Algorithmic level stream mining for Business Intelligence System Architecture building

Yang Hang, Simon Fong
Department of Computer and Information Science
University of Macau, Macau SAR
{henry.yh@gmail.com, ccfong@umac.mo}

ABSTRACT: *Datamining has potential applications in several fields including the Business Intelligence (BI) where in the datamining is useful in for understanding the trends and reacting to events. We advocate that the crucial area in dataming, the stream-mining where continuous data streams arrive into the system and get mined very quickly. It further induces the framework for creating a new type of real-time Business Intelligence architecture. The algorithmic level stream mining and digital information system architecture was addressed and applied separately. We now try to focus a single view on the real-time Business Intelligence system architecture powered by stream-mining. Besides we present a few more possible applications in which the proposed architecture can support are discussed..*

Keywords: Stream Mining, Business Intelligence, Intelligence architecture

Received: 17 November 2010, Revised 21 December 2010, Accepted 29 December 2010

1. Introduction

Nowadays businesses evolved to be more competitive and dynamic than the past, which demand for real-time BI and capability of making very quick decisions. With this new business market demand, a recent study [1], advocated that BI should be specified in four dimensions: strategic, tactical, operational and real-time.

Compared with the operational BI, real-time BI (rt-BI) shall analyze the data as soon it enters the organization. The latency (data latency, analysis latency, decision latency) shall be zero ideally. In order to establish such real-time BI systems, relevant technologies to guarantee low/zero latency are necessary. For example, operational / real-time BI data warehouse techniques are able to provide fresh data access and update. Thus operational BI can be seen as rt-BI as long as it can provide analytics within a very short time for decision making. The main approach is: system response time shall stay under a threshold that is less than the action taking time; and the rate of data processing shall be faster than the rate of data producing. However, there exist a handful of real-time data mining algorithms in theoretical fields, but their applicability and suitability towards various real-time applications are still fuzzy; so far no one has proposed a generic architecture for rt-BI with consideration of streammining. We take it as the research motivation and hence the contribution of this paper is a detailed study of such architecture.

The paper is structured as follow: Section 2 is an overview of rt-BI system; the high-level framework, system architecture, process are described. Section 3 is a discussion of how rt-BI could be applied in several typical application scenarios. A conclusion is drawn in the last section.

2. rt-BI System

2.1 Framework

rt-BI system relates to many technologies and tools evolved from strategic BI and tactical BI. A four-layer framework is proposed for rt-BI system in Figure 1. The main improvement is a real-time processing of whole knowledge discovery process.

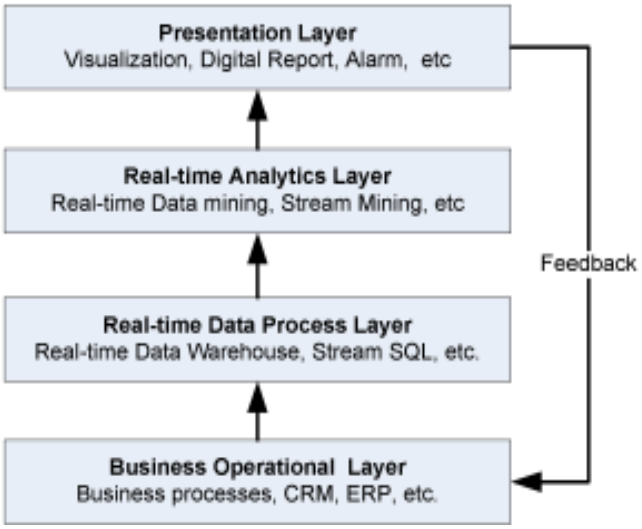

Figure 1. Four-layer Framework

2.1.1 Business Operational Layer

This layer composes of two primary functions: business activity monitoring (BAM) and real-time process tuning [2]. Azvine [3] presents the shortcoming of current BAM and process tuning technology for rt-BI: 1) current BAM can't make intelligent conclusion about the overall business process behavior; 2) and business processes changes are carried throughout initiatives manually, that is expensive and time consuming. On the other hand, the level of automation is divided into two stages: semi- and fully-automatic. Our proposed framework tackles these problems with a fully-automated process. The system is built right on the top of business operations. It shall facilitate automated mapping of existing business operations within an organization, capture the knowledge to automate process tuning, optimization and re-engineering, and monitor people and systems for process conformance.

