**Association Rule Mining (ARM)**

**Background**

Association Rule Mining is concerned with the identification of patterns in data where the records comprise binary valued attributes (yes/no). The patterns are frequently co-occurring sets of values called Frequent Item Sets. These can be used to generate probabilistic rules of the form “if A exists in the data set then B is also likely to exist in the data set”, called Association Rules (ARs). ARM has been used extensively to analyze Customer Databases to identify buying patterns. Research at Liverpool, supported by Royal Sun Insurance, was initially directed at fast algorithms for ARM. The result was two software data structures (the P-tree and T-tree) which significantly enhanced the process. A number of supporting software systems were also developed, the most significant is TFP (still used within the global data mining research community).

**Example**

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>

**Data**

- ABCD
- ABC
- AD
- ADEF
- AE
- DE

**P-tree**

- A
  - B
    - D
      - E
  - C
    - D
      - E

**T-tree (Support 50%)**

- A
  - D
    - E
  - B
  - C

**ARM for Very Large DB**

Subsequent to early work on fast ARM conducted at Liverpool further development was directed at Very Large Data Bases (VLDBs) to allow ARM to be applied to such data while still using only a single machine. A partitioning strategy was developed whereby the data was vertically partitioned (divided) into “chunks” which could be processed individually. The vertical partitioning strategy also lends itself to distributed ARM (where data is spread over several machines) and parallel ARM (where data is processed using several processors).

**Classification Association Rule Mining (CARM)**

Classification is the process of building a classifier using known “training” data to be applied to unseen data. Classification has wide application, examples include: image and document cataloguing, and the identification of medical conditions. CARM is a technique for building classifies using ARM technology. The T-tree data structure was found to be well suited to CARM. A CARM classification tool, TFPC, has consequently been developed and is in use at Liverpool and other locations.

**Current Work**

Current work on ARM is directed at enhancing the benefits that ARM can offer. Issues include:

- Standard ARM assumes all attributes are equally significant, this is of course not the case and has given rise to Utility ARM and Weighted ARM.
- So that non-binary attributes can be processed techniques exist for dividing continuously valued attributes in to ranges, this has resulted in the “crisp boundary” problem, and the concept of Fuzzy ARM to address this issue.
Text Mining

Background
The application of data mining techniques to documents was a natural progression from the early pioneering work on tabular data. Typical applications include clustering and classification of news paper articles and web pages, and emails (spam v. not spam). The primary challenge of text mining has always been not so much the data mining techniques themselves, but how to represent the input data. Pre-processing techniques include natural language parsing, key word/phrase identification, stemming, and the use of stop lists. A secondary challenge is the size of the data sets involved.

Classification Avoiding Language Dependent Features
The research team at Liverpool have directed considerable effort at developing algorithms and technique for pre-processing document collections in a way that does not require language specific techniques (such as natural language parsing and part of speech tagging). A number of approaches have been produced that use statistical techniques to identify keywords and phrases. Text classification experiments have been conducted, using Classification Association Rule Mining (CARM) techniques, on English Language and Chinese data sets with good results.

Rule Induction
Current work in Text Mining is directed at “Rule induction” systems for text classification. Classifiers, built using techniques such as CARM technique, are represented in the form of rules. Rule base classifiers offer the advantage that they are easy to understand and explain. The disadvantage of classifies built using CARM is that they produce a lot of rules (too many?). Rule induction systems partly resolve this issue but have been demonstrated to be less accurate. Current work seeks to improve the accuracy of such systems by increasing the degree of sophistication of the induced Rules (for example the inclusion of negative attributes).

Future Work
The research team are currently interested in applying their text mining knowhow to multimedia mining, and particularly questionnaire mining. A project has recently started to analyse (mine) questionnaires routinely completed by veterinary practices.

Contact:
Prof. Frans Coenen
The Department of Computer Science
The University of Liverpool
Liverpool L693BX

Tel: 0151 725 4253
Email: coenen@liverpool.ac.uk
WWW: http://www.csc.liv.ac.uk/~frans/
WWW Mining

Background
WWW mining can be divided into two categories: (i) WWW content mining and (ii) WWW usage mining. The first is not significantly different from text mining except that additional information can be included regarding URL data, WWW page “look and feel”, etc. WWW usage mining is concerned with how users of a WWW site, or a collection of sites, move around the site(s). This in turn can provide useful information regarding the design of the WWW site. The input for WWW usage mining is obtained from WWW log data. User browsing sessions are recreated from the WWW log data, using certain assumptions about the length of time a user may stay on a page. These user sessions are then encoded in such a way that pattern mining techniques can be applied. The aim being to identify frequent user patterns within the data.

