Expert Systems with Applications 39 (2012) 10544–10551

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Hybrid genetic algorithm and association rules for mining workflow best practices

Amy H.L. Lim [a], Chien-Sing Lee [a,*], Murali Raman [b]

[a] *Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia*
[b] *Graduate Institute of Management, Faculty of Management, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia*

ARTICLE INFO

Keywords:
DSS development-functionality
Development-methodology-business models
Business intelligence
Genetic algorithm
Performance measurement
E-commerce

ABSTRACT

Business workflow analysis has become crucial in strategizing how to create competitive edge. Consequently, deriving a series of positively correlated association rules from workflows is essential to identify strong relationships among key business activities. These rules can subsequently, serve as best practices. We have addressed this problem by hybridizing genetic algorithm with association rules. First, we used correlation to replace support-confidence in genetic algorithm to enable dynamic data-driven determination of support and confidence, i.e., use correlation to optimize the derivation of positively correlated association rules. Second, we used correlation as fitness function to support upward closure in association rules (hitherto, association rules support only downward closure). The ability to support upward closure allows derivation of the most specific association rules (business model) from less specific association rules (business meta-model) and generic association rules (reference meta-model). Downward closure allows the opposite. Upward-downward closures allow the manager to drill-down and analyze based on the degree of dependency among business activities. Subsequently, association rules can be used to describe best practices at the model, meta-model and reference meta-model levels with the most general positively dependent association rules as reference meta-model. Experiments are based on an online hotel reservation system.

1. Introduction

The process of finding and performing in-depth analysis on patterns/anomalies within the data is increasingly crucial as these patterns describe relationships among important decision variables which can increase the manager's understanding of the current state. The decision outcome subsequently informs the determination of strategies, which create competitive advantage. Five competitive forces that most organizations have to deal with are "*threat of substitute products, the threat of established rivals, and the threat of new entrants, the bargaining power of suppliers and the bargaining power of customers*" (Porter, 1979).

Assessment outcomes often result in adjustments to goals and tactics to outwit existing entrants in the market. Continuous assessment means that the above procedures will be carried out during specific time intervals to understand the current status of the organization and to evaluate the effectiveness of the implemented strategy and next possible actions to be performed.

This paper is an extension to our previous three works aimed at introducing approaches to assist continuous assessment and support an organization's management in decision making. *The first earlier work* highlights the need to increase the effectiveness of workflow management systems (WFMS) to allow reusability of effective and efficient workflow. This means that WFMS should be extended to provide analysis and create a pool of best practices. It can be achieved by creating a pool of best workflow practices within a repository. This is addressed with the introduction of the Weighted and Layered WF evaluation (WaLwFA) methodology (Lee & Lim, 2007). WaLwFA adopts the concept of Model-Driven Architecture (MDA) with aims to capture best practices, which can be adopted and instantiated to many domains or various information systems. In WaLwFA, the business process models are evaluated using a set of weighted criteria and sub-criteria, which are derived by averaging the assignment of weights by a group of experts. The business models with higher scores are kept within the repository to form business meta-models and reference model respectively. The repository will constantly be updated when there is a new "good" business model, thus ensuring that the repository remains updated with the latest best practice.

The *second earlier work* highlights the need to have an extensive reusable business performance measurement framework that can pinpoint causal relationships between the organization's current business performance and its future directions as well as measure the organization's workforces. To address the *second* problem, Integrated Model-Driven Business Evaluation (IMoBe) methodology is proposed (Lim & Lee, 2008). In this framework, a model-driven

* Corresponding author.
E-mail addresses: amy.lim@mmu.edu.my (A.H.L. Lim), cslee@cl.ncu.edu.tw (C.-S. Lee), murali.raman@mmu.edu.my (M. Raman).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2012.02.183

knowledge base serves as repository to store two or more commonly used business performance strategies or elements to evaluate any business organization such as Sun Tzu's 13 themes of business management strategies (Lee, Roberts, Lau, & Bhattacharya, 1998) reported in Ko and Lee (2000) as well as concepts of critical success factor (CSF) and critical barrier factor (CBF) (Niazi, Wilson, & Zowghi, 2005). IMoBe's methodology is an integration of several business performance approaches consisting of the Balanced Scorecard (BSC) (Kaplan & Norton, 1992) and the Quality Function Deployment (QFD) methodology. The selected business performance model will serve as a predictor measure and guide organizations on the next course of actions. The selected model contains criteria and they represent the rows in the House of Quality (HOQ) matrix while the columns are represented by strategic objectives identified for each perspective of the BSC. Customization of criteria is allowed where other criteria can be added to the HOQ's row in order to assist any organization in making efficient and effective actions.

The *third previous work* describes the need to improve existing evaluation methods in the first problem to extend from weighted evaluation to incorporating DSS with OLAP to improve the process of decision making where rules consisting of criteria as attributes can be hierarchically arranged and sorted according to its degree of complexity with each association rule having scores of support, confidence and correlation. For the *third* problem, a decision support system architecture consisting of our business performance methodology, namely WaLwFA, is extended to incorporate business intelligence capabilities to assist decision-making (Lim & Lee, 2010). The C4.5 decision tree algorithm (Quinlan, 1993) is used to discover significant attributes and association rule, namely the Apriori algorithm, (Agrawal, Imielinski, & Swami, 1993) is used to derive simple and complex association rules as well as to perform correlation analysis to calculate the dependency among attributes.

This paper extends from our DSS-OLAP study (Lim & Lee, 2010) to address the problems below.

2. Problem statements

Business workflow analysis has become crucial in strategizing how to create a competitive edge. Consequently, deriving a series of positively correlated association rules from workflows is essential to identify strong relationships among key business activities. These rules can subsequently, serve as best practices. In this paper, strong rules refer to rules that have correlation score of more than "1".

