KDD\(^1\) – Overcoming Massive Data Streams for Intelligence Tasks

Dr. Vera Kamp
PLATH GmbH
Gotenstrasse 18
20097 Hamburg
Germany
vera.kamp@plath.de

ABSTRACT

Actually very interesting IT systems promise to reveal connections between apparently harmless and unrelated information pieces. An article from the New York Times in February 2006\(^2\) makes clear that common data mining techniques were not successful in general. Despite huge investments, correlating data from different sources did not yield satisfactory results. Transforming low-level data by aggregation to meaningful events is nevertheless the key to building the basis for succeeding decisions in the context of situation reports.

More realistic and manageable is an approach that includes interactions with the user along with domain specific knowledge. Gaining security relevant messages should be based on an iterative multi-level process. This process represents the core element of intelligence analysis systems which play an important role for supporting decisions in management information systems\(^3\).

The following example illustrates the principal automated process for discovering communication structures in the context of radio reconnaissance: A crucial part of this process is the analysis and visualisation of communication structures, or more generally, of network information. This should be embedded in spatio-temporal data analysis with geo-oriented data access and the integration of domain-specific analysis functions.

![Figure 1: Domain-specific Analysis Function](image1)

![Figure 2: Spatial Access](image2)

\(^1\) Knowledge Discovery in Databases


**KDD Overcoming Massive Data Streams for Intelligence Tasks**

**PLATH GmbH Gotenstrasse 18 20097 Hamburg Germany**

**Approved for public release, distribution unlimited**

See also ADM002067., The original document contains color images.
The intelligent analysis of radio emission data is based on data mining techniques, cluster visualisations to validate the results, a model based communication detection (including domain-specific knowledge) and the visualisation of communications. The following use case of a simple simplex communication clarifies the problems and the applied methods. Module coupling is realised by a distributed architecture. Given are a huge amount of radio emissions which are arbitrarily distributed. Each emission is described by the attributes ID, frequency, modulation type, starting time, end time, latitude and longitude. It has to be considered that the data quality of single emissions depends on propagation conditions. Because these can vary, it can happen that single emissions or attributes are missing or on the other hand different classification level information are available. Furthermore, with a broadband collection of emissions the amount of information is extremely large and requires massive data handling which can not be processed in main memory.

1.0 USE-CASE SIMPLEX-COMMUNICATION

The use case is looking for a simplex communication chain with two stationary partners – a central station and a substation. Both are using the same constant nominal frequency and the same transmission mode. The partners are communicating alternating one after the other. The problem lies in the amount of possible communication structure instances. Although the communication can be easily described in an informal way it is necessary to find an exact, formal specification in order to perform a computer-supported analysis. It should not be realised by a specific static algorithm but should be interactively and exploratively changeable by the user. The core concept includes the following steps:

1.1 Data Mining

During the first step emissions are assigned to clusters. These subsume emissions concerning the spatial, temporal or frequency criteria. In this way significant data reduction is achieved. By spatial clustering special emitter station could be determined. Besides when processing of extremely huge data amounts the main problem to solve is how to choose the best method and parameters.

1.2 Cluster Visualisation

The next step serves the validation of the data mining results and already provides a possibility to manually discover communication structures by the user relying on the presented visualisation, for example the presentation of spatial clusters. Emission can appear as single instances or as temporal ordered parts of a cluster. It is difficult to visualise the emissions and clusters clearly arranged in order to focus on the actual interesting data. Additionally different attributes have to be integrated.

1.3 Model based communication detection

Computing communication structures from clusters is the next step. This is done by using typical communication models. A domain specific modelling language provides the possibility to represent the communication models. By this language the simplex communication can be formally specified. The model distinguishes between connection constitution and alternating communications.

Figure 3: Simplex Communication Rules
The connection constitution consists of three emissions: the central station is sending, the substation replies. The alternating sequence consists of emissions of the central and the substation. All emissions occur with the same frequency and modulation type. The distance between the emission is flexible by a delay parameter. A graphical notation of such a model is illustrated by the two adjoining pictures.

1.4 Visualisation of discovered communications

This step provides a presentation of the discovered communications and allows by this a validation of the model based communication detection. It has to deal with many composite events. A simple textual visualisation does not meet the needs. The graphical visualisation offers a better overview and manifold interaction possibilities.

The emitters are on the spatial level, time is the third dimension. The connection lines indicate the communications.

Overcoming massive data streams for intelligence tasks is a challenge which should involve the analysis process with a seamless data access and the intelligence analyst. The acceptance of the results depends on the possibility to validate the results. The sustainability of results has to be guaranteed by flexible extension of actual domain specific analysis methods.

2.0 REFERENCES


KDD
Overcoming massive data streams for intelligence tasks

Dr. Vera Kamp
Dr. Joachim Stamm
Agenda

- Motivation
- Basic Concept
- Exploration Process
- Summary
Motivation
Motivation

- ACOS delivers huge amount of data
- Intelligence requires to add value
- **Task:** Transforming low-level data by aggregation to meaningful events is the key to build the basis for succeeding decisions in the context of situation reports.
- **Problem:** “Common data mining techniques were not successful in general. Despite huge investments, correlating data from different sources did not yield satisfactory results.”
  