Case Study (Car insurance WWW site)
Staff within the Department of Computer Science have been engaged on a Knowledge Transfer Project (KTP) in collaboration with a local insurance company to build a toolkit for the generation of “bespoke” car insurance WWW sites. An integral part of this work is the WWW usage mining of a demonstration car insurance site to determine the most appropriate manner in which the site can be further developed and re-engineered to facilitate maximum usability. The aim is to proceed in a standard manner and collect usage data from WWW logs available on servers within the company. This will then be used to model user WWW sessions reflecting how users traverse/use the demonstration site. Association Rule Mining techniques were then be applied to identify frequent usage patterns.

WWW site Boundary Detection
To facilitate and speed up information retrieval, search engines, and similar systems, pre-index (tag) WWW pages. This is done by sending out pieces of software, called “Web Crawlers”, across the internet that determine, as they travel, the topic(s) of individual WWW pages. Similar WWW pages are often grouped together. A current challenge is how to determine when a Web Crawler has moved out of one site and into another. Of course this first necessitates some philosophical understanding of what a WWW site is, and this definition may vary according to the indexing application domain. However, assuming that we have such a definition, data mining techniques can be used to identify the desired boundaries so that “similar” WWW pages can be grouped.

Work at Liverpool has been directed at investigating data mining mechanisms to identify WWW site boundaries. The idea is that as the Web Crawler proceeds it can “learn” the nature (signature) of the current WWW site in which it is located and then use this knowledge to determine if the next visited WWW page is included in the site or not.

Contact:
Prof. Frans Coenen
The Department of Computer Science
The University of Liverpool
Liverpool L693BX
Tel: 0151 725 4253
Email: coenen@liverpool.ac.uk
WWW: http://www.csc.liv.ac.uk/~frans/

Finding WWW site boundaries is a challenging endeavour.
Graph Mining

**Background**
Graph mining practitioners maintain that everything can be represented as a graph (or a tree): documents, images, WWW pages, etc. As such these graphs can be mined. The objective of graph mining is to identify frequently occurring sub-graphs, either in one single large graph or across a collection of small graphs. The usual start point for graph mining is to impose a canonical representation such that, regardless of the start point, identical graphs will have an identical representation. The actual sub-graph mining normally proceeds in an Apriori style manner with one edge sub-graphs, proceeding to two edge sub-graphs and so on. Candidate generation is nontrivial as sub-graphs can be “grown” in many different ways. Further the “downward closure property” of itemsets does not hold. A number of effective sub-graph mining algorithms have been developed (e.g. gSpan and Subsume).

**Weighted Graph Mining**
Work at Liverpool has been directed at weighted sub-graph mining. The motivation is that sub-graph mining is computationally expensive (see left panel). The intuition is that, for many applications, some node/edges may be considered to be more significant than others. Consequently any sub-graph identification algorithm should be directed at the most significant nodes/edges so that the computation overhead can be reduced and larger datasets can be processed. The work has resulted in a number of weighting schemes, using both user supplied weights and inferred weights. These have been implemented in a sequence of variations of the well known gSpan algorithm. To date the weighted gSpan algorithms have been used with respect to a variety of applications domains including image mining (both artificial and real MRI scan image sets), surface representation and text mining.

**Case Study: UK Cattle Movement DB**
Research within the Department of Computer Science has investigating the application of weighted graph mining techniques to the UK’s cattle movement DB. As the name implies, this DB records all the cattle movements between locations in the UK. Each location can be viewed as a node and the movement between nodes, where it exists, as an edge in a graph which can be weighted with the number of movements. All cattle movements are time stamped thus a sequence of cattle movement graphs, one per month, can be defined. In effect the graphs represent a social network. **Social network mining** is a significant recent trend within the data mining community. By applying our weighted sub-graph mining knowhow we can identify “communities” within the cattle movement network.

**Future Work**
A programme of work is currently under in progress, at Liverpool, to investigate more effective (efficient) ways to perform graph mining. The idea is to found new frequent sub-graph mining algorithms on work carried out at Liverpool on The research team at Liverpool are also directing their efforts at the generation of debate/argument graphs. The focus is on political debates where the outcome is known (as a result of a final vote). Nodes are speakers and edges indicate supports or attacks.

