The Six Technical Gaps between Intelligent Applications and Real-Time Data Mining: A Critical Review

Simon Fong, Hang Yang
Faculty of Science and Technology, University of Macau, Macua SAR
Email: ccfong@umac.mo, henry.yh@gmail.mo

***Abstract*—Intelligent application is characterized by its ability to make decision autonomously. Some examples are distributed agents that collaboratively complete a complex task, yet each one of them is able to work and reason independently given the dynamic situation where they are in. The underlying think tank is often a collection of core mechanisms that include environment sensing, data capturing, data mining, ETL, knowledge processing, and decision making etc. All these techniques when placed and function together as a whole intelligent system, they will have to fulfill stringent deadlines imposed by the requirements of a real-time system. In the literature many research papers can be found on a wide variety of data mining techniques that enable intelligent applications operating in real-time. This report offers a critical review of the relevant literature, and contributes to the knowledge of identifying their shortcomings, so-called the technical gaps between the real-time requirements of a general intelligent system and the supporting components. A discussion follows on the possibilities of future intelligent applications that are empowered by data stream mining.**

***Index Terms*—Intelligent applications, Real-time data mining, Data stream mining.**

I. Introduction

Intelligent application is a general term that usually describes a software application program or system that possesses the inner capabilities of reasoning, thinking, and understanding, predicting and making decision autonomously based on the information collected from the external world. The term intelligent application (IA) was first conned in 1991 by Steve Weyer of Apple computers. Steve's described his concept of IA was an advanced technology equipped with a transparent user interface that is conversational and adaptive. Such advanced technology will serve as an assistant to human user or as an advisor in the task of complex decision making. One example is a multi-agent system in some collaborative environment where each agent is able to perceive data from the current situation, interact, and most importantly to decide what the rational actions are to be taken from the information gathered along the way. Other examples include but not limited to expert system, information recommender, and virtual shop assistant, etc.

What these entire IA applications share in common is that there are some elements of Knowledge discovery process (KDP) [1] either embedded as an internal part of the whole system or individually as a supporting role. KDP aims at seeking new knowledge in a targeted domain from data. As reported in [2] KDP is not just about querying from database but it sets out to explore meaningful patterns from a data repository. Data mining is the core of KDP that supports intelligent applications like an engine to a car in an analogy.

In the history of information technology industry, management information system (MIS) evolved with different names and an increasing extent of adoption of KDP for aiding making sound decisions for the executives. The research papers [3,4] described these systems namely MIS, Decision Support System (DSS), Executive System (ES), Executive Information System (EIS) and Business Intelligence (BI) systems; and they do grow in complexity level across different years. An extra dimension, as it is shown in Figure 1 is the reliance on KDP on which these systems grew upon. For instance the early MIS supported executive with database query and business reporting over their business. The techniques applied were largely database queries and statistics summation. Data mining gradually entered the evolution in supporting DSS, ES, EIS and BI. The dependence of KDP plus advances in data mining algorithms grew tremendously over the years for these systems.

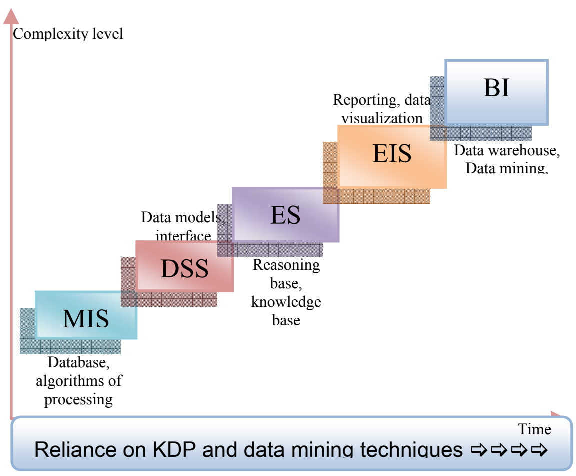

Figure 1. Development of Management Information Systems [3].

doi:10.4304/jetwi.3.2.63-73

Those management information systems such as BI and EIS have been interchangeably named as IAs in general because they behave with intelligent decision making that relies less on human manipulations but more with artificial intelligence. Intelligent decision making and automated operations alone may not be satisfying the business users nowadays. BI now which is staying at the leading edge of the information systems development is constantly under increasing demand of delivering the right information to the right user at the right place, and within the right period of time. The future generation of BI, as forecasted by researchers [5] is to leverage on real-time technology for proactively offering high quality information (or intelligence) that will fit into user's exact needs. Furthermore a business survey [6] has identified the greatest bottle-neck in business field, as "Lack of tools for doing real-time processing – with the biggest share of votes – 35%", compared to "immature technology 28%" and "Performance and scalability 24%". We can surmise that information and intelligence not only must come in the right quality and quantity, but they are demanded in real-time from the user.

A recent jargon called Real-time BI [7] has been used popularly in the industry. Real-time BI certainly has its merits in many aspects of a business; as shown in Figure 2a, the value of the action taken based on the information declines in time. This is due to the opportunity cost involved that gradually escalates and fades out the benefit anticipated by the action and the information.

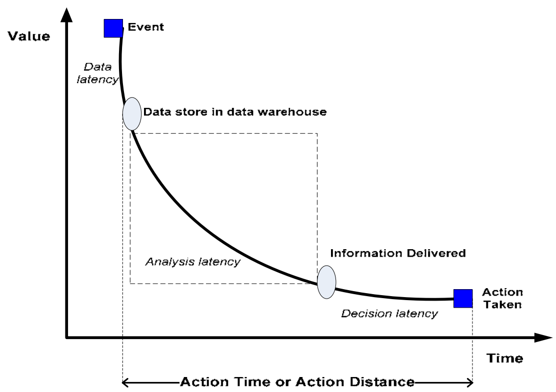

Figure 2a. The time and value in action distance.

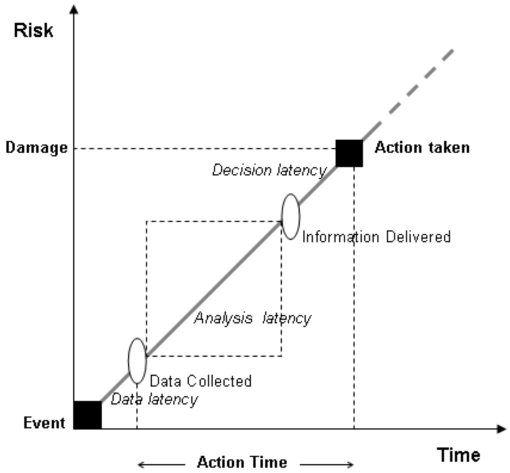

Figure 2b. The time and risk in action distance.

In Figure 2b, the relation between risk and latency is shown in emergency cases. The later the remedy action was taken, the greater the damage. For an example of radiation leak from the nuclear power plants at Fukoshima in March 2011, it was criticized that instant decision was not made on time to rectify the problem that led to subsequent disaster of the reactor melt-down and massive radiation pollution to the environment. In the emergency case in Figure 2b, timely action is even more crucial than the business case in Figure 2a, as comparing life-and-death situations to business profits and loss.

