

Ongoing Research in the DALE project Data Assistance for Law Enforcement



Universiteit Leiden

Tim Cocx
tcocx@liacs.nl

Walter Kosters
kosters@liacs.nl

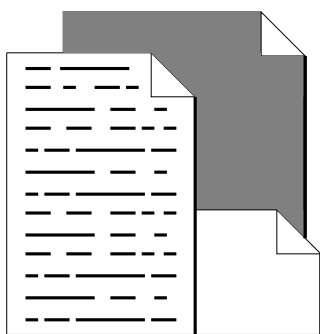
Jeroen Laros
jlaros@liacs.nl

DM & KDD

The rapid growth of available data requires new computational methods. Besides traditional statistical techniques, current research known as **Data Mining (DM)** uses modern methods that originate from research in Algorithms and Artificial Intelligence. The main goal is the quest for interesting and understandable patterns. This search has always been and will always be a critical task in law enforcement, especially for criminal investigation. Data Mining, or **Knowledge Discovery in Databases (KDD)**, can be defined as “the non-trivial extraction of implicit, previously unknown and potentially useful and understandable knowledge from data”.

Background

Databases from law enforcement applications are usually large, and contain data with varying types, including free formats, that sometimes rapidly changes. Research therefore aims at **semi-structured data**.



documents

Paradigm

Our research to this date has mainly focused on the comparison between individual investigations or cases to discover potential common perpetrators. To accomplish this task we employ a four step paradigm that transforms a collection of narrative reports and documents found on scene into a comprehensive visual image that is ready to be used by investigation teams on the job.

1 — Textmining

First we employ a commercial text miner to gain access to the specific concepts that these documents contain (cars, persons, URLs, etc.).

■		■		
		■	■	
■			■	■
	■			■
			■	
■	■			

2 — Linguistic Analysis

The table we obtain is prone to have typing mistakes or ambiguities which we try to eliminate by using a linguistic comparison engine. For this purpose we propose a binary tree search to determine the Levenshtein distance between different concepts to check if they are in fact the same.

■				
		■	■	
■				■
	■			
			■	
■				

3 — Distance Matrix

We then use a distance measure to assign numbers to couples of investigations that represent how much these investigations are alike. This measure is specifically designed to incorporate differences in size between the numerous investigations, for it is not uncommon that cases either suffer from information explosion or the lack of resources.

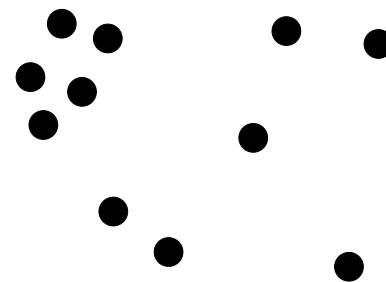
○	■	■	■	■
■	○	■	■	■
■	■	○	■	■
■	■	■	○	■
■	■	■	■	○
■	■	■	■	■

4 — Clustering & Visualization

Finally we construct a tool that employs associative array clustering to view this distance matrix, which enables the police analysts to immediately assist the personnel working on these investigations.

Other Research

Next to the testing and enhancement of the above mentioned process, future work in the DALE project will also focus on a number of other police oriented tasks. In the near future we will begin developing tools to analyze the national archive of criminal records in order to infer knowledge about criminal careers and national trends and we will try to incorporate in-house technology for DNA research in the investigational area. We will also try to employ association rules: **frequent itemset mining**, with special attention to difference mining and (timed) sequences.



clustering



The DALE project is financed in the ToKeN program of the Netherlands Organisation for Scientific Research (NWO) under grant number 634.000.430.



<http://www.dale.liacs.nl/>