

Foresight Project on Intelligent Infrastructure Systems

Research Review

# **Data mining, data fusion and information management**

John Shawe-Taylor, Tijl De Bie, Nello Cristianini

*School of Electronics and Computer Science, University of Southampton*

This discussion document is part of a series of reviews of the state of research into issues relevant to the development and implementation of an intelligent infrastructure systems (IIS), with a particular emphasis on transport. It was prepared as background material for the Foresight Project on Intelligent Infrastructure Systems. Series Co-ordinators, Professors Phil Blythe, Glenn Lyons, John Urry and Will Stewart; Series Editor, Michael Kenward

The project set out to identify the opportunities that science and technology can offer in the creation of an intelligent infrastructure system. In particular, it investigated the opportunities arising from IIS in the development of a more "intelligent" transport system that can deliver the robust, sustainable and safe services that we need.

While the Office of Science and Technology commissioned this review, the views are those of the authors and are independent of Government and do not constitute Government policy.

Further details are available at the Foresight web site: <http://www.foresight.gov.uk/>

Contact the Foresight Cognitive Systems team at:

Foresight Directorate  
Office of Science and Technology  
1 Victoria Street  
London  
SW1H 0ET

Fax: 020 7215 6715

# Contents

1	Introduction	1
2	Patterns in data	2
2.1	Pattern discovery and modelling	2
2.2	Approaches to pattern analysis	3
2.2.1	Machine learning and statistical learning theory	3
2.2.2	Multivariate statistics	4
2.2.3	Probabilistic (graphical) models	5
2.2.4	Classical data mining	6
2.2.5	Text mining and string processing	6
2.2.6	Graph mining	6
2.2.7	Artificial intelligence	7
2.2.8	Data fusion	7
2.3	A common denominator: pattern analysis	8
3	Data floods in intelligent infrastructure systems	8
3.1	Global infrastructure awareness	8
3.2	Local infrastructure awareness	8
3.3	Transportation system and user awareness	9
4	Applications	9
4.1	Intelligent traffic management	9
4.1.1	State of the art	9
4.1.2	Near future	10
4.1.3	Speculations	11
4.2	Intelligent resource allocation	11
4.2.1	State of the art	11
4.2.2	Near future	11
4.2.3	Speculations	12
4.3	Intelligent traffic flow control	12
4.3.1	State of the art	12
4.3.2	Near future	12
4.3.3	Speculations	13
5	Data mining, data fusion and information management in intelligent infrastructure systems	13
6	Conclusion	13

# 1 Introduction

Radio-frequency identification devices (RFIDs), sensor networks (SNs) and global positioning systems (GPS) are among the many ways in which we will generate and share data about many aspects of everyday life. Low cost and ubiquitous information processing and storage, as well as sensors, and flexible networking capability of these devices, allow a user to interact with an “information infrastructure” to obtain constant information about security, safety, congestion, of more traditional infrastructures such as transportation, telecommunication and distribution networks, among others. One could monitor the structural stability of a bridge, congestion in highway traffic, or the weather, at various locations. Examples are the tracking of vehicles and goods, real-time monitoring of their state, detection of global patterns and optimization of the overall system, possibly including information about weather, and other factors.

The detection and exploitation of patterns in such data hold great promise and at the same time present formidable technical challenges. Data fusion is going to be a key issue, as are the handling of large data-streams and distributed processing. Adaptive and intelligent methods for automatic pattern analysis have evolved with a series of applications from web mining to computational genomics, including monitoring the atmosphere and water quality, detecting biological attacks, early spotting of epidemics; prediction of major geological and meteorological events, earthquakes, tsunamis, storms, flooding, etc.; and of emerging trends and patterns on the Internet.

Pattern analysis, the field of information processing devoted to this kind of task, results from the fusion of statistics, machine learning, signal processing and many other fields. Pattern analysis has recently come of age due to pressure from genomics and web data. The emphasis has shifted to themes of large scale, adaptive, parallel and online processing. Further progress in this field is necessary for the development of an intelligent infrastructure based on massive data acquisition, sensor networks, etc.

This report will focus on the potential of bringing together information into an intelligent infrastructure in which pattern analysis, data fusion and information management can deliver a step change in the flexibility and effectiveness with which individuals and businesses can exploit our transport infrastructure.

Our belief is that the energy consumed in transport is poorly exploited for a number of reasons that the intelligent infrastructure analysis can significantly address:

- traffic congestion implies reduced fuel efficiency,
- the average number of people travelling per car is very low,
- public/private transport combination options are difficult to plan.

We feel that it is important not to jump to the barricades in any public/private transport debate. Our philosophy is to view all transport infrastructure and technology as a resource. Our aim is to highlight ways in which this resource could be exploited more efficiently. In some cases this might mean better linkage with public transport, but this is only realistic if it can be made more attractive than current options. In other cases it will involve enhancing the efficiency with which private transport is used, again in ways that are attractive to users.