Both cases show a typical procedure of KDP where latencies do exist in several points along the timeline. All we want is to cut them as short as possible. In a nutshell the success and efficacy of Real-time BI depends very much on the real-time capabilities of the KDP in which data mining process is the vital force, and the real-time performance of each of the components in the IA system. Practically there is no magic that converts collected raw data into usable intelligence hence action instantly. Spaces or gaps in the sense of latency are inevitable in any IA system. What it does matter would be the usefulness of the technologies we have on hand that support and operate in different parts of the IA system, and how much they can perform at a close-to-zero latency.

The objective of this paper is to identify what and where these technical gaps are from the technological perspective in an IA system. Readers would be aware of these gaps and ponder on (hopefully) eliminating them when it comes to designing an IA system, for achieving a truly Real-time BI. The paper is structured as follow. In Section 2 we probed into a data mining process that is the basis of KDP for locating the sources of latency. In particular, we compared traditional data mining and data stream mining techniques with respect to their supports of real-time operations. After we reviewed the core data mining process, in Section 3 we zoomed out our view to a system level and discussed the six technical gaps between an ideal IA and real-time data mining techniques; importantly we looked at different technologies for minimizing the latency in each gap. A brief discussion on how IAs would be developed for various industries, especially on the advantages that can be reaped if the technical gaps can be bridged by data stream mining in the respective IAs. Lastly a conclusion is made at the end.

II. Theoretical Foundations of Real-Time Data Mining

In a data mining process, if the mined result were to be delivered as useful "real-time" information, the mining speed must be in synchronization or ahead of the activity operation. Simply the basic real-time requirement can be written as that the given data source generation rate (*DR*) must not be faster than the rate of mining process (*MR*), $DR \leq MR$. The mining time has to be quick in order to catch up with the incoming fresh data. There is two ways to achieve real-time intelligence discovery:

(1) The traditional data mining methods: The mining process consists of reading in the dataset, preprocessing

the data, training (or re-training if it is not the first time) up the model, validating the model, and testing the newly arrived data for generating an output. Because of the real-time requirement, there should be a time constraint or deadline that the mining process would have to meet. That is, from the moment when an event happened, the mining processing would have to generate an output before the time expires. The value of the information output declines as time goes by, as in the relation shown in Figure 2a. In an emergency case, the delay may lead to severe damage or total loss if the wait is too long, as in Figure 2b. The total latency here for traditional data mining methods measures from the time of updating the historical dataset with the fresh data, to training and validating the whole model again. This latency must be no longer than the real-time constraint for delivering the data mining result, plus of course the condition *DR≤MR*.

(2) The stream mining method with relatively low-latency: This method takes a revolutionary approach in building up model. Without the need of reading in the whole dataset which is impossible in the scenario of flowing stream, the model is being built incrementally from scratch as new data come in. Each time a piece of new data arrives it will be used to test against the model for a prediction output, and the same data may be used to progressively train up the model based on the statistics counts stored in the model. For example if you toss a coin enough times based on how many times you had a head or a tail, you can tell the probability of having a head or tail in your next toss. The outcome of every toss in turn, contributes to building the decision model. For this type of incremental learning the mining process does not need to keep re-building the whole model whenever the dataset gets updated with fresh data. The latency is therefore much shorter. The real-time rule *DR≤MR* still applies.

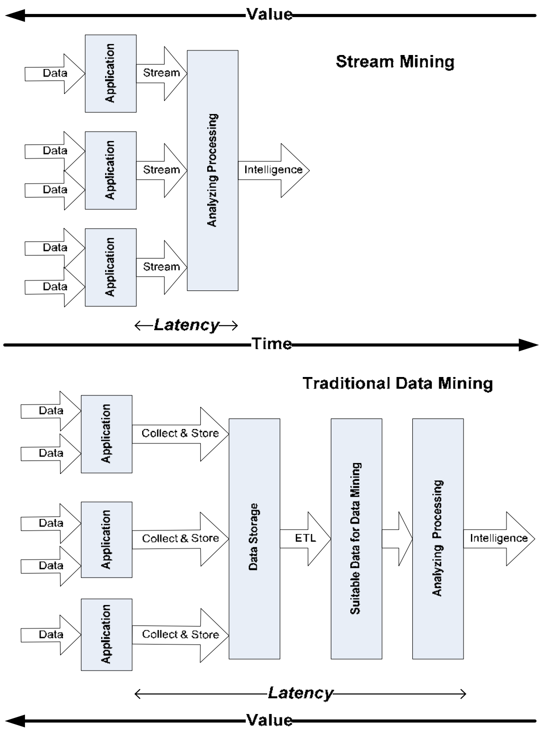

Figure 3. Two ways of real-time data mining.

A. Traditional Data Mining for Real-time Task

The latency barrier would need to be overcome when traditional data mining is used for a real-time task. The heavy latency could be streamlined in traditional data mining process by two types of solutions described in [8]: data-based and task-based.

The concept of data-based solutions is to examine only a subset of the whole dataset or to transform the data vertically or horizontally to an approximate smaller size data representation. That would help speeding up the mining process by working on a smaller training dataset. Data-based techniques have two approaches, dataset summarizing and sub-data stream analyzing. Dataset summarizing include Sampling [9], Load shedding [10, 11], and Sketching [12,13]; Sub-data stream analyzing exploits Synopsis [12,14] and Aggregation [15,16,17].

In task-based solutions, techniques from computational theory are adopted to reduce computational time and space in the hope of simplifying the mining algorithm. For the instance of mining data streams task-based techniques would try to extract meaningful structures from the models and patterns in non-stopping information streams. Some algorithms such as Approximation algorithms [18], Algorithm output granularity [19,20,21] and Sliding window [22] are invented to address the data stream computational challenges.

B. Data Stream Mining

A new generation of data mining methodology [8] has been proposed to address the challenges of processing potentially infinite volume, and real-time moving data stream. As the input data are flowing in in real-time, the output result is produced almost instantly in real-time from the freshly updated data.

As shown in Figure 4, the Data storage has become somewhat optional or even redundant in the case of data stream mining. The model changes dynamically at the same pace of data coming in; output intelligence is retrieved always from the most updated model with a very low latency. This merit is mainly inherited from incremental learning theories. Based on these theories, researchers [23,24] formulated a number of light-weight mining algorithms, which only need to perform one pass over the data. The mechanism of the process is largely based on sliding-window [25,26,27,28,29] that was used to combine with incremental learning to facilitate data stream mining. A high-level guidance to look for stream processing solutions can be found in [30] that specifies the requirements comparable to relational DBMS.