2.1.2 Real-time Data Process Layer

This layer is responsible for providing qualified data to its upper layer – analytics layer. Data come from various resources in different formats. If the data contain too much noise, it is harmful for business intelligence discovery. In this layer, the system is required to obtain the quality data within a time constraint. For this reason, preparation process should not take too long. Modern digital source has a huge volume and consist of rapidly changing data. Data stream technology [4] provides a good solution to build real-time data warehouse, with which increased refresh cycles to frequently update the data. This kind of data warehouse systems can achieve near real-time data updating, where the data latency typically is in the range of seconds.

2.1.3 Real-time Analytic Layer

Traditionally, data analyzing follows "analyst-inthe- middle" approach where human expert analysts are required to drive or configure the information with BI software and tools. This may lead to analysis latency. To this end, the analysis tools should provide a high degree of automation, which is relating to artificial intelligence technology and agent technology. Data miners serve as the kernel to build models or extract patterns from large amounts of information [5, 6, 7]. Analytics layer uses fast data mining method to interpret data to information. So far there are many real-time data mining algorithms and methodologies. The four popular types are: clustering, classification, frequency counting, and time series analysis. Stream processing engines are also used based on sliding windows technology [8].

2.1.4 Presentation Layer

This layer presents the BI to end-user in highly interactive way in order to shorten action latency. The presentations vary in formats and designs. For examples, sophisticated time-series charts show a trend, and a KPI dashboard alarms off anomalies etc. Many commercial companies provide these techniques as third-party solutions, e.g. iNetSoft, SPSS, IBM, etc.

2.2 System Architecture

Traditionally, the classic method to build a model with data mining algorithm is by a training-thentesting approach. But the

weakness is they may not suit a very large data volume and high speed data.

A Mining Model Definition Language (MMDL) is used for stream mining system [9], but it has not illustrated how to design a stream mining system in technical field. A research [10] proposed three realtime data stream processing architectures which can potentially be applied to solve high-volume low latency streaming problems, but its both Rule Engine and Stream Process Engine architectures only rely on stream data querying (SQL). Mining data streams has been studied by some researchers. Gaber [11] summarized the most cited data stream mining techniques with respect to different mining tasks, approaches and implementations. They proposed an adaptive resource-aware approach called Algorithm Output Granularity (AOG) [12, 13].