This report is structured as follows. We first discuss the current state of the art in the area of data mining, pattern analysis, text mining, and data fusion. Where possible we indicate potential relevance of the techniques discussed for transport problems. Following these sections we discuss example scenarios that demonstrate how IIS could address the weaknesses identified above. Here we bring together insights of the earlier sections to assess how realistic the proposed systems would be using current technology, and we identify relevant gaps in current technology. We discuss these systems along with an estimated time range in which we assume they can be deployed on a commercial scale.

## 2 Patterns in data

The information avalanche in today's society is immense, coming from the Internet, satellites monitoring the Earth's surface, sensors on the ground, large scale measurement systems in, for example, computational biology, telescopes screening the night sky and high-energy physics experiments. The times when a single person or at most a group of people could investigate this data flow by hand are gone for good. To cope with the flood of data, machines running sophisticated algorithms now form the interface between the data and the person (or system) using it. These algorithms serve as a filter, identifying potentially interesting relations, or patterns, in the data. Important examples where data mining plays a crucial role are the text and graph mining algorithms underlying the Google search engine, bioinformatics algorithms that allow researchers to reconstruct and analyse entire genomes and so on.

Interestingly, our confidence in automated pattern analysis is so high, that data are often generated or made available, sometimes at a significant financial cost, under the implicit assumption that we will be able to extract the interesting patterns that are assumed to be present. This confidence stems from successes in different research domains united under the heading of pattern analysis: data mining and notably frequent pattern mining, data fusion, machine learning, text mining, string processing, artificial intelligence, multivariate statistics and more. These successes are due to computational advances (following Moore's law) as well as to algorithmic and statistical insights gained in these research domains in the past few decades.

Many of these domains have a common origin in early beliefs in our ability to build intelligent machines, not unlike human beings. Yet, the inability to immediately realise these grandiose expectations caused researchers to focus on several sub-domains such as those named above. Today, there is increasing awareness that the advances made in each of the scientific communities that are part of pattern analysis are founded on the same algorithmic and statistical principles. On the one hand, diversity in approaches may be beneficial in considering the variety of practical problems encountered. On the other hand, however, recognising the common principles underlying different pattern analysis approaches seems to be increasingly fruitful in enhancing interactions and cross-fertilisations between those domains, suggesting that a unifying approach to pattern analysis should be a vital component of current and future research.

Such a joint understanding of pattern analysis techniques is required from yet another perspective. As we have said, large efforts are made to gather and store huge amounts of data without prior knowledge of the kinds of patterns to be found within them. The exclusion of any such available data source implies the wasteful abrogation of an investment made during data acquisition. Hence, the need for *data fusion* techniques that can seamlessly integrate often heterogeneous kinds of information has steadily gained importance with the shift from hypothesis driven research to data driven research that relies on pattern analysis.

Whereas it is still hard to outperform humans in their pattern finding capabilities on small data sets, computers and information management systems can now deal with enormous amounts of information, and the quality of their conclusions relies to a large extent on this abundance. Naturally, huge data sets pose non-trivial algorithmic problems. The number of potentially interesting patterns grows quickly with data-set sizes, and it is impossible to carry out exhaustive searches for interesting relations. On the statistical side, related problems can be identified: the more potential patterns that are tested, the more likely it becomes that we will pick up spurious patterns. It is well known that humans are badly placed to judge reliability of patterns: they are easily misled by an almost instinctive desire to identify patterns. Interestingly, there seem to be strong connections between the algorithmic properties and statistical properties, indicating that a problem that is algorithmically benign may also be benign in a statistical sense.

### 2.1 Pattern discovery and modelling

In the above, we used the term *pattern* in a broad meaning, as a *relation that is present in the data*. Such relations can be divided into two categories: models that describe a structure that is assumed to

be present in the data; and patterns that are unlikely to occur given such a model. In this respect there are two main tasks in pattern analysis.

The first task is the modelling task, which is concerned with finding regularities or structure in the data. Such an underlying structure may be capable of explaining a specific returning pattern in the data. For example, such a pattern may be the regular occurrence of traffic jams around rush hour, which can be explained by a model of the collective behaviour of the drivers.

The second task is commonly referred to as pattern discovery, and searches for patterns that contradict prior assumptions about the data, as specified by a model. For example, the occurrence of a traffic jam around noon due to a road accident would be an anomaly that may be a notable pattern.

Both modelling and pattern discovery are highly relevant in intelligent infrastructures. Accurate models are needed to anticipate future demands to be met by the traffic network. On the other hand, we should detect unusual events as soon as possible to enable fast and appropriate reactions to potential anomalies that deviate from the model.

We want to point out that the distinction between *patterns in modelling* and *patterns in pattern discovery* is sometimes quite subtle, and it is not always appropriate to make it. For example, imagine that our model about traffic jams is unaware of the date, i.e. we use the same model every day of the week. In that case, the absence of a traffic jam on Sunday would be noted as unlikely given the model. Then, one could either view the absence of the traffic jam as a notable and unlikely occurrence, a surprising pattern, or one can improve the model by taking the weekday into account as an additional parameter. Hence, the same pattern can either be compared with an *a priori* specified model, pattern discovery, or it can be used to build a model or update an existing model, modelling. Whether a pattern describes an anomaly or a regularity to be captured in a model often depends on the prior assumptions about the data.