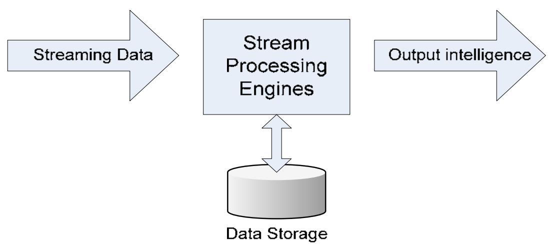

Figure 4. Stream data processing.

III. The Gaps Between Intelligent Application and Real-Time Data Mining

Is it a myth that ideal IAs do exist? An ideal IA knows all, senses all, thinks quickly, and be able to make sound decision always at the right time without fail and latency? Many I.T. vendors claimed so in their marketing propaganda. Practically and actually, we do know in reality any self-proclaimed almighty system would have limitations in one place or another. The authors however believe in I.T. system development, performance gaps certainly exist within different components of the system, and they are inevitable. It is a matter of how much or how smart the underlying technology and mechanisms can bring the gaps close. Performance gap is perceived as a distance between what the ideal or expected scenario is, and how far it falls off from there. It can be measured into some quantifiable scale for example, when an exceptionally good opportunity for purchasing arises in the market how many seconds would elapse before the management commits to procurement decision as recommended by the IA software. The performance gaps are supposed to be directly proportional to the technical gaps in the context of discussion here. The technical gap refers to the gap between the performance expectation of IA software and the activities that are actually supported by technology.

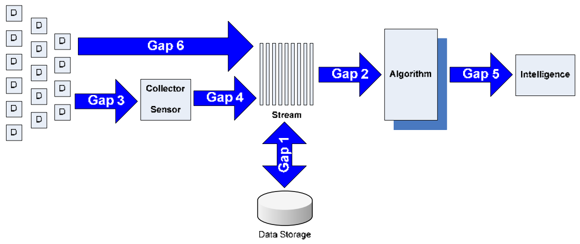

Figure 4. Six gaps that exist in an intelligent applications.

Gap 1: A reliable and efficient way to communicate with data storage

A reliable data repository is mandatory to all information systems. For IA that operates in real-time environment it is a great challenge to selectively store relevant information that is dynamic and ever-changing and to return the right piece of information upon data-query under time-constraints. Even in data stream mining environment data streams that are used mostly for one pass training-and-testing, significant events which are represented in some parts of the data streams should be stored up for future references. This may be analogous to a biological brain where both short-term memories and long-term memories would have to be stored, and be recalled whenever it is needed. Real-time BI, likewise, has two parts; one is called Operational Intelligence that collects and analyses daily or minute activities for detecting any special situation in real-time; the other part is business intelligence that concerns about how the overall business is moving along in a broader time scale.

Efficient mechanisms for recording and retrieving from/to a database/warehouse are required to support Real-time BI (that includes data mining) as well as to handle the input data in the form of data streams. SAP proposed an stream processing support for real-time BI, called MaxStream. Functional heterogeneity is a major issue for the MaxSream architecture due to a lack of standard and agreement on core feature and semantics. Stream-oriented query language (StreamSQL) is an essential all SQL extensions that incorporate a notion of a window on a stream as a way to convert an infinite stream into a finite relation in order to apply relational operators. Stream SQL with window techniques is another studied target to settle the stream communication problem. The Oracle model is based on CQL [31], in which the tuples have timestamps and have no ordering between tuples and identical timestamps. This model is a time-driven model, which uses the history input stream as the timestamps evaluation to compute the value of window. StreamBased model based on Aurora [32,33] assigns an internal rank to tuples based on their arrival order to the system and ensures that an input tuple, as well as all tuples derived from its processing as far as possible before another input tuple with a higher rank. This model is a tuple-driven model, in which each logical relation has a value as of each tuple. Because these two models have distinct basis, there are some implemented differences either time-basis or tuple-basis [34].

Gap 2: An efficient stream mining method in a distributed architecture

Intelligence discovery from business in real-time may be implemented as a distributed system, which is like a sensor network [35]. The heart of this KDP is a collection of data mining algorithms that are built into this distributed architecture. The second gap here is referred to a number of technical challenges; firstly the data mining algorithm must be fast enough to complete the mining and support real-time decision making which has already been discussed in Section 2; secondly the right data streams must be properly chosen that contribute to the targeted decision making provided that multiple data sources could be available at the same time in a distributed environment. Lastly the data mining models must be well designed because in any organization there possibly could be several types of decision to be made, hence it is not uncommon to have more than one data mining model for handling all types of decision making. Furthermore the right type of data mining algorithm and the right set of parameters must be correctly chosen for each data mining model for each type of decision. There in general is no silver bullet for this particular process because of a wide range of possibilities in the choices, and it requires a lot of human fine-tunings. Its success largely depends on the expertise and experiences of the analyst who designed the data mining models and the choice of algorithms. The data mining process in an ideal IA would have a set of correctly chosen data mining models with optimally customized parameters that match well with the target decision making requirements, and these models are supported by the right data sources that are always available. In reality at best vendors provide a user with templates, each comes with sample models and parameters for each industry. Customization then follows.

Gap 3: An well-designed collector to capture stream data in real-time

This is the direct interface between the data to be collected from the external world and the stream data ready to be mined by the IA. This interface is called a 'collector' whose job is to collect the suitable data at the information level, and to format the data into a consistent format for the miner to use at the technical level.

At this gap there are two major challenges, each at the information and technical level respectively. At the information level, the collector serves the role of information retrieval. Selectively the collector needs to find the data on a continuous basis from the external world for feeding the miner so that decision could be made based on not only the most updated and very relevant information. Most software vendors nowadays are treating this process manual; a setting function requires the users typing in a list of URL to the websites from which information could be constantly monitored. Exactly which websites are best to be monitored, and how many would be optimal? IA could only be as intelligent as what sources of the input data are to be specified. Ultimately this task rests on the user's hands. It is difficult to estimate the right amount of information such that the IA wouldn't be overloaded with too much information or starved from too little.

For the data formatting problem, it is known that a data mining process be it a traditional data mining or data stream mining, takes only one specific data format at a time for training the model. This specific format would be predefined and rarely would it be changed along the way. Typical a dataset format is defined by what the attributes are and their data types. The rigid input data format requirements at the technical level contradict with the dynamic information selection at the information level for this gap. For example we have experienced changes in cyber platforms from Web 1.0 one-way information website, to Web 2.0 socially oriented blogs and networks, Web 2.5, 3.0 and so on that have semantics and other extended features. This change impacts both the way we collect information and the format in which the information are represented. The collector must keep up with these changes.

For another example of a sensor network which is being connected to the collector to intake data, formatting the input datasets is now a matter of identifying what type of sensors they are and how many of them are well functioning. Each stream of data from a sensor may correspond to an attribute in the dataset. Sensors monitor different measurements such as temperature and humidity, the classes of normal and otherwise that are represented by these numeric values would have to be calibrated in advance [36]. Only when the data formatting complication is kept in check and the selection of the data sources is optimized IA could perform to its best with the right data for making sound decision. In reality this gap is still tainted with a lot of manual intervention and expert know-hows required for setting up the right data sources. This is on par with a known problem in KDP called preprocessing.

