DALE
Data Assistance for Law Enforcement

Project Proposal

June 2, 2004

1 The Project

1.1 Project Title

The title of the project is "Data Assistance for Law Enforcement".

1.2 Project Acronym

The project acronym is *DALE*.

1.3 Principal Investigator

The principal investigator, also contact address, is:
dr. W.A. Kosters
Leiden Institute of Advanced Computer Science (LIACS), Universiteit Leiden
P.O. Box 9512, 2300 RA Leiden, The Netherlands
phone +31–71–5277059
email `kosters@liacs.nl`, WWW `http://www.liacs.nl/home/kosters/`

2 Summary of Research Proposal

The rapid growth of available data in all regions of society requires new computational methods. Besides traditional statistical techniques and standard database approaches, 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, and more specific for the fight against terrorism.

Data Mining, sometimes also referred to as *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". Databases from law enforcement applications are usually large, and contain data with varying types, including free formats. The data is temporal, dynamic and noisy; often information is explicitly hidden.

Research therefore focuses on *semi-structured data* and *pattern bases*. In particular *association rules* will be used. The goal is to compose a framework for data mining in law enforcement, fed by problems that arise from this area and in close cooperation with domain experts. The systems developed must be such that future developments can be embodied. Emphasis is — for the moment — on text oriented databases. The approaches are however such that, e.g., multi media databases can be incorporated at a later stage. The focus is on techniques that may be used in the fight against terrorism. Research will both start from known theories and from questions arising from daily practice.

3 Classification

The research is fundamental and aims at long-term perspectives. The desired data mining framework must have a long lifespan. The research is however driven by questions arising from law enforcement practice. Classification therefore is both category 1 and category 2.

The research is centered around the themes knowledge engineering and learning. The techniques used originate from Artificial Intelligence (the learning part) and aim at the detection of interesting patterns (the knowledge engineering part). There is a visualization component present.

4 Composition of the Research Team

The following persons participate in the project:

name	affiliation	specialization(s)
prof.dr. J.N. Kok	Leiden	coordination, data mining, artificial intelligence
dr. W.A. Kosters	Leiden	data mining, artificial intelligence
dr. J.M. de Graaf	Leiden	data mining
drs. S.G.R. Nijssen	Leiden	data mining, optimization
drs. P.W.H. van der Putten	Leiden	data mining, law enforcement applications
Ph.D. researcher (OiO)	Leiden	data mining
Ph.D. researcher (OiO)	Leiden	data mining
ing. A.B.L. Holtslag MSc	Driebergen	law enforcement
ing. H.L. Willering MSc	Driebergen	law enforcement

Leiden:
Leiden Institute of Advanced Computer Science (LIACS),
Universiteit Leiden, P.O. Box 9512, 2300 RA Leiden, The Netherlands;
`http://www.liacs.nl/`
Driebergen:
Korps landelijke politiediensten (KLPD; Dutch national police),
P.O. Box 100, 3970 AC Driebergen, The Netherlands;
`http://www.klpd.nl/`
During the project contact between the groups will be further intensified. The formal Ph.D. advisor is professor Kok.

5 Research School

The Leiden group participates in the Research School IPA, the Institute for Programming Research and Algorithmics (Instituut voor Programmatuurkunde en Algoritmiek). Many contacts exist with the Research School SIKS, the School for Information and Knowledge Systems (School voor Informatie- en KennisSystemen).

6 Description of the Proposed Research

The rapid growth of available data in all regions of society requires new computational methods. Besides traditional statistical techniques (cf. [HTF01]) and standard database approaches, 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, and more specific for the fight against terrorism. Examples are the discovery of interesting links between people (social networks, see, e.g., [Kre02]) and other entities

(means of transport, modus operandi, locations, communication channels like phone numbers, accounts, financial transactions and so on). Topics range from outlier analysis to more traditional clustering and classification.

Data Mining, sometimes also referred to as *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". Some essential additions to this definition are:

- The databases under consideration are often (extremely) large.
- The databases are usually not designed for data mining. For instance, records from telephone calls are intended for accounting, but can clearly be used for many other purposes too.
- The results of the data mining process should be new and surprising. However, known facts normally will be re-discovered, thereby giving more confidence in the methods used.
- Statistical methods and standard database techniques should go hand in hand with newer, more algorithm oriented data mining techniques.
- Often data mining techniques make use of random elements. This has as a consequence that repetition of a run of an algorithm may result in a (somewhat) different result. It is not always easy to formulate explicit statements concerning statistical significance.
- Data mining research often leads to "patterns". These patterns need to be interpreted and stored in close cooperation with domain experts. In some cases mining the results is a task itself!
- Data mining is an interactive process, in which a human user is in a feedback loop with the data mining system (sometimes this is referred to as "Intelligent Data Analysis", where intelligent stands for the involvement of a human user).