## 2.2 Approaches to pattern analysis

### 2.2.1 Machine learning and statistical learning theory

Machine learning is concerned with identifying general patterns from finite samples of data. For example, given a set of emails, some of which are classified as spam and some as non-spam, develop a general method of filtering spam emails, or, given a set of images of cells some of which are labelled as cancerous, develop a system to screen for cancerous tissue. The important feature of the system is that the pattern observed in the finite sample should not just be a chance occurrence but should be evidence of an underlying property of the phenomenon that is being observed, be it emails or cells. Hence, the question to be answered is whether an observed pattern is the result of some structure that is present in the source that generated the data. If the pattern is indeed a reflection of such an underlying structure, we say that it allows us to generalise, as it allows one to make predictions on new data.

There is a tension here between the number of patterns that we are prepared to consider and the certainty with which we are able to assert the permanence of the pattern. If we are too flexible then patterns will always become apparent, but they will just be chance properties of the particular set of example data. Machine learning is concerned with designing learning systems that make the best of this trade-off, that is maintain as much flexibility as possible in the search for patterns, while still being confident about the patterns that have been identified.

The study of *generalisation* is undertaken within the framework of *statistical learning theory* which makes the assumption that there is a fixed but unknown underlying distribution generating the training data. Each data point is assumed to be independent of the others, but coming from the same distribution. This holds true in particular for the training data, on which the pattern is identified, and the test data, to which the pattern can be generalised. It is this statistical link between training and test data that enables one to generalise.

The analysis can guide the design of algorithms for many of the common tasks that are tackled using the machine learning methodology. Perhaps the most common of these is classification. *Classification*

*problems* are concerned with labelling data as being in a class or not, as for the spam and cancer problems described above. Each data item must be labelled with a binary label in the training data, and new data must be assigned to the corresponding class. If the label is a real value the task is referred to as a *regression problem*, a situation analogous to interpolation of data. If there are a number of different classes, as for example in the determination of the topic of a new article, it is known as *multi-class classification*. The problem of *ranking* refers to determining an ordering into one of a finite set of levels. All of the tasks considered so far have a label associated with each datum and are therefore often referred to collectively as *supervised learning problems*.

In contrast, *unsupervised problems* are concerned with identifying patterns in unlabelled data. *Novelty detection* is a case in point. Here a system attempts to model the 'normal' data as exemplified by the training set and hence to filter out abnormal new data that might correspond to some departure from the standard behaviour of a system. Example applications are in engine monitoring, where we hope to be able to pick up on abnormal operation, or in fraud detection, where unusual usage patterns of a mobile device may indicate the onset of illegal activity. Another well-known unsupervised task is *clustering*, in which data is automatically broken into sub-groups that in some way represent a natural structure or pattern in its source.

Before we consider where machine learning techniques might be applicable in Intelligent Infrastructure Systems (IIS), it is worth stepping back from the particular tasks we introduced above and emphasising that the key feature is in all cases the idea that a pattern of information that carries wider validity is identified from only a sample of examples.

Before the machine learning paradigm can be applied, there has to be a suitable source of data drawn from the problem domain. For transport systems, an example of such a data repository is the database of traffic density monitoring held by the relevant government departments. This extensive database could provide a rich source of information for modelling and predicting traffic flows, both under normal and abnormal conditions. Such a system could form an important component in a number of the scenarios sketched in the later sections of this report.

Where data is not readily available or has not been drawn in the right context, or where the context is continuously changing, there may be a need for further adaptation of models during the operation of the system. Adaptation to and learning from data as it is processed is referred to as *on-line learning*. This will undoubtedly form an important component of many systems as we suggest in the descriptions below.

### 2.2.2 Multivariate statistics

Virtually all data that are collected and stored are multi-dimensional. Typically, a range of features are measured at a particular time or condition and stored as a complex data object. In many cases the data comes in the form of a vector of real values, though in others one or more of the components may be discrete or may involve additional structure, such as strings, xml documents, graphs, networks etc. However, there are techniques that can render complex data objects like strings and graphs into virtual vectors, so that in virtually all cases it is possible to consider the data to be in vector format.

The processing of vector data immediately raises the question of how independent the individual features really are. In other words, is it possible to find a smaller subset of features that captures the salient aspects of the data more succinctly? Indeed, these questions merge into the problem of unsupervised learning since it is likely that different representations may better capture the structure of the data than others.

The most widely used method of performing so-called *dimensionality reduction* is known as *principal components analysis* (PCA). This approach uses the spread of the data to suggest a new basis by choosing the directions that maximise the variance of the observations. If there are significant correlations between the different features, the number of PCA features required to capture the data will be fewer than the original dimensionality.

Another situation where we can expect significant dimensionality reduction is when we have two independent representations of the phenomenon of interest. Perhaps the simplest example is a paired corpus of documents in which each English document comes together with its translation into a second language, for example French. The semantic content that is of interest is implicit in both representations but the details of the two languages mean that there are many different features that characterise the two views.

