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.

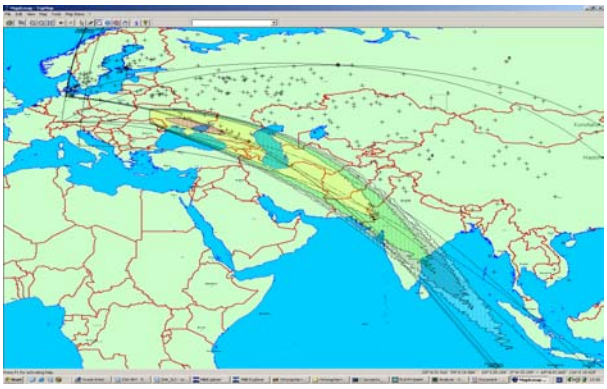

Figure 1: Domain-specific Analysis Function

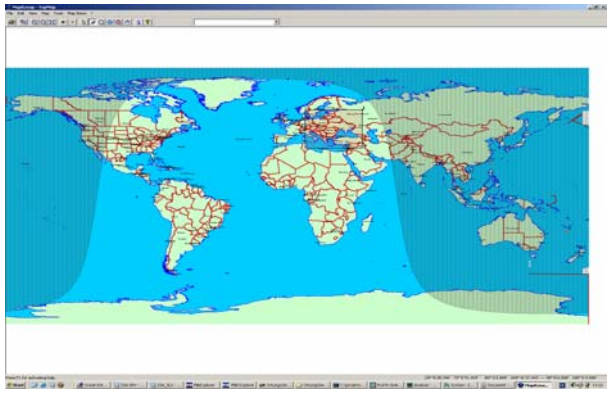

Figure 2: Spatial Access

[1] Knowledge Discovery in Databases

[2] J. Markoff, The New York Times, nytimes.com, 25.2.2006: "Taking Spying to Higher Level, Agencies Look for More Ways to Mine Data"

[3] Jan Herring, "What is Intelligence Analysis?", Competitive Intelligence Magazine, Vol. 1. No.2, July-Sep., 1998, p.14 .

Kamp, V. (2006) KDD – Overcoming Massive Data Streams for Intelligence Tasks. In *Visualising Network Information* (pp. 2-1 – 2-4). Meeting Proceedings RTO-MP-IST-063, Paper 2. Neuilly-sur-Seine, France: RTO. Available from: http://www.rto.nato.int/abstracts.asp.

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 extremly 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 interesing data. Additionally different attributes have to be integrated.

1.3 Model based communication detection

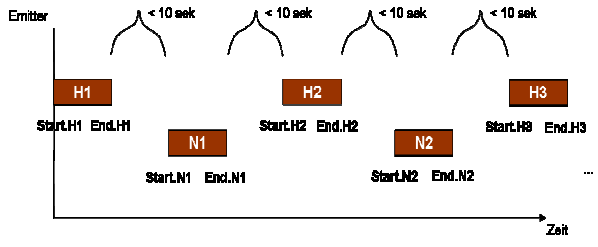

Figure 3: Simplex Communication Rules

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.

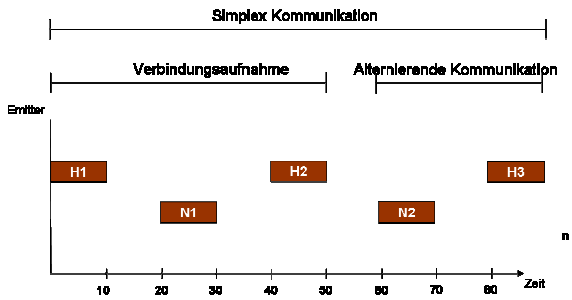

Figure 4: Simplex Communication Instance

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 sub station. 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

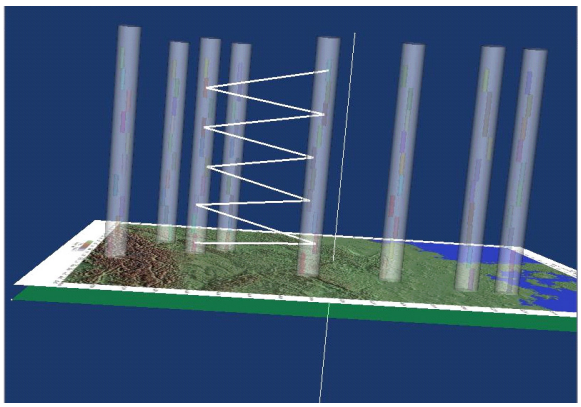

Figure 5: Discoverd Network Communications

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 acceptence of the results depends on the possiblity to validate the results. The sustainablity of results has to be guaranteed by flexible extension of actual domain specific analysis methods.

2.0 REFERENCES

[1] J. Markoff, The New York Times, nytimes.com, 25.2.2006: "Taking Spying to Higher Level, Agencies Look for More Ways to Mine Data"

[2] Jan Herring, "What is Intelligence Analysis?", Competitive Intelligence Magazine, Vol. 1. No.2, July-Sep., 1998, p. 14.