Gap 4: An efficient way to handle imperfect data stream

This Gap is on the fact that incoming data streams are far from perfect even though a well-designed collector is in place in Gap 3. Due to the distributed nature of computer networks, the delivery of the data streams may experience traffic fluctuation and data corruption at the times of congestion and system overloads. The problem may be intermittent but parts of the data streams could be impaired with either noise or missing data.

Noise data is proven to induce degradation in performance accuracy and escalate decision tree size in the run-time memory [37]. Several attempts were made to relieve the adverse effects such as using auxiliary predictors to guess what those missing data are in run time [38], and modifying the decision tree building algorithm to control the size of the decision tree in heap memory. This technique moderates on the splitting nodes probability, it works by incrementally guessing the mean values and pruning off the redundant parts of the tree due to noise data [39]. Flexible decision tree (FlexDT) [40], is an epitome approach of modifying the decision tree building algorithm, it incorporates the fuzzy logic functions into the decision tree. As a result, FlexDT has become robust to noise. The other technique called bucket cache [41] is to try smoothing the traffic data fluctuation by replenishing the missing data from the excessive data previously stored in cache.

Plenty of research efforts are still needed in handling imperfect data in data stream mining because traditional techniques for estimating missing data/noise are no more applicable in streaming environment. Data stream mining has to operate in real-time, with reading only one-pass over the data. Preprocessing for the full dataset is not possible.

In most of the research literature data stream mining is assumed to take on perfect set of data streams. This assumption of course is not valid in real-life IA where data are prone to be missing and infested with noise. This Gap may also be viewed as a reliability of the underlying hardware that is tasked to deliver input data to the data mining system. For example a malfunctioned sensor can cause a series of purported noise in the data feeding into the IA unnoticed.

So until the issues of imperfect data are solved for data stream mining, IA is vulnerable to reduction in accuracy and explosion in decision tree size as stated in this Gap.

Gap 5: Turning the output of the data mining process to an useful intelligence

Data mining algorithms come in a wide variety and so do the formats of their output results. The differences of the result formats pose a great hassle in this gap for bridging to the intelligence that can be directly put into use in an IA. IA often requires actionable intelligence that could be just a specific value, an indication/verdict or a subset of a database for further processing or further actions. For instance, we have an IA in an unmanned spy-plane performing aerial reconnaissance over a country. Its radar detected a fast-approaching object of certain size and speed. In order for the IA to decide the next course of action, specific actionable intelligence is needed. It could be a combination of images, reports, wind-speed, latest information of nearby hostile armies and others. This example shows actionable intelligence may not be derived from a singular source, and it could be composite information comprised of knowledge of the most updated situation plus that of the history (so-called experiences). Quite often in an intelligent application, multiple data mining algorithms are used together for different aspects of information. How the individual results generated from all these data mining algorithms can be combined into actionable intelligence? Are there standard formats for representing actionable intelligence? Can the intelligence be interchangeable between different IAs and/or be commonly interpreted in a collaborative group of IAs?

On the other hand, interpretability has been one of the major challenges in data mining. Results from different data mining algorithms come in various formats, below are just some common examples:

Classification → a class label, and may be a propensity;

Prediction → a class label, a propensity or a value

Clustering → groups of data items

Association rule mining → tuples of head and tail

Visual mining → selected images and reports

The results usually would be further interpreted by domain experts, and formatted into some concise management reports by the analysts. Sometimes the refined knowledge which is deemed useful would be stored in a knowledge base as cases. This gap stresses on the lack of automation that can transform and transfer the data mining results into actionable knowledge. Human intervention still remains as the biggest bottleneck. Some vendors offered ad-hoc functions for pre-configuring some connections (called threads in SPSS PASW) between the output of a data mining process and the subsequent action process. The setup needs to be predefined and it is not adaptive to the runtime situations. Unless this gap can be solved, hardly an IA can achieve a fully automated system that can make decision in real-time and can execute responsive actions in real-time.

Gap 6: Ability to adapt to environmental changes

After an IA is built, the original environment based on which the IA was designed may have changed. The environment could be referred to operational environment, political environment, business environment and even climate environment that are characterized different attributes in the dataset as well as the target class labels (in terms of decision trees in data mining). Technically this may be known as concept-drift that means a model that was initially trained to predict something will have its prediction target changed some point in time. Most of the data stream mining has the built-in algorithm that can detect and work with concept change. However if the change is too large in capacity due to environmental factors, it implies the whole data mining process will have to be reset – a model is usually trained according to a fixed format of dataset with certain attributes. When these attributes change in a large scale, the previously trained model will have to be discarded. This may not be an uncommon situation in real-life considering that an IA that was initially programmed to do something may have to slightly change its purpose or focus in a later course. For instance, a robot that was originally designed for dismantling a bomb in a ragged urban environment may be sent for damage inspection at a disaster site of nuclear fallout. The information perceived from the sensing equipment and camera would be different, because of different terrain settings, environmental hazards and mission of the tasks. Every time when this change arises the IA will have to be re-programmed to a large extent. This inadequacy hinders the usability of IA in long-range mission. To bridge the current gap, adaptive data mining models are desired that can update themselves when changes occur in the attributes or in the prediction target.

Layered conceptual views of Intelligent Application

Having discussed the six technical gaps that exist in different parts of an Intelligent Application, for the interest of software engineering the gaps can fit into different layered conceptual views of an IA. From a high to low conceptual level, the gaps ascribe to three main groups as follows:

- Interoperability and Intefacing Level (Gaps 5*, 6)
- Algorithm Level (Gaps 2,4*,5*)
- Data Level (Gaps 1,3,4*)

The mark (*) means the gap may be overlapped between two levels. Gaps 5 and 6 concern about how potentially an IA may interact and interoperate with other IAs if they were not designed to function as merely standalone systems. In particular Gap 5 is about how the output of a data mining process can be converted to some actionable intelligence that could be used by perhaps not a single IA, but a group of IAs. Some interchangeable common format that used to represent the actionable knowledge would be required. This important interface should receive much attention from the research community. Currently we have many data mining techniques and theories flourished; naturally the next step is on providing a useful format for IAs to harvest from the resultant intelligence from these data mining techniques. Intelligence should no longer be restricted to management reports for strategy formulation; but should be used to fuel the advancement of intelligent applications.

The works mentioned in Gaps 2, 4 and 5 are at algorithmic level; research progresses on these areas are anticipated, and they should help smoothing the intra-functioning of an IA. In summary, an ideal IA is one that would have the gaps tighten up across the three conceptual layers. Internally the IA has a smooth mining mechanism that can mine data of different formats, can handle data with impurity and can fully exploit data mining results of all kinds into useful knowledge, with a reliable support of data base management system and collectors at the data level. Externally the IA can possibly collaborate with other IA in real-time for complex tasks.