  (New York Times in February 2006, Taking Spying to Higher Level, …)

Domain specific approach is necessary
Motivation

- **Solution**: KDD - Overcoming massive data streams for intelligence tasks

- Domain specific approach –
  is based on reconnaissance know-how

- More realistic and manageable: interactions with the user along with domain specific knowledge.

- Gaining security relevant messages should be based on an iterative multi-level process. This process represents the core element of intelligence analysis systems which play an important role for supporting decisions in management information systems.
MIPAS-Architecture
Intelligent Evaluation Software

ICAP

COPIN
Communication Profile Intelligence

SIRES
Situation Report System

VALIT
Validation Tool

Analysis Tool
MIPAS

C4ISR System for RF-Intelligence including Scalable & Modular Evaluation Software
Basic Concepts
ICAP focus of interest

- ACOS provides for the 1st time:
  - broadband complete coverage
  - collection of all emissions on air are stored in database

- How can this information be exploited?

- What are possible interests of our customers?

- Does the data contain recurrent patterns?
  - spatio-temporal
  - communication profiles

- Is there a deviation of normal behaviour
  - frequency change
  - new net members

- Is there an indication for the outfall of expected events?
  - periodicity
  - member does not communicate
Information processing aims at adding value to information
Reduction to statements guides the development
Important topics
- Visualisation, Tracking/GIS
- DM/KDD/OLAP
- Data Warehouse
Discovery Goals

- Discovery of stationary emitters
- Discovery of mobile emitters
- Discovery of simple command structures
Anomaly

missing emission

model:
Exploration Process
Exploration Process

1. **data mining**
   emissions → cluster
   (e.g. spatial combined emissions)
2. **cluster visualisation**
3. **model based detection**
   cluster → communication structures
4. **visualisation of communication structures**

**Iterating:**
- data selection
- data mining
- cluster previsualisation
- cluster visualisation

**Iterating:**
- communication detection
- communication visualisation

**Prepare database**

**Data preprocessing**
data mining

- clustering emission data:
  - location
  - frequency
  - time

- data reduction

- determination of emitters
  - spatial clustering
  - timeseries of emissions
cluster visualisation
model based event detection

- modeling of typical communication structures
  - domain specific modeling language
  - representation of communication models
  - combination of model components

- event detection
  - translation of models in constraints
  - combinatorial analysis of possible model interpretation

- example

Domain specific model based communication detection

- central station: H
- sub station: N
- connection phase (3 emissions: H1, N1, H2)
  - $H1.end < N1.start$;
  - $(N1.start - H1.end) \leq 10$
  - $N1.end < H2.start$;
  - $(H2.start - N1.end) \leq 10$
- alternating communications (every 2 emissions: H3, N2)
  - $N2.end < H3.start$;
  - $(H3.start - N2.end) \leq 10$
model based event detection

simplex communication

connecting phase

alternating communication

emitter

central station, Cluster 20

sub station, Cluster 790

6:00:00 6:00:02 6:00:04 6:00:06 6:00:08 6:00:10 6:00:12 6:00:14 6:00:16

103413792 103413878 103414077 103414976 103415371

ACOS emission id

frequency: 9991000.0, ModType: MFSK

© Plath GmbH
We have: communications in an interesting time and space window with possibly known frequency ranges

We are looking for: changes in frequency usage for the interesting communications
visualisation of a communication structure

example: command structures

We have: known fixed emitter

We are looking for: communication partners building a special structure, e.g. A B A C A D A B A C A D ...
visualisation of a communication structure

example: free search

We have: spatio-temporal data selection

We are looking for: two partners communicating in simplex mode
example: increasing communication activity

At different times the intensity of communication may change heavily.

We have: an area of special interest
We are looking for: changes in emission occurrence
Architecture

GUI

Visualization

Disco Manager

Data-Mining Module

Database

Emissions

Emissions, Clusters, Interpretations

Emissions, Clusters

Interpretations

Models

Event Recognition Module

existing prototype
with distributed architecture
Summary
Cooperative Research

- Cooperation with computer science research institute competence fields:
  - distributed systems
  - software technique
  - intelligent systems
  - learning with new media
  - logistics simulations
  - usability and software-ergonomy
  - IT-Security
  - visualisation and interactive media

- Incorporation of the newest and best research result

- Know-How and innovation transfer from research into industry
The COPIN approach makes massive emission data streams manageable and analysable.

Domain specific orientation integrates the expertise of the intelligence analysts.

We aim at meeting the interests of our customers.

COPIN provides an exploration process to successfully add value to information:
1. data mining
2. cluster visualisation
3. model based communication detection
4. visualisation of communication structures

The model based concepts represent a flexible independent way to work with your data.

The COPIN concept is applicable also to other data collections.
Thank you very much for your attention!

Dr. Vera Kamp
Dr. Joachim Stamm

Tel: +49 40 23734-234
Fax: +49 40 23734-173
vera.kamp@plath.de

PLATH GmbH
Gotenstrasse 18
20097 Hamburg
Germany
www.plath.de