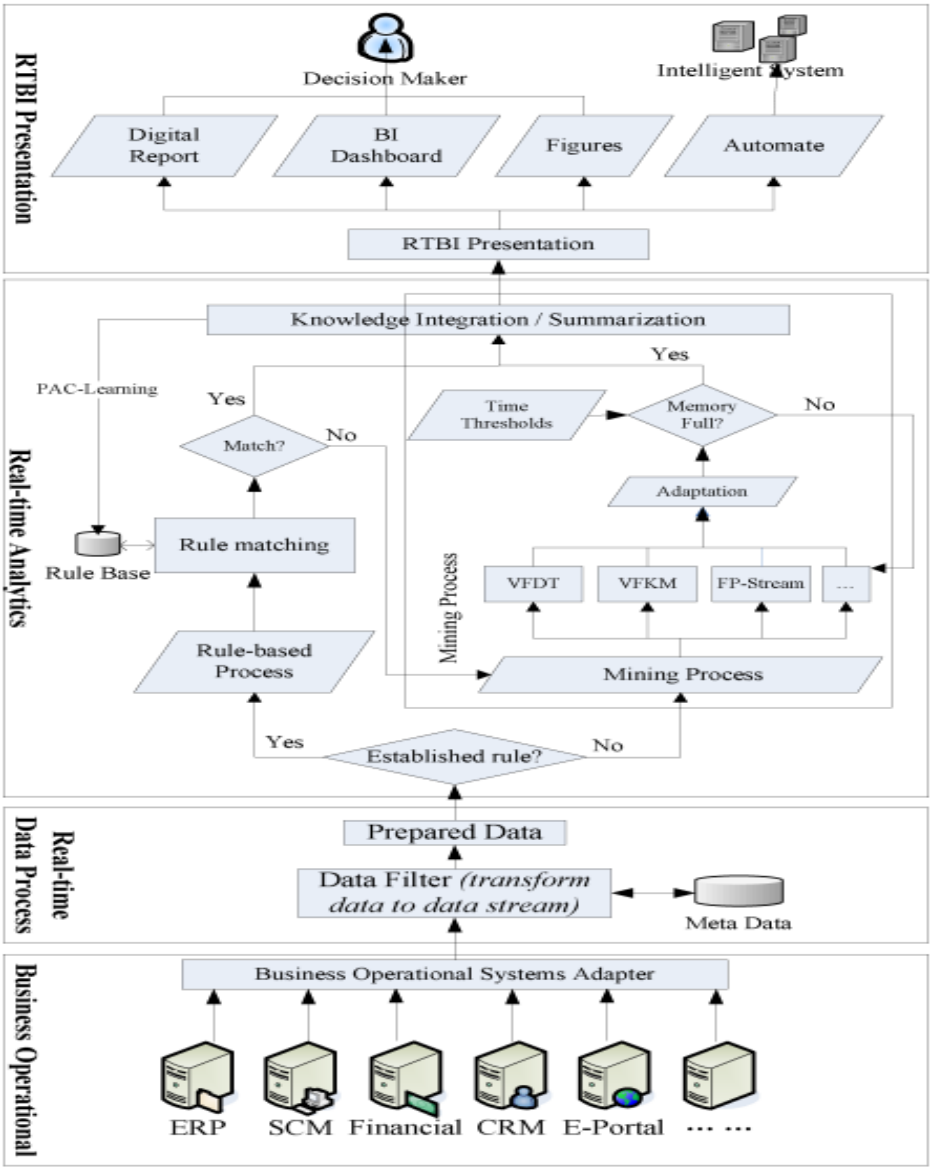

Figure 2. rt-BI System Architecture

The rt-BI system architecture described in this section is evolved from the previous research in data mining and business intelligence. Different from the previous ones, the proposed architecture concentrates on constructing a system which is able to extract potential BI and return result to end-user in real-time. Figure 2 shows a static view of rt-BI system architecture. *Firstly*, the rt-BI system collects a large amount of historical data from existing information system. *Secondly*, the system collects and monitors the new input data in real-time data process layer. If necessary it will transform the data into adequate forms. *Thirdly*, the system determines whether it relates to an established model in the real-time analytics layer. If so, the system matches it with the rules and returns BI result. Otherwise, the system runs on data-mining process in order to find new

rules and BI. A newly found model is updated to the rule-base database. *Fourthly*, the discovered information is summarized as rt-BI result and presented in appropriate formats. During this process, any mis-prediction or incorrect-pattern will be rectified and updated to the database. This process should be within a certain time threshold that the BI output is useful fordecision making (to ensure no analytics latency). By this architecture, the system collects data and generates some prediction models in real-time. The data used to discover BI is not only dependent on historical but also the new coming data.

2.3 rt-BI Analyzing Process

The analyzing kernel of an rt-BI system is the mining process. In this section, we show a dynamic process representation in Figure 3 on how to implement the data mining process in RT-BI system.