In such cases we can make use of a related technique known as *canonical correlation analysis* (CCA) to identify a common subspace that captures directions common to both representations. We can expect that this subspace representation retains the semantic content but with the features that are not common to both languages removed. Hence, the signal-to-noise ratio will be higher. This can also be seen as low level data fusion, since the two views are merged to create a common view. It should be emphasised, however, that this approach is only applicable in the case where we believe that both views contain all of the relevant information. Fusing complementary data sources requires different techniques.

Examples of paired data sets in the context of traffic information systems could range from two cameras observing the same section of road, through information about traffic movement coming from different sensors, to results of higher level inferences about the same phenomenon.

Such dimensionality reduction techniques, where a limited set of factors explaining the data is sought, are commonly referred to as *factor analysis* techniques. They are based on a model built for the data, which models the effects of the different independent factors. A related model-based technique to identify the influence of a set of variables on the data is called ANOVA (analysis of variance) or MANOVA (multivariate ANOVA). More recently, research in model-based multivariate statistics in several research communities has come together in a unified approach, under the heading of *probabilistic graphical models*.

### 2.2.3 Probabilistic (graphical) models

It makes sense to deal with (noisy) data by explicitly taking the probabilistic nature of the data into account, hence modelling the interactions between the variables as well as the noise. A large class of such probabilistic approaches to pattern analysis can be grouped under the name of probabilistic graphical models. These are based on a probabilistic model for the data, built from first principles or learned from the data itself.

The availability of such a model allows one to draw statistically well-founded and often non-trivial conclusions on probabilities of events, potentially conditioned on another (set of) events. Stated in graphical models terminology, a probabilistic graphical model allows one to make *inferences*.

While we can describe PCA, CCA, factor analysis and (M)ANOVA in terms of graphical models, this covers a much wider class of methods and makes it possible to deal with much richer data. Probabilistic graphical models not only have their merits as a unifying concept that helps to understand existing techniques, its recent popularity is also built on recent successes in bioinformatics, coding theory, web mining, among others.

An important example of a graphical model is the hidden Markov model (HMM). This is a probabilistic model appropriate for analysis of strings and sequences, notably for the analysis of genome sequences. The speech recognition community has carried a lot of work related to HMMs which form the basis of common speech-recognition systems.

We believe that such systems will be important in IIS, notably to enable safe communication between a driver of a car and their on-board instruments. Nevertheless, computational problems make some types of graphical models impractical for inference. There is often a trade off to be made between the accuracy of a model and its computational tractability. The exploration of computational frontiers in this domain represents an active line of research.

#### 2.2.4 Classical data mining

Data mining algorithms are designed to find interesting patterns in large databases, such as transaction databases in supermarkets which store the items bought by different customers, in the world wide web and in bioinformatics databases. The archetypical example is the task of finding sets of items that customers often buy together. This task is called *frequent itemset mining*, and it can be solved relatively efficiently using algorithms such as Apriori or Eclat. Knowledge of these frequent itemsets might suggest the development of specific promotional actions, or of a new store layout. In many cases more useful are *association rules*, which can be mined in a similar way as, and based on, frequent itemsets. More sophisticated algorithms can retrieve so-called *sequential patterns*, that keep track of which transactions are carried out by which customer, and infer from this sequential association rules.

Such classical data mining techniques can be of use in IIS in numerous ways, from the design of promotional campaigns for public and responsible transportation, prediction of traffic jams, which may be associated to other events in the network, and more.

While there have been enormous algorithmic achievements in the past two decades, in this domain, the statistical aspects are less well developed. However, a unified view of pattern analysis, integrating the principles of classical data mining with statistically more well-founded approaches, as can be found in machine learning, seems to be a very promising current research direction.

#### 2.2.5 Text mining and string processing

Text mining is a highly active research domain due to the massive amount of information to be mined on the web. Similarly, research on string processing algorithms has gained increasing attention in the past decade due to the present day capability to acquire genome wide sequence information at limited cost. Wherever there is textual information or symbolic sequence information in an IIS, text mining and string processing algorithms may have an important role to play.

Dealing with text can easily be reduced to dealing with vectors, which we discussed in an earlier section, by using the 'bag of words' representation of the text. While this approach is widely used and often effective, it ignores the word ordering, which is a prohibitive drawback in many other cases. The bag of words representation can be adapted to take semantic relations between different words into account. The formal grammars, and their stochastic equivalents given the more expressive representations of text. Such formal grammars were studied intensively in the second half of the 20<sup>th</sup> century, with mixed success.

More recently, suffix tree representations for text and more generally for strings have become popular, as fast algorithms can be built based on them. The most basic problem that can be addressed using suffix trees is the search for the longest common substring of a given set of strings, or for the longest frequent substring in a given long string (frequent to be specified), but more complicated questions can be addressed as well. As we pointed out above in the graphical models discussion, an alternative approach to suffix tree algorithms is based on hidden Markov models.

#### 2.2.6 Graph mining

On the logistic level, the road network represents a graph, together with the public transportation infrastructure. Shortest-path algorithms are commonly used in modern GPS systems to route the user to their destination in the shortest time or following the shortest distance. Besides their use in routing, graph algorithms make it possible to identify bottlenecks in the traffic network, important hubs and vulnerable points.