IV. Intelligent Applications in Different Domains

A. Business

The term Business intelligence (BI) has been defined generally as "a set of mathematical models and analysis methodologies that exploit the available data to generate information and knowledge useful for complex decision making processes" [42]. While the original aim of generating information and knowledge remains the same, recently the role of BI has evolved beyond the traditional batch-processing approach of reporting of analysis over static historical information and ad hoc queries. BI nowadays opts to integrate with event-driven processing and advanced data mining functions for providing real-time information to users. As an intelligent application that can proactively anticipate the needs of a market and recommend to users the best moves in business, real-time BI and its related functionalities are often embedded into the operational processes.

Real-time BI which transcends beyond traditional techniques of archiving data, supporting information access, data mining analysis and reporting, is now shifting towards the convergence of business intelligence and business processes. The corresponding IA of real-time BI therefore must have the capabilities of integrating seamlessly the operational business processes and decision making systems whose information may be disparate and distributed. In order to support real-time data mining and generate actionable intelligence, the batch approach of traditional data mining is replaced with data stream mining that can better react in continuous flow of latest data streams [43]. To achieve this, a unified view of data across the enterprise is needed, and Enterprise Application Integration (EAI) has been used as an enabling technology for this. Along with EAI, Complex Event Processing (CEP) would be needed as well that can process many events occurring across all levels of the organization, and be able to identify or monitor special events that deserve attention of the management. Before CEP as an IA can be implemented in place, the six technical gaps must firstly be made perfect while EAI helps an organization to unite the business processes.

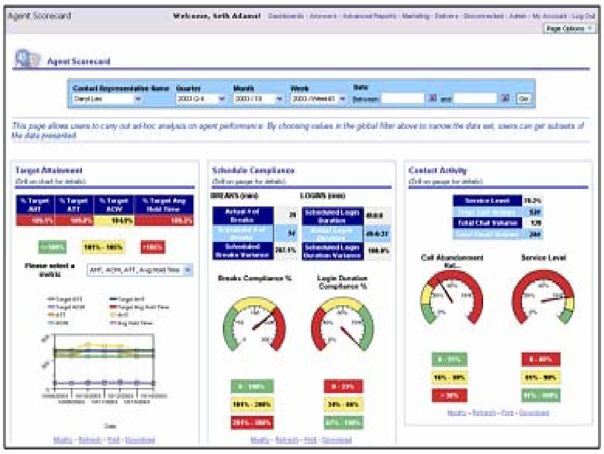

Figure 5. Sample of BI dashboards that offer analytical insights.

B. Customer Relationship Management (CRM)

Although CRM is a business methodology that aims at cultivating profitable and long-term relations with customers, IA is often a term used in the implementation of CRM systems. Data mining by tradition has been deeply used as a main 'brain' for understanding and responding to customers' needs, and improving the customer services at the front-line with value-added services such as intelligent recommender/advisor, online personalization, customers behaviors profiling and advanced sales-force-automation. Traditionally CRM was made effective with the aid of data mining techniques that classify, associate, and summarize data into patterns from various sources of data repositories for long-term strategy and customer services policies formulation. With the advent of data stream mining, many streams of related data in real-time can be simultaneously mined, and very quickly a decision can be made from the latest data.

Again this is possible only when the six technical gaps identified for the IA can be solved; thereafter, real-time CRM should enhance the following fundamental business values: *i.)* Increased customer retention by responding a timely remedy or some appropriate action to the customers who are potentially drifting away from the business; *ii.)* Increased process efficiencies by embedding the real-time BI across business processes; *iii.)* Identification and anticipation of market shifts, by aggregating the external market information into the internal business information, detect any shift of trend in real-time; *iv.)* Insights for new product development, by tapping onto customers feedbacks from forums, social-network and news of product reviews. Text mining on dynamic text streams from these resources is applied, e.g. RSS, FeedSync, Twitters etc.

C. Supply Chain Management (SCM)

Data mining can provide a further cost reductions to a manufacturing company through improving its manufacture and logistics processes. Many researches [44,45,46] have focused on the supply chain and logistics optimization to forecast with data mining techniques. Intelligent agent [47,48,49] play an significant role in supply chain management system in automation. Data mining is also a central part of a production system [50], which generally includes pattern discovery, trend detection, data dimensionality reduction and visualization. The use of data mining [51,52] aims at shortening the lead time of suppliers' delivery, e.g. i-PM algorithm, and be able to achieve more accurate demand forecasting.

As radio-frequency-identification (RFID) being a popular technology for tracing goods in SCM, real-time information on their whereabouts could be obtained in logistics management. Data stream mining techniques would be more applicable in this scenario because location information is being changed all the time as the goods are moving. The latest information would be very useful without waiting for the trained data mining models to be updated in batches. As a whole the supply chain management can be optimized more dynamically with such instant information readily available [53].

D. Healthcare

Naturally many biomedical data are produced in a streaming fashion ranging from physiological sensors to incidence records from local clinics that channel into a central data archive. This streaming data call for fast and accurate analysis, which hopefully can support effective clinical decisions as implemented as an IA or some similar clinical decision support system. The technical challenges that need to be met include but not limited to data aggregation from heterogeneous data sources (from different sensors), pattern recognition, real-time signal synchronization, incremental learning algorithms customized to learn from medical data streams, privacy-preserving issues as well as resource-aware problems in hardware – that are similar to the six technical gaps discussed earlier in this paper.

When data stream mining meets mobile computing, a new generation of mobile light-weight data stream analysis algorithms would be made possible that is resource-aware, ubiquitous and they can be implemented in compact mobile devices, such as wireless sensor networks. As an intelligent application, DSM certainly has a significant contribution to healthcare domain. A recent paper [54] shows that by DSM surely has an advantage over traditional data mining in healthcare, by collectively analyzing real-time biomedical monitoring sensors signals and patients' prognosis records in database. As a real-world scenario of healthcare monitoring, the authors in [55] adopted the DSM concept and extended it with fuzzy logic principles for modeling and reasoning about context/situations of a patient in a cell-phone prototype. Google Health (health.google.com) has recently launched as new services that allow users to store and track information about their medical history, and to support uploading of streams of medical data to an online repository from the users' sensors at home for real-time monitoring. This indicates that the computing platforms are quite ready, be it web-based ASP, customized applications on mobile phones or hospital medical system, opportunities do exist for health-care IA equipped with real-time data mining to be implemented.