Second, typical association rule algorithm (AR) such as Apriori supports only downward closure. If a set of *n*-items has reached minimum support value, the subsets of items also meet the minimum support (Liu, Hsu, & Ma, 1999). Support and confidence must be pre-specified by the user prior to applying AR.

Upward closure as mentioned by Brin, Motwani, and Silverstein (1997) is "*constructive*". Suppose we have *n*-items that are correlated, then the superset associated with *n*-items is correlated as well. The correlation score of the superset should minimally equal to the correlation score of the *n*-items. Confidence measurement does not support upward closure. This can be seen in the example provided by Brin et al. (1997) in their paper.

3. Approach

Hybridization of genetic algorithm (GA) (Holland, 1975) with association rules is used to mine interesting association rules. The efficiency of the association rule algorithm (AR) in delivering quality association rules can be enhanced through GA and through correlation in GA's fitness function. GA is chosen because it mimics the human's ability to adapt to any form of environment. Furthermore, GA can perform global search as it employs greedy algorithm and thus can help to reduce search time in non-deterministic environments.

By optimizing fitness values, it can serve as threshold to prune the list of association rules within search space in order to obtain positive association rules only. By introducing correlation in GA's fitness function, it replaces the user's role in specifying support and confidence as pruning measurements in typical AR.

4. Significance of study

First, the result from our proposed hybrid application of genetic algorithms and association rules allows the managers to view a set of association rules sorted according to rule length, allowing drill-down into many levels-of-details. In other words, we can view a set of rules consisting of parent rule, which is the most general rule having two item sets, followed by sub-parent rule and so forth. Each rule describes the level of dependency among activities within a workflow. We can regard the most general rules as reference meta-model; sub-parent rules as meta-model and complex rules as a model of best practices.

Knowing many levels of details within the sets of association rules is an advantage to the managers during the design phase of an information system. Depending on the type of information systems to be built, managers can choose to adopt any of these types of rules/models. For example, if a manager is given a project to build a new, standalone information system, it would save the manager and his team's design time if they adopt the complex rule (model) and instantiate it to a platform-specific model. If the manager is given a task of building an information system, which is consolidated into an existing information system, then adopting the most general rule (reference meta model) is sufficient, considering that the embedded system serves as a sub-module within an existing system.

In keeping with current trends, the set of association rules which represent best practices are updated by re-applying the proposed hybrid algorithms. Managers are given access to monitor and be notified of every updated best practice, so that their team can rapidly enhance the existing information system or build new ones. The result is a constantly updated, convenient and easy-to-use information system.

Second, our hybrid GA with association rule (GA-AssocRule) supports both upward and downward closure. The following explains how our solution supports upward closure. Suppose that we can find a few association rules having positive correlation score such as $A \rightarrow B$ (length = 2), $A\&C \rightarrow B$ (length = 3) and $A\&C \rightarrow B\&D$ (length = 4) with the correlation scores of 1.05 respectively. We can identify that association rule length of 3 and 4 are related to association rule length of 2 because all association rules share common basic items, i.e., *A* and *B*, and can be found in the association rule of length 2.

The ability to support upward closure allows us to sort the positively-mined correlated association rules based on rule length and allows us to see the progression and transformation of association rules from basic rules to more complex association rules of increasing rule length. In other words, we are able to identify parent association rule, sub-parent association rules and other complex rules.

The sections below cover the following: Section 4 introduces existing works on process improvement approaches and advantages of using data mining to assist the organization in strategic planning. The subsections in Section 4 describe the association rule algorithm and GA as well as application of GA for rule learning. Section 5 presents the application of hybrid GA with association rules (GA-AssocRule) to an online hotel reservation system. Subsequently, results from the application of hybrid GA and association rule are analyzed and relevance to results obtained through our IMoBe methodology is discussed. This paper ends with a conclusion and a list of references.

5. Related works

A successful organization is deemed as an organization that constantly strives to reduce risks as well as improve efficiency, effectiveness and productivity of workflows and the workforce within the organization. Several process improvement approaches have been introduced such as Capability Maturity Model integration (CMMI), IT Infrastructure Library (ITIL) and ISO SPICE (ISO/IEC-15504, 1998) to assist organizations that intend to enhance the performance of their IT projects, divisions and organization as a whole. However, tedious documentation and abstract guidelines are identified as barriers to adopting quality assurance frameworks.

Hence, Niazi et al. (2005) propose an action-based framework namely software process improvement (SPI) implementation model to replace CMMI. The SPI implementation model has reduced the existing CMMI maturity levels to three levels. Each SPI level is associated with minimally a category consisting of a set of critical success/ critical barrier factors (CSF/CBF). Niazi et al. (2005) also mention that if an organization is able to achieve a group of CSF/CBF within the category, the organization attains an SPI level that equates to a CMMI maturity level.

Introduction to these process improvement approaches highlight the importance of evaluating and tracking the organization's internal workflows to ensure that these business processes support the organization's objectives and achieve the organization's targets. Consequently, business performance management tools are incorporated in order to assess the current organization's performance and subsequently, to improve the efficiency and effectiveness of business processes. The most popular business performance measurement is the Balanced Scorecard (BSC). The BSC includes internal business processes as one of its four compulsory perspectives. Evaluation on the organization's internal business processes is crucial as it allows the organization's management to understand its current performance and better strategize how to reduce the gap between current and targeted performance.

The behaviour of people, structure, systems and resources can be captured within Workflow Management Systems (WFMS) through business process modeling. The analysis of workflow (business workflow analysis or BWA), traditionally, involves manual examination of activities within a workflow for redundancies and conflicts. Today, with the advent of the Internet, many business activities such as e-commerce along with customers' behavior are captured in server logs. Considering that the target group for e-commerce comes from a global market, we may have a huge server log that requires the use of tools/techniques to analyze the workflow. WFMS can be enhanced using data mining techniques to provide users with insight and guidance to decide on the best strategy to be executed; allowing the organization to establish its competitive advantage in the global market.