Besides the road network, graph algorithms make it possible to study social or professional interactions between the users of the IIS, which may of interest in the study of traffic problems related to specific public events, bank holidays, etc.

### 2.2.7 Artificial intelligence

Artificial intelligence is a vague term, we use it here in a somewhat reduced sense, as a set of techniques to incorporate human knowledge into an automatic system, hence simulating intelligent behaviour. There have been notable successes using *expert systems* e.g. in medicine and in geology, where it turned out to be possible to build systems that can diagnose certain diseases more accurately than trained physicians, or to identify interesting geographical locations for mining.

Expert systems use a knowledge base in which expert information is gathered, expressed in the form of rules. A *rule based* inference engine is then used for the reasoning with the information stored in the knowledge base.

Such rule based expert systems are likely to become an essential building block of recommender systems guiding users of the IIS from one place to another, taking into account various user constraints and preferences, as well as the conditions of the traffic network itself.

### 2.2.8 Data fusion

A name often used for pattern analysis algorithms that deal with different kinds of information simultaneously is data fusion. We have already discussed different approaches to data fusion. By means of the so-called kernel trick, certain machine learning algorithms as well as algorithms from multivariate statistics can deal with heterogeneous information sources, within a general framework often referred to as *kernel methods*. Probabilistic graphical models allow one to deal with various types of data in a natural way, using simple probabilistic reasoning. Furthermore, expert systems have been designed to combine confidences of certain rules in a reliable way such that interesting conclusions can be reached. For that reason, it is somewhat artificial to talk about the field of 'data fusion', as it comprises contributions from a large variety of methodologically very different fields. Data fusion is, however, interesting as a concept, emphasising a specific prevalent need in each of these domains.

Ideas from data fusion are likely to be useful in the important systems engineering aspect of the IIS. Due to the scale of the IIS, it is unavoidable that it is engineered by several different manufacturers. We do believe that distributed information processing systems should be restricted as much as possible to deal with situations that require an immediate response only, since central processing of large scale data becomes more and more feasible in real time. This is possible thanks to the advent of fast wireless communication technologies as well as the availability of high performance computing, and algorithmic techniques to analyse and exploit useful patterns in the data. However, systems as large as the IIS cannot practically be engineered as one monolithic system: they are naturally built incrementally in a modular way. For that reason, communication protocols should be established in an early stage, probably by government itself.

As the abovementioned approaches to data fusion are mostly confined to rather small scale data analysis tasks, future developments in the IIS would greatly benefit from more theoretical and applied research in data fusion and inference on large scale networks consisting of many different data sources. In particular, a deeper insight is needed concerning generalized message passing algorithms that allow the creation of emergent intelligent behaviour based on interactions between (often) locally computing entities, each alone having limited intelligence. At this moment, often ad hoc techniques are used to achieve such emergent intelligent behaviour, but the scale of the IIS calls for a clear and well-designed framework to achieve these goals.

Again we should emphasise that creating such a framework would have several advantages:

- it could clearly define a framework into which companies could design their own modules, hence ensuring compatibility and encouraging distributed development and open competition;
- it would break away from the dangers of a monolithic design and the resulting complexity challenges that have undermined a number of publicly funded software projects;
- it would provide a framework through which government could seek to influence development with appropriate incentives and calls.

## 2.3 A common denominator: pattern analysis

Two aspects always return in each of these data analysis techniques: the statistical problem of determining the significance of a surprising pattern in pattern discovery, and stability of a recurring pattern in modelling, and the algorithmic side of pattern discovery and modelling. The domain of pattern analysis is concerned with identifying fundamental common principles regarding these two aspects, in each of the abovementioned domains. The drive to integrate these different domains is very recent, and we believe it will be fruitful for our understanding of the individual disciplines, as well as in the design of large integrated pattern analysis systems that are able to integrate all sources of information available in an efficient and statistically well-founded way.

## 3 Data floods in intelligent infrastructure systems

The available data sources relevant to IIS and transportation can be divided into three levels; those for global infrastructure awareness, local infrastructure awareness, and for awareness of the vehicle or person in transport.

### 3.1 Global infrastructure awareness

Static information about the IIS is available in central databases, containing information such as time tables for public transportation, and relevant information about the users of the network.

Dynamic information concerning the status of the global infrastructure can nowadays be gathered in numerous ways, such as using satellite images, sensor networks that process information locally and send interesting information to higher levels for further processing.

### 3.2 Local infrastructure awareness

There are several motivations for using a strongly localised aspect of the IIS, among others to deal with the individual needs of the users, and to increase responsiveness in critical situations. To achieve the first goal, the network can gather information using radio-frequency identification (RFID), allowing the network to monitor who uses or requests the use of a specific service provided by the IIS. Such information can be used for road pricing, access control and more. Furthermore, as mentioned above, sensor networks monitor local traffic conditions by means of speed and flow measurement systems, such as traffic loops and cameras, and weather sensors.