E. Education

e-Learning has grown tremendously in recent years, a new trend in particular is peer-production based education where knowledge is generated via peer discussion instead of spoon-feeding from an instructor. As shown in the bibliography of a recent report [56] many authors have proposed pedagogical metrics and standardized models for learning behaviors; intensive studies have been carried out with those metrics too in various scenarios. However, analytical model for peer-production based e-learning networks that are exemplified by self-evolving data generation among peers, loosely connected network structure, implicit, distributed and scattered information derived from users' interactions and discussions, has not been formalized yet. The closest attempts are by Fong et al, [57,58] that studied about fusing elusive information and deriving trust factors in a social network, by taking Facebook as a case. As a potential project for intelligent application similar studies can be carried out pertaining to virtual e-learning social network. In fact intelligent applications that implemented personalization and recommendation systems have been proposed and studied before, as web-based educational systems [59]. But it was based on Web 1.0 platform. Likewise, Open-source education systems like Moodle are lately extended to include data mining functionalities [60]. The potential advantages of an IA with real-time data mining in an e-Learning environment are to help monitoring the students learning patterns in real-time rather than back-tracking their performance should the students have already passed or failed, any anomaly could be detected instantly, therefore early remedy could be applied.

F. Military

Battlefield information is distributed in nature where sensors and units report the latest updates as streams of data. For example, in a case of military applications, images and video from war machines send to the command center from the war zone. These footages would be analyzed in real-time so that commands of the next-course-of-actions would be given to the war fighters. As indicated in our technical gaps in earlier section, the challenge is to determine which data to analyze and which data to be overwritten should the situation be verified to have changed. The data model for military information is usually characterized as continuous, time-series, potentially infinite, fast-changing, dynamic, urgent and relatively simple (versus complex for efficient processing). This fits very well into data stream model that could be efficiently processed by one-pass DSM algorithm. Researchers in [61] studied the DSM performance for processing battlefield information with different data noise levels, number of battlefield information nodes and huge amount of data. Very Fast Decision Tree was used [61], and some improvement based on VFDT combining with Bayes classifiers were examined too [62,63,64]. A study [61] proposed an efficient algorithm BIDT (Battlefield Information Decision Tree). On the other hand, Aurora [65,66] is a stream query processing system using unique combinations of operators to filter, aggregate, discriminate, or omit certain data elements in stream mining. It solves the monitoring applications and differs substantially from conventional business data processing. In a particular case, communication network bandwidth is found to be a bottleneck and the problem continues to worsen as military evolves to mainly network centric operations with thousands of battlefield sensors. Reducing bandwidth is the principle consideration in deploying monitoring large sensor grids for homeland security and defense applications in military. For a successful IA to be implemented into military use and meet the tight performance expectations of military grade, fusion of technological advances in all levels of the system and the major six technical gaps must be solved. An IA for military use has all the features of real-time BI, such as complex event processing etc, plus certainly more.

V. Conclusions

Real-time data mining has so much to offer in various disciplines. We recognize the importance of it and observe that a new breed of real-time data mining, called data stream mining has found its suitability in many applications. In order to enable Intelligent Application that implies a software application program is possessing the capabilities to sense the environment, reason from flows of data streams, make decisions and offer instructions for the next-course-of-action fully in an autonomous fashion, still, there are some technical gaps exist in different aspects of a system waiting to be solved.

These technical gaps represent the divide between what we know must enable an ideal intelligent application and what the current technology can technically support. The gaps can be viewed laterally in different conceptual layers of an IA software program, namely Interoperability level, Algorithmic level and Data level. Higher levels depend on lower levels, so any deficiency that occurs in the low level will cause failure in high level. For example, an IA fails to communicate a wrong piece of information with another IA, so does the algorithm fails in accuracy and/or the data connection is problematic. This paper contributes in knowledge as a reference guide to designers who wish to design IA based on real-time data mining techniques. The techniques and their theoretical foundations might have matured both in academic research and commercial tools. However, the six technical gaps would have to be watched over in the IA design.

References

[1] K. J. Cios, W. Pedrycz, and R. M. Swiniarski, "Data Mining: A Knowledge Discovery Approach," Chapter 2, pp. 9-24, Springer Press, 2007

[2] Tomasz Imilienski and Heikki Mannila, "A Database Perspective on Knowledge Discovery," Communications of the ACM vol. 39, Issue 11, pp. 58-64.

[3] C. M. Ziemba, "Approach to Building and Implementing Business Intelligence Systems," *Interdisciplinary Journal of Information, Knowledge and Management*, Volume 2, 2007, pp. 135-148.

[4] L. Zeng, "Techniques, Process and Enterprise Solutions of Business Intelligence," *IEEE Conference on System, Man, and Cybernetics*, Taipei, Taiwan, October 8-11 2006, pp. 4722-4726.

[5] S. Langenfeld, "CRM and the Customer Driven Demand Chain," *Technical Report*, epiphany.com, 2004, pp. 1-4.

[6] C. White, "Building the Real Time Enterprise Report," *101 Communications Publication*, TDWI, 2003, pp. 1-35.

[7] S. Fong, Y. Hang, "Enabling Real-time Business Intelligence by Stream Data Mining," Book Chapter of *Data Mining*, Kimito Funatsu (Ed.), ISBN: 978-953-307-547-1, Intech, 05 January 2011, Vienna, Austria, pp. 83-96.

[8] M. Gaber, M. Zaslavsky, and S. Krishnaswamy, "Mining Data Streams: A Review," *SIGMOD* Rec. 34, 2, Jun 2005, pp. 18-26.

[9] P. Domingos and G. Hulten, "A General Method for Scaling Up Machine Learning Algorithms and its Application to Clustering", *The Eighteenth International Conference on Machine Learning*, 2001, Williamstown, MA, Morgan Kaufmann, pp. 160-113.

[10] B. Babcock, M. Datar, and R. Motwani, "Load Shedding Techniques for Data Stream Systems," *The 2003 Workshop on Management and Processing of Data Streams*, June 2003.

[11] N. Tatbul, U. Cetintemel, S. Zdonik, M. Cherniack, M. Stonebraker, "Load Shedding on Data Streams," *The 2003 Workshop on Management and Processing of Data Streams*, San Diego, CA, USA, June 8, 2003.

[12] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom, "Models and issues in data stream systems," The 21st ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, 2002, pp. 1-16.

[13] S. Muthukrishnan, "Data streams: algorithms and applications," *Now Publishers Inc*, ISBN-10: 9781933019147, January 10, 2005, pp. 1-136.

[14] A. C. Gilbert, Y. Kotidis, S. Muthukrishnan, M. Strauss, "One-Pass Wavelet Decompositions of Data Streams," *IEEE Transactions on Knowledge and Data Engineering*, May-June 2003, Volume: 15 Issue:3, pp. 541-554.

[15] C. Aggarwal, J. Han, J. Wang, P. S. Yu, "A Framework for Clustering Evolving Data Streams," *International Conference on Very Large Data Bases*, Volume 29, Berlin, Germany, Sept. 2003, pp. 81-92.

[16] C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A Framework for Projected Clustering of High Dimensional Data Streams," *International Conference on Very Large Data Bases*, Toronto, Canada, 2004, pp. 852-863.

[17] C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "On Demand Classification of Data Streams," *International Conference on Knowledge Discovery and Data Mining*, Seattle, WA, Aug. 2004, pp. 503-508.