5.1. Association rules

Piatetsky and Frawley (1991) describe the "*association problem*" as searching for *highly occurring* patterns in large databases. The apriori algorithm (Agrawal et al., 1993) supports the definition by Piatetsky and Frawley (1991). Until today, it remains popular. Association rule learning algorithm is taught in many schools throughout universities in the world and is implemented successfully in many application domains (Nebot & Berlanga, 2012; Sanchez, Vila, Cerda, & Serrano, 2009; Subramanian, Ananthanarayana, & Narasimba Murty, 2003). For example, research studies have shown that application of association rule algorithm helps in mining patterns that can greatly assist decision making in business such as foreign exchange market (Liao, Lu, & Lai, 2012) and Korea Composite Stock Price Index (KOSPI) (Na & Sohn, 2011).

Association rule mining is also known as market basket analysis due to its successful implementation in the business domain (Turban, Sharda, & Delen, 2011). Association rule mining aims to search for interesting links between items in large databases and forms rules consisting of antecedent and consequent (Agrawal et al., 1993). The most common association rule mining algorithm being used is the Apriori algorithm. Its detailed implementation and success stories are well-documented. Other algorithms include Eclat, FP-tree, etc. Evaluation metrics for rules lies in defining support and confidence. Given an association rule of the form $A \rightarrow B$ where A and B represent a unique set of items, support is defined as the frequency that the items exist in the same transaction (Turban et al., 2011). Confidence is denoted by conditional probability of searching for the set of items in the antecedent part of the rule given that the set of items is present in the consequent part of the rule. In mathematical formula, it is denoted by:

$$\text{Confidence}(A \rightarrow B) = P(A \cap B)/P(A) \tag{1}$$

The drawback of the typical support-confidence framework is that it requires specifying support and confidence values by the user. However, experts are faced with the challenging task of having to define the "right" settings for support and confidence. Relatively high values of support and confidence will prune even good association rules while relatively lower values will result in displaying mixed quality and many association rules.

Although the main objective of the Apriori algorithm is to display a list of association rules with high support and confidence, it does not mean that rules with lower confidence or support are insignificant association rules. There is still a chance to discover "interesting" association rules from the list of rules with lower confidence. By disregarding confidence as a measuring factor, the dependency among items in association rules can be calculated by using the correlation formula as below.

$$\text{Correlation}(A, B) = P(A \cap B)/P(A) * P(B) \tag{2}$$

where A and B are two unique sets of items, $P(A)$ = probability that item A occurs, $P(B)$= probability that item B occurs and $P(A \cap B)$= probability that both items, A and event B occur.

5.2. Genetic algorithm (GA)

GA is a heuristic method that aims at providing optimal or near-optimal solutions to a high complexity problem that requires long computations without any specific formula to solve the problem (Turban et al., 2011). Based on Darwin's Theory on *survival of the fittest*, genetic algorithm (GA) (Holland, 1975) mimics biological cell production and re-production where selection, crossover and mutation form the operations within GA. Thus, GA is believed to be an adaptive optimization technique due to its ability to use the three operators in any conditions in any environment. GA has been successfully applied in many applications such as images (Tang, Yuan, Sun, Yang, & Gao, 2011; Ware & Wilson, 2003).

GA is a recursive procedure and the possible solutions are individuals represented as genes having strings of 0's and 1's. Each individual will be evaluated via fitness function to determine its fitness value. Individuals of high fitness values form a pool of parent individuals at a generation. Parents produce children through the operation of selection, crossover, and mutation. Therefore, children with high fitness values are regarded as parents in the next generations. The GA cycle continues until the fittest individual(s) with value of 1 is found or the maximum generation has been achieved. The aim of GA is to try to build higher quality generations of individuals by selecting only the fittest surviving individuals as parents. The higher the fitness values, the more likely, the best solution can be found. Elitism is allowed for the current fittest individual to join the next generation of individuals. It is a way to maintain the current best solution that is obtained so far.

At the selection process, individuals are evaluated and those with higher fitness values have the probability of being selected for crossover and mutation processes. Crossover, also known as "mating", allows GA to randomly choose crossover points among two individuals. Swapping occurs at crossover points and the result will be two new individuals. Fig. 1 below shows an example of the application of two individuals before and after crossover operation.

Mutation is a process of randomly changing the gene of an individual in order to create variants and reduce homogeneity in each individual. Altering the gene creates a new individual. Fig. 2 shows an example of an individual before and after the mutation operation.

The general process for GA can be summarized in steps as below:

```
Step 1: Represent the content of the problems as strings of genes
Step 2: Initialize the population and generate initial number of individuals
Step 3: Evaluate the individuals' fitness values using the fitness function
Step 4: Is there any individual having fitness value = 1 (perfect solution)?
    Step 4.1:If yes, proceed to Step8.
    Step 4.2: If no, Elite individuals are automatically brought to the next generation.
    Step 4.3: Select individuals having high fitness values and proceed to Step 5.
Step 5: Apply crossover operation.
Step 6: Apply mutation operation.
Step 7: Maximum generation reached?
Step 7.1: If Yes, proceed to Step 8.
Step 7.2: If No, repeat Step 3.
Step8: Stop GA process and display result of fittest individual.
```

5.3. Application of GA for rule learning

There are two different methods to individual representation of rules for GA, namely Michigan's and Pittsburgh's. In Michigan's method, each rule is represented as an individual whereas in the Pittsburgh's method, a set of rules represents an individual.