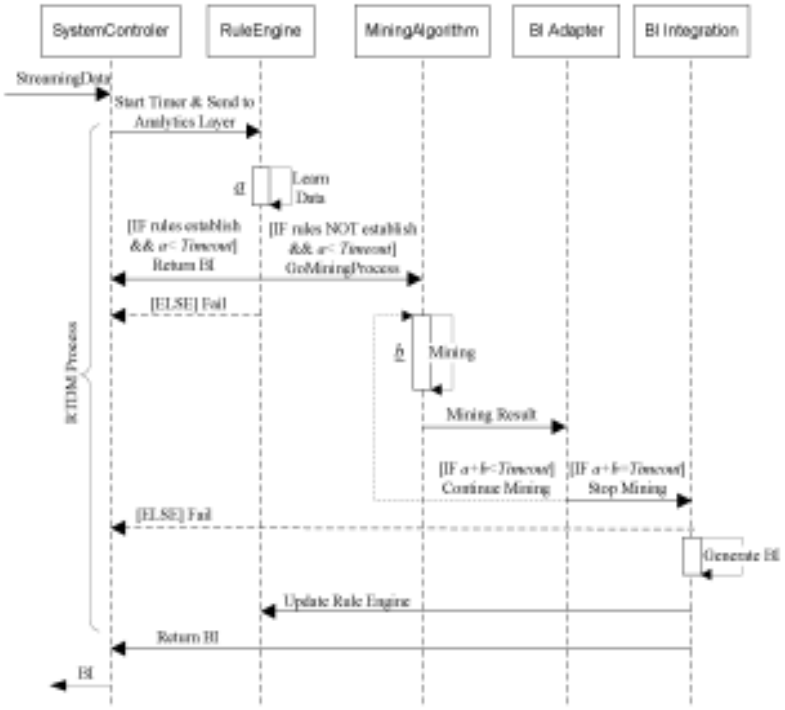

Figure 3. RT-BI Generating Workflow

The process contains two segments: rule-based matching, and new BI mining. When new data come, a timer is started to control the rt-BI running time so as to restrict analytics latency within an acceptable level. A timeout threshold is determined by the time required to make a decision, which restricts the rulebased searching/matching time as well as the BI mining time. If new arrival data are correlating to the already established rules, the rule-based matching process activates and returns the BI within the time threshold. Otherwise, the new BI mining process will trigger. The determination should be within the time threshold. If it timeouts, the rt-BI system is deemed failed to discover new BI and returns the last most updated information instead.

3. Applications of RT-BI System

The proposed system architecture can be applied in different applications. We illustrate four typical application domains. A more comprehensive comparison is presented in Appendix Table 1.

3.1 Anomaly Detection and Automated Alerts

Anomaly Detection refers to detecting patterns in a given data set that do not conform to an established normal behavior [14]. The detected patterns are called anomalies, which are also referred to as outliers, surprise, aberrant, deviation, peculiarity, etc. and often translated to critical and actionable information in several application domains. Many anomaly detection techniques have been specifically developed for certain application domains, while others are more generic. Its application domains mainly include: insurance fraud detection [15], network attack detection [16], and credit card fraud detection [17, 18], etc. A survey [19] tries to provide a structured and comprehensive overview of the research on anomaly detection, but it

doesn't give a generic design for such kind of rt-BI system. This type of applications is not only required to find the anomaly pattern from a large amount of data in realtime, but to present the result to end-user promptly and to take action efficiently.

3.2 Prediction and Suggestions Recommender

Customer Relationship Management (CRM) systems apply data mining to analyze and predict the potential customer values. Although the analysis of available information for those customers who in the past have purchased product or services based on the historical data, and the comparisons with the characteristics of those who have not taken up the offer of the enterprise, it is possible to identify the segments with the highest potential. Commercial recommender systems use various data mining techniques to provide appropriate recommendations to users during real-time online sessions. E-business transactions usually take place over online networks.For analyzing e-Portal information, rT-BI system recommends suitable suggestions to customers. A context-similarity based hotlinks assignment [20] analyzes the similarity of context between pages in order to suggest the placement of suitable hotlinks. Another real-time recommendation system based on experts' experiences is proposed in [17]. It simplifies content-based filtering through computing similarity of the keywords and recommends common users the Web pages based on experts' search histories but not the whole archive of Web pages.

Online recommender systems often use the suggested purchase items, or the items in which costumer may be also interested, as the presentation of rT-BI. These techniques widely used in call centers to investigate telephone call data streams.

