



Synchronizing various sources

- Dutch Police deploys many different systems
- Topic Maps provides a way to integrate data from these systems into a single information store
- An extended version of the OKS TMSync is used
 - Establishing identity using ‘identifying attributes’
 - Any two ‘person’ topics with equal SSN are merged using system generated PSI’s
 - History is maintained
 - A person’s living address might be removed from the source system, it’s still in the topic map, but scoped
 - The source system is included in the topic map for each characteristic
 - A user can always trace back a bit of information
- TM/XML syntax is used as intermediate syntax



Information in unstructured sources

- TopicView connects a lot but ...
 - Still lots of hidden information in unstructured sources (the 80%?)
 - E.g. email might contain additional information
- Tool to scan unstructured sources and to generate hypotheses automatically
 - Based on relevant keywords, like person names
 - But also domain keywords like ‘shotgun’ or ‘suspected’, indicating e.g. irregular behavior
- Dutch Police insists on confirming/dismissing all hypotheses by hand
 - To avoid registering the wrong ‘Alan Smithee’



Handout

Creating statements

- The system logs the statement made by confirming or dismissing a hypothesis
 - Who, what, when, based on what source
- Users create associations between topics
 - Creating a structured representation of an unstructured document
 - Constraints in the topic map
- Colleagues collaborate smoothly



Creating semantic networks

- Creating statements results in a network of semantically rich topics and associations
- New facts can be inferred iteratively from existing statements
 - All linking back to the original source file
 - The number of sources might be indicating the relevance
- Continuing on this, conclusive evidence can emerge and is stored in a structured way.





Forensic Data Analysis Use Case



Info Push/Contextualisation

Accountants, trustees in bankruptcies AND forensic accountants need an **information push to provide them with relevant facts!**

(instead of only making partial observations)

- Check for errors and irregularities
- What was cause of bankruptcy?
- Who can be hold liable? Association Roles
- Which facts regarding the administration and regulations can be found?



Handout

Use case Forensic Domain

Terabytes Project of data, tapes, archives

Bankruptcy of organisation

Years of business history

Thousands of employees

➤ 50 million of documents

➤ Structured and unstructured

– Somewhere there is also 'digital' evidence of fraud

- Question is : how to structure this for meaning based computing and establish conclusive proof ?



How to reach Meaning based Computing

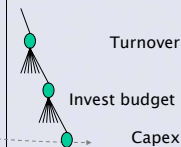
- Use the 10 - 15 % that is structured to make sense of the unstructured data
- By re-using:
 - Indexes from previous programs
 - Databases (project/employment administration) (they structure for instance also the role of an employee in a particular time frame)
 - Or the multiple role responsibilities
 - Knowledge from the domain experts (lawyers, trustees in bankruptcy, (forensic) accountants and auditors)
 - They know about the indicators of behavior
 - The allegations in the plaintiff

Handout

Allegations of plaintiff as taxonomy

Allegations

- A. Misleading investors and others
 - 1. Not publishing fact that organization had ..
 - 2. Fraudulent accounting
 - Fact : Bill H. wrote in e-mail to CFO that fact around CAPEX better could not be revealed and that revenue was to be booked in next quarter
 - 3. Engagement in undisclosed transactions to inflate reported income
 - Fact: Person Z. signed license agreement, paid fee and promised profits
- B. partner entities were owned by undisclosed related parties



Normally systems can't make any sense of CAPEX but we can infer it is related to Investment Budget and Turnover.

And do computing that it is related to Bill H. who is registered as commissioner .

- and that there are relations between CAPEX with the CFO etc..



How to find and generate keywords?

Just like in the project for Dutch Police we need software that automatically links the most relevant keywords with the taxonomies/indicators

- We need to know which ones are used frequently
- Which ones are in slang or in secret language
- Triggers upon live events



Handout

Topic maps Semantic Analysis System

- Example how we uncovered new relevant facts that could be used to strengthen the allegations of the plaintiff.
- Give me all the e-mails, documents shared between John Doe (CFO) and Sam Bend in the periods (date A and date B..) reported in separate threads in which they discussed items classified by the system as 'irregular'
 - Search process starts searching in all the e-mail boxes in which John and Sam are found in the address field of that period
 - And scans the e-mails about the relevance with the taxonomy and keywords

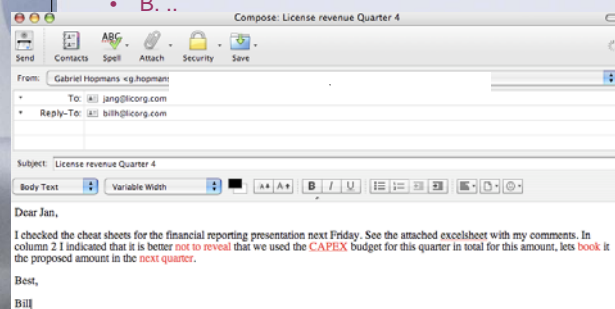


In the end .. To conclusive proof

System provides the relevant items around John Doe and classified as irregular behaviour

Allegations

- A. Misleading investors and others (15)
 - 1. Not publishing fact that organization had ..(9)
 - 2. Fraudulent accounting (1)
 - 3. Engagement in undisclosed transactions to inflate reported income (4)
- B. ..



Handout

Lessons learned & Contact

Topic Maps in Forensic Data Analysis :

- Semantic integration on legacy data
- Need for a tool that delivers semantics from the bottom-up (Iknow smart indexing is our candidate)
 - See the Arbo Unie “Knowledge Management” Presentation
- Discovery of hidden relations
- Tools for users in terabyte data world
- Support for a way of human thinking
- Thank you for your attention !
- Questions?
- Morpheus : <http://www.mssm.nl>
- SBV Forensic Data Analysis
info@sbv-forensics.com
<http://www.sbv-forensics.com>
Dordrecht, the Netherlands