Works by Fidelis, Lopes, and Freitas (2000) as well as Weiss and Hirsh (1998) adopt Michigan's method to discover prediction rules. According to Yan, Zhang, and Zhang (2005) this approach is not feasible for bigger number of attributes as it results in long string of genes for each individual which will be complex to manipulate. Yan et al. (2005) proposes association rule mining with genetic algorithms (ARMGA), consisting of hybrid GA and association rules. ARMGA uses relative confidence as fitness function in order to derive positive association rules. ARMGA allows the user to bypass the need to specify the minimum support, which according to Yan et al. (2005) is the most sophisticated task. ARMGA is further enhanced to Evolutionary ARMGA (EARMGA) (2009) to support numerical attribute values. It uses FP-tree as its association rules instead of the Apriori algorithm due to its proven search efficiency in terms of time and cost.

An example of a work using the Pittsburgh method is that Au and Chan's dAR (2002). dAR applies a combination of GA and association rules to trace changes in the dataset. Work by Anandhavalli, Suraj Kumar, Ayush Kumar, and Ghose (2009) proposes the use of GA and association rules in order to identify complex rules having two characteristics, i.e., having *negative attributes* and multi-attributes at the consequence of each rule. Its fitness consists of confidence and completeness which are a combination of true positive, true negatives, false positives and false negatives. Because it focuses on *negative* attributes (attributes not likely to be present), most of the generated rules have lower support values.

In this study, we adapt the concept of EARMGA which hybridizes GA with FP-tree algorithm. FP-tree algorithm is preferred due to its reduced search time and still inheriting downward closure property, similar to apriori algorithm.

6. Case study – online hotel reservation system

This section describes the application of our proposed GA-AssocRule on an online hotel reservation system hitherto referred to as E1 of an airline system (Lee & Lim, 2007), with emphasis on its selection process. GA is preferred due to its ability to optimize correlation as fitness function by applying a set of operators known as selection, crossover and mutation on a group of randomly generated association rules within a generation. Our aim is to identify a set of positively correlated association rules related to business processes within E1. The following paragraph describes the selection process within E1.

Initially, when a user needs to book a hotel room, he/she will insert the hotel booking details such as check-in and check-out dates, number of rooms, number of guests, the country and town in which the hotel resides and many other initial data. Based on the booking details provided by the customer, the system will generate a list of hotel names. The customer can choose a particular hotel name from the list to view and/or print the hotel information. If the customer is not satisfied with the available list of hotels, he/she can modify the booking details in order to generate a new list of hotel names.

When the customer has selected the hotel from the list, the system will display a summary of booking details and the selected hotel along with the list of available room types offered by the selected hotel. The customer can choose to view the daily break-

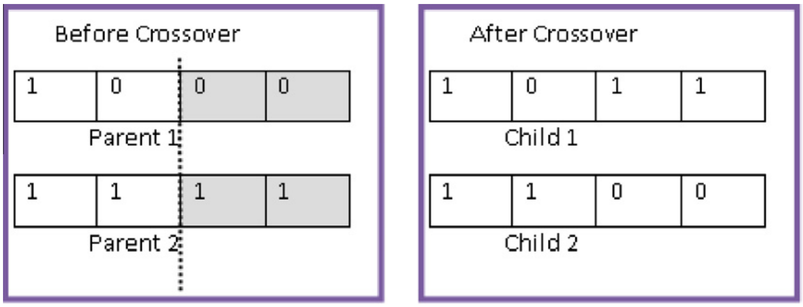


Fig. 1. Before and after crossover operation.

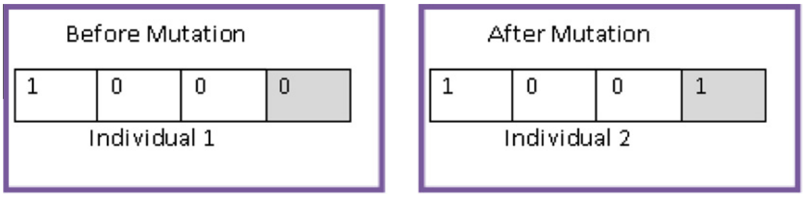


Fig. 2. Before and after mutation operation.

down of room rates. At this stage, the customer can still amend his/her selected hotel to another hotel. When the customer has selected the room type, the system will display a summary of the updated reservation and display this updated information. He/She can choose to view the breakdown of the rates and charges. The following Fig. 3 shows the proposed methodology of our experiment. Each process is further illustrated in the following sub-sections.

6.1. Represent E1 using Petri net

The workflow of E1 as described above is translated into Petri net representation. Article by van der Aalst (1998) highlights many advantages of using the Petri net representation to support modelling of intricate workflows and verification of correctness of workflows. Thus, many Petri net tools exist for the commercial and research community.

Since the audit trails of E1 remain as proprietary knowledge only to the airline company itself, we decide to choose CPN Tools because this tool allows us to represent E1 and simulate event logs. Fig. 4 shows the partial snapshot of E1 represented as Petri net in CPN Tools (<http://cpntools.org>).

6.2. Simulate event log for E1

An event log consists of event instances which are a collection of events that took place at a particular time and by particular person/department. By using CPN Tools, we can generate case ID for each instance and details of each event within each instance are simulated such as event type, timestamp and originator. A total of 1000 simulated instances form an event log.

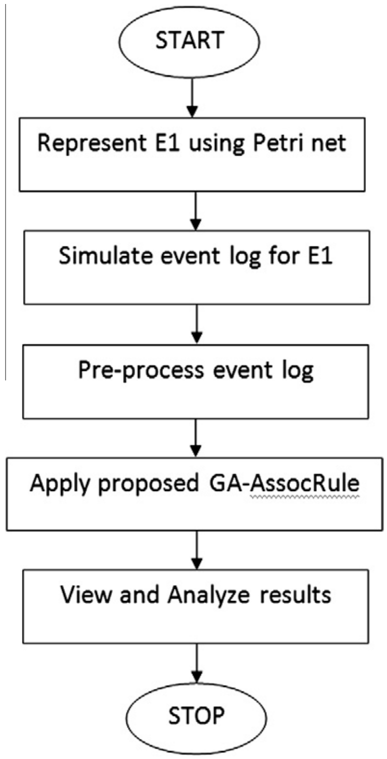

START
Represent E1 using Petri net
Simulate event log for E1
Pre-process event log
Apply proposed GA-AssocRule
View and Analyze results
STOP

Fig. 3. Proposed methodology.