3.3 Forecast and Markets Analysis

Pricing network resources is a crucial component for proper resource management and the provision of quality of service guarantees in different markets. A model used data mining to forecast the stock market with time series trends [21]. A competitive market intelligence system [22] proposes to detect critical differences in the text written about a company versus the text for its competitor. However, the intelligence system is compelled to depend on empirical performance, which requires human interaction to analyze the discovered patterns. As aforementioned, the latency is the primary constraint of operational BI and real-time BI. If the problems of data latency and data availability are solved, then businesses can react proactively to new information and knowledge in real-time.

Real-time business intelligence dashboards are used to bridge the gap between operational business intelligence and real-time business intelligence. It shall display not only historical information but also show the current status to support decision making.

3.4 Optimization and Supply Chain Management

Supply Chain Management (SCM) is one of the most important topics in e-Commerce. Online business transaction builds a dynamic pricing model that is integrated into a real-time supply chain management agent [23]. Besides the pricing strategy, real-time supply chain management in a rapidly changing environment requires reactive and dynamic collaboration among participating entities. Radio Frequency Identification (RFID) is widely used in high-tech arena. It is described as a major enabling technology for automated contactless wireless data collection, and as an enabler for the real-time enterprise. Goods are supervised while they are embedded with RFID tags. After capturing the data stream by sensors, RFID system is aware of the information of the goods, such as location and status.

The real-time supervising and gaining visibility can achieve quick responsiveness and high efficiency in business flows, if RFID technology can be applied efficiently [24].

The proposed architecture may address the challenge of processing high-volume, real-time data with requiring the use of custom code. rt-BI systems provide pattern discovery, trend detection, and visualization, controlling and improving the flow of materials and information, originating from the suppliers and reaching the end customers.

4. Conclusion

Real-time business intelligence is a new concept in knowledge discovery. rt-BI requires exploring BI from a large volume and rapidly arriving data in business operations. rt-BI system aims to achieve very short time required in data acquiring process and analysis process for decision making. We proposed a generic framework architecture for rt-BI, followed by a discussion of rt-BI applications.

References

[1] Asghar, S., Fong, S., Hussain, T. (2009). Business Intelligence Modeling: A Case Study of Disaster Management Organization in Pakistan, *In*: The 4th International Conference on Computer Sciences and Convergence Information Technology (ICCIT 2009), 24-26 November, Seoul, Korea, p.637-638

[2] McCoy, D.W. (2002). Business Activity Monitoring: Calm Before the Storm, Gartner Research, ID: LE-15- 9727.

[3] Azvine, B., Cui, Z., Nauck, D. D., Majeed, B. (2006). Real Time Business Intelligence for the Adaptive Enterprise, *In:* Proceedings of the the 8th IEEE international Conference on E-Commerce Technology and the 3rd IEEE international Conference on Enterprise Computing, E-Commerce, and E-Services. CEC-EEE. IEEE Computer Society, Washington, DC, June 26 - 29, 2006, p. 29

[4] Botan, I., Cho, Y., Derakhshan, R., Dindar, N., Haas, L., Kim, K., Tatbul, N. (2009). Federated Stream Processing Support for Real-Time Business Intelligence Applications, VLDB International Workshop on Enabling Real-Time for Business Intelligence (BIRTE'09), Lyon, France, August.

[5] Hand, D.J., Mannila H., Smyth P. (2001). Principles of data mining, MIT Press.

[6] Hand, D.J. (1999). Statistics and Data Mining: Intersecting Disciplines, ACM SIGKDD Explorations, 1, 1, June. p.16-19.

[7] Hoffmann, F., Hand, D.J., Adams, N., Fisher, D., Guimaraes, G. (eds) (2001). Advances in Intelligent Data Analysis. Springer.

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

[9] Thakkar, H., Mozafari, B., Zaniolo, C. (2008). A Data Stream Mining System, IEEE International Conference on Data Mining Workshops (ICDMW '08), p.987-990

[10] Stonebraker, M., Çetintemel, U., Zdonik, S. (2005). The 8 requirements of real-time stream processing, SIGMOD Rec. 34, 4, 2005, p. 42-47

[11] Gaber, M. M., Zaslavsky, A., Krishnaswamy, S. (2005). Mining data streams: a review, SIGMOD Rec. 34 (2) 18 26

[12] Gaber, M. M., Zaslavsky, A., Krishnaswamy, S. (2004). Towards an Adaptive Approach for Mining Data Streams in Resource Constrained Environments, *In:* The Proceedings of Sixth International Conference on Data Warehousing and Knowledge Discovery - Industry Track.