**Contact:**
Prof. Frans Coenen  
The Department of Computer Science  
The University of Liverpool  
Liverpool L693BX  
Tel: 0151 725 4253  
Email: coenen@liverpool.ac.uk  
WWW: http://www.csc.liv.ac.uk/~frans/
# Medical Image Mining

## Background
Image mining can be argued to be the current frontier with respect to data mining. The challenge of image mining is how to represent the input data set so that: (i) it can be mined, (ii) no significant information is lost, and (iii) results can be obtained in reasonable time. Representing images at the “pixel level”, however desirable, is not a realistic option. Image mining work at Liverpool is currently being undertaken in the context of two application domains: (i) MRI scan data mining, and (ii) retina image mining.

## Case Study: MRI Brain Scan Classification
Work on MRI scan mining is directed at classifying scans according to a particular feature within the scans. Currently the research team are considering the Corpus Callosum (CC), a tissue element that is highly visible in MRI scans. The CC joins the left and right sides of the brain, and it is conjectured that the size and shape of the CC is correlated to certain medical conditions (such as Epilepsy), or to certain human abilities (it has been demonstrated that the CCs of musicians and mathematicians are different).

### The Corpus Callosum is a highly visible feature in MRI scan data (marked in yellow)

### Case Study: Retina Image Screening
Retina image screening is routinely undertaken to detect conditions such as AMD (Age Related Macular Degeneration) which may lead to blindness. AMD may be identified by the presence of Drusen (fatty matter) on the retina. The image on the right hand side below includes Drusen, the left hand image does not). Work at Liverpool is investigating the use of a variety of novel techniques to classify retina images so as the automatically detect conditions such as AMD. The current approach is founded on a **histogram based representation** couple with a **Case Based Reasoning** approach. Case matching is undertaken using a time series analysis technique called Dynamic Time Warping. The research team has been greatly encouraged by the results obtained to date.

### Contact:
Prof. Frans Coenen  
The Department of Computer Science  
The University of Liverpool  
Liverpool L693BX

Tel: 0151 725 4253  
Email: coenen@liverpool.ac.uk  
WWW: http://www.csc.liv.ac.uk/~frans/
Trend Mining

Background
Trend mining, as the name suggests, is concerned with the identification of trends in data. These trends are typically temporal trends but may also be spatial trends. The Liverpool approach is directed at determining how common patterns that feature in data change with respect to time and/or space. This work is being undertaken with respect to four application domains (and four different groups of collaborators): (i) longitudinal patient data, (ii) veterinary practice data, (iii) standard customer data and (iv) the UK cattle movement DB. The fundamental challenge is how best to present the identified trends, given the large number of patterns that exist within the data sets, so that they can be utilised by decision makers.

Case Study: Veterinary Practice Data
Research work on identifying trends in veterinary data is being conducted in collaboration with the Liverpool School of Veterinary Science as part of the SAVNET project. The data source comprises consultation logs filed by Vets at a number of selected locations across the UK. The logs comprise a tabular section and a free text section. The free text section is the most challenging as this comprises ambiguities, obscure abbreviations specific to particular practices, misspellings, etc.

Case Study: Longitudinal Patient Data
Research work on trend mining in longitudinal patient data (medical records) was conducted in collaboration with the Royal Liverpool University Hospital (RLUH). Specifically in the context of the hospital’s diabetic patients. The RLUH has many years experience of treating diabetic patients, and has patient data collected over the last eighteen years. This tabular data is a rich source of information if it can be appropriately mined. What the Liverpool team were particularly interested in was finding trends in this data that will give indicators of the progress of the diabetic condition.

The particular challenge of the work was that the data is very noisy, and contains many anomalies and missing and duplicate records. To address the issues associated with the large number of identified trends that exist within the data, a system of prototype trends has been proposed to allow trends to be grouped (clustered).

Case Study: Customer Data
The identification of trends within customer data, such as that collected by eCommerce enterprises, was undertaken in collaboration with Transglobal Express, a local freight forwarding company. A sliding window approach to identify jumping patterns within the data was adopted. This was found to work extremely well when identifying customer buying patterns.

Case Study: Cattle Movement DB
Trend identification within the The UK Cattle Movement DB was undertaken in a similar manner to the longitudinal patient data study. However, in the case of the cattle movement data Self Organizing Maps (SOMs) were used and present the trend mining results and to provide cattle movement prediction information.