6.3. Pre-process event log

Further pre-processing is applied to the event log. Instances are grouped according to the number of business processes. The largest group of instances are instances having 16 business processes. They consist of 200 instances. Each instance starts with the same business process i.e., "*Insert Hotel Details*". These instances will serve as input data to be fed into the proposed GA-AssocRule simulation.

6.4. Apply proposed GA-AssocRule

The source code for proposed GA-AssocRule is adapted from Alcala-Fdez et al. (2008). Table 1 shows the parameter settings that are applied in our proposed GA-AssocRule.

Similar event log and parameter settings are applied on EARMGA (Yan, Zhang, & Zhang, 2009). The results are compared and discussed in the following subsection.

The difference between our proposed GA-AssocRule and EARMGA (Yan et al., 2009) lies in the fitness function. In EARMGA (Yan et al., 2009), the fitness function is based on relative confidence with focus on mining positive association rules. The user is allowed to specify the relative confidence score between the range of $[-1, 1]$. The score will be mapped to a value between $[0, 1]$. Given a rule of the form $X \rightarrow Y$ where X and Y are unique items, relative confidence is identified in the equation below. It also became the fitness function of EARMGA (Yan et al., 2009).

$$\text{Fitness-Function}_{\text{EARMGA}} = [\text{supp}(X \cup Y) - \text{supp}(X) * \text{supp}(Y)]/[\text{supp}(X)(1 - \text{supp}(Y))] \quad (3)$$

However, as mentioned by Brin et al. (1997), inheriting the property of confidence means that it does not support the upward and downward property of closures. An example that indicates that confidence does not support upward closure is available in the paper published by Brin et al. (1997).

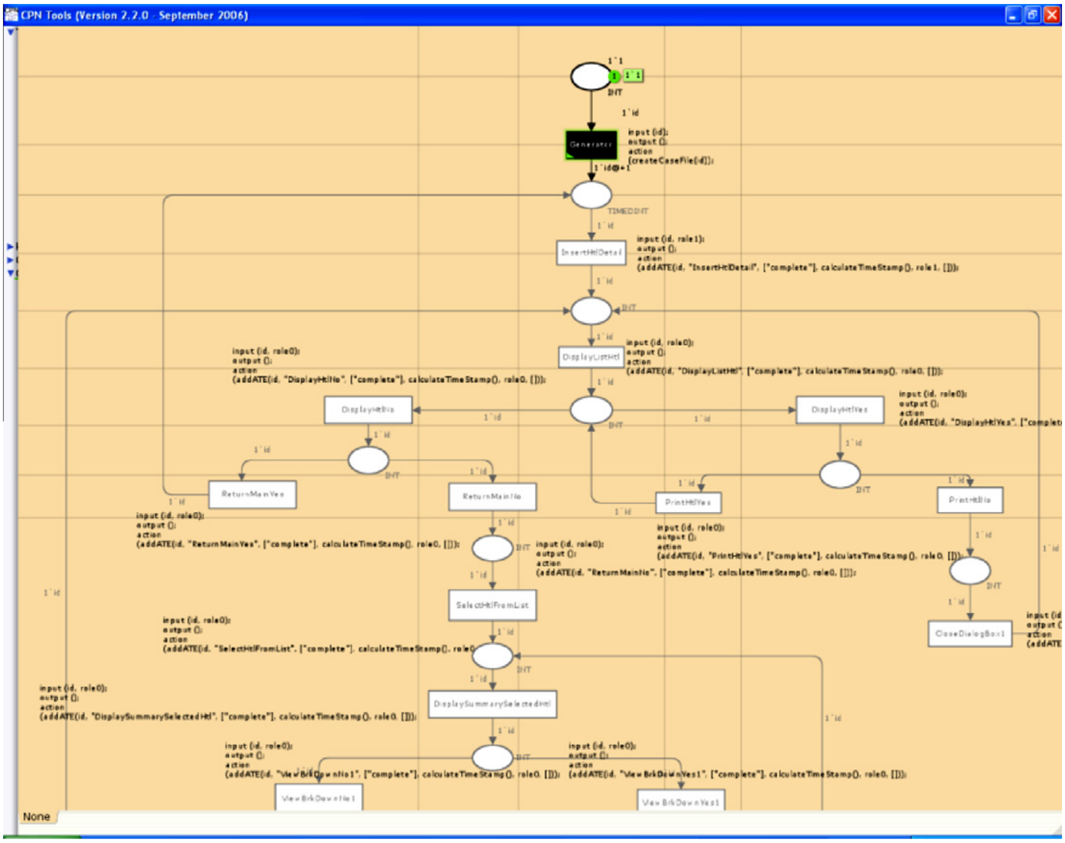
Since correlation supports upward and downward closures, correlation is used as fitness function in our proposed hybrid GA-AssocRule as indicated below:

$$\text{Fitness-Function}_{\text{GA-correlation}} = [\text{supp}(X \cup Y)]/[\text{supp}(X) * \text{supp}(Y)] \quad (4)$$

6.5. View and analyze results

In this section, we will show the association rules that are generated based on parameter settings as in Table 1. Figs. 5 and 6 show the set of association rules of different complexities, which can be obtained by applying both EARMGA and proposed GA-AssocRule on the processed event log with mutation rate of 0.05.