To achieve rapid action in the case of an accident, reliable accident detection and response systems should be implemented at a local level. The most economical solution here is most probably to put them in the transportation vehicles, sending out a radio signal whenever they are involved in an accident and warning vehicles upstream on the road. To achieve this, vehicles should communicate by dynamically forming an ad hoc network, where each vehicle is directly linked to its nearest neighbours. An alternative scheme would be that the vehicle involved in an accident communicates with sensors along the road, which subsequently propagate the information to other users of the IIS who could be affected by the accident.

Looking ahead to a 10-20 year horizon, there is a further advantage to enabling vehicles to form ad-hoc networks. We believe that as the need increases for further energy efficiency and congestion management, we can anticipate the viability and public acceptability of cars becoming largely self-driven automatically forming close convoys to reduce air drag and congestion. The ability to form ad-hoc communication links between cars forming such convoys would be an essential pre-condition for the roll-out of this technology.

### 3.3 Transportation system and user awareness

A third level of information gathering and dispatching takes place at the level of the user and their vehicle itself. It consists of the GPS, map and trajectory information supplied to the user. Furthermore, we imagine a user profile is kept – for example, stored locally in the GPS, or on a central server offering a service to the user – maintaining user preferences and needs. Lastly, the vehicle and the user may be monitored using sensors in the vehicle, to ensure optimal safety conditions.

## 4 Applications

It is reasonable to assume that within a few years time, the majority of IIS users will have access to a GPS or a similar computing device, such as a PDA, with locally stored information, such as a map for GPS. Such a device, or perhaps set of devices, will furthermore be capable of storing a user profile, containing information about the user's preferences and needs. Additionally, we assume that all users of the IIS will have access to mobile phones or other ways of communication. Given these mild assumptions – they are already fulfilled to some extent – we will discuss a few potential applications where data mining, data fusion and information management play an important role.

### 4.1 Intelligent traffic management

There are several solutions that could deal with the increasing problem of traffic congestions, road accidents and the resulting unreliability of travel. One is to stimulate users to make use of the more time and space-effective forms of public transportation. Another is to make a more optimal use of the resource that is the road network.

In any case, it is clear that it is in the user's interest to enhance the efficiency, as they would benefit from this themselves. For this reason, it is vital simply to inform the IIS's user of the options they have to reach their destination: by itself this probably represents a large part of the solution. However, an important prerequisite here is that the global optimum for the entire traffic network is the same as the local optimum for the individual user. Since we cannot expect users to take personally sub-optimal decisions for the general greater good, this is a crucial requirement, and below we will discuss a way in which this can be ensured.

#### 4.1.1 State of the art

GPS in cars, route planners on the web, time tables for public transportation in various forms are already easily accessible today from the WWW. Furthermore, radio and television inform us about the weather and traffic conditions. However, integrated resources are mostly lacking. As a consequence users do not fully exploit the information available. Furthermore, it takes too much time, and once on the road it is difficult to reschedule an itinerary. Not only are data sources not integrated, some are simply lacking – or remotely inaccessible, or at least inaccessible while travelling – such as information on the availability of parking facilities, train or bus delays, etc. Furthermore, whereas systems that report the current state of the road network have already been developed and deployed (such as [www.getmethere.co.uk](http://www.getmethere.co.uk)), there is a void when it comes to making and distributing *predictive* information, concerning the traffic conditions in the near future.

It is here that data mining and machine learning could play an important role. The absence of this information is probably due partly to the inability of the user to deal with all of it anyway. In such cases, it is the task of a computer interface to filter and fuse what is interesting and to present it to the user in an interactive and accessible way.

Ideally, the user should be able to enter their desired destination and arrival time into a local computing device (LCD), when then proposes one or several itineraries matching the traveller's criteria, potentially including itineraries that involve a mix of private and public transport. To accomplish this task, the LCD should have full access to all the available information in the complete IIS. For reasons of bandwidth restrictions of communication systems, privacy issues, and computational limitations of the LCD operated by the user, such an approach would be infeasible at this moment or in the near

future. Therefore, one should make a distinction between information locally available in the LCD, information available to a central computing device (CCD) that the LCD can query, but which remains partly invisible to the LCD for privacy and security reasons, and information dispersed throughout the IIS.

#### 4.1.2 Near future

We will explain here a system that most likely can be met technologically at this moment from a data mining, data fusion and information management point of view, and could likely be implemented within five to 10 years. It aims to provide each user with information of use to reach their goals as efficiently and reliably as possible, while keeping the interest of the global network in mind.

The LCD may be a somewhat extended version of the GPS, in that it contains some intelligence, and mostly it is capable of communicating either by itself or through a connection with a mobile phone. It is capable of storing user preferences and requirements, accessible after a user login, which is likely to be verified by using biometrics to identify the user. Through its communication interface, it can communicate with CCDs, to gather information such as that stored in time tables and traffic conditions, and can capture information dispersed over the traffic network in its neighbourhood, distributed over an ad hoc network.

The CCDs keep track of the global state of the network, and perform the computational tasks needed to manage the network and predict traffic loads in the near future. By centralising relevant information, they can provide the LCDs with information about the time and cost needed to follow a certain route. They obtain information from several sources.