Contact:
Prof. Frans Coenen
The Department of Computer Science
The University of Liverpool
Liverpool L693BX
Tel: 0151 725 4253
Email: coenen@liverpool.ac.uk
WWW: http://www.csc.liv.ac.uk/~frans/
Multi-Agent Data Mining (MADM)

Background
Agents are small autonomous software systems. Multi-Agent Systems (MAS) are collections of such autonomous software systems designed to collaborate (without centralized control) to resolve problems. As such MAS are seen to have a lot to offer in many IT domains. Data Mining (DM) is one such domain. Multi Agent Data Mining (MADM) systems are directed at realizing the advantages offered by MAS with respect to DM. Work in this area has been typically directed at a particular DM task or a particular application area. Work at Liverpool has been focused on the concept of a generic MADM framework. A demonstration generic MADM was initially developed, EMADS (the Extendible Multi-Agent Data mining System), followed by a more sophisticated system directed as Multi-Agent Based Clustering (MABC).

EMADS
The fundamental insight behind EMADS was that the domain of DM is so large that the creation of an all-encompassing framework is not feasible. Instead EMADS provided the tools that facilitated easy extension of the system to allow it to grow in an “organic” manner as more and more end users contributed more and more data and DM tools. The research team conducted a sequence of experiments to determine the support provided by EMADS for: (i) meta Association Rule Mining (meta ARM), (ii) parallel/distributed ARM, and (iii) the generation of a best classifier given a large choice potential classifier generators. Extendibility in EMADS was facilitated by a collection of Wrapper Classes that allowed new data sources and DM algorithms to be “wrapped up” to become EMADS agents.

The MABC Framework
Communication between EMADS agents comprised a set of bespoke primitives. Thus there was no requirement for EMADS users to be aware of the communication mechanism, this was taken care of by the wrappers. MABC has taken a different approach, namely the creation of a generic, but extendible, “performatives”. For contributors to add agents to MSBC they simply have to adhere to the restrictions of the proposed set of performatives. The work was directed specifically at multi-agent based clustering. One of the main challenges was determining how an agent can know that it has a good cluster configuration. The current version of the MABC framework, as in the case of EMADS, has been implemented in JADE (the Java Agent Development environment). Some interesting results have been produced prompting further investigation of the MADM paradigm.

Contact:
Prof. Frans Coenen
The Department of Computer Science
The University of Liverpool
Liverpool L693BX
Tel: 0151 725 4253
Email: coenen@liverpool.ac.uk
WWW: http://www.csc.liv.ac.uk/~frans/
Data Mining and Arguing From Experience

Background
Argumentation is the study of how humans reason about issues and come to decisions regarding courses of action. Computer scientists are interested in modeling this process so that computers can reason about issues, an example application is auction systems. Argumentation also has a role to play in Multi-Agent Systems (MAS), computer systems where many relatively small, autonomous, computer programmes collaborate to solve problems. Work within the Department of Computer Science has sought to combine the theory of argumentation together with work on Multi Agent Data Mining (MADM). Two system have been developed: PADUA and PISA. Both are directed at providing the classification of “cases” in a distributed setting. The exemplar application is benefits claims. Claims are resolved by considering past experience (in the form of individual data bases of past cases) distributed across two (PADUA) or more (PISA) benefit offices. Arguments are structured by each “player” (agent) mining their own databases using Association Rule Mining (ARM) techniques.

PISA
PISA is a multi-player argumentation framework. In the context of classification each player advocates their own class, there are thus as many players as there are classes. The multi-player model features a number of issues that are not of significance in the two player model (PADUA). The first issue is the turn-taking regime. The second is the storage of the arguments presented so far. The multi-player model also allows various strategies to be employed. Players can elect to attack the strongest argument proposed so far, or eliminate weaker arguments first. Alternatively players can elect not to contribute an argument until they think they are in a winning position. Players can also form temporary alliance to defeat another player’s argument (it has been found that by “pooling” information stronger arguments can be derived). PISA has been evaluated using many “bench-mark” data sets. During the experimentation it was found that PISA produced good classification results. It was also found that PISA was very good at operating with noisy data (much more so than the other classifier generators tested).

PADUA
PADUA is a two player arguing from experience software system. This is essentially a simplified version of the N player (PISA) system which does not feature any of the multi-player issues associated with the N player game. The players (agents) take it in turns to argue about a case. Each player proposes arguments intended to refute the other players’ arguments, or promote their own arguments, by mining their own data sets. This process continues until one of the players cannot generate any more arguments, in which case the remaining player is declared the winner (the winning classification is adopted).