The KDD process (see [Mei02]) is usually split into data selection, cleaning (e.g., de-duplication, domain consistency), enrichment (e.g., data fusion, see [vdPKG02]), coding, the real data mining and reporting. In this view data mining is the "discovery" step from this process. However, often the term Data Mining covers the whole process; in this project the term is used as such: KDD and DM are considered as synonyms. Keywords include *clustering* and *classification*. Investigation needs to go beyond classical data mining tasks like prediction and clustering, and should also include functionalities like link analysis and network mining, integrated with search and navigation. Techniques used are among others: decision trees, neural networks, Bayesian networks, temporal difference learning, evolutionary algorithms (see [RN03, BH03, HTF01] as a general reference) and association rules (see, e.g., [AMS$^+$96]). The reporting step is very important: how interesting a result may be, it should be clearly presented. This usually means that the information, the patterns perhaps, must be visualized in an adequate and convincing way. Finally, note that the whole process can be highly automated. This has as a consequence that also non technical users will be able to make use of the tools, and even better: their input and influence are requested for a proper handling. In some cases however autonomous agents can deal with (parts of) the process. As an example, they could be used to monitor cold cases.

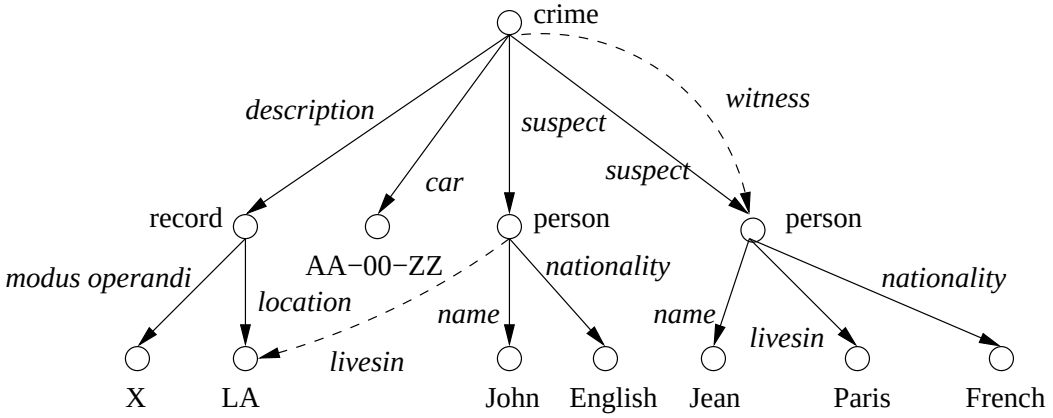
In the area of law enforcement the need for these techniques is apparent in view of the enormous load of information that is (and can be made) available nowadays (see, e.g., the link in Section 9). We list some of the problems:

- Data is usually spread among many different data sets stored in different places.
- Data uses different — and sometimes even free or changing — formats; the data in law enforcement is often *semi-structured* (see [WL00, AAK$^+$02], the link in Section 9 and the Intermezzo below): the source or the environment does not impose a rigid structure.
- Data is of a very diverse nature (ranging from low-level sensor data, through multi-media data, all the way to formal text documents).

- Data is often noisy, and may contain discrepancies and anomalies.
- Data can be temporal: for example, if a case is closed, data should perhaps be removed.
- Data is usually dynamic; for instance, the whereabouts of a person vary continuously.
- Data is intentionally altered in order to conceal valuable information.
- As mentioned above, in contrast with classical data mining other tasks than clustering and prediction are likely to be relevant. In particular we are looking for methods that discover associations, networks and links.
- Ideally mining methods should not only lead to abstract general patterns, but rather provide guidance to find interesting links to individual persons or objects. In other words, the analysis should be combined with navigation and search.
- Privacy and security issues need to be taken into account. This point is of particular interest in relation to the ways in which results will be published, see Section 7.

Intermezzo

In order to get some idea of semi-structured databases we examine a simplified crime database (the example is inspired by [WL00]). Suppose that a *Crime* object has a *Car*, *Person*s and a crime *Record*. Every *Person* has a (not necessarily unique) *Name*, usually a *Nationality* and a home *Town*. A *Record* has (among other things) a *Modus operandi* and a *Location*. All objects can have more data elements, such as pictures. Other databases contain extensive information on the *Person*s, and so on. Real time information is also available, for instance about the whereabouts of the *Person*s.