[18] S. Muthukrishnan, "Data streams: algorithms and applications," The fourteenth annual ACM-SIAM symposium on discrete algorithms, ISBN:0-89871-538-5, Philadelphia, USA, 2003, pp. 413 – 413.

[19] M. Gaber, S. Krishnaswamy, A. Zaslavsky, "On-board Mining of Data Streams in Sensor Networks", *Advanced Information and Knowledge Processing*, 2005, Part II, Springer Verlag, pp. 307-335.

[20] M. Gaber, A. Zaslavsky, and S. Krishnaswamy, "A Cost-Efficient Model for Ubiquitous Data Stream Mining", *The Tenth International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, Perugia Italy, July 4-9.

[21] M. Gaber, A. Zaslavsky, and S. Krishnaswamy, “Towards an Adaptive Approach for Mining Data Streams in Resource Constrained Environments,” *The Sixth International Conference on Data Warehousing and Knowledge Discovery – Industry Track (DaWak 2004)*, Zaragoza, Spain, 30 August – 3 September 2004, Lecture Notes in Computer Science (LNCS), Volume 3181/2004, Springer Verlag, pp. 189-198.

[22] G. Dong, J. Han, L.V.S. Lakshmanan, J. Pei, H. Wang and P.S. Yu. “Online mining of changes from data streams: Research problems and preliminary results,” *The 2003 ACM SIGMOD Workshop on Management and Processing of Data Streams. In cooperation with the 2003 ACM-SIGMOD International Conference on Management of Data*, San Diego, CA, June 8, 2003

[23] M. Gaber, S. Krishnaswamy, and A. Zaslavsky, “Cost-efficient mining techniques for data streams”, *The Second Workshop on Australasian information Security, Data Mining and Web intelligence, and Software internationalisation* - Volume 32 (Dunedin, New Zealand). J. Hogan, P. Montague, M. Purvis, and C. Steketee, Eds. ACM International Conference Proceeding Series, vol. 54. Australian Computer Society, Darlinghurst, Australia, pp. 109-114.

[24] H. Yang, S. Fong, “An Experimental Comparison of Decision Trees in Traditional Data Mining and Data Stream Mining,” *The 2nd International Conference on Data Mining and Intelligent Information Technology Applications (ICMIA 2010)*, 30 November - 2 December 2010, Seoul, Korea, pp.442-447.

[25] J. H. Chang, and W.S. Lee, “Finding Recent Frequent Itemsets Adaptively over Online Data Streams,” *ACM SIGKDD Conference*, 2003, pp. 487-492.

[26] C.K.-S. Leung, Q.I. Khan, “DSTree: a Tree Structure for the Mining of Frequent Sets from Data Streams”, *IEEE ICDM*, 2006, pp. 928–932.

[27] J. Li, D. Maier, K. Tuftel, V. Papadimos, P.A. Tucker, “No Pane, No Gain: Efficient Evaluation of Sliding-window Aggregates over Data Streams,” *SIGMOD* Record 34 (1) (2005) pp. 39-44.

[28] C.-H. Lin, D.-Y. Chiu, Y.-H. Wu, A.L.P. Chen, “Mining Frequent Itemsets from Data Streams with a Time-sensitive Sliding Window,” Advances in Web and Network Technologies, and Information Management, Lecture Notes in Computer Science, 2007, Volume 4537/2007, Springer-Verlag, pp. 62-73.

[29] B. Mozafari, H. Thakkar, C. Zaniolo, “Verifying and Mining Frequent Patterns from Large Windows over Data Streams,” *ICDE*, 2008 pp. 179–188.

[30] M. Stonebraker, U. Çetintemel, and S. Zdonik, “The 8 Requirements of Real-time Stream Processing,” *SIGMOD* Rec. 34, 4 (Dec. 2005), pp. 42-47

[31] A. Arvind, S. Babu, J. Widom, “The CQL Continuous Query Language: Semantic Foundations and Query Execution”, *VLDB Journal*, Vol. 15, No. 2, June, 2006, pp. 121-142.

[32] Hua-Fu Li, “Mining Top-k Maximal Reference Sequences from Streaming Web Click-sequences with a Damped Sliding Window,” *Expert Systems with Applications*, Volume 36, Issue 8, October 2009, pp. 11304-11311.

[33] Y. S. Kim, “Streaming Association Rule (SAR) Mining with a Weighted Order-dependent Representation of Web Navigation Patterns,” *Expert Systems with Applications*, Volume 36, Issue 4, May 2009, pp. 7933-7946

[34] R. Cooley, B. Mobasher, and J. Srivastava, “Data Preparation for Mining World Wide Web Browsing Patterns,” *Knowledge and Information Systems*, Vol. 1, No. 1, pp. 5-32, 1999

[35] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins, “The Web as a Graph: Measurements, Models and Methods,” *Computing and Combinatorics, Lecture Notes in Computer Science,* Springer-Verlag, 1999, Volume 1627, 1999, pp. 1-17.

[36] J. Sun, S. Papadimitriou, and C. Faloutsos, “Distributed Pattern Discovery in Multiple Streams,” *The Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2006, pp. 713-718.

[37] H. Yang, S. Fong, “The Impacts of Data Stream Mining on Real-Time Business Intelligence,” *The 2nd International Conference on IT and Business Intelligence (ITBI 2010)*, IMTCJ, ISSN: 2229-6743, 12-14 November 2010, Nagpur, India, pp. 9-19

[38] H. Yang, S. Fong, “Aerial Root Classifiers for Predicting Missing Values in Data Stream Decision Tree Classification”, *2011 SIAM International Conference on Data Mining (SDM 2011)*, April 28-30, 2011, Mesa, Arizona, USA, Accepted for Publication

[39] H. Yang, S. Fong, “Optimized Decision Tree Algorithm in Stream Mining for Time-critical Applications,” *The 37th International Conference on Very Large Data Bases*, Aug 29-Sep 3, Seattle, USA, Under review.

[40] M. Balabanovic and Y. Shoham, “Learning Information Retrieval Agents: Experiments with Automated Web Browsing,” *AAAI Spring Symposium on Information Gathering from Heterogenous, Distributed Resources*, 1995, pp. 13-18

[41] H. Yang, S. Fong, “Stream Mining Over Fluctuating Network Traffic at Variable Data Rates,” *The 2nd International Conference on Data Mining and Intelligent Information Technology Applications (ICMIA 2010)*, 30 November - 2 December 2010, Seoul, Korea, pp.436-441

[42] C. Vercellis, “Business Intelligence Data Mining and Optimization for Decision Making”, Chapter 1, pp.3, Published by *Wiley & Sons*, 2009

[43] H. Yang, S. Fong, “Real-time Business Intelligence System Architecture with Stream Mining,” *5th International Conference on Digital Information Management (ICDIM 2010)*, July 2010, Thunder Bay, Canada, pp.29-34

[44] D. A. Lax, J. K. Sebenius, “The Manager as Negotiator: Bargaining for Cooperation and Competitive Gain,” *The Free Press*, New York, NY, 1986

[45] D. F. Muñoz, “A Bayesian Framework for Modeling Demand in Supply Chain simulation Experiments”, *The 35th Conference on Winter Simulation*, 2003, pp.1319-1325.