Contact:
Frans Coenen  
The Department of Computer Science  
The University of Liverpool  
Liverpool L693BX  
Tel: 0151 725 4253  
Email: coenen@liverpool.ac.uk  
WWW: http://www.csc.liv.ac.uk/~frans/
3-D Surface Mining

Background
There are many data mining applications where we wish to gather information concerning the nature of 3D surfaces. The application that researchers at Liverpool have been directing their efforts at is the prediction of the errors that may result from the application of sheet metal forming processes. If the errors can be predicted it may be possible to define a shape that takes these errors into account prior to production. Two approaches are being followed at Liverpool: (i) a tabular approach and (ii) a graph mining approach based on the concept of Vertex Unique Labeled Sub-graphs (VULS).

3-D Surface Mining Using Tabular Data
Work on tabular approaches to 3-D surface analysis has been directed at: (i) Local Geometry Matrices (LGMs), (ii) distance from edge representations and (iii) time series approaches. A LGM is a $n \times n$ matrix with the point of interest at the center (the idea is founded on the concept of local binary patterns used in texture analysis). The geometry of any surface can therefore be expressed as a collection of LGMs. The distance from edge representation is founded on the observation that, in the context of sheet metal forming, distortion is greater away from edges. In this case each point in a 3-D surface is represented by the distance from its nearest edge. The time series technique is founded on the concept of representing local geometries using a linear curve.

Case Study: AISF Sheet-metal Forming
As part of a large European project (Framework 7) researchers at Liverpool have been looking at the distortion induced in titanium and steel as a result of the Application of Asymmetric Sheet-metal Forming (AISF) processes. A process that results in distortions being introduced into the manufactured parts called “springback”. Researchers at Liverpool have been able to predict the degree of springback to the extent that corrections can be made to the desired shape’s specification prior to applying the AISF process. Production experiments have found that a more “accurate” shape is produced.

An AISF Machine

3-D Surface Mining using VULS
The idea behind the use of VULS for surface representation and mining is that a collection of points representing a surface can be thought of in terms of a lattice where each point constitutes a node which in turn is connected to each of its four neighbours. Each edge then represents the absolute difference in height between two nodes. The nodes have errors (distortion) values associated with them. Sub-graphs within this lattice thus describe local geometries. If, using a training set, we can identify sub-graphs whose nodes have a unique error labelling associated with them (VULS) these can then be used as generic geometry descriptions with which to predict errors in new “unseen” shapes that we may wish to manufacture.
Integrating Data Mining and Agent Based Simulation

**Multi Agent Based Modeling and Simulation**

Multi Agent Based Simulation (MABS) is concerned with using the advantages offered by Multi-Agent Systems (MAS) based technology to produce realistic simulations. MAS technology is particularly well suited to simulations that involve a number of individuals. Each individual can be conceived of as an agent with its own behaviours. A standard agent platform, such as JADE (Java Agent Development Environment), can then be used to realise the simulation. Researchers at Liverpool have been working on various aspects of MABS for sometime and are particularly interested in how data mining (machine learning) techniques can be used to allow agents to learn behaviours.

**Case Study: Mouse Behaviour Simulation**

Mice cause considerable damage to crops. To combat mouse infestations scientists would like to gain a deep understanding of their behaviour. Researchers at Liverpool have developed a demonstration Mouse MABS that attempts to model the way that mice behave in various scenarios. The majority of these scenarios are variations on a “mouse in a box scenario”. This work is on-going, but encouraging results have been obtained to date.

**Data Mining for MABS**

One of the difficulties with respect to creating realistic simulations using the MABS concept is that the behaviours must be “hard coded”. Currently each agent’s actions are based on a behaviour tree, one per agent. The behaviour trees may all be identical or all be different or the simulation may feature a mixture of these two extremes. At any fixed moment of time during a simulation each agent will be located at a node within its behaviour tree which will dictate its next possible moves. Currently these behaviour trees have to be hand constructed based on advice and observations of domain experts. This is a time consuming process prone to mistakes and the introduction of inaccuracies. To address this issue researchers at Liverpool have been looking at ways in which these behaviour trees can be automatically generated using machine learning (data mining) techniques. The idea is that we suspend a video camera over a “mouse on a box scenario” and film the behaviour of the mouse (or mice) in question. The resulting video can then be analysed using some appropriate learning technique so that the desired behaviour tree can be generated automatically. More specifically individual mice in the video will be tagged and their movements tracked and recorded in tabular format which can then be analysed. As a result a behaviour tree can be automatically extracted that serves to capture mouse movements in a realistic manner.

Contact:
Prof. Frans Coenen
The Department of Computer Science
The University of Liverpool
Liverpool L693BX

Tel: 0151 725 4253
Email: coenen@liverpool.ac.uk
WWW: http://www.csc.liv.ac.uk/~frans/