The tree above represents a sample crime from a database. Note that it is also possible to have a graph-like structure (including the dotted arrows), where locations are shared between criminal records and persons, and a suspect can appear as witness too. Locations may contain pointers to other locations, and analogous for persons. The crime record may contain comprehensive descriptions.

Now an interesting pattern may be the presence of crimes with many local suspects, or the search for the most often occurring modus operandi. And how many crime records in LA have an above average number of French witnesses? At first sight this may only look like answering SQL-queries, but it rather requires the discovery of the interesting ones.

This example is adapted from the NWO project *MISTA* (project number 612.066.304; granted July 17, 2003), that deals with semi-structured data, and is carried out by members of the same research group as the current project, in cooperation with Utrecht University.

Semi-structured data arises when the source or the environment does not impose a rigid structure on the data and when data is combined from several heterogeneous sources. More and more data sets do not fit in the rigid relational model because the individual data items do not have the same structure completely. Rather, the data items share only partly the same structure. In a semi-structured database, there is no fixed database schema: conceptually the data is stored in a graph-like structure (like XML) which contains both information about the data as well as the data itself. Since in normal databases all data items share the same structure, the structure of the data items plays no role in standard data mining. For semi-structured databases, however, the structure of an individual data item encodes an important part of its semantics. Most data mining algorithms are not designed for semi-structured data and should at least be adapted in order to deal with such data. In particular, we want to look at semi-structured data in the domain of law enforcement.

The data might contain multi media streams too, such as sound and vision. This may on the one hand be considered as highly structured data, but for our purposes it is definitely not: it requires a lot of preprocessing before it can be used in a data mining context. Future research beyond this project may include information from other modalities like speech; the framework, tools and algorithms will be kept open for this purpose. For the moment we assume that the necessary data extraction from these inputs has already been adequately performed. The general feeling is that

- there is much more information and knowledge in the data than is currently being extracted and used,
- the data flood becomes too large to be handled by the users themselves.

The *DALE* project focuses on the introduction of new data mining and data handling techniques in the area of law enforcement, where the input is a combination of existing practical needs and of expected problems and possibilities from the near future. New frameworks need to be developed, but also current questions should be dealt with. The scientific challenge is to be able to deal with and combine the specific types and properties of data in this area, so that it will be possible to handle the data flood with computer aided means, and to use data mining tools to extract information and knowledge that can be used in law enforcement.

The project clearly has very practical aspects. For instance, the data has to be gathered and prepared (the first and second step in the KDD process). One of the two Ph.D. researchers is supposed to take care of this part. He or she deals with the specific problems that might occur in what may be called the preprocessing phase.

Association rules have been studied in much detail, see the extensive literature (e.g., [AMS$^+$96, DT01, KMO99, BMS97, HSW00, NK01, dGKW01, AY01, HH99, KPP03, KP03]). Focus was first on fast implementations, but now this issue has been solved effectively, the time has come to investigate the interpretation and the applicability in more complicated situations, e.g., trees (the search for frequent subtrees, see [NK02]). The first Ph.D. researcher is supposed to concentrate on association rules, and their use in law enforcement. In particular, interestingness of rules, fuzziness, ordered sequences, hierarchies and different types of databases will be investigated. It is expected that the existing algorithms with ample work can be adapted to the new situation. New methods need to be developed to incorporate data from different databases that change in time.

The second Ph.D. researcher starts from the problem side. Problems originate from the fight against terrorism (see, e.g., [Kre02]). One particular aspect is the search for cell organised terrorist groups. The above mentioned data mining techniques can and will all be used for this purpose. It might be interesting to strive for a minimalistic approach, in the sense that as little information as possible should be used in order to obtain the desired results. It is clear that actual events and needs can influence and steer the direction of the research. On the one hand results will be general ("there is a connection between activities X and Y"), but on the other hand they will be person specific ("person Z has a more than average probability of committing fraude").

The research will concentrate on a framework that incorporates data currently used in law enforcement, and offers sufficient flexibility to accommodate for the near future. In close cooper-

ation with domain experts mining tools will be developed that answer current questions, reveal patterns in the data, and provide a general understanding. The tools should be usable in practical situations.

We will focus on:

- Methods for handling the large data volumes, including ways of integrating different types of data such that they become amenable for data mining.
- Intelligent query and visualization of data.
- Adaptation of data mining tools to deal with the specific types of data in law enforcement.
- Detection, storage and retrieval of patterns ("fingerprints") in the data, resulting in pattern bases.