[13] Gaber, M. M. (2009). Data Stream Mining Using Granularity-Based Approach, Studies in Computational Intelligence, Foundations of Computational, V. 6, p.47-66

[14] Hodge, V.J., Austin, J. (2004). A Survey of Outlier Detection Methodologies", Artificial Intelligence Review, Kluwer Academic Publishers, p.85–126

[15] Phua, C., Alahakoon, D., Lee, V. (2004). Minority report in fraud detection: classification of skewed data, SIGKDD Explorer, News l.6, 1 Jun, p.50-59

[16] Zhengbing, H., Zhitang, L., Junqi, W. (2008). A novel Network Intrusion Detection System (NIDS) based on signatures search of data mining, *In*: Proceedings of the 1st international Conference on Forensic Applications and Techniques in Telecommunications, information, and Multimedia and Workshop, Adelaide, Australia, January 21 - 23, 2008, ICST, Brussels, Belgium, p.1-7

[17] Quah, J. T., Sriganesh, M. (2008). Real-time credit card fraud detection using computational intelligence, *Expert Systems Applications*, 35 (4) 1721-1732.

[18] Whitrow, C., H.,, D. J., Juszczak, P., Weston, D., Adams, N. M. (2009). Transaction aggregation as a strategy for credit card fraud detection", Data Mining Knowledge Discovery, 18 (1) 30-55

[19] Chandola, V., Banerjee, A., Kumar, V. (2009). Anomaly Detection: A Survey, *ACM Computing Surveys*, .41(3) Article 15, July.

[20] Antoniou, D., Garofalakis, J., Makris, C., Panagis, Y., Sakkopoulos, E. (2009). Context-similarity based hotlinks assignment: Model, metrics and algorithm, Data & Knowledge Engineering, In Press, Corrected Proof, Available online 4 May 2009.

[21] Dietmar, H., Dorr, Anne., Denton, M(2009). Establishing relationships among patterns in stock market data, *Data & Knowledge Engineering*, 68 (3) 318-337

[22] Weiss, S. M., Verma, N. K. (2002). A system for real-time competitive market intelligence, *In:* Proceedings of the Eighth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining, KDD '02, ACM, New York, NY, p.360 365.

[23] Ku, T., Zhu, Y., Hu, K (2008). A Novel Complex Event Mining Network for RFID-Enable Supply Chain Information Security, *In*: Proceedings of the 2008 international Conference on Computational intelligence and Security - Volume 01, CIS., IEEE Computer Society, Washington, DC, p.516- 521

[24] Gonzalez, H., Han, J., Li, X (2006). Mining compressed commodity workflows from massive RFID data sets, *In:* Proceedings of the 15th ACM international Conference on information and Knowledge Management, CIKM '06, ACM, New York, NY, p.162-171
[25] Mukhopadhyay, A., Chatterjee, S., Saha, D., Mahanti, A., Sadhukhan, S.K. (2006). e-Risk Management with Insurance: A Framework Using Copula Aided Bayesian Belief Networks", System Sciences, 2006. HICSS '06. Proceedings of the 39th Annual Hawaii International Conference on , V. 6, Jan. 2006, p.04-07
[26] Sun, J., Yu, X., Wu, Z., Li, X. (2008). Real Time Recommendation Utilizing Experts' Experiences, *In:* Proceedings of the 2008 Fifth international Conference on Fuzzy Systems and Knowledge Discovery - Volume 05, FSKD., IEEE Computer Society, Washington, DC, p.379-383
[27] Dorr, Dietmar H., Denton,.Anne M (2009). Establishing relationships among patterns in stock market data, *Data & Knowledge Engineering*, 68 (3) p.318-337