Based on Figs. 5 and 6, we can see that both algorithms successfully obtain the sets of positive correlated association rules of var-



```
input (id);
output ();
action
(createCaseFile(id));

input (id, role1);
output ();
action
(addATE(id, "InsertHtlDetail", ["complete"], calculateTimeStamp(), role1, []));

input (id, role0);
output ();
action
(addATE(id, "DisplayListHtl", ["complete"], calculateTimeStamp(), role0, []));

input (id, role0);
output ();
action
(addATE(id, "DisplayHtlNo", ["complete"], calculateTimeStamp(), role0, []));

input (id, role0);
output ();
action
(addATE(id, "DisplayHtlYes", ["complet

input (id, role0);
output ();
action
(addATE(id, "ReturnMainYes", ["complete"], calculateTimeStamp(), role0, []));

input (id, role0);
output ();
action
(addATE(id, "ReturnMainNo", ["complete"], calculateTimeStamp(), role0, []));

input (id, role0);
output ();
action
(addATE(id, "PrintHtlYes", ["complete"], calculateTimeStamp(), role0, []));

input (id, role0);
output ();
action
(addATE(id, "SelectHtlFromList", ["complete"], calculateTimeStamp(), role0

input (id, role0);
output ();
action
(addATE(id, "DisplaySummarySelectedHtl", ["complete"], calculateTimeStamp(), role0, []));

input (id, role0);
output ();
action
(addATE(id, "ViewBrkDownNo1", ["complete"], calculateTimeStamp(), role0, []));

input (id, role0);
output ();
action
(addATE(id, "ViewBrkDownYes1", ["complete"], calculateTimeStamp(), role0, []));

input (id
output (
action
(addATE
```

Fig. 4. Partial snapshot of E1 in CPN Tools.

Table 1
Parameter settings.

Parameters	Value
Population size	100
Number of generations	100
Probability of crossover	0.95
Probability of mutation	0.95

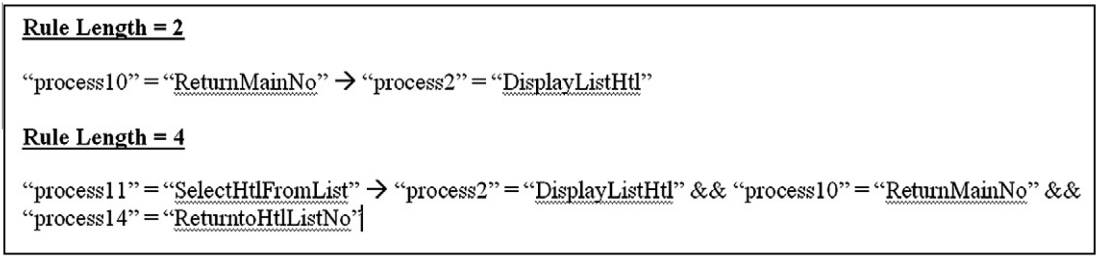
iable lengths. The number associated with each process represents the ordering of the business process within E1. In other words, if the attribute is "process2" it represents the second business process within E1. Based on Fig. 5, we can get two strong association rules which are association rules of length 2 and 4. However, we are unable to obtain the association rule of length 3 which has relation between association rules of length 2 and 4. In other words, EARMGA fails to derive a set of related association rules that can be arranged to show the progression of rules' complexity from basic association rules to complex rules.

Although the result from the application of EARMGA does not show a set of correlated and related association rules of varying length, the two association rules are still valid rules. In Fig. 5, there is a strong relationship between process10 and process2. Process10 represents the activity of not returning to the main page of the online hotel reservation system while process2 displays the list of hotels. This association rule is valid because if the user does not return to the main page of the online reservation system to re-key the hotel booking details, it means that the user is satisfied with his/her hotel booking details and interested to know the possible list of hotels that match his/her booking criteria. In Fig. 5, the association rule of length 4 is not directly an association rule that stems from the association rule of length 2 but of a closer match. Here, process11 is activated based on a group of three processes which are process2, process10 and process14. In other words, if the user is currently at process11 of "*selecting hotel from list*" it means that the user has been given a list of hotels based on his/her hotel booking criteria in the previous step, does not return to the main page of the online hotel reservation system to change his/her hotel booking criteria and does not return to the list of hotels to re-select another hotel. Although EARMGA fails to display a set of correlated and related association rules, EARMGA manages to derive two valid association rules that have support and confidence scores of one respectively.

In Fig. 6, our proposed hybrid GA-AssocRule successfully displays a set of association rules that shows the progression from the most general association rule of length 2 to sub-parent association rule of length 3 and subsequently complex association rule of length 4. All the above three association rules contain two common items, which form the most general association rule of length 2. In the most general association rule, there is a strong dependency between process5 and process8. It is a valid and strong association rule because in order for the user to select a particular hotel from the list, there is a need to display the list of hotels to the user.

The general association rule can be further extended to include another process without affecting the correlation score of the basic rule. The sub-parent rule extends the basic association rule to incorporate the initial-process of inserting hotel booking details. This is also a strong association rule because process5 displays the list of hotels and subsequent processes stemming from the initial-process by the user who keys in his/her hotel booking details.