More specifically, we will mine for structure in the semi-structured data warehouse. We want to use inductive databases (also called *pattern bases*). Inductive databases are databases that, in addition to data, also contain generalizations, i.e., patterns, or models extracted from the data. Within the inductive database framework knowledge discovery is modelled as an interactive process in which users can query both data and patterns/models to gain insight about the data. To this aim so-called inductive query languages are used. Very often inductive queries impose constraints on the patterns/models of interest. Within the framework of inductive databases, knowledge discovery is considered as an extended querying process: from the user point of view, there is no such thing as real discovery, just a matter of the expressive power of the query languages. (For more information, consult the link in Section 9.) We want to lay the foundations for such pattern bases in the domain of law enforcement, in particular we want to

- represent data and patterns in a uniform formalism,
- design query languages that allow to query the patterns and the data,
- represent the answers in an intuitive way,
- have an efficient implementation, based on existing methods (for example based on the relational database model).

7 Description of the Proposed Plan of Work

The two researchers are supposed to deliver a Ph.D. thesis within four years. During the course of the project, results will be published by means of the usual conferences. Due to the nature of the research, care has to be taken about what to publish; in close cooperation between the groups involved decisions have to be made on this matter. In most cases however, the scientific research delivers such results that they can be communicated freely without doing any harm to the real application.

The first three months are reserved for literature study and for getting acquainted with law enforcement practices and problems. The rest of the first year for the first Ph.D. researcher is devoted to the development of ideas about association rules in the case of law enforcement applications. The second Ph.D. researcher should make an inventory of existing questions and the possible techniques that can be used; he or she can also prepare the data. In the second year attention is given to implementation and experimentation. During this year it is decided where to proceed in the third year, taking into account the input from the other Ph.D. student, or more general the results of the project so far, and other new developments. For both Ph.D. students, the fourth and final year will be devoted to the production of their theses.

The work programme is summarized in the following table:

year	first Ph.D. student	second Ph.D. student
1; 3 months	literature	literature
1; 9 months	association rules	inventory
		data preparation
2	implementation	implementation
	and experimentation	and experimentation
3	generalization	generalization
4	thesis	thesis

The researchers are expected to design and develop prototypes, that can be implemented and tested on synthetic and real life databases: the different approaches need to be verified for their practical implications. We plan some Master's projects to accompany the current research. In such a project the Master's student should implement and validate a technique on selected databases.

Both researchers are supposed to take part in the educational programmes of the Research School IPA (Institute for Programming Research and Algorithmics). Both IPA and SIKS (see Section 5) offer a wide range of basic and advanced courses for Ph.D. students. The supervisors will make a selection of relevant courses from the course programmes, together with the researchers. They should plan to visit these activities together. Furthermore there is the possibility of supervised self-study, participation in relevant courses from our Master programmes, and visits to European summer schools.

8 Expected Use of Instrumentation

The hosting organisations supply hardware and software.

9 Literature

Some interesting weblinks:

- Homepage for Mining Structured Data, Universiteit Leiden
  `http://hms.liacs.nl`
- Law Enforcement & Public Safety
  `http://www.aaai.org/AITopics/html/lawenf.html`
- Consortium on discovering knowledge with Inductive Queries
  `http://www.cinq-project.org/`

The five main contributions of the research team (in connection with the current project) are: [NK02], [dGKW01], [KPP03], [vdPKG02] and chapters in [Mei02].

References

[AAK+02] T. Asai, K. Abe, S. Kawasoe, H. Arimura, H. Sakamoto, and S. Arikawa. Efficient substructure discovery from large semi-structured data. In R. Grossman, J. Han, V. Kumar, H. Mannila, and R. Motwani, editors, *Proceedings of the Second SIAM International Conference on Data Mining (SDM2002)*, pages 158–174, 2002.

[AMS+96] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A.I. Verkamo. Fast discovery of association rules. In *Advances in Knowledge Discovery and Data Mining*, pages 307–328. AAAI/MIT Press, 1996.

[AY01] C.C. Aggarwal and P.S. Yu. Mining associations with the collective strength approach. *IEEE Transactions on Knowledge Discovery and Data Engineering*, 13:863–873, 2001.

[BH03] M. Berthold and D.J. Hand. *Intelligent Data Analysis, An Introduction*. Springer, second edition, 2003.

[BMS97] S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: Generalizing association rules to correlations. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD'97)*, pages 265–276, 1997.