Sensor networks along the IIS measure traffic flow and road conditions, and perform local low-level pattern analysis, in order to decide what is interesting enough to broadcast to the CCD. Such sensor networks might well be quicker and more accurate than satellites in detecting and revealing congestion and accidents. Also, users themselves, by querying the system, reveal information about their intentions and hence provide predictive information about traffic loads in different locations. Lastly, the CCD can rely on external databases and parallel systems containing information such as time tables, car park availability and more. The amount of dispersed information should probably be kept as small as possible, as an integrated approach is generally preferable whenever this is possible. However, to deal with acute local issues, such as traffic accidents, dynamic ad hoc networks will remain indispensable, as they will for some of the longer horizon tasks mentioned above.

In such a system, it should be possible for the CCD to provide the LCDs querying it with reliable estimates of travel times along several routes, from door to door, including parking, waiting and walking times. The LCDs may then inform the user of all reasonable options, taking into account preferences stored in a user profile.

To prevent user opportunism adversely affecting the system, the *game theoretic optimum* for each of the users should be equal to the *global optimum* of the system. For example, the system should avoid approaches where users are all directed via a small road that is not (yet) congested. This can be achieved using the predictive power built into the CCDs, based on advanced machine learning techniques. For example, CCDs can keep track of how many queries the LCDs have made for different roads that it is managing, and based on this it can adapt its travel time estimates to return to new queries by other LCDs. Effectively, in such an approach, by querying the system users get a virtual ticket for a road. Every additional ticket sold will then be 'sold' at a higher time cost. A similar principle can ensure a parking spot at the point of destination.

A billing system should be set up to make a reservation for a parking spot. Authentication of the user and safety of the communication channel then becomes important. Note however that java enabled mobile phones will soon be capable of more secure wireless communications. Upon arrival, an RFID system or licence plate reader would grant permission to the parking spot. Similar to this, while not a requirement, one may introduce a financial cost associated with the use of a certain road, of which the user must be informed prior to choosing their route.

### 4.1.3 Speculations

Pattern analysis methods make it possible to mine large data sets in the search for interesting patterns that make it possible to reliably predict traffic conditions in the future, taking special events, such as soccer games, festivals, holidays and so on, into account, as well as weather forecasts, and so on. The data sets to achieve this can be constructed based on past satellite images and measurements by traffic loops, and databases on weather conditions and event calendar information. The resulting models are likely to prove extremely useful in a definite reduction of traffic problems through the dissemination of predictive information to the users, or to their LCDs, who can then make informed decisions based on accurate predictions of traffic loads.

## 4.2 Intelligent resource allocation

An important problem in traffic networks is the inefficient use of available resources. The marginal cost associated with the use of a car, once bought, is too small for users to refrain from using it for short distances with few passengers. At the same time, public transportation is not exploited to its full extent, because of its lack of flexibility and accessibility. A more rational use of resources should make it possible to reduce the number of cars on the road, and increase the efficiency of public transportation.

To enable the efficient use of the available resources, there should be systems that allow people to find out what modes of transportation are available in a short time frame, as we explained in the previous section. Here we want to go a step further, and propose to break down the distinction between public and private transportation to a large extent. The goal we wish to reach in this way is an increased availability and flexibility of conventional public transportation, as well as a well organised sharing system of private transportation.

### 4.2.1 State of the art

Commercial computer packages for car sharing already exist, such as <http://www.carshare.uk.com/>. However, they are limited to car sharing within one or a few companies. Furthermore, they match users based on limited user information – journey times and geographic locations, and a limited number of other constraints – while disregarding other relevant information that may make the journey more or less enjoyable. For example, safety concerns are often a major issue preventing people from use such services. Online services can be used, such as <http://www.car-pool.co.uk> and <http://www.taxistop.be> in Belgium. However, the availability of lifts is usually rather low, and the critical mass has not yet been reached.

### 4.2.2 Near future

There are a few challenges to meet: for each driver and for each passenger there should be a user profile, containing preferences (smoking or non-smoking, cost charged or willing to pay for joining a ride, relevant aspects of personality, interests ...), feedback from previous passengers and drivers, besides the objective specificities concerning the journeys themselves. Matching user profiles based on various information types is an interesting task that can be solved reliably using established pattern analysis techniques.

Nowadays, queries for a ride have to be made online or into a computer system of the company: drivers have to make seat availability known well in advance of the journey. This is clearly too inflexible to enable a wide applicability of the system. Instead, it should be possible to make queries using a (mobile) phone and automatic speech recognition to find a match. After a match has been found, which can happen in real time, the caller is offered one or a few options. They then express their agreement, ideally in an authenticated way. A text message on the user's mobile phone then automatically reports when the driver arrives, or prior to their arrival, as reported by their GPS.

Probably in the first stage, drivers should enter their information in a computer system as is now the case. However, to increase the number of participants, it would be desirable if the on-board GPS alerted the driver that one of their routes would lend itself well to car sharing.

Also professional taxi drivers can be subscribed to the system, as a highly reliable and available resource, while probably more expensive.