The sub-parent association rule can be further expanded to association rule of length 4 to include another process namely process9 (displaying a summary of selected hotels). Process9 is enabled only after the user has gone through a series of processes



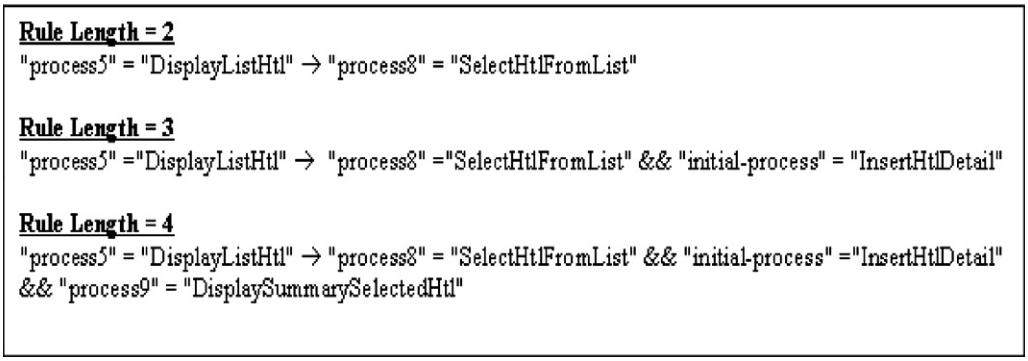
Rule Length = 2

“process10” = “ReturnMainNo” → “process2” = “DisplayListHtl”

Rule Length = 4

“process11” = “SelectHtlFromList” → “process2” = “DisplayListHtl” && “process10” = “ReturnMainNo” && “process14” = “ReturntoHtlListNo’

Fig. 5. Set of rules from application of EARMGA.



Rule Length = 2

"process5" = "DisplayListHtl" → "process8" = "SelectHtlFromList"

Rule Length = 3

"process5" ="DisplayListHtl" → "process8" ="SelectHtlFromList" && "initial-process" = "InsertHtlDetail"

Rule Length = 4

"process5" = "DisplayListHtl" → "process8" = "SelectHtlFromList" && "initial-process" ="InsertHtlDetail" && "process9" = "DisplaySummarySelectedHtl"

Fig. 6. Set of rules from application of proposed GA-AssocRule.

from the initial process of inserting his/her hotel booking details, displaying a list of hotels and selecting a hotel from the list.

Although additional items, namely processes, are added that results in incremental length of formation of association rule from length of 2 to 4; it does not affect the correlation and support scores of each association rule. Referring to Fig. 6, we are able to obtain three positively correlated association rules having high support scores of 0.9729 respectively while the support values for the two rules in Fig. 5 are valued at 1 respectively. Although mathematical comparison shows that rules with support value of 1 in Fig. 5 are better compared to rules in Fig. 6 having support values of 0.9729, rules in Fig. 6 are also strong rules which can be considered as best practices.

As mentioned in the third paragraph of Section 3, ‘‘*the correlation score of the superset should minimally equal to the correlation score of n-items*.’’ The three association rules have correlation scores of 1.0278 respectively representing positive association rules where both sub-parent association rule and complex rule have minimally the same correlation scores as the most general association rule. Positively correlated association rules contain at least two processes which exists within the event log that are highly dependent on one another. These processes are popularly occurring set of processes within the workflow which managers should adopt when designing or enhancing any information systems.

7. Relation between results in Section 6.5 and IMoBe framework

As mentioned in Lim and Lee (2008), IMoBe is a comprehensive business performance framework consisting of BSC and QFD methodology with model-driven KB inclusive of CSF/CBF and Sun Tzu’s strategies. To achieve the objectives in each BSC perspective requires managers to identify and decide the extent to which the relevant CSF/CBF or Sun Tzu’s strategies should be executed. Although these outcome-based models (CSF/CBF and Sun Tzu’s strategies) help organizations to achieve the perspectives in BSC, each factor or strategy requires a certain form of support such as materials, libraries, methods or others during execution to ensure that the factor/strategy can be successfully executed.

When IMoBe (2008) is applied on the same airline company using CSF/CBF as model within HOQ matrix, internal business perspective is identified by managers as the main focus of the organization as improvement in this perspective will have ‘‘domino effect’’ leading to improvement in other perspectives within BSC. To improve this perspective, *review* is an example of the most important CSF/CBF that should be executed. By definition, *review* on internal business perspective refers to constant assessment on the organization’s workflow and executing this factor requires availability of selected supporting components.

The first supporting component is the availability of our results from the application of data mining on server log in Section 6.5, which can serve as a pool of best practices to different levels of details. Having libraries of best practices will assist managers to review or compare existing internal workflows with the best practices. With different levels-of details of best practices, managers can choose to enhance the existing workflows depending on demand or system architecture.

The second supporting component is the availability of means to evaluate and update the best practice to ensure that during the review activity, the latest best practices can be compared with existing workflow. In this paper, hybridizing genetic algorithm and correlation-based association rules is proposed because genetic algorithm mimics human cells of selection crossover and mutation, making this algorithm adaptive like humans and therefore works well in dynamic business environments. Its adaptability means it can avoid subjective assignment of support and confidence values in association rules algorithms and focus on deriving positively correlated items (events). Furthermore, hybrid GA-Association rule can be executed at anytime or periodically on the server log although it keeps increasing in size, ensuring that managers always have the latest best practices based on trends when reviewing and evaluating the performance of the organization's internal business operations.

In the above paragraph, two supporting components for *review* have been identified. Work by Niazi et al. (2005) classifies *review* as the only CSF/CBF in the "Support" category. Niazi et al. (2005) mapped "Support" as front-end category and back-end categories as "Awareness" and "Organizational". Achieving both categories equates to achieving CMMI Level 4 Maturity Level (Optimising). Back-end categories consist of lower-end practices, which are linked to CMMI Level 2 and Level 3 respectively. Hence, back-end categories are important and successful execution of CSF/CBF in Back-end categories with supporting components for the *review* factor will ensure that *review* can be executed successfully. Subsequently, successful execution of *review* indicates successful implementation of front end category practices. A collection of successful implementation of back-end and front-end categories indicates that the organization has achieved CMMI Level 4; a very commendable industry achievement.