[dGKW01] J.M. de Graaf, W.A. Kosters, and J.J.W. Witteman. Interesting fuzzy association rules in quantitative databases. In L. De Raedt and A. Siebes, editors, *Proceedings of the Fifth European Conference on Principles of Data Mining and Knowledge Discovery (PKDD 2001)*, volume 2168 of *Lecture Notes in Artificial Intelligence*, pages 140–151. Springer, 2001.

[DT01] L. Dehaspe and H.T.T. Toivonen. Discovery of relational association rules. In N. Lavrac and S. Dzeroski, editors, *Relational Data Mining*, pages 189–212. Springer, 2001.

[HH99] R.J. Hilderman and H.J. Hamilton. Knowledge discovery and interestingness measures: A survey. Technical Report CS 99–04, Department of Computer Science, University of Regina, 1999.

[HSW00] H. Hofmann, A.P.J.M. Siebes, and A.F.X. Wilhelm. Visualizing association rules with interactive Mosaic Plots. In R. Ramakrishnan, S. Stalfo, R. Bayardo, and I. Parsa, editors, *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2000)*, pages 227–235, 2000.

[HTF01] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001.

[KMO99] W.A. Kosters, E. Marchiori, and A. Oerlemans. Mining clusters with association rules. In D.J. Hand, J.N. Kok, and M.R. Berthold, editors, *Proceedings of the Third Symposium on Intelligent Data Analysis (IDA-99)*, volume 1642 of *Lecture Notes in Computer Science*, pages 39–50. Springer, 1999.

[KP03] W.A. Kosters and W. Pijls. Apriori: A depth first implementation. In B. Goethals and M.J. Zaki, editors, *Proceedings of FIMI'03, the first Workshop on Frequent Itemset Mining Implementations, Melbourne, Florida, USA (CEUR Workshop Proceedings)*, 2003. http://CEUR-WS.org/Vol-90/.

[KPP03] W.A. Kosters, W. Pijls, and V. Popova. Complexity analysis of depth first and FP-growth implementations of apriori. In P. Perner and A. Rosenfeld, editors, *Proceedings of MLDM 2003 (Machine Learning and Data Mining in Pattern Recognition)*, volume 2734 of *Lecture Notes in Artificial Intelligence*, pages 284–292. Springer, 2003.

[Kre02] V.E. Krebs. Uncloaking terrorist networks. *First Monday*, 7, issue 4, April 2002. http://www.firstmonday.dk/issues/issue7_4/krebs/.

[Mei02] J. Meij, editor. *Dealing with the Data Flood: Mining Data, Text and Multimedia*. STT/Beweton, Den Haag, 2002.

[NK01] S. Nijssen and J.N. Kok. Faster association rules for multiple relations. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI'01)*, pages 891–896, 2001.

[NK02] S. Nijssen and J.N. Kok. Tree sets: Towards a set-oriented view on multi-relational data mining. In H. Blockeel and M. Denecker, editors, *Proceedings of the Fourteenth Belgium-Netherlands Artificial Intelligence Conference (BNAIC 2002)*, pages 219–226, 2002.

[RN03] S.J. Russell and P. Norvig. *Artificial Intelligenec: A Modern Approach*. Prentice Hall, second edition, 2003.

[vdPKG02] P. van der Putten, J. N. Kok, and A. Gupta. Why the information explosion can be bad for data mining, and how data fusion provides a way out. In R.L. Grossman, J. Han, V. Kumar, H. Mannila, and R. Motwani, editors, *Proceedings of the Second SIAM International Conference on Data Mining*. SIAM, 2002. http://www.siam.org/meetings/sdm02/proceedings/sdm02-08.pdf.

[WL00] K. Wang and H. Liu. Discovering structural association of semistructured data. *IEEE Transactions on Knowledge and Data Engineering*, 12:353–371, 2000.

10 Budget

The requested budget is as follows. We apply for two Ph.D. researchers (OiO's), both for a period of 4 years. The total costs, including benchfee, amount to:

$$2 \times (157,683 + 4,538) = 324,442 \text{ euro.}$$

Matching is created by (cf. Section 4):

name	amount
prof.dr. Kok	$0.05 \times 454,948 = 22,747$
dr. Kosters (UD)	$0.15 \times 265,647 = 39,847$
dr. de Graaf (UD)	$0.05 \times 265,647 = 13,282$
KLPD (UD level)	$0.15 \times 265,647 = 39,847$
total	$115,723$ euro

In this table 0.05 and 0.15 denote fractions of fulltime jobs ("fte's").