Similar approaches could be useful to fully exploit lorry space. Today, all too often, lorries return empty after having made a delivery, though systems are being deployed to tackle this problem.

#### 4.2.3 Speculations

In a next stage, the GPS of drivers who enabled the option to do so could send information about the driver's probable trajectory to the CCD, after which it enters the pool of potential cars to give people a ride along the way. Such a system could provide highly flexible and fast car-sharing opportunities, through communication between the mobile phone of the person seeking a ride, a CCD performing the matching, and the GPS of the driver. Obviously, the more limited the freedom of the driver, the easier it will be to carry out the matching. Hence, the question remains to what extent drivers will be ready to give up the freedom they have today in deciding which route they follow, but there will of course be a significant financial incentive. We come back to this in the next example.

### 4.3 Intelligent traffic flow control

Besides an optimal distribution of the traffic on the global network (sending users the right way in order to minimise travel time and cost), some form of local optimisation is necessary in order to improve the safety and speed of a journey.

#### 4.3.1 State of the art

There are already several ways for automatic traffic flow control. *Green waves* are synchronised traffic lights, such that at least one of both driving directions is guaranteed a green light when driving with a certain speed. Interestingly, currently pre-timed traffic lights often outperform adaptive traffic lights in optimising the flow.

Another approach used mostly on highways is the system where *computer controlled speed limit signs* vary according to the traffic intensity, to optimise the total flow. A way to enforce this speed is by *pace cars* driven by police officers.

#### 4.3.2 Near future

While adaptive systems, such as green waves and computer controlled speed limit signs, currently respond to the current situation, there is a need for proactive systems that anticipate reliable predictions about future traffic loads. If on a highway the traffic load is predicted to increase soon, the maximum speed should be reduced before the problem manifests itself. The absence of a reliable prediction of future traffic loads may well be by pre-timed traffic lights are better than adaptive traffic signals. Predicting future traffic loads is a non-trivial problem, but one that pattern analysis techniques from data mining and machine learning can tackle.

Besides optimising the traffic flow under normal conditions, systems are being developed and investigated to increase travel safety by incident detection and reporting. Since speed is often critical in detecting and reporting an incident, and since an incident should initially be reported only to a rather small environment around it, it makes sense to propagate the relevant information along an ad hoc network formed by the LCDs of different road users. Passing the signal through a CCD would introduce unnecessary delays. The challenge here is to implement dynamically changing ad hoc networks, and to guide the propagation of the information on the incident to users for whom it is relevant. An effective information filtering step in the LCDs is necessary to limit the number of false alarms, based on the actual incident reported, and on information concerning the local infrastructure, the road network, as stored in the LCD. Similar systems are under development to facilitate highway entries and lane merging by warning affected road users.

### 4.3.3 Speculations

In the long run, one can envision a totally coordinated movement of vehicles that are globally guided to their individual destinations, and locally steered by their direct environments. In such a system the user's influence would be limited to a speech driven interaction with their LCD, which then makes the most appropriate decisions for 'their own best', as well as for 'the general good'. Such a system would integrate a traffic management system discussed above, as well as a resource allocation system and most prominently, and probably most challenging, the traffic flow and vehicle control mechanisms based on local ad hoc networks.

## 5 Data mining, data fusion and information management in intelligent infrastructure systems

The traffic management, resource allocation and traffic flow control systems are examples of the need for data mining, data fusion and information management techniques in IIS. We encountered a multitude of pattern analysis tasks related to predictions of traffic loads, matching users with each other, speech recognition systems, recommendation systems for trajectories or for car sharing, shortest path algorithms on graphs, and pattern analysis algorithms to filter interesting information in a large information flow reaching a sensor or network hub. Many of the algorithms needed to solve these problems are already within our capabilities. However, there is much work to be done before we can achieve some of the more ambitious goals of IIS on a large scale.

Probably one of the most vital, and challenging, specific tasks for pattern analysis to tackle, will be to make reliable predictions of traffic load. The availability of accurate traffic-load predictions is a prerequisite for many of the other objectives: it is required for efficient traffic management, finding the fastest itinerary, as well as local traffic flow control, intelligent speed signs, green waves, etc..

## 6 Conclusion

Pattern analysis, with data mining, machine learning and data fusion as its main constituents, represents an indispensable enabling set of techniques for present day and future intelligent infrastructure systems. The amount of data generated in the IIS is ever growing: efficient systems to filter the useful information from the irrelevant are vital for its exploitation.

The number and the diversity of pattern analysis tasks, and the scale of the IIS as a whole, make it inconceivable that just one or a few market players can reliably engineer and implement them. Therefore, a good strategy for a well-organised deployment is probably to encourage industrial participation by a system of bidding for contracts offered by the Government, or using other strategies to facilitate the market entrance of systems developed by industry. The final IIS is likely to be highly modular, such that a crucial ingredient of a strategy to expedite industrial participation is the early establishment of communication protocols and detailed task specifications for the different players of the IIS. We believe that this will flow from a sound framework for integrated reasoning with disparate and distributed information. In this respect, it is important to have further research in the development and understanding of pattern analysis and data fusion techniques that can be implemented on modular and dispersed systems.