[46] L. Aburto, R. Weber, “Improved Supply Chain Management based on Hybrid Demand Forecasts,” *Applied Soft Computing Archive*, Volume 7, Issue 1, 2007, pp.136-144.

[47] Y. Hang, S. Fong and Z. Yan, “Applying Pareto-Optimal and JIT Techniques for Supply Chains,” *The 8th International Conference on Hybrid intelligent Systems, (HIS 2008)*, IEEE Computer Society, Washington, DC, September 10 - 12, 2008, pp. 308-313.

[48] H. S. Kim, J. H. Cho, H. R. Choi, S. G. Hong, M. H. Kang, “Optimal Supply Chain Formation using Agent Negotiation in a SET Model-based Make-To-Order”, *International Conference on Electronic Commerce*, 2006, pp.579-583.

[49] K. Dorer, M. Calisti, “An Adaptive Solution to Dynamic Transport Optimization”, *International Conference in*

Autonomous Agents, the International Workshop on Agent Theories, 2006, pp. 45-51.

[50] A. Kusiak, M. Smith, "Data Mining in Design of Products and Production Systems," *Annual Reviews in Control*, Volume 31, Issue 1, 2007, pp. 147-156.

[51] H.C.W. Lau, G.T.S. Ho, Y. Zhao, N.S.H. Chung, "Development of a Process Mining System for Supporting Knowledge Discovery in a Supply Chain Network," *International Journal of Production Economics*, Volume 122, Issue 1, Transport Logistics and Physical Distribution; Interlocking of Information Systems for International Supply and Demand Chains Management; ICPR19, November 2009, pp. 176-187.

[52] Lau et al., H.C.W. Lau, G.T.S. Ho, K.F. Chu, W. Ho and C.K.M. Lee, "Development of an Intelligent Quality Management System Using Fuzzy Association Rules," *Expert Systems with Application* 36 (2009), pp. 1801–1815.

[53] H. Yang, S. Fong, Z. Yan, "CSET Automated Negotiation Model for Optimal Supply Chain Formation," *World Review of Science, Technology and Sustainable Development (WRSTSD)*, Vol. 7, Nos. 1/2, 2010, pp.67-78

[54] H. Yang, S. Fong, A. Ip, S. Mohammed, "Case-based and Stream-based Classification in Biomedical Application," *The Eighth IASTED International Conference on Biomedical Engineering (Biomed 2011)*, 16-18 February 2011, Innsbruck, Austria, pp. 207-214.

[55] P. D. Haghighi, B. Gillick, S. Krishnaswamy, M. Gaber and A. Zaslavsky, "Situation-Aware Adaptive Visualization for Sensory Data Stream Mining," *Knowledge Discovery from Sensor Data*, Lecture Notes in Computer Science, Springer-Verlag, 2010, Volume 5840/2010, pp. 43-58.

[56] V. Hinze-Hoare, "HCI and Educational Metrics as Tools for VLE Evaluation", *CoRR Journal*, 2006, pp. 1-10.

[57] C. Wei, S. Fong, G. Kim, "Multi-Collaborative Filtering Trust Network Model for Web 2.0 Recommender", *2011 SIAM International Conference on Data Mining (SDM 2011)*, April 28-30, 2011, Mesa, Arizona, USA, To be published.

[58] C. Wei, S. Fong, "Social Network Collaborative Filtering Framework and Online Trust Factors: a Case Study on Facebook", *The 5th International Conference on Digital Information Management (ICDIM 2010)*, July 2010, Thunder Bay, Canada, pp. 266-273.

[59] S. Ha, S. Bae, S. Park, S, "Web Mining for Distance Education," *IEEE international conference on management of innovation and technology*, 2000, pp. 715–719.

[60] C. Romero, S. Ventura, E. Garcia, "Data Mining in Course Management Systems: Moodle Case Study and Tutorial," *Computers & Education*, Volume 51, Issue 1, August 2008, pp. 368-384

[61] N. Wang, T. Wang, "An Efficient Method for Battlefield Information Data Stream Mining," *International Joint Conference on Computational Sciences and Optimization*, Vol. 1, 2009, pp.723-725.

[62] G. Hulten, L. Spencer, P. Domingos, "Mining Time-changing Data Streams," *KDD 2001*, ACM Press, New York, NY, pp. 97–106.

[63] J. Gama, R. Fernandes, R. Rocha, "Decision Trees for Mining Data Streams," *Intelligent Data Analysis*, Vol. 10, 2006, pp. 23-45.

[64] F. Li, Q. Liu, "An Improved Algorithm of Decision Trees for Streaming Data Based on VFDT," *International Symposium on Information Science and Engineering*, 2008, Vol. 1, pp.597-600.

[65] D. Carney, U. Centintemel, M. Cherniack, C. Convey, S. Lee, G. Seidman, M. Stonebraker, N. Tatbul, S. Zdonik, "Monitoring Streams – A New Class of Data Management Applications," The 28th international conference on Very Large Data Bases, Hong Kong, August 2002, pp. 215 – 226.

[66] M. Cherniack, H. Balakrishnan, M. Balazinska, D. Carney, U. Cetintemel, Y. Xing, and S. Zdonik, "Scalable Distributed Stream Processing," The First Biennial Conference on Innovative Database Systems (CIDR'03)," Asilomar, CA, January 2003

Simon Fong graduated from La Trobe University, Australia, with a 1st Class Honours BEng. Computer Systems degree and a PhD. Computer Science degree in 1993 and 1998 respectively. Simon is now working as an Assistant Professor at the Computer and Information Science Department of the University of Macau. He is also one of the founding members of the Data Analytics and Collaborative Computing Research Group in the Faculty of Science and Technology. Prior to joining the University of Macau, he worked as an Assistant Professor in the School of Computer Engineering, Nanyang Technological University, Singapore. Prior to his academic career, Simon took up various managerial and technical posts, such as systems engineer, IT consultant and e-commerce director in Melbourne, Hong Kong and Singapore. Some companies that he worked before include Hong Kong Telecom, Singapore Network Services, AES Pro-Data and United Oversea Bank, Singapore. Dr. Fong has published over 120 international conference and journal papers, mostly in the area of E-Commerce technology, Business Intelligence and Data-mining.

Hang Yang a PhD candidate in the University of Macau. He obtained a MSc (First Honor) in Electronic Commerce Technology in the University of Macau in 2009; and a bachelor degree of economics in Electronic Commerce in Guangdong University of Foreign Studies (China) in 2007. He worked for the companies of Fortis Insurance, China Petrol in Hong Kong and Beijing. He is supervised by Dr. Simon Fong, with the research interest covering data mining, business intelligence, electronic commerce, web intelligence. So far, he has over 20 publications, including journals and conference papers.