8. Conclusion

We have enhanced the existing WaLwFA, which currently supports DSS and OLAP during the decision-making process. We have also improved on the DSS_OLAP component in the IMoBe framework. We have introduced GA as hybrid to the existing association rule algorithm to achieve two objectives. The first objective is to complement the existing association rule algorithm to support upward closure through the introduction of correlation to replace support and confidence as measures in association rules and fitness function in GA. The second objective is to introduce correlation as fitness function in GA. This serves to allow optimization and reduce search time and search space in order to get highly positive correlated association rules. It also serves to bypass the difficult challenge faced by users in specifying the support and confidence parameter values. Positively-correlated association rules consisting of business processes as items represent high dependency among business processes. Such association rules can serve as best practices to refine strategic planning models and to refine and enhance the design of decision support systems. Experimental outcomes have confirmed the viability of our approach.

References

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD conference on management of data* (pp. 207–216).

Alcala-Fdez, J., Garcia, S., Berlanga, F. J., Fernandez, A., Sanchez, L., del Jesus, M. J., & Herrera, F. (2008). KEEL: A data mining software tool integrating genetic fuzzy systems (pp. 83–88).

Anandhavalli, M., Suraj Kumar, Sudhanshu, Ayush Kumar & Ghose, M. K. (2009). Optimized association rule mining using genetic algorithm. *Advances in Information Mining, 1*(2), 01–04.

Au, W., & Chan, K. (2002). An evolutionary approach for changing patterns in historical data. In *Proceedings of 2002 SPIE* (pp. 398–409).

Brin, S., Motwani, R., & Silverstein, C. (1997). Beyond market baskets: generalizing association rules to correlations. In *Proceeding of the 1997 ACM SIGMOD international conference on management of data*.

Fidelis, M., Lopes, H., & Freitas, A. (2000). Discovering comprehensible classification rules with a genetic algorithm. In *Proceedings of the 2000 Congress of Evolutionary Computation* (pp. 805–810).

Holland, J. H. (1975). Adaptation in natural and artificial systems. Ann Arbor: The University of Michigan Press.

ISO/IEC-15504, 1998. Information Technology-software process assessment. Technical Report – Type 2.

Kaplan, R.S., Norton, D.P., 1992. The balanced scorecard: Measure that drives performance, Harvard Business Review.

Ko, S. O., & Lee, S. F. (2000). Implementing the strategic formulation framework for the banking industry of Hong Kong. *Managerial Auditing Journal*, 469–477.

Lee, C.-S., & Lim, A. H. L. (2007). Layered and weighted approach to workflow evaluation. *International Journal of Electronic Business, 5*(3), 380–400.

Lee, S. F., Roberts, P., Lau, W. S., & Bhattacharya, S. K. (1998). Sun Tzu's The Art of War as business and management strategies for world-class business excellence evaluation under QFD methodology. *Business Process Management Journal, 4*(2), 96–113.

Liao, S.-H., Lu, S.-L., & Lai, Y.-W. (2012). Mining the hedge and arbitrage of the Taiwan foreign exchange market. *Expert Systems with Applications*, 3197–3206.

Lim, A. H. L., & Lee, C.-S. (2008). Integrated model-driven business evaluation methodology for strategic planning. *International Journal of Business Information Systems, 3*(4), 333–355.

Lim, A. H. L., & Lee, C.-S. (2010). Processing Online Analytics with Classification and Association Rule Mining. *Knowledge-Based Systems.*.

Liu, B., Hsu, W., & Ma,Y. (1999). Mining association rules with multiple minimum supports. In *The 1999 international conference on knowledge discovery and data mining* (pp. 337–341).

Na, S. H., & Sohn, S. Y. (2011). Forecasting changes in Korea Composite Stock Price Index (KOSPI) using association rules. *Expert Systems with Appplications*, 9046–9049.

Nebot, V., & Berlanga, R. (2012). Finding association rules in semantic web data. *Knowledge-Based Systems, 25*(1), 51–62.

Niazi, M., Wilson, M., & Zowghi, D. (2005). A maturity model for the implementation of software process improvement: An empirical study. *Journal of Systems and Software*, 155–172.

Piatetsky, G., & Frawley, W. (1991). *Knowledge discovery in databases*. AAAI/MIT Press.

Porter, M. E. (1979). How competitive forces shape strategy. harvard business review. March/April 1979.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann Publishers.

Sanchez, D., Vila, M., Cerda, L., & Serrano, J. (2009). Association rules applied to credit card fraud detection. *Expert Systems with Applications, 36*(2), 3630–3640.

Subramanian, D. K., Ananthanarayana, V. K., & Narasimba Murty, M. (2003). Knowledge-based association rule mining using AND–OR taxonomies. *Knowledge-Based Systems, 16*(1), 37–45.

Tang, K., Yuan, X., Sun, T., Yang, J., & Gao, S. (2011). An improved scheme for minimum cross entropy threshold selection based on genetic algorithm. *Knowledge-Based Systems, 28*(8), 1131–1138.

Turban, E., Sharda, R., & Delen, D. (2011). Decision Support and business intelligence systems. Pearson International Edition.

van der Aalst, W. M. P. (1998). The application of Petri nets to workflow management. *Journal of Circuits, Systems and Computers, 8*(1), 21–66.

Ware, J. M., & Wilson, I. D. (2003). A knowledge based genetic algorithm approach to automating cartographic generalization. *Knowledge-Based Systems, 16*(5–6), 295–303.

Weiss, G., & Hirsh, H. (1998). Learning to predict rare events in event sequences. In *Proceedings of the 4th international conference on knowledge discovery and data mining* (pp. 359–363). AAAI Press.

Yan, X., Zhang, C., & Zhang, S. (2005). Armga: Identifying interesting association rules with genetic algorithms. *Applied Artificial Intelligence, 19*(7), 677–689.

Yan, X., Zhang, C., & Zhang, S. (2009). Genetic algorithm-based strategy for identifying association rules without specifying minimum support. *Expert Systems with Applications, 36*, 3066–3067.